OXFORD

Oxford Textbook of Global Public Health

The Scope of Public Health

SEVENTH EDITION

EDITED BY
Roger Detels
Quarraisha Abdool Karim
Fran Baum
Liming Li
Alastair H. Leyland

Free personal online access for 5 years

The copy of *Oxford Textbook of Global Public Health* you have purchased entitles you to free personal online access for 5 years.
Customers outside North and South America, please visit:
https://subscriberservices.sams.oup.com/token to set up access.
Customers in North and South America, please visit:
https://ams.oup.com/order/OTPUBHEA5YRSCRIP to set up access.
Please note: the URL is case-sensitive.
To register, enter the code on this page. Please keep the number for future reference as, in the event of a query, you may be asked for it.
Please note that as part of the registration process you will be informed of our terms and conditions, and will be asked to accept these to confirm your online access.

Unique code:

The online access provided free with this book is for individuals who have purchased a personal copy only, and this will be verified as part of the access process. Users of library copies should ask their librarian about access in their institution.

CUSTOMER SUPPORT

Customers outside North & South America Tel: +44 (0) 1865 353705
Email: accesstokens@oup.com
Customers in North & South America
Tel: 1-800-334-4249 ext. 6484
Email: oxfordonline@oup.com

Oxford Textbook of

Global Public Health

Oxford Textbook of

Global Public Health

SEVENTH EDITION

Volume 1: The Scope of Public Health Volume 2: The Methods of Public Health Volume 3: The practice of public health EDITED BY

Roger Detels

Distinguished Research Professor of Epidemiology and Infectious Diseases, Schools of Public Health and Medicine, University of California, Los Angeles, USA Quarraisha Abdool Karim Associate Scientific Director, CAPRISA, Professor in Clinical Epidemiology, Columbia University, Pro Vice-Chancellor (African Health), University of Kwazulu-Natal, South Africa Fran Baum
Matthew Flinders Distinguished Professor of Public Health and Director of the Southgate Institute of Health, Society and Equity, Flinders University, Australia Liming Li
Professor of Epidemiology, School of Public Health, Peking University Health Science Center, Beijing, China

Alastair H. Leyland

Associate Director and Professor of Population Health Statistics, MRC/CSO Social and Public Health Sciences Unit, University of Glasgow, UK

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Great Clarendon Street, Oxford, OX2 6DP, United Kingdom

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship, and education by publishing worldwide. Oxford is a registered trade mark of Oxford University Press in the UK and in certain other countries © Oxford University Press 2022

The moral rights of the authors have been asserted First edition published in 1984
Second edition published in 1991
Third edition published in 1997
Fourth edition published in 2002 (reprinted in paperback 2004, 2005 twice) Fifth edition published in 2009 (reprinted in paperback 2011) Sixth edition published in 2015
Seventh edition published in 2022

Impression: 1

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, without the prior permission in writing of Oxford University Press, or as expressly permitted by law, by licence or under terms agreed with the appropriate reprographics rights organization. Enquiries concerning reproduction outside the scope of the above should be sent to the Rights Department, Oxford University Press, at the address above
You must not circulate this work in any other form and you must impose this same condition on any acquirer Published in the United States of America by Oxford University Press 198 Madison Avenue, New York, NY 10016, United States of America British Library Cataloguing in Publication Data Data available

Library of Congress Cataloguing in Publication Data Data available
ISBN 978–0–19–887166–8
eISBN 978–0–19–254877–1
DOI: 10.1093/med/9780198816805.001.0001

Oxford University Press makes no representation, express or implied, that the drug dosages in this book are correct. Readers must therefore always check the product information and clinical procedures with the most up-to-date published product information and data sheets provided by the manufacturers and the most recent codes of conduct and safety regulations. The authors and the publishers do not accept responsibility or legal liability for any errors in the text or for the misuse or misapplication of material in this work. Except where otherwise stated, drug dosages and recommendations are for the non-pregnant adult who is not breast-feeding Links to third party websites are provided by Oxford in good faith and for information only. Oxford disclaims any responsibility for the materials contained in any third party website referenced in this work.

Preface to the seventh edition

Since the publication of the 6th edition of the Textbook in 2015, there have been many events and changes that have impacted public health and which public health must address. Although we have made progress in many areas globally new challenges, often unanticipated, have emerged or re-emerged that compromise known and effective public health interventions. Public health now operates in a world of rapid communications and social media which, though very influential, is not monitored for veracity and, thus, can be counterproductive for public health goals (see Chapter 4.3). The rise in machine learning and artificial intelligence are redefining personal and public health practice globally and could be an important bridge to address health disparities particularly in settings where health infrastructure and resources are limited or lack of affordability could sow greater divisions even in industrialized countries.

Several new infectious disease agents have emerged for which we must develop innovative interventions (Chapter 8.11). These include Zika virus, the Middle Eastern Respiratory Virus (MERS), acute flaccid paralysis (AFP), for which an aetiologic agent is currently unknown, and SARS-CoV-2 (Chapter 8.17). Further there has been a resurgence of measles and mumps, diseases which we thought we had largely under control through high immunization coverage. The resurgence of these diseases especially in industrialized countries are not the result of genetic or biologic changes in the agents or poverty, but of changes in the social and political environment that include challenging the use of vaccines. Thus, infectious disease control persists as an important global problem (Chapters 8.13–8.17, 11.3, and 11.4). Mathematical strategies for modelling control of infectious diseases are being increasingly used (Chapter 5.17).

For the first time in recorded history human lifespan is not expanding in many parts of the world. The reasons for this slow down are not fully understood but may be related to issues surrounding access to healthcare,

unhealthy lifestyles (including lack of exercise), the drug epidemic (precipitated by overprescription of painkillers), and an increase in suicide and obesity (Chapters 8.4, 8.5, and 9.2).

While in industrialized countries understanding of and models of healthcare for their ageing populations continue to evolve, in Africa there is a youth bulge resulting from reduced life expectancy in adults primarily as a result of large numbers of adult AIDS-related deaths in the 1990s and early 2000s and the success of the post-2000 PMTCT efforts. These demographic transitions between countries have important ramifications for social cohesion within and between countries.

Politically there has been a shift to conservatism. For example many of the regulations safeguarding the environment are being reversed, particularly in the United States (Chapters 2.2 and 7.1). The impact of these roll backs will exacerbate the problem of climate warming, perhaps currently the major public health issue facing earth (Chapter 2.9). Increasing warming of the earth will impact food supplies, habitable areas, and cause major shifts in the distribution of infectious agents and the vectors which carry many of them. The outcomes of climate change are already apparent in the increasing number of floods and fires plaguing many countries. Unfortunately the issue of climate change has become politicized to the detriment of the global population.

Although the global economy is currently doing well, on average, the disparity between the rich and the poor is widening as the rich get disproportionally richer (Chapter 2.4). This disparity is reflected in the increasing number of individuals who are homeless. Finding shelter and counselling and support services for them has become a political problem stemming from the attitude ‘not in my back yard!’ Inequitable distribution of income which has resulted in increasing levels of poverty remains the major cause of ill-health globally limiting access to healthcare, academic achievement, and a healthy lifestyle (Chapters 2.3, 2.4, 2.5, and 3.3).

Although safe working conditions are still an issue in high-income countries, transnational corporations have shifted manufacturing from high-income countries to low-income countries, where costs are lower. The problem is greatest in low- and middle-income countries in which regulations to promote worker safety and health are minimal because of the need to keep labour costs low. Governments in these countries are reluctant to implement expensive regulations which increase cost, making their

goods less competitive (Chapters 7.4 and 7.5). Worker safety adds to the cost of manufacturing, thus occupational health and safety legislation gets relatively little government attention in poor countries.

Migration from low- and middle-income and/or countries where conflict is high has increased as reflected in the number of individuals seeking asylum in high-income countries. Unfortunately the destination countries have been reluctant to accept the migrants. Underlying this problem is racial bias. As a result many individuals have died and families been disrupted. Violence and hate crimes are increasing, often targeting individuals of different religions or ethnicity. War in the eastern Mediterranean countries has persisted creating refugee camps which promote infectious, social, and psychologic problems. The problem of violence against women persists. Intimate partner violence affects as many as 35–40% of women in some regions of the world. Stigma has complicated disease control programmes, particularly for HIV/AIDS (Chapters 8.13, 8.18, 9.5, 9.6, and 10.2).

Increasing resistance to drugs especially for tuberculosis and gonorrhoea, is now a major problem (Chapters 8.12 and 8.14). We are out of drugs for multidrug resistant tuberculosis and increasingly gonorrhoea. The origin of this growing resistance to drugs is overprescription and commercial use of drugs (e.g. prescribing antibacterial drugs for viral infections) and the widespread use of drugs in the food industry).

Technology has facilitated many interventions but has also created many unanticipated public health problems. The internet and social media are potentially powerful tools for health education and dissemination of public health interventions (Chapter 4.3). However, they can also be used to undermine public health programmes. There are effective vaccines for measles, mumps, and many other infectious diseases. Unfortunately, social media in particular has been used to disseminate false information asserting that vaccination of children can cause autism and other health problems. As a result, many mothers are refusing to vaccinate their children, thus lowering the prevalence of immune individuals below the level of herd immunity required to stop transmission. For example, measles and mumps had been virtually eliminated in North America and Europe by the beginning of the twenty-first century. Due to parents' reluctance to vaccinate children and assumptions that these diseases were no longer a problem, thousands of cases of measles are occurring in North America,

Europe, and elsewhere with a significant number of deaths, especially in Samoa and Ukraine where both cases and deaths have soared. Although California and several other states have mandated immunization as a requirement for entry into school, mothers are putting pressure on physicians to write waivers for their child. Many physicians do not recognize their public health responsibility not to accede to these requests explaining to the mothers that they may actually be putting their children at greater risk of disease and more serious complications if infected at an older age. Other problems have blocked the programme to eradicate polio in the last remaining countries in which they persist. The killing of vaccinators in remote areas of Pakistan and Afghanistan has posed a new challenge to achieving sufficient levels of immune individuals to halt spread and achieve eradication. Thus, it has been necessary to use other strategies, such as mass vaccination days, to achieve a high level of herd immunity in these areas.

Positive mental health is essential for an individual to achieve a healthy productive life and is the foundation for the effective functioning of a community (Chapter 8.7). Yet, it has been estimated that in some populations as high as 25% of total disability is due to mental health disorders and is a major cause of suicide. Among adolescents, in developed countries the proportion diagnosed with mental problems may be as high as 50%. However, the amount and proportion of health dollars spent on mental health is low, particularly in lower- and middle-income countries. The lack of funding translates into a paucity of health professionals capable of treating mental illness.

The advancement of technology in the field of genetics and genomics has resulted in new opportunities for studying the genetic correlates of disease (Chapter 2.6). However, advanced technology such as the altering of genes in human foetuses has also introduced complex ethical issues about the future direction of genetic analysis and intervention (Chapter 3.2).

The rapid rise of the internet and social media presents a serious challenge to the effective implementation of public health interventions (Chapter 4.3). For example, the development of websites to promote commercial and recreational sexual activities plays a role in the current epidemic of several sexually transmitted infections. Future generations of public health professionals will need to develop effective strategies using the internet and social media to counter misinformation being disseminated by them and to promote public health.

The Sustainable Development Goals provide an opportunity to enable full development of healthy nations and communities from birth throughout life encompassing communicable, non-communicable diseases; other causes of morbidity and mortality; gender, economic, and racial disparities and emerging and re-emerging issues; environmental and non-environmental drivers of ill-health; spatial planning, and a basic human rights approach to health and well-being. The recent expansion and growth in social movements globally reflect that citizens are becoming more engaged and want more influence in decisions and actions that shape and influence their lives.

As the 7th edition of the *Oxford Textbook of Global Public Health* goes to press the world is experiencing a pandemic caused by SARS-CoV2 which, only 4 months after it was first recognized in Wuhan, China in early January 2020, has already caused over 3.25 million cases and over 226,000 deaths. Thus far, the only intervention that has been demonstrated to mitigate the pandemic has been social distancing resulting in global quarantining of the population. The impact on the global economy has been devastating. It is clear that the SARS-CoV2 pandemic will have a major impact on public health, the economy, and the way society functions globally in the future (Chapter 8.17).

As in previous editions we have divided the 7th edition of the *Oxford Textbook of Global Public Health* into three volumes—the scope of public health, the methodologies utilized by public health professionals, and the practice of public health. As in previous editions of the *Oxford Textbook of Global Public Health*, the intention of the 7th edition is to provide insights into the range and character of health problems challenging public health and effective strategies to address them as well as to anticipate what the challenges to public health will be in the future. There will be a diversity of views about how to achieve this goal, but hopefully agreement on what the goals are. It is the nature of science to consider divergent views and to select those most likely to achieve success. It is our hope that the 7th edition will continue to contribute to the success of that process.

The target audience of the 7th edition remains public health professionals, particularly those entering into the field, who wish to learn about the scope and diversity of the field and strategies to improve the health of the public. The ultimate goal of the 7th edition is to convey the

excitement and the power of public health to promote ‘health for all’! (Chapter 3.3).

Free personal online access for five years

Individual purchasers of this book are also entitled to free personal access to the online edition for 5 years via oxfordmedicine/otglobalpublichealth7. Please refer to the access token for instructions on token redemption and access.

Accessing this content online allows you to print, save, cite, email, and share content; download high-resolution figures as PowerPoint® slides; save often-used books, chapters, or searches; annotate; and quickly jump to other chapters or related material on a mobile-optimized platform.

During the production of this publication, new public health challenges were presented in the form of COVID-19. This has been addressed where possible in related chapters, but for a full chapter on the topic, please access the online version of this book at Oxford Medicine Online (oxfordmedicine/otglobalpublichealth7).

Preface to Volume 1: *The scope of public health*

This opening volume of the *Oxford Textbook of Global Public Health* showcases the breadth of public health in terms of the range of topics that are covered and the number of disciplines on which public health now draws. The scope of public health has continued to evolve and a strong public health base will be a key to meeting the United Nations' Sustainable Development Goals that relate to both the need for health improvement and for a reduction in inequalities in health. The contributions in this volume provide an introduction to the development of public health as a discipline in countries at different stages of development and give a detailed overview of areas of critical importance. It does this through three sections updated from the previous edition: the development of the discipline of public health, determinants of health and disease, and public health policies, law, and ethics.

Section 1 begins by considering the scope and concerns of public health in the twenty-first century, including an overview of many of the challenges facing contemporary public health, many of which are related to poverty. The next two chapters cover the history and development of public health in developed countries and in low- and middle-income countries. Chapter 1.4 then describes the development of public health in transition, focusing on middle-income countries. The section concludes with an account of how the increased interconnectedness of human societies, or globalization, is impacting on global public health both positively and negatively. This chapter discusses the shift in global health governance from international to global collective action and attempts to address the impacts of globalization through regulatory frameworks.

Section 2 provides in-depth coverage of the determinants of health and disease. A solid understanding of these determinants is necessary before any response to disease can be formulated and public health put into practice. The first chapter in this section presents an overview of biological,

behavioural, social, and environmental determinants of health, but gives particular emphasis to the social determinants of health because of their role in creating health inequalities. Governments (at all levels) have a role in the improvement of population health and the reduction of health inequalities through the creation of healthy public policy. This leads into a chapter on the politics of public health, new to this edition, which describes how politics sit alongside evidence in the formulation of policy. An evaluation of how the role of politics influences the adoption of public health policy is dependent on defining the problem and evidence, understanding the roles and players, understanding motivations, and understanding strategies and tactics. The authors use a number of case studies illustrating the interaction of evidence and politics including reproductive health rights for women in the United States. Chapter 2.3 then covers poverty, justice, and health, discussing the relationship between poverty and health and how there is often reference to human rights in justice theories that argue for the need to reduce or eliminate poverty to promote health. If we are to create a fairer society, public policy needs to be based on both equality of opportunity and equality of outcome. The following chapters discuss the determinants of socioeconomic inequalities in health in high-income and developing countries and what is known about how best to reduce them. Socioeconomic inequalities have been well described in high-income countries for a range of indicators of socioeconomic position and different health outcomes. We know much less about how to reduce health inequalities, and although a number of policies and interventions have been identified as having the potential to reduce inequalities, making progress in the reduction of health inequalities is dependent on international cooperation. In low- and middle-income countries, socioeconomic inequalities are often associated with high poverty levels but, despite the magnitude of such inequalities, there are few examples of the reduction of health inequalities becoming key targets of public policy. Agenda 2030 and the Sustainable Development Goals provide a useful multisectoral approach to address both health problems and their social determinants, and as such show a potential route to the reduction in health inequalities in low- and middle-income countries. Chapter 2.6 considers the advances that have been made in genomics and the opportunities and challenges that these present for public health. These include pharmacogenomics testing—

examining the effects that a person's genes have on their response to treatment by drugs—and the use of epigenetic therapy.

The following three chapters are devoted to long-standing public health concerns which remain central to the practice of public health: water and sanitation, food and nutrition, and climate change and human health. The lack of clean and safe drinking water and of basic sanitation and the diseases caused by these deficiencies are largely problems of low- and middle-income countries, and Chapter 2.7 details the major diseases and their aetiologies and considers interventions to prevent these. The following chapter provides an overview of the important role played by food systems and the need for cross-sectoral approaches to address undernutrition and overnutrition. This double burden of malnutrition has a common cause: a lack of access to affordable, quality sufficient foods, and examples are given from high-income and low- and middle-income countries of programmes attempting to address overnutrition and undernutrition, respectively. The main effects of climate change on health will be through the exacerbation of disease and injury from familiar causes. Since the populations that are most likely to be seriously affected by climate change are those that already face high burdens of disease it is important to improve baseline health in preparation. The overall response required will involve both coping with the health impacts of changes to weather patterns while halting emissions and considering ways to reduce existing pollutants in the atmosphere. The next chapter covers behaviours, the determinants of behaviours, and the many ways in which behaviours can influence health. Given that the environment can constrain or influence health-related behaviour, there is a clear need to develop interventions that target such environmental aspects. The final chapter in section two looks at the ways in which access to healthcare affects population health, covering issues including the roles of health policy, financing, and workforce planning.

The final section in this volume turns to public health policies, law, and ethics. The first chapter in this section looks at leadership in public health. Chapter 3.2 details how public health ethics originated, the issues that distinguish ethics for public health from other disciplines, and the ethical issues raised by genetic research. The following chapter considers the right to health and how this aligns with the goals of public health globally. The authors provide examples and case studies illustrating how human rights support global public health. Chapter 3.4, on law and the public's health,

considers the difficulty of balancing the promotion of health and protection of the population against individual liberties. Clearly there is a real need for global governance to combat disease and its vectors, and this means stepping beyond the health sector to encompass other areas affecting health including trade, food, migration, and the environment. The final two chapters of this volume cover health policy in developing and developed countries respectively. The focus of health policy varies between settings, but increased globalization leads to some commonalities between high-income countries and low- and middle-income countries. The reader is furnished with recommended strategies to strengthen health policy in developing countries along with examples of the failure of policy in developed countries.

In summary this volume details how broad the scope of public health has become as the twenty-first century develops. The contributors to this volume have been drawn from a range of disciplines and are all specialists in their fields. A common theme throughout is the need for the public health workforce to work collectively, across all disciplines, sectors, and countries, in the struggle to overcome emerging challenges and to improve global public health and reduce health inequalities.

AL

Preface to Volume 2: *The methods of public health*

Volume 2 of the seventh edition of the *Oxford Textbook of Public Health* presents the methods which bring scientific rigour to the public health endeavour. With a firm and broad grounding in the methods of public health, students and practitioners will ensure that their research and practice is based on robust evidence a critically important consideration in informing decision-making. The methodology utilized is a key for assuring the strength and validity of evidence for decision-making. Randomized controlled trials, systematic reviews, meta-analyses, and economic evaluations are accepted methods for providing clinical evidence but there are difficulties in applying these in a public health context where there are large target populations and long timescales for the emergence of outcomes. While evidence emanating from the application of these techniques has become more common to inform public health decision-making, there is increasing appreciation of the complexity of public health interventions and the need to use a range of disciplinary approaches to develop and evaluate these.

The range and complexity of health problems facing populations is also increasing, and public health methodologies must evolve and expand to meet this challenge and ensure effective responses. With globalization of emerging and re-emerging infections, and demographic, nutrition, and physical activity transitions, the methods of public health must be adapted to meet the new challenges. Since the sixth edition, the effects of globalization on health including the increasing importance of global environmental changes and notably climate change, have burgeoned. These new challenges are more difficult to study than the traditional concerns of infectious and non-infectious diseases and local environmental issues. The adaptation of old methods and the development of new ones are key to ensure the continued relevance and robustness of public health research. Public health practitioners have always had to face the dilemma of

balancing delays in action emanating from concern for methodological rigour with the need to act expeditiously. This dilemma is becoming increasingly complex and challenging as some of the new and emerging public health challenges are less amenable to traditional public health approaches and in some instances, are associated with ethical considerations that add to the complexity of the issue.

All the chapters in Volume 2 have been extensively revised and updated, and several new chapters added. This volume is organized into four sections: information systems and sources of intelligence; epidemiological and biostatistical approaches; social science techniques and environmental and occupational health sciences. Information systems are the foundation of all public health research and action. The lack of good information is still a barrier to effective action in much of the world. Basic information on births and deaths is not routinely available for most low-income and many middle-income countries. This gap remains a major impediment to tracking progress towards the Sustainable Development Goals. Fortunately, a concerted effort is now being made to close the gap, especially for maternal and child statistics with the support of several philanthropic foundations. Three chapters in this section (Chapters 4.1–4.3) examine the contrasting challenges facing information systems in both high-income and low- and middle-income countries. Chapter 4.2 by Zunyou Wu and Jennifer McGoogan captures the substantial advances being made in information systems and community diagnosis in low- and middle-income countries. Chapter 4.3 by Patrick Sullivan provides an insightful and contemporary overview of the impact of, and rapidly evolving new communication streams on public health.

Epidemiology and biostatistics are the core sciences of public health. Public health practice requires a firm connection to the priority health needs of populations. Epidemiological research is almost always required to establish this connection, and exceptions are few; some major acute outbreaks or overwhelming catastrophes do not allow for serious epidemiological investigation before the response is required. However, epidemiological study is still needed to assess the scope of the problem and the effectiveness of the response. Public health methods to ensure the appropriateness of inferences (Chapter 5.13) drawn from public health investigations—perhaps the most difficult and certainly the most contentious aspect of public health science—continue to evolve. This

process often requires a systematic approach to ensuring that all data from all studies—published and unpublished—are synthesized into a useable summary assessment (Chapter 5.14). A critical and expanding methodological area deals with interventions and their effectiveness at a community level by Kathy Baisley et al. (Chapter 5.8). Technological advances—computers and the internet—are changing the scope of public health and opening new possibilities, from data collection and analysis for the early identification of disease outbreaks to the use of modelling of disease transmission to predict the future trends and needs (Chapters 5.3 and 5.17). The methods and special issues facing clinical epidemiology are discussed in Chapter 5.10. The final chapter in this section stresses the essential importance of surveillance to monitor public health problems and the effectiveness of intervention programmes (Chapter 5.18). A new chapter in this section on qualitative research (Chapter 5.6) by Jennie Popay and Fran Baum covers both methodological approach and underscores the importance of understanding the voices of potential beneficiaries and Indigenous knowledge systems in the knowledge generation process.

Increasingly a more systems-based approach is being utilized to understand disease and risk. Rona Campbell provides an excellent update on new developments in the implementation of multicomponent interventions (Chapter 6.5). The completion of the human genome project is increasingly being translated to novel approaches to diseases including the use of gene therapy and leading to precision public health approaches. Of note are two new contemporary chapters in this section: Lifecourse epidemiology and analysis (Elizabeth Rose Mayeda) and Natural and quasi-experiments by Peter Craig (Chapter 5.10). The methodologies for measuring burden of disease Chapter 5.17 are particularly important as the implementation of these new methodologies is being used to provide the basis for both global and local rationing of resources and priority setting.

Social science techniques are assuming even greater importance to the practice of public health with the recognition that epidemiological information alone is not sufficient for the development and implementation of effective public health policies and programmes (Chapter 6.1). Demography is another underappreciated basic science of public health; the ageing of all populations, especially in low- and middle-income countries, will be a critical public health issue for the economic survival of these countries in the twenty-first century (Chapter 6.3). Health economics and

the use of cost-effectiveness analysis (Chapter 6.6) have expanded the audience for public health research to sectors outside health, especially the finance and development sectors, nationally and globally. Health promotion expands the focus of public health from a primary concern with disease prevention and control towards an understanding of the underlying determinants of health (Chapter 6.4). These and other social science tools, including management of the health programmes (Chapter 6.8), are key for the development and implementation of effective public health policy in all countries eloquently and comprehensively covered by Wafaa el Sadr and colleagues in a new chapter on Implementation Science and translational public health (Chapter 6.9). The HIV pandemic has brought renewed interest and focus on sexuality and health that Parker et al. cover eloquently in Chapter 6.2. A novel and increasingly used approach to programme and treatment adherence is incentivizing and supporting desired behaviours which is captured in the chapter on Behavioural Economics and Health (Chapter 6.7) by Harsha Thirumurthy et al.

Environmental and occupational health sciences cover traditional public health issues, as well as the even more difficult global health challenges that are discussed in Volume 1. This section deals with both traditional and emerging environmental and occupational health hazards, many of which have been exacerbated by globalization. (Chapters 7.1–7.3). The increasingly important issues of risk assessment and management, and risk perception and communication are covered in Chapters 7.4 and 7.5. The chapter on urbanization and health by Jason Coburn highlights how increasing urbanization is impacting public health.

The importance of methodological advances in public health is illustrated by the way in which many chapters in other volumes of this Textbook consider methodological issues in considerable detail—for example, the chapters on measuring the global burden of diseases and responding to global environmental challenges. The chapters in this section illustrate the evolution and breadth of public health methods as its scope continues to expand. No doubt, this process will continue well into the future and remains a good marker of the growth and evolution of the public health sciences as sound and rigorous strategies to address emerging and ongoing public health challenges.

QAK

Preface to Volume 3: *The practice of public health*

> *Public health is what we, as a society, do collectively to assure the conditions for people to be healthy. This requires that continuing and emerging threats to the health of the public be successfully countered … through effective, organised and sustained efforts led by the public sector. Institute of Medicine (1988)*[1]

The practice of public health has never been more vital. The social, economic, political, and environment threats to human health are accumulating in the twenty-first century. Volume 3 of the *Oxford Textbook of Global Public Health* provides an up-to-date account of how the practice of public health can contribute to reducing these threats through 49 chapters, each written by leading experts. The contents of this volume identify public health as a subject of local, national, and global concern, and a key responsibility of governments and public agencies, with the need to regular private industry in order to promote the health of the public more pressing. The chapters in this volume concern the practice of public health. They demonstrate the broad focus of public health and how more than ever public health is a global concern and requires effective global, national, regional, and local governance to ensure that life expectancy continues to increase and the reversals in life expectancy seen in some countries is reversed. This volume focuses on the application of population science methods to the major challenges being addressed through public health interventions and approaches. The chapters in this edition provide the evidence for public health's strategies, reassess priorities when required, recognize the evolving burden of disease and context in which they occur, and recommend new strategies for intervention.

Volume 3 begins with a discussion of the major groups of non-communicable disorders, including cardiovascular and respiratory diseases and cancer (Section 8). This section includes separate chapters on obesity, physical inactivity, and diabetes, reflecting growing concern at the increasing prevalence and growing impact of these conditions on global health. The chapter on obesity contains a crucial discussion on the

commercial determinants of health showing very clearly how marketing and free trade have made unhealthy high fat and sugar foods and beverages more accessible. Section 8 continues with a series of chapters on communicable diseases, with chapters on tuberculosis, malaria, hepatitis, and emerging and re-emerging infections and the threat of bioterrorism. Together these conditions represent major priorities: tuberculosis has been the cause of death for billion people since the 1900s and there were 6.3 million newly notified cases of tuberculosis (TB) worldwide in 2016; malaria accounts for nearly half a million death annually and the earlier success in reducing this number has stalled since 2015; while 328 million people are chronically infected with hepatitis viruses B and C, with one million deaths annually. As well as outlining problems and their causes, each of these chapters discusses potential solutions, including national and international strategies for disease control and prevention. Without significant process in each disease area the Sustainable Development Goals for 2030 set by the United Nations will not be achievable.

The emphasis on intervention for prevention of disease and promotion of health is continued in the volume's second section (Section 9) on public health hazards including tobacco, alcohol, drug abuse, injuries and violence Proposed intervention strategies encompass both population-based approaches; including, for example, regulation, the use of deterrents and incentives, and public education; and strategies targeted at individuals at high risk through healthcare services. It remains clear, however, that the greatest risks are generally found among those groups for whom interventions and environments conducive to good health are least accessible.

Section 10 considers the public health needs of different population groups paying particular attention to groups that for a range of reasons are vulnerable to public health hazards, disease risks, and increasingly emphasized in this revision is the inequitable impact of climate change on health status. Separate chapters outline the needs of families, women, children, adolescents, older people, ethnic minorities, people with disabilities, and forced migrants and displaced populations. While significant progress has been made in preventing child deaths, a staggering 5.4 million under 5-year-olds died in 2017. The largest ever cohort of adolescents now exists and their health and well-being is increasingly shaped by the global forces of urbanization, migration, global

communications, economic development, and the sustainability of planetary ecosystems. The chapter on people with disabilities notes that 15% of the world's population has a disability and that public health must frame responses within a social determinants framework to tackle underlying issues such as poverty and unemployment. The chapters on Indigenous peoples and prisons and public health similarly stress the importance of social determinants and demonstrate that the impacts of European colonization reverberate across the centuries and have causes multiple health issues for Indigenous populations worldwide. Chapter 10.7 notes that as the world faces the health issues of globalization and climate change, Indigenous knowledges and ways of being may offer important insights into managing these challenges. Collectively, these chapters emphasize the importance of the public health role in analysing the health needs of these often marginalized populations and advocating for their right to health and the living conditions that create health. Together the chapters provide clear directions on how to reduce health inequities within and between countries.

Section 11 presents an analysis of the core public health skills required for improving population health and reducing inequalities in health. This section begins with a chapter on the concept of need, which shows that technical assessments of the capacity to benefit from health intervention can rarely be separated from underlying assumptions concerning the justification and rationale for societal intervention. Action between and across sectors is usually vital and this is exemplified by chapters on current strategies for control of non-communicable diseases and infectious diseases. The chapter on the political economy of non-communicable disease also demonstrates the political nature of public health and the importance for public health practitioners to understand the implications of this for their practice. A series of chapters then outlines opportunities for intervention through the health sector including healthcare services, population screening, and environmental health practice. The section on the public health workforce includes chapters on training of community health workers, and public health professionals, to meet public health needs. There are also chapters on planning and responding to public health emergencies, including chemical and radiological emergencies.

In the final section of the book, Quarraisha Abdool Karim and Roger Detels question the assumption that intervention on public health problems must be led by the public sector. They describe an enhanced role for

powerful private sector advocates of public health intervention, while at the same time drawing attention to some of the challenges of this approach and the importance of the stewardship and governance role of governments. The closing chapter, by Tedros Adhanom Ghebreyesus (Director General of the World Health Organization), comments on several key issues that have been raised in earlier chapters of the book, analysing the pressing challenges now facing the public health community globally, and emphasizing that public health is a powerful driver for positive change to dramatically improve the health and lives of all people. In particular he notes how air pollution has become an even more crucial issue and requires a cross-sectoral response. He notes that the World Health Organization (WHO) expects approximately 250,000 additional deaths per year between 2030 and 2050 to be attributed to the health effects of climate change.

This volume shows the dynamic nature of the changing world in which we live and how old and new threats to health and well-being interact to challenge public health to respond creatively.

FB

---

[1] Institute of Medicine (1988). *The Future of Public Health.* National Academies Press, Washington DC. Page 19.

Brief Contents

Volume 1 *The scope of public health*

SECTION 1
The development of the discipline of public health

SECTION 2
Determinants of health and disease

SECTION 3
Public health policies, law, and ethics

Volume 2 *The methods of public health*

SECTION 4
Information systems and sources of intelligence

SECTION 5
Epidemiological and biostatistical approaches

SECTION 6
Social science techniques

SECTION 7
Environmental and occupational health sciences

Volume 3 *The practice of public health*

SECTION 8
Major health problems

SECTION 9
Prevention and control of public health hazards

SECTION 10
Public health needs of population groups

SECTION 11
Public health functions

Contents

Volume 1 The scope of public health

Abbreviations
Contributors

SECTION 1 The development of the discipline of public health

1.1 The scope and concerns of public health
Roger Detels and Chorh Chuan Tan

1.2 The history and development of public health in developed countries
Simon Szreter

1.3 The history and development of public health in low- and middle-income countries
Than Sein

1.4 Public health priorities in countries undergoing economic transition: the middle-income countries
Stephen Tollman and Jessica Price

1.5 Globalization
Kelley Lee

SECTION 2 Determinants of health and disease

2.1 Determinants of health: overview
Matthew Fisher, Belinda Townsend, Patrick Harris, Ashley Schram, and Fran Baum

2.2 Politics of public health
Kaitlyn B. McBride and Linda Rosenstock

2.3 Poverty, justice, and health
Ronald Labonté, Fran Baum, and David Sanders

2.4 Socioeconomic inequalities in health in high-income countries: the facts and the options
Frank J. van Lenthe and Johan P. Mackenbach

2.5 Reducing health inequalities in developing countries
Romulo Paes-Sousa, Paulo M. Buss, and Mauricio L. Barreto

2.6 Genomics and public health
Veron Ramsuran and Tulio de Oliveira

2.7 Water and sanitation
Thomas Clasen

2.8 Food and nutrition
Roger Shrimpton, David Sanders, and Anne Marie Thow

2.9 Climate change and human health
Alistair Woodward and Alex Macmillan

2.10 Behavioural determinants of health and disease
Lawrence W. Green, Kristin S. Hoeft, and Robert A. Hiatt

2.11 How access to healthcare affects population health
Yvonne Inall, Rachel Lamdin Hunter, Stephen Leeder, and Angela Beaton

SECTION 3 Public health policies, law, and ethics

3.1 Leadership in public health
Kevin A. Fenton

3.2 Ethical principles and ethical issues in public health
Nancy Kass, Amy Paul, and Andrew Siegel

3.3 The right to health supports global public health
Carmel Williams, Alison Blaiklock, and Paul Hunt

3.4 Law and the public's health
Lawrence O. Gostin

3.5 Health policy in developing countries
Thein T. Htay, Yu Mon Saw, James Levinson, S.M. Kadri, Ailbhe Helen Brady, Cecilia S. Acquin, and Aung Soe Htet

3.6 Public health policy in developed countries
John Powles † and Hebe Gouda

Index

Volume 2 The methods of public health

Abbreviations
Contributors

SECTION 4 Information systems and sources of intelligence

4.1 Information systems in support of public health in high-income countries
Tjeerd-Pieter van Staa and Liam Smeeth

4.2 Community diagnosis and health information systems in low- and middle-income countries
Zunyou Wu and Jennifer McGoogan

4.3 New communication technologies, social media, and public health
Patrick S. Sullivan, Aaron J. Siegler, and Lisa Hightow-Weidman

SECTION 5 Epidemiological and biostatistical approaches

5.1 Epidemiology: the foundation of public health
Roger Detels

5.2 Cross-sectional studies
Manolis Kogevinas and Leda Chatzi

5.3 Principles of outbreak investigation
Sopon Iamsirithaworn, Panithee Thammawijaya, and Kumnuan Ungchusak

5.4 Case–control studies
Noel S. Weiss

5.5 Cohort studies
Alvaro Muñoz and F. Javier Nieto

5.6 Qualitative research imagination
Jennie Popay and Fran Baum

5.7 Methodological issues in the design and analysis of cluster randomized trials

Kathy J. Baisley, Richard J. Hayes, and Lawrence H. Moulton

5.8 Community intervention trials in high-income countries
John W. Farquhar and Lawrence W. Green

5.9 Natural and quasi-experiments
Peter Craig

5.10 Clinical epidemiology
Fiona F. Stanaway, Naomi Noguchi, Clement Loy, Sharon Reid, and Jonathan C. Craig

5.11 Validity and bias in epidemiological research
Sander Greenland and Tyler J. VanderWeele

5.12 Causation and causal inference
Katherine J. Hoggatt, Tyler J. VanderWeele, and Sander Greenland

5.13 Systematic reviews and meta-analysis
Nandi Siegfried and Lawrence Mbuagbaw

5.14 Statistical methods
Gail Williams and Robert S. Ware

5.15 Measuring the health of populations: the Global Burden of Disease study methods
Theo Vos, Christopher J.L. Murray and Alan D. Lopez

5.16 Mathematical models of transmission and control of infectious agents
Alex Welte and Cari van Schalkwyk

5.17 Public health surveillance
Nguyen Tran Hien, James W. Buehler, and Ann Marie Kimball

5.18 Life course epidemiology and analysis
Elizabeth Rose Mayeda, Alexandra M. Binder, and Lindsay C. Kobayashi

SECTION 6 Social science techniques

6.1 Sociology and psychology in public health
Stella R. Quah

6.2 Sexuality and public health
Richard Parker, Jonathan Garcia, Miguel Muñoz-Laboy, Marni Sommer, and Patrick Wilson

6.3 Demography and public health
Emily Grundy and Michael Murphy

6.4 Health promotion, health education, and the public's health
Fran Baum

6.5 Development and evaluation of complex multicomponent interventions in public health
Rona Campbell and Chris Bonell

6.6 Economic appraisal in public healthcare: assessing efficiency and equity
David Parkin, Stephen Morris, and Nancy Devlin

6.7 Behavioural economics and health
Alison Buttenheim and Harsha Thirumurthy

6.8 Governance and management of public health programmes
Zhiyuan Hou and Na He

6.9 Implementation science and translational public health
Wafaa M. El-Sadr, Judith Wasserheit, Bryan Wiener, Andrea Howard, Catherine Hankins, Patricia J. Culligan, and Katherine Harripersaud

SECTION 7 Environmental and occupational health sciences

7.1 Environmental health methods
Chien-Jen Chen and San-Lin You

7.2 Radiation and public health
Leeka Kheifets, Adele Green, and Richard Wakeford

7.3 Occupational health
David Koh and Wee Hoe Gan

7.4 Toxicology and environmental risk analysis
David Koh and Ro-Ting Lin

7.5 Risk perception and communication
Baruch Fischhoff and Tamar Krishnamurti

7.6 Urbanization and health
Jason Corburn

Index

Volume 3 The practice of public health

Abbreviations
Contributors

SECTION 8 Major health problems

8.1 Epidemiology and prevention of cardiovascular disease
Nathan D. Wong and Wenjun Fan

8.2 Cancer epidemiology and public health
Paolo Boffetta, Zuo-Feng Zhang, and Carlo La Vecchia

8.3 Chronic obstructive pulmonary disease and asthma
Craig M. Riley, Jessica Bon, and Alison Morris

8.4 Obesity
Anna Peeters and Tim Lobstein

8.5 Physical activity and public health
Nyssa T. Hadgraft, Neville Owen, and Paddy C. Dempsey

8.6 Diabetes mellitus
Farah Naz Khan, Nida Izhar Shaikh, K.M. Venkat Narayan, and Mohammed K. Ali

8.7 Public mental health and suicide
Danuta Wasserman and Kristian Wahlbeck

8.8 Dental public health
Amira S. Mohamed and Peter G. Robinson

8.9 Musculoskeletal disorders
Lope H. Barrero and Alberto J. Caban-Martinez

8.10 Neurological diseases, epidemiology, and public health
Walter A. Kukull, Kumeren Govender, and James Bowen

8.11 Infectious diseases and prions
Davidson H. Hamer, Amira Khan, and Zulfiqar A. Bhutta

8.12 Sexually transmitted infections
Noah Kojima and J.D. Klausner

8.13 Acquired immunodeficiency syndrome (AIDS)

Quarraisha Abdool Karim, Urisha Singh, Cheryl Baxter, and Salim S. Abdool Karim

8.14 Tuberculosis
Roxana Rustomjee

8.15 Malaria
Frank Baiden, Keziah L. Malm, and Fred Binka

8.16 Viral alcoholic and fatty liver diseases
Ehud Zigmond and Daniel Shouval

8.17 Emerging and re-emerging infections
David L. Heymann and Vernon J.M. Lee

8.18 Bioterrorism
Peter Katona

8.19 Genetic epidemiology
Elizabeth H. Young and Manjinder S. Sandhu

SECTION 9 Prevention and control of public health hazards

9.1 Tobacco
Tai Hing Lam and Sai Yin Ho

9.2 Substance use and misuse: considerations on global public health
Giang Le Minh and Steve Shoptaw

9.3 Alcohol
Robin Room

9.4 Injury prevention and control: the public health approach
Corinne Peek-Asa and Adnan A. Hyder

9.5 Interpersonal violence
Rachel Jewkes

9.6 Collective violence: war
Barry S. Levy

SECTION 10 Public health needs of population groups

10.1 The changing family
Ann Evans and Gavin W. Jones

10.2 Women, men, and health

Diane Cooper and Hanani Tabana

10.3 Child health
Tyler Vaivada, Amira Khan, Omar Irfan, and Zulfiqar A. Bhutta

10.4 Adolescent health
George Patton, Peter Azzopardi, Natasha Kaoma, Farnaz Sabet, and Susan Sawyer

10.5 Intersectional and social epidemiology approaches to understanding the influence of race, ethnicity, and caste on global public health
Jennifer Beard, Nafisa Halim, Salma M. Abdalla, and Sandro Galea

10.6 The health of Indigenous peoples
Papaarangi Reid, Donna Cormack, Sarah-Jane Paine, Rhys Jones, Elana Curtis, and Matire Harwood

10.7 People with disabilities
Anne Kavanagh, Marissa Shields, and Alex Devine

10.8 Health of older people
Samir K. Sinha and Brittany Ellis

10.9 Forced migrants and other displaced populations
Catherine R. Bateman Steel and Anthony B. Zwi

10.10 Prisoners: a wicked problem for public health
Tony G. Butler and Peter W. Schofield

SECTION 11 Public health functions

11.1 Health needs assessment
Michael P. Kelly, Jane E. Powell, and Natalie Bartle

11.2 The political economy of non-communicable diseases: lessons for prevention
Anne Marie Thow, Raphael Lencucha, and K. Srinath Reddy

11.3 Immunization and vaccination
Eleonora A.M.L. Mutsaerts and Shabir A. Madhi

11.4 Principles of infectious disease control
Robert J. Kim-Farley

11.5 Medical screening: theories, methods, and effectiveness
Tang Jin-ling and Li Li-ming

11.6 The practice of environmental health in an era of sustainable development
Yasmin E.R. von Schirnding and Lynn R. Goldman

11.7 Strategies and structures for public health interventions
Sian Griffiths and Kevin A. Fenton

11.8 Strategies for health services
Chien Earn Lee and Fran Baum

11.9 Training of public health professionals in developing countries
San Hone and Roger Detels

11.10 Transformative learning for health professionals in the twenty-first century for the future health workforce
Wanicha Chuenkongkaew and Suwit Wibulpolprasert

11.11 Humanitarian emergencies
Craig Spencer and Les Roberts

11.12 Principles of public health emergency response for acute environmental, chemical, and radiation incidents
Virginia Murray, Thomas Waite, and Paul Sutton

11.13 Private support of public health
Quarraisha Abdool Karim and Roger Detels

11.14 Global health in the era of sustainable development
Fiona Fleck

Index

Abbreviations

AAAQ	availability, accessibility, acceptability, quality
ABI	ankle brachial index
ACA	Affordable Care Act
ACASI	audio-computer assisted self-interviewing
ACCORD	Action to Control Cardiovascular Risk in Diabetes
ACF	active case finding
ACGIH	American Conference of Governmental Industrial Hygienists
ACHPR	African Commission on Human and Peoples' Rights
ACLED	Armed Conflict Location and Event Data
ACT	artemisinin-based combination therapy
ACTG	AIDS Clinical Trials Group
AD	Alzheimer's disease
ADA	American Diabetes Association
ADAPT	Alzheimer's disease anti-inflammatory prevention trial
AED	antiepileptic drug
AF	attributable fraction
AFRINEST	African Neonatal Sepsis Trial
AGVP	African Genome Variation Project
AHA	American Heart Association
AHRQ	Agency for Healthcare Research and Quality
AI	artificial intelligence
AIDS	acquired immunodeficiency syndrome
AIIR	airborne infection isolation room
ALL	acute lymphoblastic leukaemia
ALSPAC	Avon Longitudinal Study of Parents and Children
AMD	advanced molecular detection
AMI	acute myocardial infarction
AMR	antimicrobial resistance
ANISA	Aetiology of Neonatal Infection in South Asia
ANOVA	analysis of variance
ANUG	acute necrotizing ulcerative gingivitis
APHA	American Public Health Association
ARA	American Relief Administration
ARDS	acute respiratory distress syndrome
ARI	acute respiratory infections
ARIC	Atherosclerosis Risk in Communities

ART	antiretroviral therapy
ART	atraumatic restorative treatment
ASD	autism spectrum disorder
ASEAN	Association of Southeast Asian Nations
ASFR	age-specific fertility rate
ASH	Action on Smoking and Health
ASMR	age-specific mortality rate
ATS	amphetamine-type stimulants
ATSDR	Agency for Toxic Substances and Disease Registry
BACH	Boston Area Community Health
BAL	British anti-Lewisite
BEI	Biological Exposure Index
BFP	Bolsa Familia Programme
BMGF	Bill & Melinda Gates Foundation
BMI	body mass index
BMT	buprenorphine maintenance therapy
BNP	brain naturietic peptide
BOD	Burden of Disease
BRFSS	Behavioural Risk Factor Surveillance System
BSE	bovine spongiform encephalopathy
BV	bacterial vaginosis
CABG	coronary artery bypass grafting
CAD	coronary artery disease
CBA	cost-benefit analysis
CBF	cerebral blood flow
CBR	community-based rehabilitation
CBR	cost-benefit ratio
CBS	community-based surveillance
CBT	cognitive behavioural therapy
CCA	cost-consequences analysis
CCM	community case management
CCT	conditional cash transfer
CDC	Centres for Disease Control
CDCP	Centers for Disease Control and Prevention
CDSR	Cochrane Database of Systematic Reviews
CEA	cost-effectiveness analysis
CEAC	cost-effectiveness acceptability curve
CEDAW	Committee on the Elimination of Discrimination Against Women
CER	cost-effectiveness ratio
CETP	cholesterol ester transferase protein
CFIR	Consolidated Framework for Implementation Research
CHD	coronary heart disease
CHEW	Checklist of Health Promotion Environments at Worksites
CHF	congestive heart failure
CHNA	Community Health Needs Assessments

CHP	Centre for Health Protection
CHS	Cardiovascular Health Study
CHW	community health workers
CI	confidence interval
CIDARS	China Infectious Disease Automated-alert and Response System
CIHR	Canadian Institutes of Health Research
CIMT	carotid intima medial thickness
CIOMS	Council for International Organizations of Medical Sciences
CLTS	community-led total sanitation
CM	contingency management
CMA	cost-minimization analysis
CMO	context-mechanism-outcome
COMEST	Commission on the Ethics of Scientific Knowledge and Technology
CONSORT	Consolidation of Standards for Reporting of Trials
COP	Conference of the Parties
COPD	chronic obstructive pulmonary disease
CORTIS	Correlate of Risk Targeted Intervention Study
CPAP	continuous positive airway pressure
CPE	carbapenemase-producing Enterobacteriaceae
CPG	clinical practice guidelines
CQG	cost per QALY gained
CRC	Convention on the Rights of the Child
CRE	carbapenem-resistant Enterobacteriaceae
CRE	Centre of Research Excellence
CRED	Centre for Research on the Epidemiology of Disasters
CRFA	common risk factor approach
CRIMS	Comprehensive Response Information Management System
CRISPR	clustered regularly interspaced short palindromic repeats
CRM	cross-reacting material
CROI	Conference on Retroviruses and Opportunistic Infections
CRS	congenital rubella syndrome
CRT	cluster randomized trials
CSD	community socioeconomic deprivation
CSDH	Commission for the Social Determinants of Health
CT	computed tomography
CTC	Communities that Care
CTE	chronic traumatic encephalopathy
CTS	carpal tunnel syndrome
CUA	cost-utility analysis
CVD	cardiovascular disease
DACA	Deferred Action on Childhood Arrivals
DAG	directed acyclic graphs
DAH	development assistance for health
DALE	disability-adjusted life expectancy
DALY	disability-adjusted life year

DASH	Dietary Approaches to Stop Hypertension
DBM	double burden of malnutrition
DCEA	distributional cost-effectiveness analysis
DDD	digital disease detection
DDD	Doing Development Differently
DDT	dichloro-diphenyl-trichloroethane
DHF	dengue haemorrhagic fever
DHS	Demographic and Health Survey
DiD	difference in differences
DLB	dementia with Lewy bodies
DMD	Duchenne muscular dystrophy
DOTS	directly observed treatment short
DR	drug-resistant
DRIP	Declaration on the Rights of Indigenous Peoples
DSA	demographic surveillance area
DSS	dengue shock syndrome
DTA	Declaration of Territorial Asylum
DTA	diagnostic test accuracy
DTP	diphtheria-tetanus-pertussis
EA	economic appraisal
EAE	experimental autoimmune encephalomyelitis
EASD	European Association for the Study of Diabetes
EBM	evidence-based medicine
ECDC	European Centre for Disease Prevention and Control
ECEA	extended cost-effectiveness analysis
ECHA	European Chemicals Agency
ECRHS	European Community Respiratory Health Survey
ED	erectile dysfunction
EF	error factor
EGAPP	Evaluation of Genomic Applications in Practice and Prevention
EHR	electronic health record
EIU	Economist Intelligence Unit
ELISA	enzyme linked immunosorbent assay
EMR	electronic medical records
EMRO	Eastern Mediterranean Regional Office
ENOC	essential obstetric and newborn care
EOS	early onset sepsis
EPA	European Psychiatric Association
EPHF	essential public health functions
EPI	Expanded Programme on Immunization
EPIET	European Programme for Intervention Epidemiology Training
EPOC	effective practice and organization of care
EPODE	Ensemble Prévenons l’Obésité Des Enfants
EPPI	Evidence for Practice and Policy Information
ERC	Emergency Risk Communication

ERM	emergency risk management
ERS	Economic Research Service
EU	European Union
EVD	Ebola virus disease
EWAS	epigenome-wide association studies
FAO	Food and Agriculture Organization
FAP	Food Acquisition Programme
FCA	Framework Convention Alliance
FCAC	Framework Convention on Alcohol Control
FDA	Food and Drug Administration
FDC	fixed-dose combination
FELTP	Field Epidemiology and Laboratory Training Programme
FETP	Field Epidemiology Training Programme
FH	familial hypercholesterolemia
FPG	fasting plasma glucose
FRR	familial relative risk
FVC	forced vital capacity
GA	General Assembly
GAM	generalized additive regression models
GAPPD	Global Action Plan for Pneumonia and Diarrhoea
GAR	Global Alert and Response
GATS	Global Adult Tobacco Survey
GATT	General Agreement on Tariffs and Trade
GAVI	Global Alliance for Vaccines and Immunization
GBD	Global Burden of Disease
GBS	Group B streptococcus
GBV	gender-based violence
GCM	global coordination mechanism
GCS	Glasgow Coma Scale
GDP	gross domestic product
GDPR	General Data Protection Regulation
GEE	generalized estimating equations
GETT	Genetic Testing Evidence Tracking Tool
GHO	Global Health Observatory
GHPSS	Global Health Professions Student Survey
GI	GINI Index
GIS	geographic information systems
GIV	generic inverse variance
GLASS	Global Antimicrobial Resistance Surveillance System
GM	genetically modified
GMF	global monitoring framework
GNP	gross national product
GOARN	Global Outbreak Alert and Response Network
GP	General Practitioners
GPHIN	Global Public Health Intelligence Network

GPP	Good Participatory Practice
GPS	Global Positioning System
GRADE	Grading of Recommendations Assessment, Development, and Evaluation
GRID	gay-related immune deficiency
GRR	gross reproduction rate
GSPS	Global School Personnel Survey
GST	glutathione S-transferase
GTR	Genetic Testing Registry
GTSS	Global Tobacco Surveillance System
GVAP	Global Vaccine Action Plan
GWAS	genome-wide association study
GYTS	Global Youth Tobacco Survey
HAART	highly active antiretroviral therapy
HAC	Humanitarian Action for Children
HAT	Human African trypanosomiasis
HAV	hepatitis A virus
HBOC	Hereditary Breast and Ovarian Cancer
HBSC	Health Behaviour in School-aged Children
HBV	hepatitis B virus
HCQI	Health Care Quality Indicator
HCV	hepatitis C virus
HCW	healthcare workers
HDI	Human Development Index
HDL	high-density lipoprotein
HDT	host-directed therapies
HIA	health impact assessment
HiAP	Health in All Policies
HibCV	Hib polysaccharide-protein conjugate vaccine
HIC	high-income countries
HICPAC	Hospital Infection Control Practices Advisory Committee
HIS	health information systems
HITECH	Health Information Technology for Economic and Clinical Health
HIV	human immunodeficiency virus
HLA	HLA allele
HMN	Health Metrics Network
HPS	hantavirus pulmonary syndrome
HPTN	HIV Prevention Trial Network
HPV	human papilloma virus
HRDAG	Human Rights Data Analysis Group
HRQOL	health-related quality of life
HSE	Health and Safety Executive
HTA	health technology assessment
HUS	Haemolytic uraemic syndrome
HWE	healthy worker effect
HWTS	household water treatment and safe storage

IACHR	Inter-American Commission on Human Rights
IACtHR	Inter-American Court of Human Rights
IARC	International Agency for Research on Cancer
IASC	Inter-Agency Standing Committee
IASFM	International Association for the Study of Forced Migration
IAVI	International AIDS Vaccine Initiative
ICC	International Criminal Court
ICC	intracluster correlation coefficient
ICCPR	International Covenant on Civil and Political Rights
ICD	International Classification of Diseases
ICESCR	International Covenant on Economic, Social and Cultural Rights
ICHD	International Classification of Headache Disorders
ICRC	International Committee of the Red Cross
ICT	information and communication technologies
IDF	International Diabetes Federation
IDM	intensified disease management
IDMC	Internal Displacement Monitoring Centre
IDP	internally displaced persons
IDS	integrated diseases surveillance
IDU	injection drug use
IFG	impaired fasting glucose
IFRC	International Federation of Red Cross
IGRA	Interferon-gamma release assays
IGT	impaired glucose tolerance
IHD	ischaemic heart diseases
IHME	Institute for Health Metrics and Evaluation
IHR	International Health Regulations
ILAE	International League Against Epilepsy
ILI	influenza-like illness
ILO	International Labour Organization
IMC	International Medical Corps
IMF	International Monetary Fund
IMMANA	Innovative Methods and Metrics for Agriculture and Nutrition Actions
IMNCI	integrated management of neonatal and childhood illness
IMR	infant mortality rates
INCLEN	International Clinical Epidemiology Network
IOF	Implementation Outcomes Framework
IOM	Institute of Medicine
IPAQ	International Physical Activity Questionnaire
IPCW	inverse probability of censoring weighted
IPD	invasive pneumococcal disease
IPEC	International Programme on the Elimination of Child
IPT	intermittent preventive treatment
IPT	isoniazid preventive therapy
IPV	inactivated polio vaccine

IPV	intimate partner violence
IRB	institutional review boards
IRS	indoor residual spraying
ISAAC	International Study of Asthma and Allergies in Childhood
ITC	International Tobacco Control
ITS	interrupted time series
ITU	International Telecommunication Union
JE	Japanese encephalitis
JEE	joint external evaluation
JEV	Japanese encephalitis virus
JSNA	Joint Strategic Needs Assessment
KCR	Kosova Cancer Registry
KPI	key performance indicator
LASI	Longitudinal Study of Aging in India
LAV	lymphadenopathy-associated virus
LBP	low back pain
LBW	low birth weight
LDL	low-density lipoprotein
LGBT	lesbian, gay, bisexual, and transgender
LGBTI	lesbian, gay, bisexual, transgender, and intersex
LIC	low-income countries
LMIC	low and middle-income countries
LOS	late onset sepsis
LRI	lower respiratory infections
MACS	Multicenter AIDS Cohort Study
MAI	Multilateral Agreement on Investment
MAM	moderate acute malnutrition
MAPT	microtubule-associated protein tau
MAT	medication-assisted treatment
MAV	Municipal Association of Victoria
MCAD	medium chain acyl CoA dehydrogenase
MCI	mild cognitive impairment
MCS	Millennium Cohort Study
MCU	maternal and child undernutrition
MDD	major depressive disorder
MDG	Millennium Development Goals
MDS	Model Disability Survey
MENA	Middle East and Northern Africa
MERS	Middle East respiratory syndrome
MET	metabolic equivalents of task
MI	motivational interviewing
MI	myocardial infarction
MIGS	Mediterranean Institute of Gender Studies
MIHL	minimum income for healthy living
MkV	MEMA kwa Vijana

MMA	methylated into monomethylarsonic acid
MMR	maternal mortality ratio
MMR	measles, mumps, and rubella
MMT	methadone maintenance therapy
MMT	methadone maintenance treatment
MMV	Medicines for Malaria Venture
MMWR	Morbidity and Mortality Weekly Report
MNS	mental, neurological, and substance
MOH	Ministry of Health
MOST	multiphase optimization strategy implementation trial
MOUD	medications for opioid use disorder
MPI	multidimensional poverty index
MRI	magnetic resonance imaging
MRSA	methicillin-resistant Staphylococcus aureus
MS	multiple sclerosis
MSA	multiple system atrophy
MSF	Médecins Sans Frontières
MSI	Management Systems International
MSM	men who have sex with men
MSUD	maple syrup urine disease
MTBE	methyl tert-butyl ether
MTCT	mother-to-child transmission
MTO	Moving to Opportunity
MUAC	mid-upper-arm circumference
MUSP	Mater University of Queensland Study of Pregnancy
MVPA	moderate-to-vigorous intensity physical activity
NAAT	Nucleic acid amplification tests
NAFTA	North American Free Trade Agreement
NAT	nucleic acid-based testing
NBER	National Bureau of Economic Research
NCBI	National Center for Biotechnology Information
NCD	non-communicable disease
NCEP	National Cholesterol Education Program
NCSEM	National Centre for Sports and Exercise Medicine
NDIS	National Disability Insurance Scheme
NDNS	National Diet and Nutrition Survey
NEP	needle-exchange programmes
NERC	National Ebola Response Committee
NFI	non-food item
NGO	non-government organization
NGS	Next-generation whole-genome sequencing
NHANES	National Health and Nutrition Examination Survey
NHS	National Health Service
NHSP	National Healthy Schools Programme
NIH	National Institutes of Health

NIHL	noise-induced hearing loss
NIOSH	National Institute for Occupational Safety and Health
NIPT	non-invasive prenatal testing
NK	natural killer
NMA	network meta-analysis
NMSC	non-melanoma skin cancer
NNRTI	non-nucleoside RT inhibitor
NOS	Newcastle-Ottawa Scale
NP	neck pain
NPS	new psychoactive substances
NPV	negative predictive value
NRR	net reproduction rate
NRT	nicotine replacement therapy
NRTI	nucleoside reverse transcriptase inhibitor
NSCH	National Survey of Children's Health
NSP	needle and syringe programmes
NTD	neglected tropical diseases
NZHTA	New Zealand Health Technology Assessment
OAS	Organization of American States
OAU	Organization of African Unity
OCHA	Office for Coordination of Humanitarian Affairs
ODA	official development assistance
ODF	open defecation-free
OECD	Organisation for Economic Cooperation and Development
OGTT	oral glucose tolerance test
OHCHR	Office of the High Commissioner for Human Rights
OHS	occupational health and safety
OIHP	Office International d'Hygiène Publique
OOP	out-of-pocket
OPV	oral polio vaccine
OR	odds ratio
ORS	oral rehydration salts
ORS	oral rehydration solution
ORT	oral rehydration therapy
OSD	occupational skin diseases
OSHA	Occupational Safety and Health Administration
OST	opioid substitution therapy
OTI	Office of Transition Initiatives
P&I	pneumonia and influenza
PACES	Prevent Anxiety in Children through Education in Schools
PACK	Practical Approach to Care Kit
PAF	population attributable fraction
PAHO	Pan American Health Organization
PATH	Programme for Appropriate Technologies in Health
PCA	principal components analysis

PCI	percutaneous coronary interventions
PCP	pneumocystis carinii pneumonia
PCR	polymerase chain reaction
PCT	preventive chemotherapy and transmission
PCV	pneumococcal conjugate vaccine
PCV	protein conjugate vaccine
PCV	protein-conjugated polysaccharide vaccines
PDNA	post-disaster needs assessments
PEPFAR	President’s Emergency Plan for AIDS Relief
PF	prevented fraction
PHAC	Public Health Agency of Canada
PHCA	primary oral healthcare approach
PHE	Public Health England
PHEIC	Public Health Emergency of International Concern
PHM	People’s Health Movement
PHN	public health nutrition
PHR	public health research
PI	protease inhibitor
PIVOT	Prostate Cancer Intervention Versus Observation Trial
PM	particulate matter
PMI	President’s Malaria Initiative
PMTCT	prevention of mother-to-child transmission
POC	Point-of-care
POR	prevalence odds ratios
PPE	personal protective equipment
PPROM	Preterm premature rupture of membranes
PPV	pneumococcal polysaccharide vaccines
PPV	positive predictive value
PreP	pre-exposure prophylaxis
PRIO	Peace Research Institute in Oslo
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
ProMED	Programme for Monitoring Emerging Diseases
PSA	prostate-specific antigen
PSBI	possible serious bacterial infections
PSHE	personal, social, and health education
PTSD	post-traumatic stress disorder
PURE	Prospective Urban Rural Epidemiology
PVP	predictive value positive
PWID	people who inject drugs
PYD	positive youth development
QALY	quality-adjusted life year
QIF	Quality Implementation Framework
QUAC	Quaker Upper Arm Circumference
QUADAS	Quality Assessment of Diagnostic Accuracy Studies
R&D	research and development

RA	rheumatoid arthritis
RCS	respirable crystalline silica
RCT	randomized controlled trial
RCV	rubella-containing vaccines
RD	regression discontinuity
RDD	random digit dialling
REACH	Registration, Evaluation, Authorisation, and Restriction of Chemicals
REACT	Randomised Evaluations of Accepted Choices in Treatment
REP	Rochester Epidemiology Project
RHIS	routine health information system
ROC	receiver operator characteristic
ROI	return on investment
RP	restorative practice
RPA	rapid participatory appraisal
RR	rate ratio
RR	relative risk
RSV	respiratory syncytial virus
RT	reverse transcriptase
RTW	return to work
RVF	Rift Valley fever
SAHNA	South Asian Health Needs Assessment
SAM	severe acute malnutrition
SAR	structure activity relationship
SARI	severe acute respiratory infection
SCC	squamous cell carcinoma
SCD	sickle cell disease
SCIH	Swiss Centre for International Health
SD	standard deviation
SDC	Swiss Agency for Development and Cooperation
SDG	sustainable development goal
SDH	social determinants of health
SDIL	Soft Drinks Industry levy
SDOH	social determinants of health
SE	standard error
SES	socioeconomic status
SEYLE	Saving and Empowering Young Lives in Europe
SF	spotted fever
SG	standard gamble
SHS	second-hand smoke
SIS	Statistical Information System
SIV	simian immunodeficiency virus
SMA	spinal muscle atrophy
SMART	Sequential multiple assignment randomized implementation trial
SMD	standardized mean difference
SMR	standardized mortality ratio

SNP	single nucleotide polymorphism
SPRINT	Systolic Blood Pressure Intervention Trial
SRH	sexual and reproductive health
SRNT	Society for Research on Nicotine and Tobacco
SSB	sugar-sweetened beverage
STEM	science, technology, engineering, and mathematics
STH	soil-transmitted helminth
STI	sexually transmitted infections
SUD	substance use disorders
SUDEP	sudden unexplained death in people with epilepsy
SUDI	sudden unexpected death in infancy
SUN	scaling up nutrition
TALENS	transcription activator-like effector nucleases
TasP	Treatment as Prevention
TBA	traditional birth attendant
TBI	traumatic brain injury
TCV	typhoid conjugate vaccine
TDF	tenofovir disoproxil fumarate
TDF	Theoretical Domains Framework
TEDS	Treatment Episode Data Set
TFR	total fertility rate
THS	third-hand smoke
TIA	transient ischaemic attack
TIPS	Treatment and Identification of Psychosis Study
TLV	threshold limit values
TNT	Treat to New Targets
TPFR	total period fertility rate
TST	Tuberculin skin test
TTCV	tetanus-toxoid-containing vaccines
TTO	time trade-off
TTP	trusted third party
TTS	temporary threshold shift
UAV	unmanned aerial vehicles
UDHR	Universal Declaration of Human Rights
UHC	universal health coverage
UIS	UNESCO Institute for Statistics
UK	United Kingdom
UKPDS	United Kingdom Prospective Diabetes Study
UMV	unmanned maritime vehicles
UN	United Nations
UNCRC	United Nations Convention on the Rights of the Child
UNGA	UN General Assembly
UNHCR	United Nations High Commissioner for Refugees
UNICEF	United Nations Children’s Fund
UNISDR	United Nations International Strategy for Disaster Reduction

UNMEER	UN Mission for Ebola Emergency Response
UNOCHA	United Nations Office for Coordination of Humanitarian Affairs
UNODC	United Nations Office on Drugs and Crime
UR	uncertainty range
US	United States
USAID	United States Agency for International Development
USCRI	US Committee for Refugees and Immigrants
USD	United States dollars
USDA	US Department of Agriculture
VA	verbal autopsy
VADT	Veterans Affairs Diabetes Trial
VAS	visual analogue scales
VAW	violence against women
VFD	Veterinary Feed Directive
VGDF	vapours, gases, dusts, and fumes
VIP	ventilated improved pit
VLGA	Victorian Local Governance Association
VR	vital registry
VZV	varicella zoster virus
WEA	work-exacerbated asthma
WFP	World Food Programme
WFS	World Fertility Survey
WGS	whole genome sequencing
WHA	World Health Assembly
WHAS	Worcester Heart Attack Study
WHO	World Health Organization
WRA	work-related asthma
WRMSD	work-related musculoskeletal disorders
WRSD	work-related skin diseases
WSP	Water and Sanitation Programme
WTCCC	Wellcome Trust Case Control Consortium
WTO	World Trade Organization
YLD	years lived with disability
YPLL	years of productive life lost
YSP	Youth Smoking Prevention
ZFN	zinc fingers nucleases

Contributors

Salma M. Abdalla Research Fellow, Department of Epidemiology, Boston University School of Public Health, Boston, MA, USA **Salim S. Abdool Karim** Centre for the AIDS Programme of Research in South Africa (CAPRISA), Durban, South Africa, and Department of Epidemiology, Columbia University, New York, NY, USA **Cecilia S. Acquin** Senior Scientist, Human Nutrition, International Rice Research Institute (IRRI), Los Baños, Philippines; Associate Professor, University of the Philippines, Los Baños, Philippines **Mohammed K. Ali** Associate Professor of Global Health, Epidemiology, and Family and Preventive Medicine; Associate Director, Georgia Center for Diabetes Translation Research; Vice Chair (Research), Department of Family and Preventive Medicine **Peter Azzopardi** Associate Professor and Co-Head of Adolescent Health, Burnet Institute, Australia

Frank Baiden Associate Professor of Clinical Epidemiology, London School of Hygeine and Tropical Medicine, and Manager of Ebola Vaccine Studies, University of Sierra Leone-LSHTM, Sierra Leone **Kathy J. Baisley** London School of Hygiene and Tropical Medicine, University of London, Bloomsbury, London, UK

Lope H. Barrero Professor of Human Factors, Ergonomics, Occupational Epidemiology, Pontifical Xavierian University, Bogotá, Colombia **Mauricio L. Barreto** The Oswaldo Cruz Foundation (FIOCRUZ), Rio de Janeiro, RJ, Brazil

Natalie Bartle Freelance Writer

Catherine R. Bateman Steel Staff Specialist and Deputy Director, South Eastern Sydney Local Health District Public Health Unit, Australia **Fran Baum** Matthew Flinders Distinguished Professor, Southgate Institute for Health, Society & Equity, Flinders University,

South Australia **Cheryl Baxter** Centre for the AIDS Programme of Research in South Africa (CAPRISA), Durban, South Africa

Jennifer Beard Clinical Associate Professor, Department of Global Health, Boston University School of Public Health, Boston, MA, USA **Angela Beaton** Deputy Chief Executve, Delivery and Academic, Te Pūkenga, Waikato, New Zealand

Zulfiqar A. Bhutta Robert Harding Inaugural Chair in Global Child Health & Policy at the Centre for Global Child Health, Toronto and Distinguished University Professor and Founding Director of its Institute for Global Health & Development, TheAga Khan University and Medical Center, Department of Pediatrics, Karachi, Pakistan **Alexandra M. Binder** University of Hawaii Cancer Center, Honolulu, Hawaii, USA; University of California, Los Angeles Fielding School of Public Health, Los Angeles, CA, USA **Fred Binka** Professor of Clinical Epidemiology, School of Public Health, University of Health and Allied Sciences (UHAS), Ghana **Alison Blaiklock** Honorary Senior Clinical Lecturer, Department of Public Health, University of Otago, Wellington, New Zealand **Paolo Boffetta** Associate Director for Gloval Oncology, Tisch Cancer Institute, Icahn School of Medicine, Mount Sinai, New York, NY, USA **Jessica Bon** Assistant Professor of Medicine (PACCM), University of Pittsburgh, Pennsylvania, USA

Chris Bonell Professor of Public Health Sociology, London School of Hygiene & Tropical Medicine, London, UK

James Bowen Assistant Clinical Professor, Department of Health Research Methods, McMaster University, Ontario, Canada **Ailbhe Helen Brady** Warrington and Halton Teaching Hospitals NHS Foundation Trust, Lovely Lane, Warrington, UK

James W. Buehler Clinical Professor in the Department of Health Management & Policy at Drexel University Dornsife School of Public Health, Philadelphia, USA **Paulo M. Buss** The Oswaldo Cruz Foundation (FIOCRUZ), Rio de Janeiro, RJ, Brazil

Tony G. Butler Professor, School of Population Health, Faculty of Medicine, University of New South Wales, Sydney, Australia **Alison Buttenheim** Associate Professor of Nursing and Health Policy, University of Pennsylvania School of Nursing, Philadelphia, USA **Alberto J. Caban-Martinez** Assistant Professor, Department of Public

Health Sciences, University of Miami Health System, Florida, USA **Rona Campbell** Professor of Public Health Research, Bristol Medical School, Bristol, UK

Leda Chatzi Department of Preventive Medicine, Keck School of Medicine, University of Southern California, USA

Chien-Jen Chen Professor, Graduate Institute of Epidemiology, National Taiwan University College of Public Health, Taipei, Taiwan **Wanicha Chuenkongkaew** Professor, Department of Ophthalmology, Faculty of Medicine, Siriraj Hospital, Mahidol University, Bangkok, Thailand **Thomas Clasen** Professor in the Faculty of Environmental Health and Hubert Department of Global Health, Rollins School of Public Health, Emory University, Atlanta, USA **Diane Cooper** Professor, School of Public Health, University of Western Cape; Honorary Professor, School of Public Health and Family Medicine, University of Cape Town, South Africa **Jason Corburn** Professor, School of Public Health, Department of City & Regional Planning; Director, Institute of Urban & Regional Development, Center for Global Healthy Cities, UC Berkeley, USA **Donna Cormack** University of Auckland, Auckland, New Zealand

Peter Craig MRC/CSO Social and Public Health Sciences Unit, University of Glasgow, Glasgow, Scotland

Jonathan C. Craig College of Medicine and Public Health, Flinders University, Adelaide, Australia

Patricia J. Culligan Professor of Civil Engineering, Department of Civil Engineering and Engineering Mechanics, Columbia University, New York, USA **Elana Curtis** University of Auckland, Auckland, New Zealand

Paddy C. Dempsey MRC Epidemiology Unit, School of Clinical Medicine, University of Cambridge, Cambridge, UK

Roger Detels Distinguished Research Professor, Fielding School of Public Health, University of California, Los Angeles, CA, USA **Alex Devine** Research Fellow in Disability Research Policy, Melbourne School of Population and Global Health, Melbourne, Australia **Nancy Devlin** Professor of Health Economics, University of Melbourne, Parkville, VIC, Australia

Brittany Ellis Assistant Professor in Emergency Medicine, Department of Emergency Medicine, University of Saskatchewan, Canada **Wafaa M. El-Sadr** University Professor of Epidemiology and Medicine, Columbia University, Mailman School of Public Health, New York, USA **Ann Evans** Professor in Family Demography, School of Demography, Australian National University, Canberra, Australia **Wenjun Fan** Department of Medicine, Heart Disease Prevention Program, Division of Cardiology, University of California, USA **John W. Farquhar** Professor Emeritus of Medicine and Health Research Policy, Stanford University School of Medicine, USA **Kevin A. Fenton** Professor and Public Health England National Director for Health and Wellbeing, Public Health England, UK

Baruch Fischhoff Howard Heinz University Professor, Department of Engineering and Public Policy, and Institute for Politics and Strategy, Carnegie Mellon University, Pittsburgh, USA **Matthew Fisher** Professor of Fungal Disease Epidemiology, School of Public Health, Imperial College London, London, UK

Fiona Fleck WHO Press, World Health Organization, Switzerland

Matthew Flinders Distinguished Professor, Director, Southgate Institute for Health, Society and Equity, College of Medicine and Public Health, Health Sciences Building, Flinders University, Adelaide, Australia **Sandro Galea** Dean and Robert A. Knox Professor, Boston University School of Public Health, Boston, MA, USA

Wee Hoe Gan Head of the Department of Occupational and Environmental Medicine, Singapore General Hospital, Singapore **Jonathan Garcia** Assistant Professor, College of Public Health and Human Sciences, Oregon State University, Corvallis, USA **Lynn R. Goldman** Professor of Environmental and Occupational Health, George Washington University Milken Institute School of Public Health, Washington, DC, USA **Lawrence O. Gostin** O'Neill Chair in Global Health Law and Director, WHO Collaborating Center on Global Health Law, Georgetown University, Washington, DC, USA **Hebe Gouda** Adjunct Associate Professor, School of Public Health, The University of Queensland, Australia

Kumeren Govender Nuffield Department of Medicine, University of Oxford, Oxford, UK

Lawrence W. Green Professor Emeritus of Epidemiology and Biostatistics, Department of Epidemiology and Biostatistics, University of California, San Francisco, California, USA **Adele Green** Honorary Professor, the School of Public Health, QIMR Berghofer Institute of Medical Research, Brisbane, Australia **Sander Greenland** University of California, Los Angeles, CA, USA

Sian Griffiths Emeritus Professor, Chinese University of Hong Kong, and Advisory Member of the Board of Public Health England and Chair of the PHE Global Health Committee, UK

Emily Grundy Professor of Population Science, University of Essex, Colchester, UK

Nyssa T. Hadgraft Post-doctoral Research Fellow, Centre for Urban Transitions, Swinburne University, Australia

Nafisa Halim Research Assistant Professor, Department of Global Health, Boston University School of Public Health, Boston, MA, USA **Davidson H. Hamer** Professor of Global Health, Boston University, School of Public Health, Boston, USA

Catherine Hankins Assistant Professor, Department of Global Health and Amsterdam Institute for Global Health and Development, University of Amsterdam, Amsterdam, The Netherlands, and Department of Epidemiology, Biostatistics, and Occupational Health, Faculty of Medicine, McGill University, Montreal, Quebec, Canada **Katherine Harripersaud** Institute of Health Policy Management and Evaluation, Dalla Lana School of Public Health, University of Toronto, Toronto, Canada **Patrick Harris** University of New South Wales, Sydney, Australia

Matire Harwood University of Auckland, Auckland, New Zealand

Richard J. Hayes London School of Hygiene and Tropical Medicine, University of London, Bloomsbury, London, UK

Na He Professor of Epidemiology, School of Public Health, Fudan University, Shanghai, China

David L. Heymann Professor of Infectious Disease Epidemiology, London School of Hygiene & Tropical Medicine, London, UK and Head of the Centre on Global Health Security at Chatham House, London, UK

Robert A. Hiatt Professor, Department of Epidemiology and Biostatistics and Associate Director of Population Sciences, Helen Diller Family Comprehensive Cancer Center, University of California, San Francisco, California, USA **Nguyen Tran Hien** Senior Researcher, National Institute of Hygiene and Epidemiology, Hanoi, Vietnam

Lisa Hightow-Weidman Professor in the Division of Infectious Diseases, School of Medicine, University of North Carolina, North Caroline, USA **Marcia Hills** Professor and Associate Director of Research and Scholarship, University of Victoria, British Columbia, Canada **Kristin S. Hoeft** Assistant Professor, Department of Preventive & Restorative Dental Sciences, University of California, San Francisco, California, USA **Katherine J. Hoggatt** San Francisco VA Health Care System, University of California, San Francisco, CA, USA

San Hone Department of Public Health, Ministry of Health and Sports, Yangon, Myanmar

Zhiyuan Hou Associate Professor of Social Medicine, School of Public Health, Fudan University, Shanghai, China

Sai Yin Ho Associate Professor, Division of Community Medicine and Public Health Practice, University of Hong Kong, China **Andrea Howard** Director, Clinical and Training Unit, ICAP Global Health, University of Columbia, New York, USA

Thein T. Htay Former Deputy Minister, Ministry of Health and Sports, Nay Pyi Taw, Myanmar

Aung Soe Htet International Relations Division, Ministry of Health and Sports, Nay Pyi Taw, Myanmar; Institute of Health and Society, Faculty of Medicine, University of Oslo, Oslo, Norway **Paul Hunt** Human Rights Centre, Department of Law, Essex University, Colchester, UK

Rachel Lamdin Hunter Centre for Health and Social Practice, Waikato Institute of Technology, Hamilton, New Zealand **Adnan A. Hyder** Professor and Senior Associate Dean for Research, George Washington University Milken Institute of Public Health, Washington, USA **Sopon Iamsirithaworn** Department of Disease Control

Yvonne Inall Sessional Lecturer, University of Sydney, Australia; Project Officer, Australian Archaeological Institute, Athens, Greece,

Australia **Omar Irfan** Healthcare Management Consultant, Toronto, Canada

F. Javier Nieto Dean and Professor of Epidemiology, College of Public Health and Human Sciences, Oregon State University, Corvallis, OR, USA **Rachel Jewkes** Honorary Professor at the University of the Witwatersrand School of Public Health, Pretoria, South Africa **Tang Jin-ling** Professor, Division of Epidemiology, The Jockey Club School of Public Health and Primary Care, The Chinese University of Hong Kong, Hong Kong, China **Gavin W. Jones** Emeritus Professor, School of Demography, Research School of Social Science, Australian National University, Canberra, Australia **Rhys Jones** University of Auckland, Auckland, New Zealand

S.M. Kadri Assistant Director/State Surveillance Officer, Disease Control, Directorate of Health Services, Kashmir, India **Natasha Kaoma** Youth Commissioner, Lancet Standing Commission on Adolescent Health and Wellbeing, Lusaka, Zambia; Copper Rose, Lusaka, Zambia **Quarraisha Abdool Karim** Associate Scientific Director: CAPRISA, PI: KwaZulu-Natal Clinical Trials Unit, Professor in Clinical Epidemiology: Columbia University, UNAIDS Special Ambassador for Adolescents and HIV, Pro Vice-Chancellor African Health: University of KwaZulu-Natal, Nelson R Mandela School of Medicine, Durban, South Africa **Nancy Kass** Deputy Director for Public Health and Phoebe R. Berman Professor of Bioethics and Public Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, USA **Peter Katona** Clinical Professor of Medicine (Infectious Diseases); Adjunct Professor of Public Health (Epidemiology), UCLA Schools of Medicine and Public Health, Los Angeles, CA, USA **Anne Kavanagh** Chair in Disability and Health, Melbourne School of Population and Global Health, University of Melbourne, Carlton, Australia **Michael P. Kelly** Department of Public Health and Primary Care, University of Cambridge, UK

Farah Naz Khan Clinical Assistant Professor of Medicine, University of Washington, Seattle, WA, USA

Amira Khan Centre de Recherche Inserm, Team-Physiologie de la Nutrition & Toxicologie, Université de Bourgogne Franche-Comté (UBFC)Dijon, France **Leeka Kheifets** Professor of Epidemiology,

Department of Epidemiology, University of California School of Public Health, Los Angeles, California, USA **Ann Marie Kimball** Senior Consulting Fellow, Chatham House (Royal Institute of International Relations) London, UK; Professor Emerita, School of Public Health, University of Washington, Seattle, WA, USA **Robert J. Kim-Farley** Professor-in-Residence, Epidemiology and Community Health Sciences, UCLA Fielding School of Public Health, Los Angeles, CA, USA **J.D. Klausner** Professor of Medicine and Public Health, UCLA David Geffen School of Medicine and Fielding School of Public Health, Los Angeles, CA, USA **Lindsay C. Kobayashi** University of Michigan School of Public Health, Ann Arbor, MI, USA

Manolis Kogevinas Scientific Director, Cancer, Water Pollution, Non-communicable Diseases and Environment, ISGlobal, Barcelona, Spain **David Koh** Saw Swee Hock School of Public Health, National University of Singapore, Singapore, Singapore

Noah Kojima Department of Internal Medicine, University of California, Los Angeles, CA, USA

Tamar Krishnamurti Assistant Professor of Medicine and Clinical and Translational Science, Division of General Internal Medicine, University of Pittsburgh, Pittsburgh, USA **Walter A. Kukull** Professor of Epidemiology, National Alzheimer Coordinating Center (NACC), Department of Epidemiology, University of Washington, Seattle, USA **Ronald Labonté** University of Ottawa, Ottawa, Canada

Tai Hing Lam Chair Professor of Community Medicine, School of Public Health, The University of Hong Kong, Hong Kong, China **Kelley Lee** Simon Fraser University, Burnaby, BC, Canada

Vernon J.M. Lee Associate Professor in Preventative Medicine, Saw Swee Hock School of Public Health and Associate Professor of Public Health National University of Singapore and National University Health System, Singapore **Chien Earn Lee** Deputy Group CEO of Singhealth Regional Health System, and Adjunct Professor at Saw Swee Hock School of Public Health, National University of Singapore, Singapore **Stephen Leeder** Emeritus Professor of Public Health and Community Medicine, The University of Sydney, Menzies Centre for Health Policy and Economics, Faculty of Medicine and Health, Sydney School of Public Health, Australia **Raphael Lencucha** Associate

Professor, School of Physical and Occupational Therapy, McGill University, Montreal, QC, Canada **James Levinson** Former Director, International Food and Nutrition Center Tufts University School of Nutrition Science and Policy, Boston, MA, USA **Barry S. Levy** Adjunct Professor of Public Health and Community Medicine, Tufts University School of Medicine, Sherborn, USA **Li Li-ming** Research Associate at the ESRC Centre for Society and Mental Health and Department of Global Health & Social Medicine, King's College London, UK

Ro-Ting Lin Associate Professor, Department of Occupational Safety and Health, College of Public Health, China Medical University, China **Tim Lobstein** Policy Director for the World Obesity Federation based in London, UK, and Adjunct Professor at the Public Health Advocacy Institute of Western Australia, Curtin University, Perth, Australia **Frances Lehman Loeb**Professor of Epidemiology T.H. Chan School of Public Health, Harvard University, Boston, USA **Alan D. Lopez** University of Melbourne, Parkville, VIC, Australia

Clement Loy Sydney School of Public Health, Faculty of Medicine and Health University of Sydney, Sydney, Australia **Johan P. Mackenbach** Department of Public Health, Erasmus MC, Rotterdam, the Netherlands

Alex Macmillan Associate Professor, Department of Preventive and Social Medicine, University of Otago, Dunedin, New Zealand **Shabir A. Madhi** Professor of Vaccinology, South African Medical Research Council, Vaccine and Infectious Diseases Analytical Research Unit, University of the Witwatersrand, Johannesburg, South Africa **Keziah L. Malm** Programme Manager, The National Malaria Control Program, Ghana Health Service, Accra, Ghana

Elizabeth Rose Mayeda University of California, Los Angeles Fielding School of Public Health, Los Angeles, CA, USA **Lawrence Mbuagbaw** Associate Professor, McMaster University, Hamilton, ON, Canada

Kaitlyn B. McBride University of California, Los Angeles, CA, USA

Jennifer McGoogan Chinese Center for Disease Control and Prevention, Beijing, China

Giang Le Minh Chair, Department of Global Health, Hanoi Medical University, Vietnam

Amira S. Mohamed Department of Pediatric Dentistry and Dental Public Health, Faculty of Dentistry, Alexandria University, Alexandria, Egypt **Stephen Morris** RAND Professor of Health Services Research, University of Cambridge, Cambridge, UK

Alison Morris Health, Safety, and Governance Manager, NHS, UK

Lawrence H. Moulton Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA

Alvaro Muñoz Professor of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA

Miguel Muñoz-Laboy Associate Professor, School of Social Welfare, Stony Brook School of Social Welfare, New York, USA **Michael Murphy** Professor of Demography at the London School of Economics and Political Science, London, UK

Christopher J.L. Murray University of Washington, Seattle, WA, USA

Virginia Murray Consultant in Global Disaster Risk Reduction for Public Health England, and Visiting Professor in Health Protection, Imperial College and King's College London, UK

Eleonora A.M.L. Mutsaerts Medical Research Council, Respiratory and Meningeal Pathogens Research Unit, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa **K.M. Venkat Narayan** Ruth and O.C. Hubert Professor of Global Health and Epidemiology, Hubert Department of Global Health, Emory University, Atlanta, GA **Naomi Noguchi** Sydney School of Public Health, Faculty of Medicine and Health University of Sydney, Sydney, Australia **Tulio de Oliveira** KwaZulu-Natal Research Innovation and Sequencing Platform (KRISP), University of KwaZulu-Natal, South Africa and Department of Global Health, University of Washington, Seattle, WA, USA **Neville Owen** NHMRC Senior Principal Research Fellow and Distinguished Professor of Health Sciences, the Centre for Urban Transitions, Swinburne University, Australia **Romulo Paes-Sousa** The Oswaldo Cruz Foundation (FIOCRUZ), Rio de Janeiro, RJ, Brazil

Sarah-Jane Paine University of Auckland, Auckland, New Zealand

Richard Parker Professor Emeritus in Sociomedical Sciences and Anthropology, Mailman School of Public Health, Columbia University, New York, USA **David Parkin** Honorary Visiting Professor of Economics, University of London, London, UK; Senior Visiting Fellow, Office of Health Economics, London, UK

George Patton Professor at the Centre for Adolescent Health, Royal Children's Hospital and Murdoch Children's Research Institute, Melbourne, VIC, Australia; Department of Paediatrics, University of Melbourne, Melbourne, VIC, Australia; Nossal Institute, School of Population and Global Health, University of Melbourne, Melbourne, VIC, Australia **Amy Paul** Strategy and Research Team, Centre for Digital Development, US Agency for International Development, USA **Corinne Peek-Asa** Associate Dean for Research, Department of Occupational and Public Health, University of Iowa, Iowa, USA **Anna Peeters** Director, Institute of Health Transformation, School of Health and Social Development, Deakin University, Victoria, Australia **Jennie Popay** Distinguished Professor of Sociology and Public Health, Division of Health Research, Faculty of Health and Medicine, Lancaster University, Lancaster, UK

Jane E. Powell School of Health and Social Wellbeing, University of the West of England, UK

John Powles† Cambridge Institute of Public Health, Cambridge, UK

Jessica Price Co-Founder, the Rhodes Incubator, University of Oxford, Oxford, UK

Stella R. Quah Adjunct Professor, Health Services and Systems Research, Duke-NUS Medical School, National University of Singapore, Singapore **Veron Ramsuran** Senior Lecturer, College of Health Services, University of KwaZulu-Natal, Durban, South Africa

Sharon Reid Sydney School of Public Health, Faculty of Medicine and Health, University of Sydney, Sydney, Australia **Papaarangi Reid** University of Auckland, Auckland, New Zealand

Craig M. Riley Division of Pulmonary, Department of Medicine, Allergy and Critical Care Medicine, University of Pittsburgh,

Pennsylvania, USA **Les Roberts** Professor, Population and Family Health, Columbia University Medical Centre, New York, USA

Peter G. Robinson Emeritus Professor, Bristol Dental School, University of Bristol, Devon, UK

Robin Room Distinguished Professor, Centre for Alcohol Policy Research, La Trobe University, Melbourne, Australia; Professor, Centre for Social Research on Alcohol and Drugs, Department of Public Health Sciences, Stockholm University, Stockholm, Sweden **Linda Rosenstock** Dean Emeritus, UCLA Jonathan and Karin Fielding School of Public Health, Professor of Health Policy and Management, Environmental Health Sciences, and Medicine, University of California, Los Angeles, CA, USA **Roxana Rustomjee** Senior Vice President, Research and Development, Sabin Vaccine Institute, Washington, USA

Farnaz Sabet Centre for Adolescent Health, Royal Children's Hospital, Parkville, Victoria, Australia

David Sanders

† University of Western Cape, Cape Town, South Africa

Manjinder S. Sandhu Reader in Global Health and Population Science, University of Cambridge, Cambridge, UK

Yu Mon Saw Associate Professor, Nagoya University Graduate School of Medicine, Nagoya, Japan; Nagoya University Asian Satellite Campuses Institute, Nagoya, Japan **Susan Sawyer** Geoff and Helen Handbury Chair of Adolescent Health, Department of Paediatrics, The University of Melbourne, Victoria, Australia **Cari van Schalkwyk** Stellenbosch University, Stellenbosch, South Africa

Peter W. Schofield School of Medicine and Public Health and Hunter New England Local Health District, Neuropsychiatry Service, University of Newcastle, Callaghan, Australia **Ashley Schram** Australian National University, Canberra, Australia

Than Sein Chairman at People Health Organisation, Myanmar

Nida Izhar Shaikh Assistant Professor of Nutrition, Byrdine F. Lewis College of Nursing and Health Professions, Atlanta, USA **Marissa Shields** Centre for Health Equity, Melbourne School of Population and Global Health, The University of Melbourne, Victoria, Australia **Steve**

Shoptaw Professor in Family Medicine and Psychiatry and Biobehavioral Sciences and Vice Chair for Research in Family Medicine, University of California, Los Angeles, California, USA **Daniel Shouval** Professor Emeritus of Medicine, Liver Unit, Hadassah-Hebrew University Hospital, Jerusalem, Israel **Roger Shrimpton** Adjunct Professor, Department of Global Community Health and Behavioural Sciences, Tulane School of Public Health and Tropical Medicine, New Orleans, USA **Andrew Siegel** Department of Urology, Hackensack University Medical Center, Hackensack, New Jersey, USA

Nandi Siegfried Chief Specialist Scientist and Associate Professor, South African Medical Research Council, Tygerberg, WC, South Africa **Aaron J. Siegler** Associate Professor, Rollins School of Public Health, Department of Behavioral Sciences and Health Education, Emory University, Atlanta, USA **Urisha Singh** African Health Research Institute, Durban, South Africa

Samir K. Sinha Peter and Shelagh Godsoe Chair in Geriatrics, Director of Geriatrics, Sinai Health System and the University Health Network Hospitals; Associate Professor of Medicine, Health Policy, Management and Evaluation, University of Toronto; Assistant Professor of Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, USA **Liam Smeeth** Professor of Clinical Epidemiology, Dean, Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, UK

Marni Sommer Department of Sociomedical Sciences, Columbia University Mailman School of Public Health, New York, USA **Craig Spencer** Assistant Professor, Department of Population and Family Health, Mailman School of Public Health, Columbia University, New York, USA **K. Srinath Reddy** Professor and President of the Public Health Foundation of India, India

Fiona F. Stanaway Sydney School of Public Health, Faculty of Medicine and Health, University of Sydney, Sydney, Australia **Patrick S. Sullivan** Professor, Department of Epidemiology, Rollins School of Public Health, Emory University, Atlanta, USA **Paul Sutton** Interim Director, EPRR, Public Health England, England, UK

Simon Szreter University of Cambridge, Cambridge, UK

Hanani Tabana Senior Lecturer, School of Public Health, University of the Western Cape, Cape Town, South Africa; Professor, SSH School of Public Health and YLL School of Medicine, National University of Singapore, Singapore **Chorh Chuan Tan** Executive Director, Ministry of Health, Singapore

Panithee Thammawijaya Bureau of Epidemiology, Department of Disease Control, Ministry of Public Health, Bangkok, Thailand **Harsha Thirumurthy** Associate Professor, Department of Medical Ethics and Health Policy University of Pennsylvania, Philadelphia, USA **Anne Marie Thow** Associate Professor, Public Policy and Health, Menzies Centre for Health Policy, The University of Sydney, Director of Academic Titles, School of Public Health, Deputy Academic Lead, Research Education, School of Public Health **Stephen Tollman** Research Professor and Head, Division of Health and Population , School of Public, Health University of the Witwatersrand, South Africa **Belinda Townsend** Australian National University, Canberra, Australia

Kumnuan Ungchusak Department of Disease Control, Ministry of Public Health, Thailand

Tyler Vaivada Centre for Global Child Health, The Hospital for Sick Children, Toronto, Ontario, Canada

Frank J. van Lenthe Erasmus University Medical Center, Rotterdam, the Netherlands

Tjeerd-Pieter van Staa Professor of Health eResearch, School of Health Sciences, Faculty of Biology, Medicine and Health, The University of Manchester, Manchester, UK

Tyler J. VanderWeele John L. Loeb and Frances Lehman Loeb Professor of Epidemiology, T.H. Chan School of Public Health, Harvard University, Boston, USA **Carlo La Vecchia** Professor of Epidemiology, Department of Clinical Sciences and Community Health, University of Milan, Milan, Italy **Yasmin E.R. von Schirnding** Adjunct Assistant Professor in Environmental Health Sciences, Columbia University Medical Centre, New York, USA **Theo Vos** University of Washington, Seattle, WA, USA

Kristian Wahlbeck Research Professor at the Finnish Institute for Health and Welfare, National Institute for Health and Welfare, Helsinki, Finland **Thomas Waite** Public Health Wales, Cardiff, UK

Richard Wakeford Honorary Professor, Centre for Occupational and Environmental Health, Institute of Population Health, The University of Manchester, Manchester, UK

Robert S. Ware Menzies Health Institute Queensland, Griffith University

Judith Wasserheit Professor of Global Health, University of Washington, Seattle, Washington, USA

Danuta Wasserman Professor of Psychiatry and Suicidology, National Centre for Suicide Research and Prevention of Mental Ill-Health (NASP), Karolinska Institute, Stockholm, Sweden **Noel S. Weiss** Professor, Department of Epidemiology, University of Washington, Seattle, WA, USA

Alex Welte Stellenbosch University, Stellenbosch, South Africa

Suwit Wibulpolprasert Public Health Specialist, Bureau of Epidemiology, Department of Disease Control, Ministry of Public Health, Nonthaburi, Thailand **Bryan Wiener** Professor of Global Health and Professor of Health Services, University of Washington, Washington, USA **Carmel Williams** Executive Editor, Health and Human Rights Journal, FXB Center for Health and Human Rights, Harvard University, Boston, MA, USA **Gail Williams** Professor, School of Public Health, The University of Queensland, Australia

Patrick Wilson Associate Professor and Director of SPHERE, Columbia University, New York, USA

Nathan D. Wong Professor and Director, Heart Disease Prevention Program, Division of Cardiology, University of California, Irvine, CA, USA **Alistair Woodward** Professor, Section of Epidemiology and Biostatistics, University of Auckland, Auckland, New Zealand **Zunyou Wu** National Center for AIDS/STD Control and Prevention, Chinese Center for Disease Control and Prevention, Beijing, China **San-Lin You** School of Medicine, Fu-Jen Catholic University, Taiwan

Elizabeth H. Young Senior Staff Scientist, Sandhu Group of Global Health and Populations, Wellcome Sanger Institute, UK

Zuo-Feng Zhang Distinguished Professor of Epidemiology, Associate Dean for Research, Fielding School of Public Health, University of California, Los Angeles, USA **Ehud Zigmond** Gastroenterology and Hepatology Institute, Tel Aviv Sourasky Medical Center, Sackler School of Medicine, Tel-Aviv University, Tel Aviv, Israel **Anthony B. Zwi** Professor of Global Health and Development, School of Social Sciences, University of New South Wales, Australia

The SARS-CoV-2 Pandemic

Quarraisha Abdool Karim, Fran Baum, and Roger Detels

Summary

The SARS-CoV-2 pandemic is the most recent and challenging pandemic facing humanity in the twenty-first century. While it is a reminder of our global interconnectedness and shared vulnerabilities, it has also highlighted the stark inequities within and between countries that render some more vulnerable than others. The disruptions to the economy and loss of lives unmasked yet again the intersection of race, gender, displacement, and poverty as well as the differing vulnerabilities of those who could not pivot to remote working, access health services, ensure food and income security, or implement public health measures, and those who could.

In the words of António Guterres, United Nations Secretary-General at the 2021 High Level Political Meeting on Accelerating progress on the 2030 Sustainable Development Goals (SDGs):

> *This crisis is a wake-up call to do things differently. We need a recovery based on economic and social justice since response gaps in pandemics, whether HIV or COVID-19, lie along the fault lines of inequality.*

Country responses to this pandemic, identified in early 2020, provide important insights on political leadership and commitment in the face of an emergent crisis and balancing or prioritizing saving lives, livelihoods, and interrupting chains of viral transmission. The returns on investments in epidemic preparedness including modernized health information systems, access to healthcare, critical care capacity, investments in science and technology, and/or bulk manufacturing and distribution capacity enabled some countries to respond more rapidly. Trust in government continues to influence individual-level decisions on the adoption of public health and

social measures. The rate of knowledge generation has been unprecedented, for example, the identification of the causative organism within weeks; the development of diagnostic assays; development of six highly efficacious vaccines in 2020; and the importance of oxygen and dexamethasone for survival in the face of severe disease. Many political leaders have had their responses guided by the science despite substantial scientific uncertainty. Others have ignored scientific advice causing their countries to suffer rapid transmission, rampant infections, and death overwhelming their healthcare systems. This ongoing pandemic provides opportunities to reflect on the strengths and weaknesses of global and country responses to SARS-CoV-2 and the need to respond to multiple parallel pandemics. The hope emanating from vaccines informed by mimicking natural recovery from infection has been dampened by inequitable access as a result of vaccine nationalism. The ongoing emergence of new variants of concern of SARS-CoV-2, even in the face of high vaccine coverage rates in some countries, highlights the need to better understand viral immune escape and redefine what epidemic control needs to look like. At the time of writing (August 2021) it is evident from the emergence of the Delta variant, the most transmissible variant of concern, that the pandemic is far from over. The reluctance of many to accept vaccination; the third and fourth waves/surges of infections in mid-2021; the growing cohorts of long COVID-19 patients are sobering reminders of how little we know about SARS-CoV-2 and COVID-19, and how uncertain the future is in terms of when countries can safely implement their post-COVID-19 social and economic recovery plans. This chapter highlights key elements of the biology of SARS-CoV-2 infection; effective responses to SARS-CoV-2; the interaction between SARS-CoV-2 and HIV; and the opportunity and importance of control of existing non-communicable diseases and other serious communicable diseases in the face of the pandemic. It illustrates the application of basic and more advanced epidemiological concepts of pandemics; public health responses; communication strategies, case identification, isolation, and quarantine; incidence and prevalence; principles of screening in use of diagnostic assays; herd immunity; transmission dynamics; molecular surveillance and use of genomics and meta-data. The pandemic underscores the challenges of responding effectively to global emergencies and disasters and the need for solid investments in epidemic and pandemic preparedness. SARS-CoV-2 provides an opportunity to rethink global health, the importance of

partnerships and collaborations between countries and between governments and the public; revisiting intellectual property rights in relation to the public good in the context of global emergencies and disasters; and opportunities for accelerating progress on the UN 2030 Sustainable Development Goals through addressing the root causes of inequalities.

1. The emergence of a new pandemic

On 27 December 2019, a physician in Wuhan, China reported seeing an unusual pneumonia to the Jiangshan District Center for Disease Control. Within four days an epidemiologic alert had been raised by the local health authority, and the Chinese Center for Disease Control initiated an investigation. Initially suspicion involved a seafood market as the source of cases and the Huanan Seafood Market in Wuhan was closed down immediately. But by then, secondary cases without a history of contact with the market were already occurring. By 3 January 2020, China notified the World Health Organization (WHO) and other countries that an epidemic of an unknown disease was occurring in the Wuhan area (1). By 7 January 2020, a corona virus, named by WHO, the severe acute respiratory coronavirus 2 (SARS-CoV-2), was identified as the cause of the corona virus disease 2019 syndrome (COVID-19) (2). Remarkably, within three days the genetic sequence of the virus was elucidated and shared with WHO and, through open-source access, facilitated the development within days of a diagnostic polymerase chain reaction test in Germany to identify infected individuals (3).

By 17 January 2020, most countries had started to implement some form of port of entry (thermal and/or symptom) screening for COVID-19. By 21 January and 24 January 2020, the first cases were reported in the United States and Europe, respectively. On 20 January 2020, COVID-19 was added to the WHO list of notifiable diseases and three days later, on 23 January 2020, Wuhan was placed on lockdown, cases isolated, and contacts quarantined. On 30 January 2020, WHO declared the SARS-CoV-2 pandemic a Public Health Emergency of International Concern that was followed by similar national declarations in several countries around the world (4).

The early action by the government of China successfully slowed the spread of SARS-CoV-2 in China but at the cost of limiting civil liberties

through rapidly enforced isolation of cases, quarantine of contacts, and forceful implementation of 'lockdown' regulations. These important milestones and timelines are captured in Figure 1.

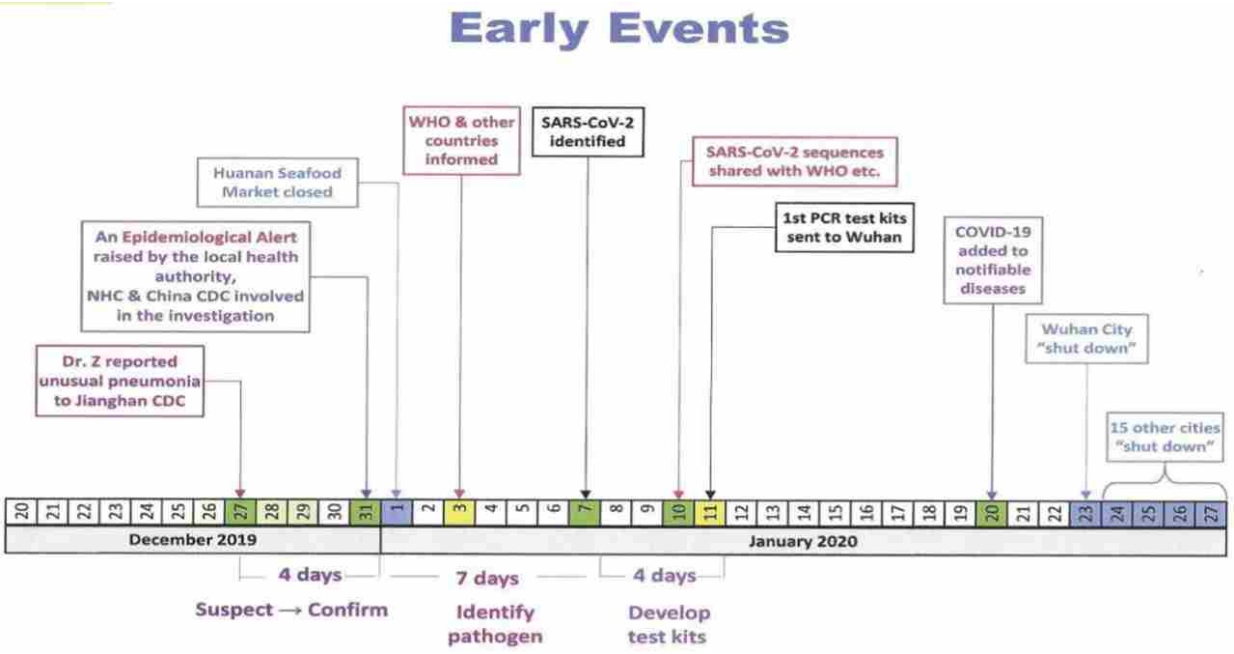


Figure 1 Timeline for the emergence of a new pandemic in China and key milestones.

2. The first 18 months of the pandemic

As the virus started to spread beyond China, countries started to implement national programmes to respond to the emerging epidemics within their countries, some guided by the Chinese government response. There were wide variations in responses that were influenced by, among others: political leadership; public fear and anxiety of the unknown; economic and human rights considerations that had implications for the epidemic trajectory within a country as well as for the global trajectory of SARS-CoV-2 infections (5). From mid-January 2020 to mid-March 2020, return from travel to countries with identified COVID-19 cases was viewed as the main risk factor for infection and diagnostic testing was restricted to symptomatic individuals who had recently returned from a country with COVID-19 and their close contacts. In contrast to China, Singapore, New Zealand, and South Korea,

most governments spent days and weeks debating issues like isolation, quarantine, banning or restricting large gatherings, criteria for eligibility for diagnostic testing; mandatory wearing of masks, social distancing, and lockdown regulations which, combined with a lack of national leadership, allowed the virus to spread largely unchecked within communities. Delays in the implementation of prevention strategies; lack of unity in action; poor preparedness of health facilities for hospitalization of severe cases; supply chain challenges for personal protective equipment; access to reliable diagnostic tests and/or reagents; capacity to identify, isolate, and quarantine cases contributed to the success or failure to contain the spread of the virus and the COVID-19 mortality rates at a country level. By April 2020 most countries had established special COVID-19 Science Advisory Groups or COVID-19 Task Forces, instituted declarations of COVID-19 as a public health emergency and introduced restrictions against large gatherings; school closures; curfews, and were implementing public health and social measures such as mask wearing; social distancing and isolation of confirmed cases. The goal of most country responses was to slow down the spread of SARS-CoV-2 and prepare health facilities to manage the care of severely ill COVID-19 patients.

The United States of America (USA), United Kingdom (UK), and Italy suffered among the highest incidence and mortality rates globally (6, 7). After an initial delay, the UK subsequently implemented strong responses and Neil Ferguson is quoted as saying ‘institution of a national lockdown one week earlier would have reduced the COVID-19 mortality rates by 50% in the UK’ (8). The anticipated high rates of COVID-19 in most low-and-middle-income countries (LMICs) did not unfold in the first half of 2020. Over the past 18 months we have seen the emergence of waves/surges of infections every 90–120 days with the most recent (fourth wave) occurring even in countries with high vaccination coverage rates. By early 2021, as the high death rates in the USA started to decline we saw substantial increases in cases in Brazil, India, and Indonesia. As of mid-August 2021, globally there were 207 million cases of COVID-19 and 4.3 million COVID-19-related deaths and most countries have been experiencing a new surge of infections (6, 7).

As the virus has continued to spread unrelentingly and with high speed in the past 18 months, our knowledge and understanding of this new pandemic

has grown in parallel and continues to grow as described by some as analogous to ‘building the ship while it is sailing’.

3. Biology of SARS-CoV-2

Coronaviruses are a very diverse family of enveloped, single-stranded ribonucleic acid (RNA) viruses (9, 10). Coronaviruses, first discovered in the 1960s, are found in birds and mammals, especially in bats, civets, camels, and rats. Coronaviruses constitute the subfamily *Orthocoronavirinae*, family *Coronaviridae*, and order *Nidovirales*, which is further subdivided into four genera (Alpha-, Beta-, Gamma-, and Delta-). SARS-CoV-2 is a beta-coronavirus (β-CoV) and is the seventh coronavirus known to infect humans. Notwithstanding restricted geospatial transmission, SARS-CoV-1 (Asia) in 2003 and MERS-CoV (Middle East) in 2012 caused serious disease. Coronaviruses HKU1, NL63, OC43, and 229E are associated with mild symptoms and represent about a fifth of ‘common colds’ in most parts of the world. SARS-CoV-1 caused a much smaller pandemic and ultimately has disappeared globally. MERS-CoV has caused localized epidemics in the eastern Mediterranean area, Korea, and Thailand and is still present in camels in the Middle East. MERS has a case fatality ratio of over 30%. Both SARS-CoV-1 and MERS are thought to have originated in bats, which are immunologically very permissive, and gone through an intermediate animal host—for SARS-CoV-1, civets and for MERS, camels. SARS-CoV-2 is thought to have undergone a similar process originating in bats but an intermediate animal host has not been identified (9, 10).

The inability to identify a likely animal intermediary host has resulted in considerable controversy about the origin of the virus, much of which involves international politics (11). The true origin of SARS-CoV-2 in humans remains to be established. Phylogenetic analysis (the study of evolutionary relationships among biological entities, often species, individuals, or genes, where viruses isolated from humans and or other animal species are sequenced and compared to identify common ancestry or where differences exist between species) suggests that bats are the most likely original host. Phylogenetic analyses show that the pangolin coronavirus is genetically related to the SARS-CoV-2 and other bat coronaviruses. It carries the same receptor binding domain as does SARS-

CoV-2 but the rest of the virus is only 90% genetically similar. Pangolins may therefore not be the intermediary host. SARS-CoV-2 binds with high affinity to human angiotensin-converting enzyme-2 (ACE-2) as its cellular receptor and uses it as an entry point to invade target cells in humans. ACE-2 receptor expression is much higher in the lower lobes of our lungs compared with the upper respiratory tracts (12).

SARS-CoV-2 has four structural proteins: spike (S), membrane (M), envelope (E), and nucleo-capsid (N) (9, 10). The spikes on the surface of SARS-CoV-2 virus particles each comprise three identical subunits, each of which, in turn, is made up of two functional components; the S_1 part is responsible for binding to the host cell receptor and the S_2 part is for the fusion of the viral and cellular membranes. The spike protein binds to the cellular ACE-2 receptor of the cell by which it gains entry into a cell through a process of fusion to the cell membrane and uncoating of the virus in the cytoplasm of the cell. Unlike retroviruses, the intracellular process of replication of coronavirus does not involve entry of viral material into the nucleus and genetic DNA of the host cell. Once in the cytoplasm of the cell the infecting virus is subject to a series of complex steps resulting in the expression and replication of messenger RNA and genomic RNA to produce full-length copies that are incorporated into the newly produced viral particles which are secreted and released by exocytosis.

Despite the RNA proofreading function of the virus, the high incidence of infections in humans provides opportunities for mutations in the SARS-CoV-2 virus genome and the development of variants (distinct viruses with a specific combination of mutations) that is discussed further in Section 10.3.

4. Transmission of SARS-CoV-2

Transmission in most instances takes place via inhalation or contact with mucosal surfaces of respiratory droplets (generated through coughing and sneezing) or the much smaller and lighter respiratory aerosols (generated by normal breathing, talking, or singing), and by direct transfer of virus particles from contaminated surfaces to the mouth, nose, or eyes (13). Airborne transmission is most likely to occur at short range, but longer-range transmission can occur in poorly ventilated spaces. Transmission also occurs through direct contact with still-active virus left on surfaces in the immediate

environment of, or on objects used by, actively infected persons. The stability of the virus in air and on various surfaces varies widely. For example, on copper it is 4 hours, 24 hours on cardboard, 48 hours on stainless steel, and 72 hours on plastic.

The virus is usually detectable in the nasopharynx/oropharynx 1–3 days before the onset of symptoms (13). The viral load peaks around the time of symptom onset. Viral loads are generally higher in sputum specimens than in nasopharyngeal or oro-pharyngeal swabs. The virus is detectable by PCR in the nasopharynx/oropharynx for ≥7 days after the onset of illness in most cases, but culturable virus appears to be rare after day 7. Pre-symptomatic transmission is an important contributor to community transmission, estimated to be up to 50%. Transmission occurs from symptomatic cases in the first few days of symptoms. Transmission after 7 days of illness is likely to be rare. The diagnostic test positivity rate is recommended by WHO as a proxy marker of transmission rates and test positivity rates of <5% is considered as a marker of low levels of transmission (14).

Two studies provided important insights on viral transmission very early on in the pandemic (15, 16). The first are data from a cruise ship, the Diamond Princess, that showed that for every 100 symptomatic cases, there were an additional 19 asymptomatic cases. The overall test positivity rate was 19% in the sample of the total population on the cruise ship. In contrast, data from the US aircraft carrier, Theodore Roosevelt, demonstrated that there could be as many as 120 asymptomatic cases for every 100 symptomatic cases. The test positivity rate was 15% in this cohort. An important confounder in the data from these two ships is age, with the Diamond Princess having many more older people in contrast to the younger cohort of military personnel on the Theodore Roosevelt.

Transmissibility can be described with the basic reproduction number (R_0), which was estimated at between 2 and 3 for the original Wuhan strain of SARS-CoV-2 (and is now estimated to be 6–8 for the Delta variant, highlighting the increased transmissibility of the Delta variant) (17). This can be contrasted to seasonal influenza with an R_0 around 1.3. R_0 simply expresses how many people on average will be infected by each patient, and it is calculated with the product of the probability of transmission, the number of contacts per unit time, and the duration of infectiousness (18). Thus, interventions aimed at reducing the rate of spread to less than 1 need to influence one or more of these variables. Typically for herd immunity to

be reached, R0 has to be <1, or in other words, each person infected is unlikely to spread infection to anyone else. To create a time-dependant reproductive number, a new approach was adopted called $R_{t.}$ $R_{t.}$ is important as it takes into account the lag between cases and provides a more accurate estimate of the quantity of viral spread. R_t cannot be measured but it can be estimated for any period of time, with 7 and 14 days being the time periods of interest for Sars-CoV-2 (19, 20). Further $R_{t,}$ takes into account vaccination coverage rates and adoption of other prevention measures.

The virus spreads easily and remains viable outside the body for several hours or days depending on the way it leaves and the material of any surfaces it lands on (13). Stopping spread is complicated as the virus is highly infectious and spreads even before an infected person has symptoms (pre-symptomatic spread) and, further, the majority of exposed and infected individuals remain asymptomatic. The transmission rate is high, on average one person infects between two and three others; but, in the circumstances in which spreading is most effective (closed, unventilated spaces with many people not wearing masks and talking or singing, etc.), superspreading can occur with tens or hundreds of people being infected in a single event (21). It is estimated that about 80% of transmissions take place at such superspreading events from less than 10% of infected individuals. Even when substantial reductions in new infections have been achieved, there can readily be a resurgence of infection. The virus replicates and spreads rapidly within a matter of days. This rapid spread and surge of infections can easily overwhelm health services and divert attention from other serious health problems, resulting in higher death rates (22). During winter it is challenging to differentiate between flu and COVID-19.

5. Temporal trends in SARS-CoV-2 infections and national responses

The global numbers of SARS-CoV-2 cases and COVID-19 deaths masks wide variations within and between countries in terms of magnitude of epidemics, severity of disease presentation and mortality rates (Figure 2) (6, 7). For example in Africa, while there have been 6.2 million cumulative cases representing 3.2% of global cases by mid-August 2021, it has a high case fatality rate of 2.5% (158,697 deaths) translating to approximately 3.8%

of global deaths in 15% of the world's population (6, 7). By the beginning of August 2021, most countries had experienced at least three surges or waves of infections with brief respites of between 90–120 days between surges. Some countries are now experiencing a fourth wave of new infections (6, 7).

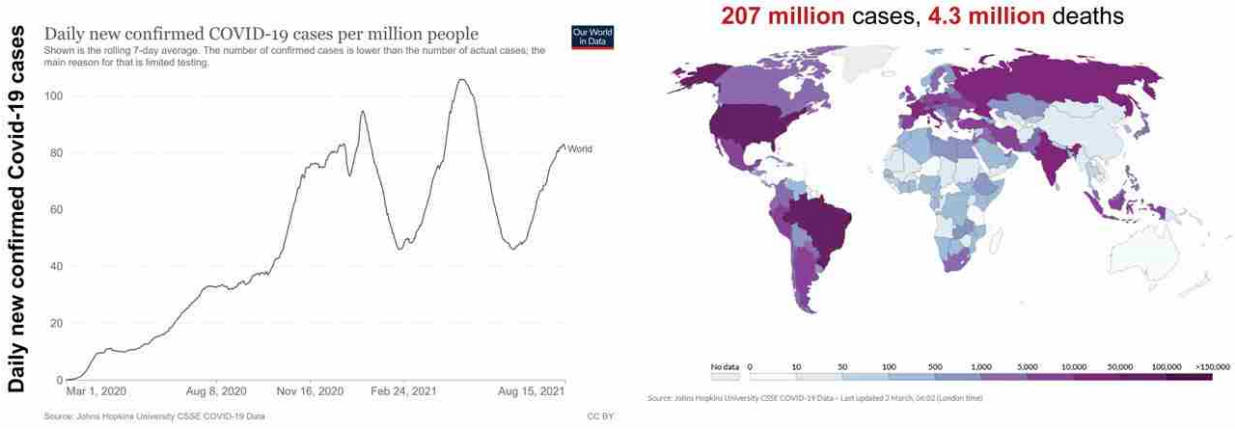


Figure 2 Temporal trends in global distribution of COVID-19 infections.

National responses to the pandemic have varied considerably and with the ongoing pandemic and waves of resurgences at a national level, continued to evolve in the first six months of the pandemic. Those countries, such as Australia, New Zealand, China, South Africa, and South Korea, that implemented strict national lockdown measures and closed ports of entry, were able to avoid widespread transmission whereas other countries, like the United States and Brazil, where limiting personal freedom was avoided on political grounds, experienced severe epidemics (23).

In September 2019, the Global Health Security Index (24) predicted that the world in general was not well prepared for the pandemic but did not predict individual country preparedness. In contrast to other pandemics and most diseases, in the first year of the SARS-CoV-2 pandemic rich countries suffered a higher burden of disease and death than most middle- and low-income countries (25). An analysis of the factors that appear to affect country epidemic trajectories and resilience indicates that the following factors were important:

The extent of globalization and geography: Most rich countries were more highly globalized and linked. Geography itself was a determinant such that island countries and those with only one border were more protected.
Health system capacity: Nations with universal publicly funded health systems that were not financially distressed and had strong public health capacity, appear to have been relatively well prepared for COVID-19, including Australia and New Zealand. Conversely, the pandemic highlighted the weaknesses of fragmented systems relying on for-profit healthcare providers, such as the USA.
Political leadership: Those countries with leaders who denied the scientific advice (including Brazil and the USA) fared badly. Others that acted on scientific advice fared better (e.g. South Africa, Australia, and New Zealand).
A country's wealth was not the only or main determinant of health security. Some lower-income countries made a conscious decision to allocate their scarce resources to prioritize their COVID-19 response including taking out loans to cover the increased debt resulting from the pandemic measures implemented (26, 27).
The extent of inequalities within countries also affected how well a country managed the pandemic. Those with high levels of inequities appear particularly vulnerable, as is discussed next.
Social security provisions to protect people from losing their jobs and homes was a vital part of pandemic responses (28)
An active civil society also played an important role in improving a country's pandemic responses. social solidarity built on civil society engagement offered protection even where trust in government was weak. For example, in South Africa, community action networks worked to both ameliorate the consequences of lockdown and reduce local transmission prevention (29). Using social media, they built local relationships based on trust and created a collective consciousness for responses to COVID-19 which has challenged divisive individualism.

6. Preventing the spread of SARS-CoV-2

Prior to the development of efficacious vaccines in November 2020, the only strategies effective in slowing the transmission of SARS-CoV-2 were the classical public health strategies of identification and isolation of cases and

quarantine of contacts. The collection of prevention measures to minimize the spread of SARS-COV-2, also sometimes referred to as the ‘COVID-19 prevention toolbox’ includes**:** symptom screening**;** staying home if symptomatic; hand sanitizing; regular cleaning of frequently touched surfaces and objects; use of face masks (cloth or medical masks) when interacting with others or in closed spaces; and maintaining social distance (>2 metres) (30). All healthcare workers need personal protective equipment (PPE), which includes combinations of respirators, surgical masks, face shields, goggles, gloves, plastic gowns or aprons depending on the risk of exposure in care provision.

Early in the pandemic, national lockdowns were an extreme form of social distancing used by some countries to slow down the spread of SARS-CoV-2 and utilize the time to prepare healthcare facilities to cope with the anticipated increase in patients with severe disease. Another form of social distancing adopted was remote working or schooling that was also referred to as voluntary ‘stay at home’ strategies. Restrictions on movement included policies against social gatherings including political rallies, sports events, and religious gatherings, entertainment venues, closing of shopping malls, and other circumstances or places in which people would come in close contact with others. In other situations where exposure to other people could not be avoided, especially in health facilities and other essential service points, masks and protective gear were strongly advised or mandated. The adoption of these prevention measures nationally or for individuals became a political issue; for example, in the United States of America the wearing of a mask was considered by many to be an expression of political affiliation (31). The implementation of any level of lockdown caused severe economic hardship, with many persons losing employment and many businesses closing down (32). In other settings, the enforcement of mandatory national lockdowns with curfews resulted in severe violations of human rights including loss of lives. Regardless of the measures adopted, the COVID-19 pandemic has resulted in the most severe economic and social disruptions and has resulted in societal divisions in many settings (33). Twenty months into this pandemic, implementation of national lockdowns is considered very parsimoniously by most governments and in fact the focus now is on striking a better balance between saving lives and livelihoods and living with the virus in our midst, implemented through varying degrees of restrictions based on magnitude of surges or declines of new infections. Notwithstanding

the public health rationale, we continue to see public protests, particularly in industrialized countries, when any form of lockdown measures are instituted.

The only way to identify someone who is infected is through viral diagnostic testing (21). The early development of diagnostic tests facilitated identification of cases. There are currently two types of diagnostic tests available for COVID-19 viral tests—polymerase chain reaction assays and rapid antigen tests. The timing of testing is critically important; testing too early after exposure or after the peak stage of infection, yields a negative test result (21) (Figure 3). The molecular real-time polymerase chain reaction (RT-PCR) test detects the virus's genetic material collected from respiratory specimens, such as nasopharyngeal, nasal, or oropharyngeal swabs, or sputum. False-positive results can occur with PCR tests, because non-viable genetic material may linger in the body long after recovery from an infection. False-negative results from antigen tests may range from 20 to 30%. Importantly, even in symptomatic individuals and their contacts a negative viral test only indicates that in that individual the virus is not detectable and repeat testing is necessary to accurately measure infection and contain further spread.

An antibody test confirms past infection and provides important individual or population level data of prior infection. It can take 1–3 weeks post-infection for detectable antibodies to emerge. Some serological assays differentiate between different classes of immunoglobulins with IgM detected earliest during infection followed by IgG. IgA is detectable at mucosal surfaces (26). Given that SARS-CoV-2 is a new virus, we are still learning about immune response and the durability of protection including vaccine-induced immunity. While it was initially thought that immunity from natural infection offered protection against re-infection for several months, the repeated waves/surges of infection and the emergence of variants of concern even in countries with high vaccination coverage rates indicate a more complex interplay between host immune responses and the evolution of viral immune escape. Anti-spike antibody tests are used to detect vaccine-induced immunity and are different from antibody tests used to detect natural infection.

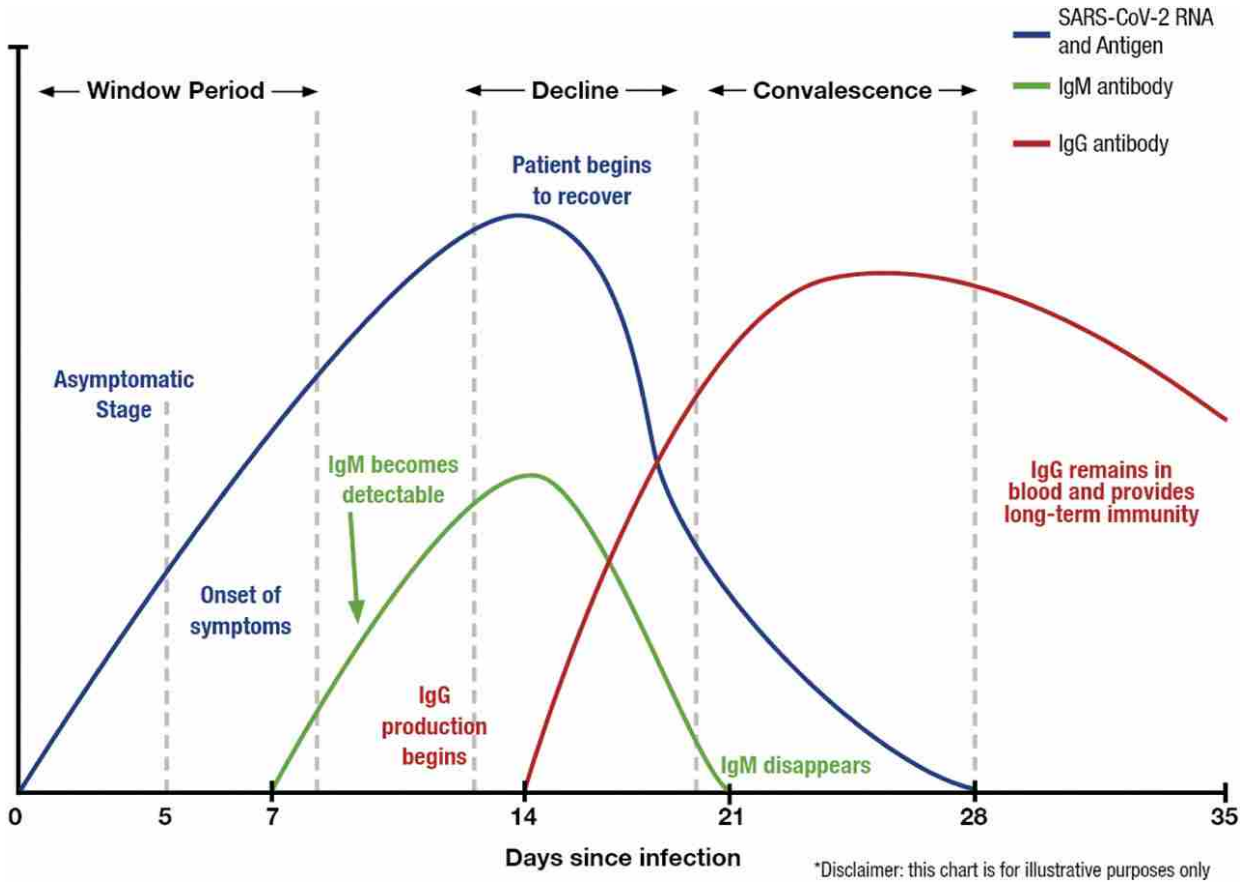


Figure 3 Detecting SARS-CoV-2 infection—viral and antibody kinetics.
Source: Senthuraman N, Jeremiah SS, Ryo A. Interpreting diagnostic tests for SARS-Cov-2. *JAMA* 2020, 323(22):2249–51.

Several factors contribute to the challenges of containment of SARS-CoV-2. These include:

Pre-symptomatic infectiousness where individuals unknowingly are spreading the virus prior to developing symptoms;
Asymptomatic infectiousness where individuals are unknowingly spreading the virus without ever knowing they have the infection;
Rapid spreaders/superspreaders where an infected person infects more than one other person or at events, such as large gatherings, where several hundred or thousand people can get infected from a few infected persons; for example, patient 31 in South Korea who through her social interactions infected over 1,000 people in a few days;
Repeated waves or resurgence of infections where after the initial epidemic is controlled a new surge of infections occurs creating an ever-present risk of new epidemics;
Emergence of new variants of concern that are able to escape natural or vaccine-induced immune responses.

Efforts to test both symptomatic and asymptomatic individuals through contact tracing and community testing in most countries were constrained by logistics, capacity, and delays in undertaking large numbers of RT-PCR tests; the need for repeated testing, the high cost of testing; supply chain issues and access to test kits due to global demand and travel/flight restrictions. In many settings, contact tracing had to be abandoned as diagnostic kits were correctly prioritized for confirmatory testing of hospitalized COVID-19 patients (34). The emergence of rapid, point-of-care diagnostic viral tests has assisted with rapid screening of large numbers of individuals leading to better surveillance and containment of viral spread in some settings. Measures of voluntary isolation of confirmed cases and a 10-day quarantine of potential contacts is commonly instituted to assist with viral containment efforts.

7. The natural history and disease progression of COVID-19

Infection with the SARS-CoV-2 results in a range of clinical manifestations ranging from asymptomatic infection to death. Figure 4 is a schematic representation of the natural history of COVID-19, from onset to recovery or death (35). There are basically three stages or phases in the natural history of COVID-19: the first phase is related to the onset of the disease and is generally characterized by the development of influenza-like symptoms from

mild to moderate. Some individuals recover and some progress to the second phase. Depending on the severity of phase 2, patients can improve or worsen with the necessity of intubation and mechanical ventilation. Phase 3 is characterized by hyperinflammation and sepsis of lungs. Patients with severe disease require treatment in an intensive care unit (ICU) and most of them, unfortunately, cannot overcome the infection and may eventually die.

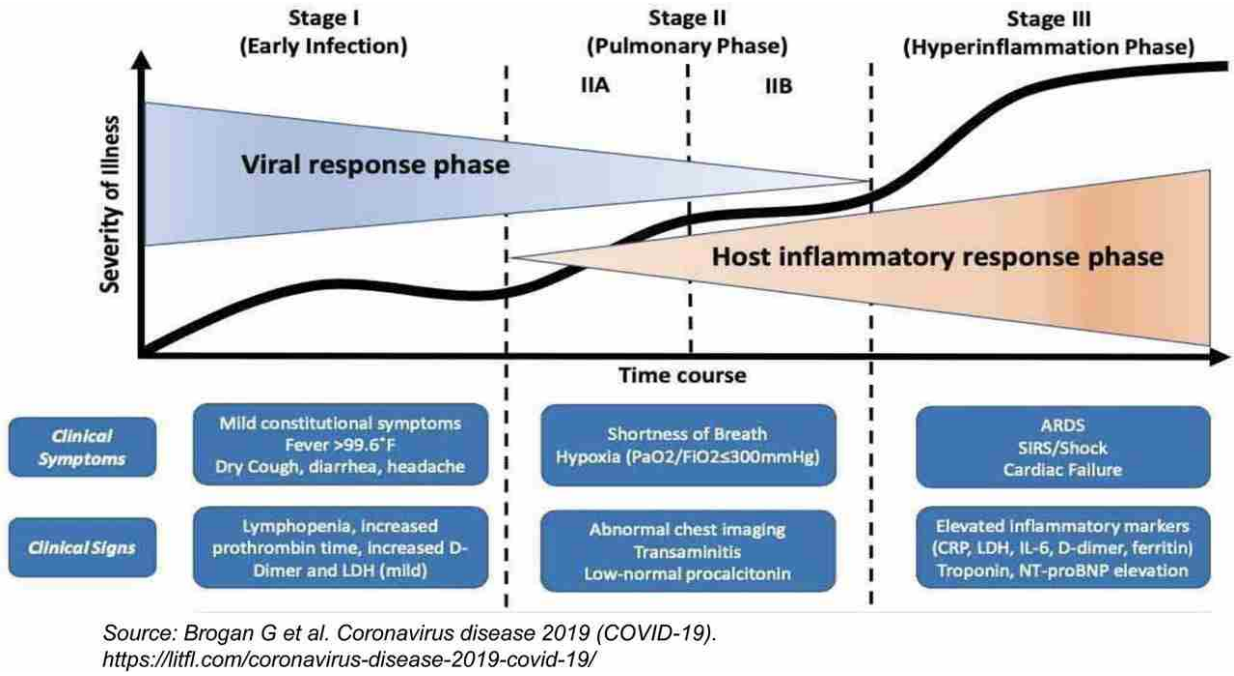


Figure 4 Natural history of COVID-19 infection.
Source: Brogan G et al. Coronavirus disease 2019 (COVID-19). https://litfl.com/coronavirus-disease-2019-covid-19/.

In the asymptomatic or pre-symptomatic stage, the individual has a positive nucleic acid amplification test but no symptoms consistent with SARS-CoV-2 infection. The individual may either resolve the infection or progress to a more serious stage of infection. An individual in the asymptomatic or pre-symptomatic stage can transmit the virus complicating surveillance and contact tracing. Some infected individuals progress to the mild stage, manifested by a range of symptoms including fever, cough, nausea, vomiting, malaise, headache, and diarrhoea and in some individuals loss of smell (anosmia) and/taste (ageusia) (Table 1) (36). Once the individual has developed shortness of breath, dyspnoea, and signs of lower

respiratory disease with an oxygen saturation level of 94% or higher, they are considered to be in the moderate stage of disease. Once their oxygen saturation level falls below 94% and they have a breathing frequency greater than 30 breaths per minute they are considered to have severe disease. In the critical stage patients progress to respiratory failure, septic shock, and death. The likelihood of serious disease is greater in individuals 65 years or older and those with underlying conditions including obesity, metabolic syndrome, type 2 diabetes, hypertension, cardiovascular disease, lung disease, smoking, cancer, kidney disease, or if immunosuppressed. The majority of SARS-CoV-2-infected individuals, particularly, younger individuals, recover completely. A proportion of infected persons go on to suffer from long COVID that manifests as varying persisting signs and symptoms of infection for weeks to months after infection. Long COVID disease can follow either subclinical or clinical disease (37).

Table 1 Severity of disease presentation

Severity of disease presentation	Signs and symptoms
Mild	Fever, cough, fatigue, anorexia, shortness of breath, myalgia. Other non-specific symptoms, such as sore throat, nasal congestion, headache, diarrhoea, nausea and vomiting, anosmia, and ageusia
Moderate	Above symptoms with clinical signs of pneumonia and SpO_2 >90%
Severe	Respiratory distress with SpO_2 <90%
Critical	Acute respiratory distress syndrome (ARDS)

Adapted from: WHO clinical guidelines of grading, 2020.

While not everyone who gets infected gets sick; not everyone who gets ill develops acute respiratory distress syndrome (ARDS) requiring mechanical ventilation. While bilateral multilobular lung involvement was very common in the early cases of COVID-19, we now know that COVID-19 is a multisystem disease affecting human beings from head to toes (Table 2) (38). SARS-CoV-2 is capable of infecting and affecting any organ in which its receptor is present. Angiotensin-converting enzyme 2 (ACE-2) embedded

in the surface membrane of many cells is the functional receptor for SARS-CoV-2 to gain access to human cells. The ACE-2 receptor is highly expressed in lung epithelial cells, kidneys, gastrointestinal tract, liver, vascular endothelial cells, and arterial smooth muscle cells. Thus, all of these organs and systems with high expression of ACE-2 receptors are targets for SARS-CoV-2 infection. COVID-19 may induce a widespread inflammatory response involving all these organs. (39)

Table 2 Multisystem involvement of COVID-19: a disease from head to toe

Organ system	Signs and symptoms
Neurologic	Headaches, dizziness, encephalopathy, Guillain–Barre, ageusia, myalgia, anosmia, stroke
Thromboembolism	Deep vein thrombosis, pulmonary embolism, catheter-related thrombosis
Renal	Acute kidney injury, proteinuria, haematuria
Cardiac	Takotsubo cardiomyopathy, myocardial injury/myocarditis, cardiac arrhythmias, cardiogenic shock, myocardial ischaemia, acute cor pulmonale
Hepatic	Elevated aminotransferases, elevated bilirubin
Endocrine	Hyperglycaemia, diabetic ketoacidosis
Gastrointestinal	Diarrhoea, nausea/vomiting; anorexia, abdominal pain
Dermatological	Petechiae, livedo reticularis, erythematous rash, urticaria, vesicles, pernio-like lesions

Adapted from: Feng, Z., et al *Nature Communications*, 2020, 11(1), pp.1–9.

7.1. Predictors of severe disease and mortality: age and comorbidities

There is growing evidence that COVID-19 can also itself either cause or exacerbate metabolic syndrome (pre-diabetic condition) and hypertension (40). Clinical management for hospitalized patients with COVID-19 is focused on supportive care for complications, including supplemental oxygen and advanced organ support for respiratory failure, septic shock, and multiorgan failure. Most mortality attributable to SARS-CoV-2 infection

occurs through the development of viral pneumonia-induced ARDS. While the exact mechanisms through which SARS-CoV-2 causes ARDS and how certain host factors confer an increased risk of developing severe disease remains unclear, one factor has emerged as a dominant predictor of disease severity and risk of mortality—which is age (41). In an analysis of more than 44,672 confirmed cases by the Chinese Centre for Disease Control and Prevention, the case fatality rate was <1% among <60 years old; 3.6% for those 60–69 years; 8.0% in people 70–79 years; and 14.8% among those 80 years or older (41). Other predictors of disease progression include malnutrition (both underweight and overweight/obesity), and non-communicable diseases (NCDs), including hypertension and diabetes (40, 42).

Infection with SARS-CoV-2 triggers an inflammatory process that can lead to direct or indirect insults to the lung. About 5–15% of patients with COVID-19 infection require intensive care monitoring and ventilatory support, mainly for severe hypoxemia (43, 44). The goal of treatment is maintenance of oxygen saturation >90%. That means supplementation of oxygen via several ways. There are no clear evidence-based guidelines for the ideal time to proceed to mechanical ventilation in patients with COVID-19. Availability of ventilators, intensive care capacity, considerations of palliative care, and end-of-life resources as well as individual patient characteristics all play a role in decisions to institute mechanical ventilation. There are several strategies for delivering oxygen as presented in Table 3.

Table 3 Strategies for delivering oxygen in moderate to severely ill COVID-19 patients

Strategies for delivering oxygen	Indication
Nasal cannula	Effective for mildly hypoxic patients, supplemental oxygen delivered by nasal cannula can induce significant dispersion of exhaled air, even at low flow rates. Supplemental oxygen by nasal cannula provides up to about 5–6 L/min
Simple face mask	Oxygen supplementation via simple mask is only marginally greater than that of nasal cannula. Simple masks provide supplemental oxygen with flow rates up to approximately 5–10 L/min
Continuous positive airway pressure (CPAP)	Delivers pressure through all phases of the respiratory cycle. The patient continues to breathe spontaneously both with and against the positive airway pressure
Mechanical ventilation	Endotracheal tube for patients may be necessary for patients with frank respiratory failure or multisystem organ dysfunction. The machine takes over completely or supports the patients spontaneous breathing
High flow oxygen system	High flow nasal oxygen (HFNO) includes high flow nasal cannula and high velocity nasal insufflation. High flow oxygen systems provide oxygen-rich heated humidified gas to the patient's nose at flow levels sufficient to deliver a constant, precisely set high FiO_2

7.2. Clotting and strokes

Strokes occur when the brain's blood supply is interrupted, usually by a blood clot. COVID-19 causes complications to younger patients being hospitalized for, and sometimes dying from, serious strokes (45, 46). It is unknown if coronavirus itself stimulates blood clots to form, or if they are a result of an overactive immune response to the virus. These strokes are found in patients who test positive for coronavirus but who do not have any traditional risk factors for stroke. They tend to have no COVID-19 symptoms, or only mild symptoms. The type of stroke occurring in these patients typically occurs in much older patients. COVID-related strokes occur because of a body wide increase in blood clot formation, which can damage any organ, not just the brain. A blood clot in the lungs is called

pulmonary embolism and can cause shortness of breath, chest pain, or death; a blood clot in or near the heart can cause a heart attack; and blood clots in the kidneys can cause kidney damage requiring dialysis.

7.3. Long COVID

Fatigue, muscle weakness, insomnia, anxiety, loss of smell or taste persist beyond several weeks for a significant proportion of those infected, including those who were mildly symptomatic. Persistent cardiac symptoms have also been observed. A Norwegian (47) study found at 6 months, 61% (189/312) of all patients had persistent symptoms. Given the short experience (less than two years) with SARS-CoV-2 disease there is still uncertainty about the frequency, manifestations, and severity of persisting COVID-19 disease/long COVID or its management (37). Given the growing cohorts of persons with long COVID, this is becoming an area of increasing concern and requires urgent attention.

7.4. Children and COVID-19

To date, data on COVID-19 in children and adolescents remain scarce. Young people can be asymptomatic or may present with fever, respiratory, and/or gastrointestinal symptoms. Recent studies have demonstrated that COVID-19 is generally a mild disease in children when compared with older patients. It is not yet known what the reasons for the difference in severity of disease are, but three possibilities exist (48): i. Children may have lower viral loads even if they get COVID-19. There may be a correlation between the severity of COVID-19 and the viral loads (or the duration of virus-shedding period); ii. Children have lower expression levels of ACE2, when compared to adults, and, therefore, have fewer receptor sites for the SARS-CoV-2 to bind to; iii. There may be a simultaneous presence of other viruses in the mucosa of the lungs and airways, common in young children, and these viruses may compete with SARS-CoV-2 virus and limit its growth.

A small proportion of children develop a rare, COVID-19 associated multisystem inflammatory syndrome which appears to develop after the infection rather than during the acute stage of COVID19. The clinical features of these paediatric cases are both similar and distinct from other well-described inflammatory syndromes in children, including Kawasaki

disease, Kawasaki disease shock syndrome, and toxic shock syndrome. This COVID-19associated multisystem inflammatory syndrome in children and adolescents is referred to interchangeably as paediatric inflammatory multisystem syndrome temporally associated with SARS-CoV-2 (PIMSTS) or multisystem inflammatory syndrome in children (MISC) associated with COVID19 (49). Symptoms of MIS-C in children include: conjunctivitis; loss of appetite; red, cracked lips or red tongue; enlarged lymph nodes; prolonged fever over several days; diarrhoea and/or vomiting; urticaria; abdominal pain; swollen hands and feet. Of greater growing concern for children is the impact of the death of parents and the increased numbers of orphaned children.

7.5. Treatment

The priority in treatment has been saving the lives of those with severe illness. Currently there are a limited number of drugs that have clinical trial level evidence against SARS-CoV-2. A comparison of the WHO clinical care guidelines (50) to the NIH panel recommendations for treatment of SARS-CoV-2 (Table 4) (51) highlights huge disparities in what drugs can be accessed where leading to huge differences in COVID-19 mortality rates between industrialized countries and LMICS.

The recommendations from the NIH Panel on Treatment for SARS-CoV-2 in terms of the treatment strategy and recommended drugs is dependent on the stage of disease (51). Most infected individuals will have few symptoms and will require primarily symptomatic treatment including rest, fluid replacement, and pain management with non-opioids such as acetaminophen. They do not require hospitalization but may require monoclonal antibodies if at risk of progression. Patients developing respiratory distress and shortness of breath, chest pain, new onset confusion, inability to stay awake, a bluish caste to their face and other signs of hypoxia require hospitalization. Treatment of individuals with more severe disease is based on physician judgement and it may be necessary to put patients on artificial ventilators and oxygen therapy.

Table 4 NIH panel recommendations for treatment of SARS-Cov-2

Patient condition	Treatment recommendation
Mild to moderate symptoms, not requiring hospitalization or supplemental oxygen but at risk of progression	Anti-SARS-CoV-2 monoclonal antibody products such as casirivimab plus imdevimab or sotrovimab. Do not use dexamethasone or other systemic glucocorticoids
Discharged from hospital despite requiring new or increasing need for supplemental oxygen	6 mgs of dexamethasone once daily for duration of supplemental oxygen but less than 10 days
Hospitalized but does not require supplemental oxygen	No corticosteroids. If at high risk for progression may use remdesivir
Hospitalized and requires supplemental oxygen	Remdesivir, remdesivir plus dexamethasone if increasing need for oxygen, dexamethasone if remdesivir not available or contraindicated
Hospitalized and requires oxygen through high flow ventilator	Dexamethasone or dexamethasone plus remdesivir. If systemic inflammation, add baricitinib or tocilizumab
Hospitalized and requires intensive mechanical ventilation or extracorporeal membrane oxygenation	Dexamethasone. If within 24 hours of admission plus tocilizumab

Adapted from NIH guidelines (https://www.covid19treatmentguidelines.nih.gov).

Given the burgeoning of the antiviral drug field in recent years, it was natural for researchers to turn to proven agents successful against other viruses to attempt to treat COVID-19. Data from small observational studies of chloroquine and hydroxychloroquine (CQ and HCQ) looked promising but the UK's adaptive multiarmed randomized controlled trial, RECOVERY (52), and the WHO's multicountry Solidarity treatment trial demonstrated that neither HCQ nor lopinavir/ritonavir had any effect in reducing mortality in COVID-19 (53).

The RECOVERY trial did have a COVID-19 treatment breakthrough in demonstrating that dexamethasone reduced mortality in hospitalized patients on supplemental oxygen. Patients requiring ventilation had a 33% reduction in death. Dexamethasone or its equivalent, e.g. prednisone, is now standard of care for severe, hospitalized COVID-19 patients, worldwide. Hopes for remdesivir, a repurposed antiviral nucleoside analogue first used in Ebola patients, were also high following a US study that showed that it

significantly reduced recovery time if people were treated early (11 days vs. 15 for untreated) but was less effective in severely ill patients (54). The Solidarity therapeutics trial demonstrated no effect of remdesivir on mortality rates in severe COVID-19. There have also been reports of success in reducing the death rate among in severely ill patients by use of infused convalescent plasma, which prompted the FDA in the USA to issue an 'emergency use authorisation (EUA)' to allow hospitals to use it (55).

Monoclonal antibodies are recommended for mild-to-moderate disease but cost and access remain barriers to widespread access. The RECOVERY trial also showed mortality benefit for REGEN-CoV in hospitalized adults (56), RECOVERY (57), and REMAP-CAP (58) have shown benefit of tocilizumab.

There remains controversy about drug use and at what stage of disease they should be used. Of concern is the promotion through social media platforms and by some clinicians of several unproven drugs notably ivermectin and hydroxychloroquine.

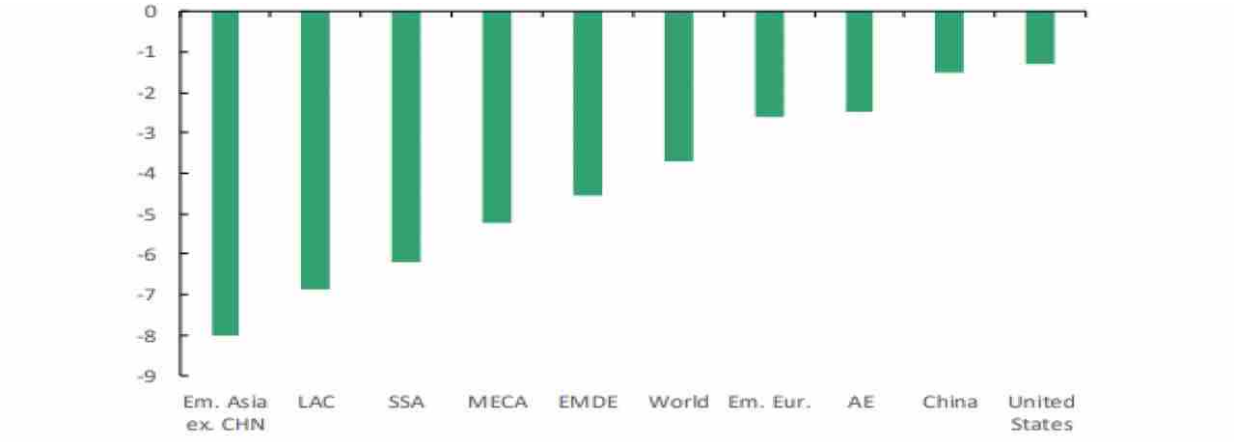
As data from rigorously and robustly conducted trials emerge, the current WHO and NIH treatment guidelines will change. The challenge for benefit will be access. Importantly, in contrast to the substantial investments and advances made in vaccine development (Section 10.2), treatment has had less attention or investments.

8. Social and economic impact of the pandemic

The dichotomy of this pandemic lies in the struggle between saving lives and saving livelihoods. With 90% of the world's countries in recession there has been an unprecedented economic downturn, possibly exceeding even that of the great depression in the 1930s. Some indication of the severity of the social and economic impact of COVID-19 can be gleaned from this quote from the Organisation for Economic Co-operation and Development (OECD): '*The COVID-19 pandemic has triggered the deepest economic recession in nearly a century, threatening health, disrupting economic activity, and hurting well-being and jobs. Extraordinary policies are required to walk the tightrope towards recovery, which will shape the economic and social prospects of the coming decade*' (59).

The COVID-19 pandemic has created an unprecedented economic and public health crisis. The economic impacts have resulted from reduced

productivity, loss of life, business closures, trade disruption, interruption of global supply chains and decimation of the travel and tourism industry (60). The US Congressional Research Service (61) estimates that '… the virus reduced global economic growth in 2020 to an annualized rate of –3.4% to –7.6%, with a recovery of 4.2% to 5.6% projected for 2021. Global trade is estimated to have fallen by 5.3% in 2020 but is projected to grow by 8.0% in 2021'. Further, it notes that the recovery is likely to remain uneven among countries with differences driven by the extent of SARS-CoV-2 vaccine rollout, ability to provide fiscal and monetary support, the extent of hard-hit sectors (especially tourism) and predict that emerging economies will face the most problems because they do not have access to COVID-19 vaccines and have lower resources, providing some indication of how COVID-19 will entrench and deepen existing global economic inequity (Figure 5). While virtually every country covered by the International Monetary Fund (IMF) posted negative economic growth in 2020, the downturn was more pronounced in the poorest parts of the world (62, 63).



Source: IMF-WEO Apr-2021. Note: AE = Advance economies; Emerging Asia ex. CHN = emerging and developing Asia excluding China; EM. Eur = Emerging and developing Europe; LAC = Latin America and the Caribbean; MECA = Middle East and Central Asia; SSA = sub-Saharan Africa.

Figure 5 Global GDP growth 2020.

A McKinsey survey of executives in June 2021 (64) showed that those countries with 'developed' economies were more optimistic about the risk posed by the SARS-CoV-2 pandemic to their domestic growth than those countries with 'emerging' economies—an about turn from their projections

in June 2020. This is most likely because more OECD countries have higher vaccination rates than LMICs.

Yeyati and Filippini (65) note that the pandemic triggered an unprecedented fiscal response in terms of speed and magnitude. Globally, the fiscal support reached nearly USD 16 trillion (around 15% of global GDP) in 2020. They point out that the capacity of countries to implement such measures varied significantly and noted the different capacities country have to support households and firms which depended on access to international financial markets. They also noted that the capacity of states to rapidly and efficiently implement policies to support household and firms, requires substantial state capacity and well-developed tax and transfer infrastructures, which many LMIC do not have.

The report goes on to note that the fiscal and monetary policies of government meant the 2020 economic turndown was not as severe as forecast. These policies included support for businesses to retain workers and additional support for those who were or became unemployed. For instance in March 2021, the UK Government announced a £65 billion stimulus package over two years to revive the UK economy through various business incentives and worker income support measures to be followed by tax increases starting in 2023 (66). At the same time the US Congress passed a $1.9 trillion economic stimulus bill, designated the American Rescue Plan (67). Such moves demonstrated the ways in which the pandemic upended the neoliberal consensus that has been in ascendance since 1990 when UK Margaret Thatcher and US President Regan instituted neoliberal policies which drastically cut public spending and ‘rolled back the state’. This illustrates to what extent a public health crisis can have a dramatic impact on the economic policy of a country.

A congressional report also noted in July 2021 that new variants of the COVID-19 virus and a surge in diagnosed cases in countries like India and Indonesia and resistance to vaccinations among some populations in developed economies means it was hard at that point to predict the speed and strength of an economic recovery.

During the pandemic there were passionate debates about whether the economic sacrifice necessary to impose public health control measures would create more health problems than the pandemic (68). Such dilemmas are not new in public health where those who perceive a loss of profits often oppose public health constraints. The experience of countries where public

health measures were not taken (e.g. Brazil and the USA in 2020) demonstrate that the result is a very high death toll and causes extreme pressure on health systems. Ultimately the decision of whether to protect health rather than profits is a moral one. Public health ethics dictate that health is the primary consideration.

The COVID-19 pandemic has laid bare the many inequities that existed before it started. The Independent Panel for Pandemic Preparedness and Response report in May 2021 (69) captures this as follows: 'The lack of planning and gaps in social protection have resulted in the pandemic widening inequalities with a disproportionate socioeconomic impact on women and vulnerable and marginalized populations, including migrants and workers in the informal sector'. These inequities concern those within as well as between countries and reflect structural drivers of ill-health. The inequities between countries have been most noticeable in the vaccine distribution, drugs, and equipment necessary to diagnose, prevent, and treat COVID-19; this is discussed in greater detail next. Vaccine inequities, for example, will lead to a range of social and economic inequities affecting countries' ability to return to more normal life, thereby exacerbating the existing inequities between these countries.

Within countries inequities were also apparent according to: ethnicity and race; being a migrant or refugee; employment and economic status; gender. While limited data are available largely from small studies with several confounders the intersection of race, gender, and deprivation or marginalization is not surprising.

8.1. Ethnicity and race

Globally, COVID-19 has shown a strong tendency to affect minority populations more than other groups. Table 5 shows that in the US while Black, Indigenous peoples, and Hispanic or Latino populations were only slightly less likely to get COVID-19, they were much more likely to be hospitalized or die from it. The Centers for Disease Control and Prevention (CDC) notes that race and ethnicity are risk markers for other underlying conditions that affect health, including socioeconomic status, access to healthcare, and exposure to the virus related to occupation, e.g. frontline, essential, and critical infrastructure workers. In the UK the ethnic and racial differences have been reported as being even greater, as illustrated in a

prospective cohort study that demonstrated a more than 4-fold higher risk of COVID hospitalization for Blacks vs. a 2-fold increased for Asians and other non-white groups compared to whites, after adjusting for age and sex (70).

Indigenous peoples are at greater risk of harm from COVID-19 primarily because of their existing poorer health status, and because 'they have been made fragile by the legacy and persistence of colonial practices'. In Brazil, where the overall death rate was already one of the highest in the world, for Indigenous people the death rate was even higher (71). It is notable that Aboriginal Community Controlled Health Services in Australia have been very successful in protecting the health of Aboriginal people, especially those living in remote areas through developing a range of resources and a National Management Plan for Aboriginal and Torres Strait Islander Peoples (72). This suite of measures to protect population groups with many vulnerabilities has been demonstrably effective.

Table 5 Risk for COVID-19 infection, hospitalization, and death by race/ethnicity—United States, July 2021

Race and ethnicity are risk markers for other underlying conditions that impact health				
Rate ratios compared to white, non-Hispanic persons	**American Indian or Alaska Native, Non-Hispanic persons**	**Asian, non-Hispanic persons**	**Black or African American, non-Hispanic persons**	**Hispanic or Latino persons**
Cases	1.6×	0.7×	1.1×	2.0×
Hospitalization	3.3×	1.0×	2.9×	2.8×
Death	2.4×	1.0×	2.0×	2.3×

Source: Centre for Disease Control and Prevention, COVID-19 web page https://www.cdc.gov/coronavirus/2019-ncov/covid-data/investigations-discovery/hospitalization-death-by-race-ethnicity.html (accessed 18 July 2021).

8.2. Migrants and refugees

Refugees and migrants have been disproportionately affected by the SARS-CoV-2 pandemic. WHO reported that the pandemic has triggered many different adaptation policies related to international migration that affect

public health, directly or indirectly. They note that the overall picture is ‘highly complex and disparate, as national reactions to the global pandemic vary considerably in both nature and scale, impairing, in turn, a truly coordinated and consistent public health response’ (73). In addition to experiencing a range of health conditions including NCDs, mental health, and a high rate of injuries, they also may also experience highly insecure living conditions and lifestyles. Very often they also suffer discrimination, poverty, poor housing and education, poor employment practices, often without access to essential health and other services including to COVID-19 vaccines. WHO stresses that better international coordination is required to safeguard the human rights of refugees and migrants and this is particularly the case given the additional stress of the COVID-19 pandemic.

8.3. Employment and economic status

The International Labour Organisation reported in 2021 that the COVID-19 pandemic has caused unparalleled disruption to employment and livelihoods worldwide. Workers whose labour market position was disadvantageous prior to the crisis—women, young people, migrants, informal workers, and workers in lower-skilled occupations—suffered disproportionately (74). People working in casual and informal employment have been more likely to be exposed to COVID-19. Some forms of work pose multiple risks (see Figure 6 for the example of meat workers). The case of meatworkers highlights the ways in which a range of vulnerabilities—casual work, migrant status, low wages, home environment, low occupational health and safety conditions—converge to make a particular group of workers very vulnerable to COVID-19.

- Significant Covid-19 outbreaks have affected meat workers in Brazil, the UK, USA, Ireland, France, Australia and Germany. (75)
- Under “normal” circumstances, working conditions in slaughterhouses cause workers mental, psychological and physical harm (76) and Covid-19 intensifies these existing health risks.
- Many meatworkers workers are immigrants and refugees who may be undocumented and at risk of deportation and so are reluctant to unionize or challenge exploitative labour practices (77-79)
- Additionally, workers are often dependent on their employers for housing and sometimes transport (80, 81)
- Their relatively low wages, and the housing discrimination that racial minorities, low-income and immigrant workers face, constrain their ability to seek alternative housing.
- The physical configuration of slaughterhouses (i.e. working in close proximity, the need to shout to be heard) and communal housing and transport make social distancing near impossible (82)

Figure 6 Meatworkers and susceptibility to COVID-19

8.4. Gender

A study in China found (83) that while men and women have the same prevalence, men with COVID-19 are more at risk for worse outcomes and death, independent of age. A global data base (84) reveals that females are more likely to get tested (57% of those tested are female), equally likely to be diagnosed with COVID-19 (50% of cases), less likely to be hospitalized (48% of hospitalizations), less likely to be admitted to an ICU (31% of admissions), and less likely to die (43% of deaths) compared with males. In terms of social determinants (which will affect health in the longer term) the impacts of pandemic-induced social and economic change appear more severe for women. Nearly 40% of employed women work in the hardest hit industries including the informal sector, which is dominated by women workers (85). There are approximately 740 million women working in the informal sector and their incomes saw a 60% reduction in the first month of the pandemic (86). Other affected female-dominated sectors include arts and entertainment, accommodation and food services, wholesale and retail trade and manufacturing (garment industry). Women also provide three-quarters of unpaid care work that has increased due to closure of schools and childcare services, increased care needs for aged and overwhelmed health systems (87). A UN Women report (88) highlights the intensification of violence against women as a result of COVID-19 containment measures and as the economic impact starts to unfold. Of concern is that globally 243 million women and girls aged 15–49 years have been subjected to sexual or physical violence perpetrated by an intimate partner and describe this surge in

violence against women and girls as a ‘shadow pandemic’ linked to COVID-19 and call for national COVID-19 responses to include services to mitigate this pandemic.

8.5. Healthcare workers

Health workers are one group of workers who are at high risk of contracting COVID-19. The Independent Panel for Pandemic Preparedness and Response (69), notes that at least 17,000 health workers died of COVID-19 in the first year of the pandemic and said this underscored the need for countries to do much more to support and protect them. Of note is that 70% of those employed in health and social work are women and they often hold lower-skill and lower-paying jobs. Their working conditions put them at greater risk of infection, especially in LMICs (89). Community health workers (CHWs) are mostly women, and often from rural and lower socioeconomic backgrounds, paid irregular salaries/stipends/incentives, especially in LMICs. CHWs have been an essential part of the pandemic response and they have played a crucial role in public health systems. They have undertaken COVID-19 surveillance, contact tracing, monitoring quarantine and isolation, along with their regular tasks. Despite the high risk, they often work without adequate PPE or training (90, 91, 92).

8.6. Children

UNICEF estimates 42–66 million children could fall into extreme poverty as a result of the pandemic adding to 386 million children already in extreme poverty in 2019 (93). With the second, third, and fourth surges mortality rates have increased in 40–60-year-old people who are typically the main breadwinners and likely have young children. The impact of premature loss of lives on child orphans remains to be quantified.

8.7. Digital divides

When many are confined to the home and many workplaces are forced to convert employees to working from home, the divide between those with internet access and the necessary technology to continue their work remotely and those that don’t becomes even more pronounced. Under the

circumstances of the COVID-19 pandemic, individual access to information is critical for coping both with work for breadwinners, and with education for young children looking to further their prospects in life. Roughly 60% of the global population has internet access however, in sub-Saharan Africa, 82% of students do not have internet access and 90% don't have access to a household computer (94). Putting percentages into more direct numbers, there are more than twenty-one million people in South Africa alone who do not have internet access. (95).

8.8. Mitigating the impact of COVID-19

A Congressional report (96) estimates that 95 million people may have entered into extreme poverty in 2020 with 80 million more undernourished compared to pre-pandemic levels. The International Labour Organization (ILO) (97) identifies four pillars of action designed to lessen the impact of COVID-19 on businesses, jobs, and people in vulnerable circumstances: i. Stimulation the economy and jobs; ii. Supporting enterprises, employment, and incomes; iii. Protecting workers in the workplace; iv. Using social dialogue between government, workers, and employers to find solutions. The ILO views these pillars as vital to using the response to COVID-19 as a chance to ensure decent working conditions.

9. The colliding pandemics of HIV and SARS-CoV-2

SARS-CoV-2 is one of several pandemics facing humanity today. Here we focus on HIV noting that advancing HIV disease is closely intertwined with Tuberculosis and sub-Saharan Africa bears a disproportionate 70% of the global burden of HIV infection (see Chapter 8.13, Acquired immunodeficiency syndrome [AIDS], for more details). Importantly, in recent times, the global response to the HIV pandemic set a precedent in responding to pandemics, for example, the establishment of UNAIDS (the joint United Nations Programme for HIV/AIDS). This new global solidarity to responding to global health crises supplemented with substantial multilateral and bilateral collaborations and capacity building efforts enabled global equity for access to affordable life-saving anti-retroviral treatment that transformed AIDS from an inevitably fatal condition to one that is chronic and manageable. Substantial increases in life expectancy in excess

of 12–15 years particularly in countries in Africa have been achieved since the establishment of the Global Fund for AIDS, TB, and Malaria and the USA President's Emergency Program for AIDS Relief (PEPFAR) (98). In 2020 there were approximately 38 million people living with HIV, 690 thousand deaths, 1.5 million new infections, 32 million HIV-infected individuals on treatment and transmission rates to infants reduced to <2% (99). Despite predictions early in 2020 that sub-Saharan Africa would be devastated by SARS-CoV-2 that did not happen during the first surge of infections (6, 7). However, since the latter half of 2020, with each new surge of SARS-CoV-2 infections the impact on sub-Saharan African countries has worsened. The ongoing HIV pandemic in the midst of the emerging and evolving SARS-CoV-2 pandemic provides important insights on addressing parallel pandemics and the potential cost and short-sightedness of abandoning work on an existing pandemic in favour of a new pandemic.

9.1. The use of existing resources

The technology, infrastructure, and systems developed to control and treat HIV served as a springboard to accelerate the response to COVID-19 in Africa. For example, diagnostic equipment such as PCR and gene-expert platforms have become commonplace in primary healthcare clinics and other health settings for HIV viral load testing and enabled the rapid rollout of SARS-CoV-2 RT-PCR diagnostic testing. The community outreach programmes developed for HIV and TB contact tracing and case management enabled the rapid mobilization of hundreds of thousands of healthcare workers in Africa to assist in COVID-19 education, screening, testing, contact tracing, isolation, and quarantine efforts. Similarly, the overall strengthening of healthcare delivery systems and human resource capacity to enable massive rollout of anti-retroviral (ARV) treatment for AIDS patients facilitated rapid pivoting in health facilities to deal with the large numbers of severely ill COVID-19 patients. Clinical staff already trained in infectious diseases meant that there was a critical mass of skilled personnel who could and did provide expert advice to guide national and local responses to COVID-19 across Africa. The same genetic sequencing platforms designed to monitor trends in ARV drug resistance and understand HIV viral transmission dynamics (100) enabled a faster and more reliable identification of variants of SARS-CoV-2 and rapidly strengthen molecular

surveillance efforts including for outbreak investigations (101, 102). Clinical trial infrastructure developed to evaluate HIV prevention technologies, including potential HIV vaccines, pivoted rapidly to evaluate COVID-19 vaccine trials and establish COVID-19 vaccine rollout platforms (103).

9.2. The complications of two concurrent pandemics

In many countries in Africa, interventions to prepare for the anticipated increase in hospitalization for severe COVID-19 cases or to slow the spread of SARS-CoV-2 such as national lockdowns had unanticipated negative effects on responses to HIV. For example, testing and treatment for both HIV and TB dropped substantially as patients were apprehensive about visiting a clinic or hospital where they faced the risk of acquiring SARS-CoV-2 or where access to these services were restricted. This increases the risk of new infections, and/or the emergence of drug-resistant variants of HIV because of ARV treatment interruptions. While there is limited data to quantify the magnitude of these effects, a study undertaken in South Africa during national lockdown showed that weekly testing for HIV viral loads dropped by 22%, and CD4+ cell count monitoring dropped by 33% (104). In addition to this impact on HIV patients in care in public sector health facilities, the number of new HIV tests being done each month declined nearly 50% during lockdown, as did the number of antiretroviral treatment (ART) initiations negatively (105) impacting the country's efforts to reach the '90–90–90' targets towards achieving the UN 2030 goal of achieving epidemic control of HIV (106). Notably there were no marked changes in ART collections suggesting that those patients already on medication remained adherent (105). Access to other essential services such as pre-natal care, access to sexual reproductive health services, and other chronic conditions were also severely restricted. However, innovations of multi-month dispensing, community collection points for medication, and community-based buddy/peer adherence support clubs ensured continuity of care and minimized treatment interruptions for HIV and other chronic conditions (107).

9.3. Does HIV enhance the risk of acquiring SARS-CoV-2?

Given that HIV attacks the immune system of infected individuals, this is an important question. The answer to this question was a bit more complicated for several reasons. Firstly, in contrast to Africa, industrialized countries have low HIV infection rates that are concentrated in key populations (99). Therefore, in the first half of 2020 little data was available on whether HIV was a risk factor for SARS-CoV-2 infection. Secondly, 65% of the population in Africa is under 35 years old and for most of 2020 for SARS-CoV-2 infection, older age was regarded as an important risk factor and this partly explained why the number of infections in Africa remained relatively low (6, 7). However, as SARS-CoV-2 infection became more established in Africa, data on HIV risk has started to emerge (108, 109, 110). These data suggest that HIV positive individuals (PLHIV) who are virally suppressed as a result of being an ART and who have immune reconstitution viz a CD4 count >500 have a similar risk of acquiring SARS-CoV-2 as HIV uninfected persons. However, PLHIV who are not on ARV treatment or have treatment interruptions, or have low CD4 counts (i.e. severely immunocompromised) and high HIV viral loads were more likely to acquire SARS-CoV-2 infection; shed SARS-CoV-2 for longer durations and experience a more severe clinical course of COVID-19 (108). Specifically, PLHIV with low CD4 cell counts, or untreated HIV infection have higher mortality rates, and those with CD4 counts<200/ml have higher rates of ICU admissions that require mechanical ventilation. A retrospective study conducted in the USA (109) found a 2.9-fold higher death risk for patients living with HIV, with a 2–4-fold confidence interval. Of note in this study was that individuals in this cohort who were Black, male, and economically deprived, had 4.3-fold higher mortality rate compared to the 1.8-fold higher mortality rates noted for other HIV infected persons (109). At least 8% of COVID-19 deaths are attributed to co-infection with HIV (111, 112, 113).

9.4. Uncontrolled HIV infection fuelling the evolution of SARS-CoV-2 variants of concern and compromising vaccination efforts

Individuals with advanced HIV and ART failure or who are on treatment but still immune-compromised (i.e. CD4<500) are likely to have prolonged COVID-19 infection. This extended COVID-19 infection can be reduced with effective HIV ARV treatment and once immune reconstitution reaches

CD4 counts >500 it results in clearance of SARS-CoV-2 infection (114). However, while HIV infection persists, SARS-CoV-2 replicates freely and mutations constantly emerge within those immunocompromised patients. This is clearly demonstrated in the early emergence of the N501Y transmissibility substitution and escape mutations such as the E484K substitution HIV infected patients (114). A HIV stratified analysis in a matched case-control study demonstrates that substantially lower concentrations of both IgG and pseudo-virus neutralizing antibody titres after SARS-CoV-2 infection in PLHIV, that might explain the diminished immune response to SARS-CoV-2 infection (115).

What are the implications of a compromised immune response in HIV infected for vaccine efficacy? Preliminary data from a longitudinal analysis of a severely compromised HIV infected patient with uncontrolled viraemia (VL >800,000 copies; CD4< 20 cells/ml) showed no response after two doses of the BNT162b2 vaccine (116). The patient had no IgG reaction against the spike protein, no SARS-CoV-2 specific neutralization, and no quantifiable spike protein specific T cells. Given the high burden of HIV infection in sub-Saharan Africa, this has important implications for COVID-19 vaccination efforts suggesting that an additional vaccine dose is needed in this group and should be considered in the ongoing responses to the SARS-CoV-2 pandemic. Of note is more recent data demonstrating that with well-controlled HIV, similar immune responses are observed as in HIV uninfected persons (117, 118) underscoring the importance of strengthening efforts to increase HIV testing, treatment initiation in those infected with HIV, and ensuring viral suppression rates in those initiated on treatment as we increase SARS-CoV-2 vaccination efforts.

10. Ending SARS-CoV-2: An evolving endgame

As many countries have pondered their post-COVID-19 recovery plans, the public health goal has shifted from herd immunity, that is, sufficient numbers of individuals being exposed to SARS-Cov-2 and who have recovered; to high vaccination coverage rates combined with natural immunity to stop community spread; to natural immunity, vaccination, and slowing the emergence of variants of concern. Each of these key concepts are elaborated on next in relation to COVID-19. For now, the virus appears to have the upper hand but we need to continue our efforts to get ahead of it. With

pandemics in the past such as smallpox or polio, global efforts with vaccine availability have focused on elimination of the presence of the organism in humans. In the case of SARS-CoV-2, will the goal be epidemic control as with the UN2030 HIV goal, i.e. the numbers of infection reach sufficiently low levels for it to no longer pose a public health threat? The repeated surges of infection globally usually triggered by the emergence of a variant of concern, each more infectious than the last, suggests that we are dealing with greater complexity of transmission dynamics than we have ever known.

10.1. Role of herd immunity

Herd immunity is an inflection point in an epidemic or pandemic where sufficient people in a country or globally are protected from a disease and the virus can no longer spread within a population. This can result from natural recovery from infection and/or from vaccination and/or from the adoption of other public health and social interventions as described in section 4, earlier. Early in the pandemic some countries such as Sweden (119) used the concept of natural recovery from infection to define their SARS-CoV-2 response. A challenge with natural immunity or vaccine-induced immunity is that we don't yet know for how long that protection lasts. In mid-2021 some reports of re-infection with COVID-19 started to emerge (120). We now know that there is wide variability in SARS-CoV-2 transmission depending on the environment and R_0 only applies to a population that has no natural or vaccine-induced immunity or widespread adoption of other prevention measures such as mask wearing; physical distancing; and hand washing that reduce the risk of getting infected and the rate at which people get infected. In 2020 using the highest transmission rate estimate of 3 for COVID-19; a herd immunity level of 67% was estimated to be needed to slow down viral spread within the community (121). But, because viruses evolve over time, have different levels of infectiousness and new variants emerge, and we don't have vaccines that are 100% efficacious or prevent infection, R_0 is not static, and the herd immunity level will need to be higher than 67%, some estimate closer to 100%.

10.2. Vaccines

Based on prior experience with pandemics, the development and rollout of an efficacious vaccine was regarded as the most effective approach to control the SARS-CoV-2 pandemic (122). During the early stages of the pandemic, many expressed scepticism that an effective vaccine would soon become available, pointing out that time period from initial vaccine development through a series of trials and finally to approval often takes over a decade. Yet despite the many obstacles, several highly efficacious and safe vaccines have emerged in less than a year (123, 124, 125). A major reason for this is that pharmaceutical companies, in many instances with substantial financial support from governments, have compressed the timeline by overlapping many of the phases of this process through novel adaptive study designs. To facilitate access, the infrastructure for producing vaccine in massive quantities was put in place even before the success of clinical trials was demonstrated, and actual production in mass quantities of vaccine preceded approval and, in some cases, the final results of clinical trials. From a financial point of view, such a compressed timeline entails considerable extra risk, in case the vaccine fails. But many governments were willing to subsidize these risks given the considerable social and economic losses that result from a continuation of the pandemic.

The accelerated development for COVID-19 vaccines was also made possible because several scientists had already been working either with related coronavirus vaccine candidates, for example the Middle Eastern respiratory syndrome coronavirus (MERS-CoV) or the original severe acute respiratory syndrome coronavirus (SARS-CoV-1) and were able to rapidly replace inserts of SARS-CoV-2 genes or proteins; or had well-developed vaccine ‘backbones’ like those for measles and flu vaccines, making it easy to adapt them for COVID-19 (126). These factors, and the willingness to adapt the traditional clinical evaluation pathway to allow concurrent Phase 1/2 and 2/3 testing, meant that development time could be reduced to as little as 10 months to a year. Thus, within less than a year of the global report of the pandemic, several effective vaccines had been developed by several corporations in collaboration with the US, the UK, and other government agencies globally. Over 150 vaccines were under development in 2021 (127).

Several approaches have been used by different producers including: **i. Genetic vaccines**, consisting only of RNA or DNA encoding the spike protein, or the main virus particle outer protein; **ii. ‘Killed’ or inactivated**

whole virus particle vaccines; iii. Subunit, or nanoparticle, vaccines, which consist only of the spike protein made in a recombinant expression system; **iv. Live virus vaccines**, such as several different kinds of adenoviruses or measles vaccine or vesicular stomatitis virus (VSV, also used for Ebola) that carry the gene for spike protein; and **v. Attenuated vaccines**, with the live modified SARS-CoV-2 virus itself.

Table 6 presents vaccine efficacy by vaccine type. The AstraZeneca and Johnson & Johnson (J&J) vaccine efficacy varied by trials and country in which the trials were conducted (128, 129).

Table 6 Vaccine type, producer, and efficacy

Vaccine type	Company (vaccine)	Efficacy (pre-existing variants)
mRNA (123, 124)	Pfizer BioNTech (BNT162)	92%
	Moderna (mRNA-1273)	>90%
Viral Vector (129, 130, 125)	J&J (Ad26.COV2-S)	64–72%
	Gamaleya (Sputnik V rAd5, rAd26)	92%
	Oxford-AstraZeneca (AZD1222)	67–76%
	CanSinoBio (Ad5-nCoV)	65%
Spike Protein	Vecttor Institute (EpiVacCorona)	Unknown
	Novavax (NVX-CoV2373)	89%
Inactivated Virus	Sinopharm	78%
	Sinovac	51–84%
	Sinopharm-Wuhan	73%
	Bharat Biotech (BBV152)	78%

The level of efficacy for all the trials completed to date and reported earlier is based on protection against severity of COVID-19 clinical disease rather than preventing infection with SARS-CoV-2.

10.2.1. Vaccination and global equity in access to vaccines

The increasing number of safe and efficacious vaccines are an important starting point to control of the pandemic. However the potential of vaccines can only be realized by global equity in access and ensuring the sufficient numbers of individuals globally are vaccinated.

Notwithstanding the many formidable challenges of developing, testing, and licensing a vaccine we now have a growing number of safe and efficacious vaccines (Table 1). Continuing large-scale manufacture and distribution requirements pose new and equally formidable challenges depending on the types of vaccines. Those merely requiring a live virus to be grown, such as the adenovirus-based candidates, are probably the easiest to produce as the volume of material to be produced is less than for subunit vaccines, which typically require much more material to be made and purified as they are not infectious. Killed vaccines are one of the quickest ways to make the product, but this requires that a very large amount of live, virulent SARS-CoV-2 be grown. This is problematic due to the stringent safety requirements imposed when working with live unmodified virus (131). The production volumes required for the non-live vaccines are such that the world may simply not have the capacity to make the required amounts in less than a couple of years to protect the majority of persons globally (132).

Distribution is also potentially a problem, as certain of the vaccines require refrigeration to as low as –80° C for storage and transport, which may pose a serious obstacle for rolling these out to remote locations especially in LMICs. Of note is that with the exception of the J&J vaccine, which requires only a single dose, most of the other available efficacious vaccines require at least two doses for full protection. It remains to be seen how manufacturing capacity is scaled up and the geopolitical issues of vaccine distribution are handled (133).

The rapid dissemination of the vaccines, primarily in wealthy countries, has had a dramatic effect in reducing the transmission of SARS-CoV-2 and has caused the lifting of lockdowns in many areas. An expected surge in transmission accompanied the loosening of restrictions has occurred but the surge is higher in communities with low vaccination rates. Although the level of persons vaccinated has not reached a level to achieve ‘herd immunity’, the dissemination of vaccinations has had a dramatic effect at reducing the transmission of SARS-CoV-2. Underscoring the importance of

being vaccinated, 99% of persons dying in the United States in mid-2021 had not received a vaccination (134). Data from the UK demonstrates substantial reductions in mortality rates in the current wave of infections compared to the second wave that correlates with their high vaccination coverage rates (Figure 7) (135). If a vaccinated person gets infected, while they may not experience severe disease they may unknowingly transmit the virus, especially if they are asymptomatic. Some emerging data underscore the importance of vaccination for breaking chains of transmission, for example, in a study from Israel that included 253,564 individuals in 65,624 households there was a 79% reduction in infection in those exposed to vaccinated household contacts compared to unvaccinated infected household contacts (136). At a community level there is evidence that with each 20% increase in vaccination there is a 50% decline in test positivity rate in unvaccinated people (137). These data underscore the importance of enhancing access to vaccines and high vaccination coverage rates in our efforts to curb the spread of SARS-CoV-2 and have some return of ‘normality’ globally.

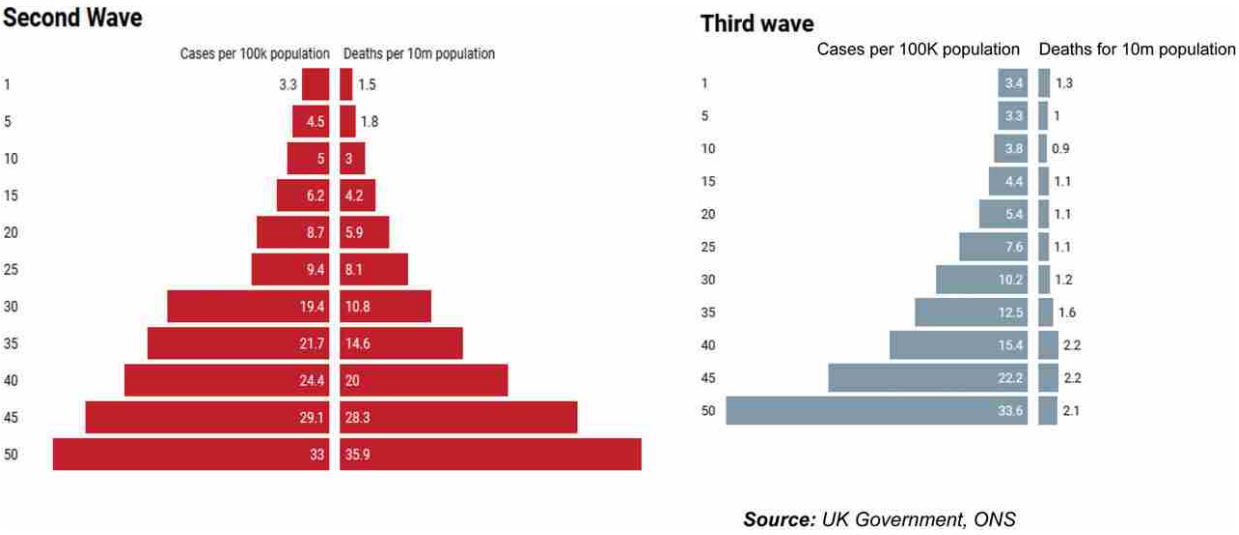


Figure 7 Impact of vaccination on hospitalization rates in the UK with circulation of Delta variants. Adapted from: UK Open National Statistics.

Unfortunately, in some countries where vaccines are available, a substantial proportion of the population is reluctant to be vaccinated. Some of the reluctance stems from hesitancy emanating from the rapid production

of the vaccine (138). However, as increasing numbers have been safely vaccinated confidence in their safety is growing and many have become vaccinated. Various forms of incentives are also having an impact on those who are hesitant. In some settings such as workplaces or travel to other countries, evidence of vaccination is becoming a requirement. While no country has yet made vaccination mandatory, an increasing number of sectors have started to institute vaccine mandates.

Of greater concern is the 'anti-vaxxer' movement whose anti-vaccine stance preceded SARS-CoV-2 and has now become more complex (139). At heart are underpinnings of mistrust in government, science, supremacy, and interpretation of individual rights. While their presence was initially observed in industrialized countries the dissemination of misinformation about vaccines and vaccinations spread through the social media platforms and conservative news stations is expanding to LMICs (140). Of concern are 'anti-vaxxer' rallies and marches occurring in many countries with a disregard for any of the public health preventative measures particularly use of masks and social distancing.

Politics and panic have caused major disruptions in the distribution of vaccines to where they are needed with vaccine nationalism at the forefront. Vaccine nationalism involves: buying more vaccines than necessary; striking deals with pharma to redirect doses away from equity groups; negotiating bilateral deals and queue skipping by paying more and thus pushing prices higher; rushed regulatory approval of vaccines despite inadequate data available; and bans on export of vaccine by the manufacturing countries. These actions and the diversion of vaccines away from LMICs will cost rich countries the most. Review of the global vaccine rollout in mid-2021 reveals that just 10 countries have received 75% of the vaccine doses that have been administered, while the remaining 170 countries share only 25% of global vaccine doses (141) (Figure 8). Only a small fraction of the world's population has been vaccinated, even in the affluent developed countries. The fraction of people having received at least one vaccine dose in the USA and the UK are 49.6% and 41.8%, respectively, whereas in Africa it is <2% (142). By 15 August 2021, 4.7 billion vaccine doses were administered and 1.9 billion people were fully vaccinated (143).

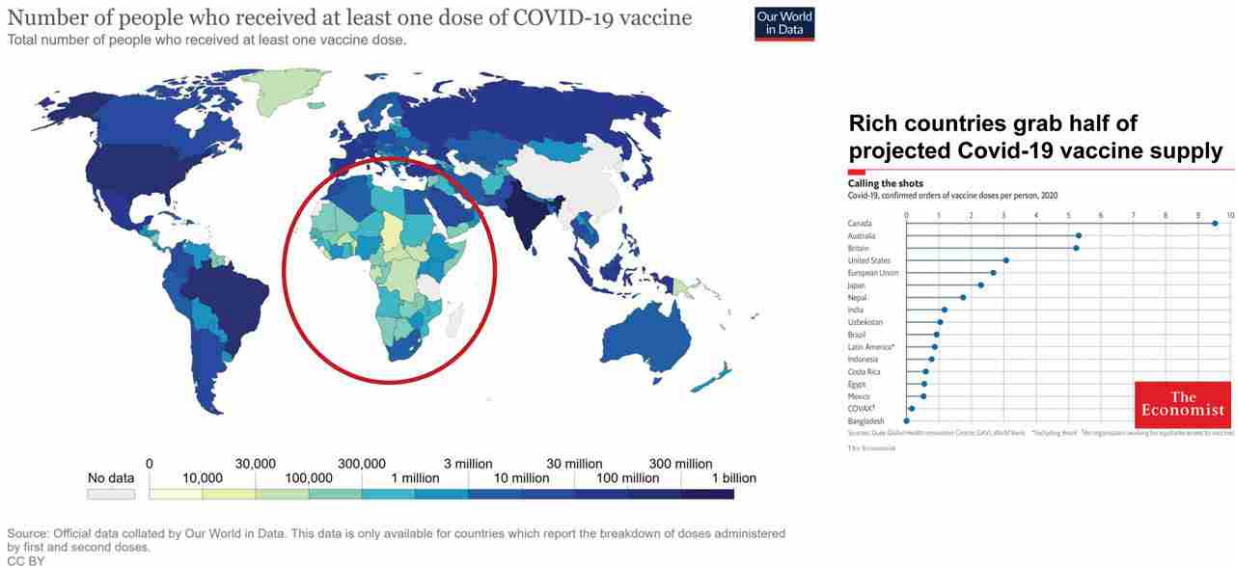


Figure 8 Vaccine nationalism defining who has vaccine access
Source: Official data collated by Our World in Date. This data is only available for countries which report the breakdown of doses administered by first and second doses.

Civil society groups around the world and some LMIC countries have been campaigning for a waiver to the World Trade Organization's intellectual property rights through a waiver of the Trade-Related Aspects of Intellectual Property Rights (TRIPS) agreement (144). A major boost was given to their campaign when President Biden announced that the USA would support the waiver and more recently his commitment to distributing over 200 million doses of vaccines to LMICs through Covax are important contributions to increase vaccine access to LMICs. Yet the world continues to be starkly divided in terms of equal access to vaccines and civil society groups continue to advocate for more equitable access. On Twitter, the hashtag #vaccineequity has been trending demonstrating the moral concern about the maldistribution of the vaccine. Laudable access efforts include the European Union's commitment to building vaccine production capacity in Africa in partnership with the African Union through the Africa CDC bringing some hope for future pandemic preparedness and represents an important commitment for building global capacity and strengthening local resilience. Notably it highlights recognition by wealthy countries of their global responsibilities (145). Notwithstanding, bilateral and public-private agreements, a recent *New York Times* report that of 90% of J&J vaccine produced in South Africa was shipped to Europe contrary to pre-production commitments made to the South African government in their export permit

applications that indicated that export permit waivers were needed to facilitate vaccine distribution in LMICs through Covax (146). This type of corporate greed contributes to distrust in the public and LMIC governments in industry.

Notwithstanding efforts to bridge global disparities in vaccine access, the hope and optimism from vaccines as a gateway to return to pre-COVID-19 normality is being further compromised by the emergence of new and more potent variants of concern that are infecting fully vaccinated individuals, which is described next.

10.3. Variants of concern

For viruses, mutations occur naturally during replication. The majority of these mutations are minor and do not impact viral functioning or structure. However, sometimes the mutation can result in a substantive structural or functional change to the virus that enhances transmissibility, virulence, or alternatively leads to extinction of the virus.

With the rapid transmission of SARS-CoV-2 resulting in millions of replication events many variants have evolved in several areas of the world. For example, the Alpha variant first described in the UK, the Beta variant first described in South Africa, the Gamma variant first described in Brazil, the Delta variant first described in India, and the Epsilon variant first described in California (147, 148, 149). Variants can occur via multiple pathways. Naturally SARS-CoV-2 has a slow genetic drift of 1–2 mutations a month, occasionally this same genetic shift will incorporate multiple mutations and thus create a variant. Persistent long-term active infection in immunocompromised individuals is an important source of new variants wherein antibody treatment or vaccination can create variants with immune escape. The majority of new variants have no functional advantage and so die off before spreading widely, though occasionally a new variant does have a functional advantage such as higher transmissibility that allows it to spread faster and become the dominant variant. These new variants may also be more pathogenic for humans.

Variants have been classified into two categories (150): i. variants of interest (VoI), and ii. variants of concern (VoC). A VoI can be raised to a variant of concern if it spreads faster and causes more serious cases or deaths. Typically, identification of VoI is based on a laboratory identification

of biological mutation that may potentially causes infectiousness or severe disease in a small number of patients. A VoI can be recategorized as a VoC when it spreads more widely within populations; spreads faster, or causes more cases of death, and/or infects large proportions of vaccinated people.

Variants of interest have been identified in the USA, the UK, Brazil, and India. Variants of concern have been identified in the UK, South Africa, and Japan, and Brazil. The range of reductions in neutralization activity to convalescent sera relative to the original Wuhan strain have ranged from <2-fold to as high as 12-fold (151). The vaccine efficacy against individual variants has ranged from 10% (AstraZeneca) to 90% (Pfizer) depending on the variant, the country reporting, and the outcome being measured, e.g. mild–moderate or severe disease (152). The current vaccines are thought to provide an adequate level of protection against the currently known variant strains although there is concern about the need for ongoing surveillance for variants that may escape in vaccinated individuals as well as immunocompromised individuals (153).

In November 2020, scientists in South Africa identified the 501Y.V2, B.1.351, or the Beta variant that coincided with the emergence of the second surge of SARS-CoV-2 infections in South Africa that was worse than the first wave of infections. The E484K mutation that resulted in the Beta variant occurs at the binding domain and helps the virus to evade antibodies produced by the body's immune system. This immediately raised questions around whether current vaccines will work to protect people from the Beta variant and its implications for continued spread. The Beta variant can significantly evade immunity from previous infection thereby resulting in many cases of re-infections and increased household transmission resulting in higher rates of hospitalizations from severe disease. In other words, the rise of new variants poses a problem for reaching herd immunity, as the more infectious the virus becomes, the more people need to be protected in order to stop the virus from spreading. There are also implications for vaccine-induced immunity. This was, for example, seen in the trials for the AstraZeneca and J&J vaccines that were undertaken while the Beta variants were in circulation in South Africa (154). The rapidity of the spread of the Beta variant across Africa within a matter of weeks also highlight our interconnectedness and vulnerability and why we need equity in access to and coverage rates of interventions such as vaccines. Only a handful of COVID vaccine manufacturers have released results from their clinical trials

and received emergency use approval in countries such as the United States, Israel, and the UK. All of these vaccines rely on our bodies identifying the spike protein on the surface of the coronavirus as a way of producing an immune response. We are starting to see second- and third-generation variants emerge, e.g. the Delta variant is a variant to the Beta variant.

From March to June of 2021, a severe outbreak of the Delta variant in India resulted in a substantial surge in India followed by noticeable increases in infections globally even in countries with high vaccination coverage rates. The new surges in infection experienced globally in 2021 are being driven by the Delta variant (6, 7). In the USA for the first time since February 2021, there were more than 100,000 cases per day reported in August 2021. Similar scenarios are emerging in Australia, New Zealand, China, and several countries in Africa. There was initial evidence suggesting that the variant reduces vaccine efficacy. Preliminary data from England shows that the effectiveness of both the Pfizer and AstraZeneca vaccines had been lowered after one dose from 51.1% to 33.5% that was offset following the second dose—the AstraZeneca vaccine efficacy was reduced from 66.1% to 59.8% and for the Pfizer vaccine it dropped from 93.4% to 87.9% (155). These data suggest that for the two dose vaccines, there is sufficient protection against the Delta variant for those who are fully immunized. The rapidity of the emergence and transmission of VoC underscores the importance of higher global vaccination coverage rates globally and that existing non-pharmaceutical interventions such as wearing masks and physical distancing need to continue. There is also some evidence of reduction in vaccine protection against mild–moderate symptomatic disease, but high levels of protection against severe disease and death seem to be maintained.

Should herd immunity be dismissed now, what is the role of vaccines, should we be giving booster doses of vaccines, are we likely to see more VoC and how many more? What we know is that the interplay between natural immunity, vaccine-induced immunity, and VoC is complex and we have much to learn. We know that there are individual level and population level benefits from vaccines but that the population level benefit is dependent on many factors, particularly R_0 and the extent to which infected vaccinated individuals transmit the virus. Vaccination coverage for herd immunity varies and we do not currently know what levels are needed to reach this goal. In terms of viral transmission we are seeing similar viral

loads in vaccinated and unvaccinated cases (156) [median Ct values: 21.9 (n = 80) vs. 21.5 (n = 65)]. New variants have changed the game with their higher transmissibility and/or immune escape, making reaching herd immunity much more difficult and we don't know what the next variants will look like (157).

Several rich countries, e.g. Israel, France, and Germany, are giving third–dose boosters when many poor people have not yet received their first dose of vaccine, spurring WHO to place a moratorium on boosters based on global disparity in vaccine access. We see corporate greed manifestations through vaccine manufacturing companies promoting the administration of booster doses of vaccines using arguments such as inadequate vaccine responses, waning antibody levels (homologous boosts), and variants escaping vaccine immunity need for heterologous boosts. To date, we have no correlate of immunity, nor do we know how much antibodies are required for protection. We do know from prior experience—for example, older people who had Spanish flu—that antibody levels decline over time but that memory B cells in lymph nodes can regenerate antibodies very quickly and therefore a homologous boost adds little value. Thus far, all mRNA vaccines remain effective against all current variants, displacing the argument that a heterologous vaccine boost is needed to enhance vaccine immunity against new mutated forms of the virus. There is no clinical trial evidence currently to justify booster doses. The three-dose vaccination dose schedule for persons with autoimmune disease or who are severely immunocompromised to enhance immune responses to date is the only clear justification for an additional dose of vaccination.

Finally, the hybrid immunity resulting from natural immunity with vaccine-induced immunity appears to be higher and more durable than either on its own (158) (Figure 9), providing a strong rationale for rapidly increasing vaccine coverage rates globally to limit the emergence and transmission of VoC and achieving some level of epidemic control.

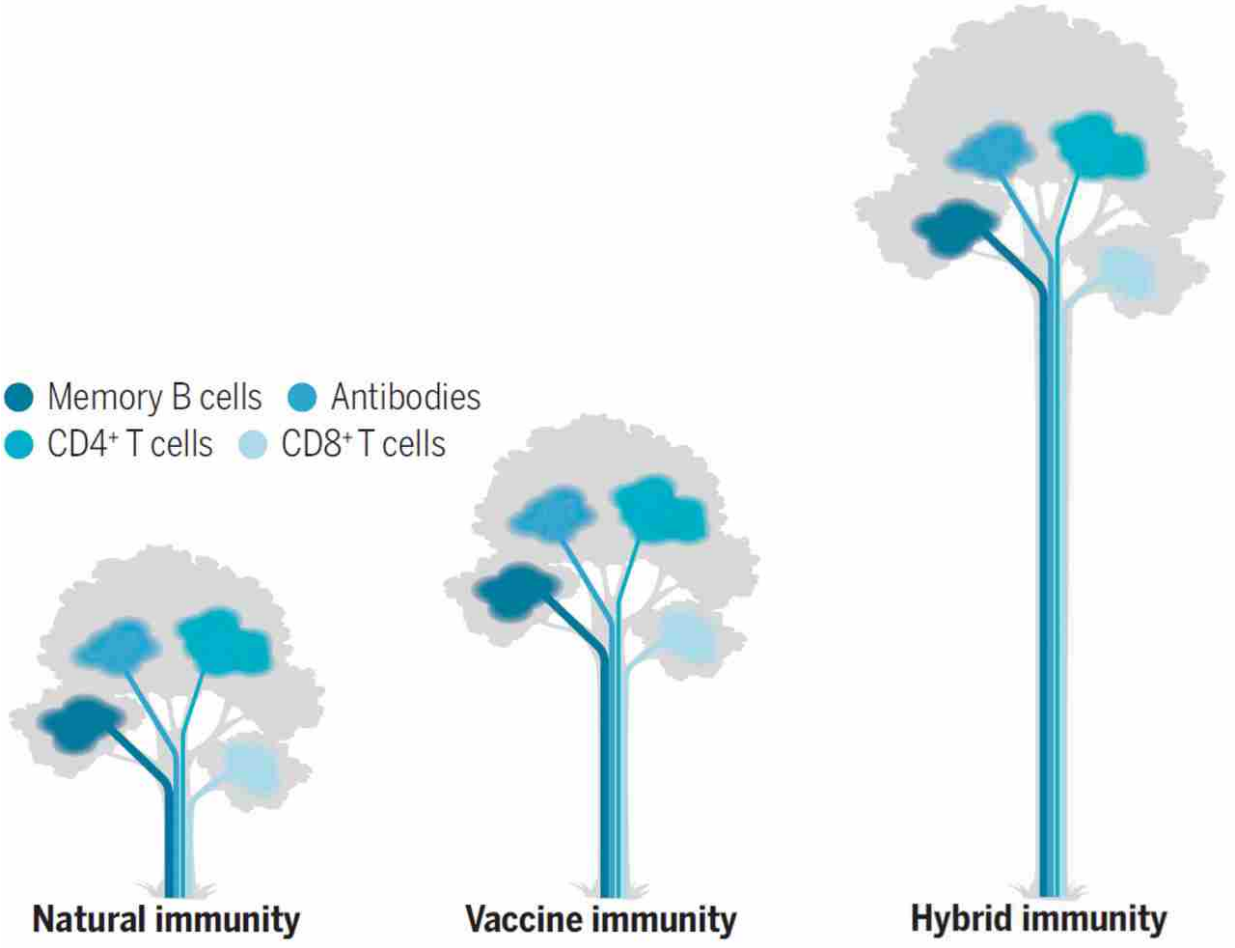


Figure 9 Hybrid immunity is much better than natural immunity.
Source: Crotty S, Science 2021.

11. Lessons learned and the way forward

The SARS-CoV-2 pandemic has affected virtually every aspect of life on Earth: physical and mental health economies, national well-being, social interactions, political fortunes, and exacerbated global, national, and local inequalities. The pandemic has underscored the importance of public health principles and applications and the interconnectedness of all persons whether in wealthy or poor countries or conditions. We know that attaining a high level of community immunity to SARS-CoV-2 will reduce the threat and spread of the infection. Due to the reluctance of many persons to get vaccinated and the failure of the world's institutions to provide vaccine equity, the pandemic continues. In rich countries where people have the

opportunity to have a vaccine some people, choose not to be vaccinated. Thus, the pandemic is now a pandemic of the unvaccinated who refuse to acknowledge their community responsibility to protect their fellow humans. Their argument that their 'human rights' are being violated ignores the human rights of their fellow humans to not be infected. Further, the human rights of people in LMICs are denied because they have not been given the chance to be vaccinated because of the system of 'intellectual property' laws which protects the profits of corporations at the expense of lives and health.

The pandemic has exacerbated existing inequalities that intersect with race, gender, and social and economic marginalization. In all countries, people living in economically deprived circumstances have been more affected by COVID-19. Crowded housing, insecure employment, care work, work in the retail and manufacturing sectors, being a migrant or Indigenous person have meant more exposure to the virus and less ability to cope with its impact. Strong healthcare delivery systems with universal health coverage and access have been shown to be most effective in protecting people from the impact of COVID-19.

The impact of the pandemic has been mitigated to some extent by our experience with the previous epidemic of HIV/AIDS, which through global solidarity ensured access to the scientific advances that has resulted in for example AIDS being transformed from an inevitably fatal condition to a chronic one. That solidarity emanated from the synergy of years of activism and advocacy from affected communities and scientists, strong public engagement and communication, and demand for accountability from politicians. Collaborations between scientists across different regions and countries resulted in the development of new methods of diagnostic testing, prevention, and treatment and infrastructure to undertake cutting edge research and clinical trials in most affected populations. The implementation of these advances strengthened health systems, improved health information systems, and enabled innovation in delivery of interventions.

The rapid pace of knowledge generation with COVID-19 was possible because of prior investments in science and technology. For example, the decades of investigations from several research centres on other coronavirus epidemics enabled an accelerated pathway to vaccine development and evaluation into the arms of people in less than a year. This unprecedented rapid response to SARS-CoV-2 needs to be applied to existing challenges and provides important lessons for confronting future pandemics.

The pandemic also demonstrated the importance of strong political leadership in ensuring that responses are guided by science, as uncertain and evolving as it may be. We have seen the dual facets of public health, namely science and social justice, become more politicised on ideological grounds by both politicians and some of the public. Public health practitioners have a responsibility to inform decision-makers, political leaders, and the public and to ensure science-based policy promulgation and implementation. Ongoing public communication and engagement has become even more important in the current era of modern formal and informal rapid communication and the rapid expansion of social media platforms. Public health must develop strategies to counter misinformation and negative messages. Parallel to the COVID-19 pandemic, a pandemic of misinformation and distrust has flourished and undermined effective responses to COVID-19.

We can be confident that there will be future pandemics, some possibly involving coronavirus, and likely other disease agents. To mutate a virus must replicate. The more people who are infected the more replications that will occur, each with a probability of mutating. Although the majority of these mutations will be non-pathogenic, some will increase their adaptability to humans and will demonstrate even greater transmissibility and pathogenicity. Thus, ongoing molecular surveillance and laboratory predictions of new and mutating agents such as Sarbecovirus is important and key for adequate pandemic preparedness and expanding the protective benefits of current vaccines.

Above all, the SARS-CoV-2 pandemic has demonstrated the complexities of disease transmission and control in a rapidly changing world. To confront the next pandemic, as well as to have a greater impact on the current pandemic it is necessary to muster all our resources: public health, political, scientific, and social, and to recognize our interconnectedness. We hope that this chapter will help understanding of and responses to the current pandemic and provide a path forward for future pandemics.

We would like to end with an exhortation from Pope Francis:

> *The pandemic has exposed the paradox that while we are more connected we are also more divided … To come out of this crisis better we have to recover the knowledge that as a people we have a shared destination. The pandemic has reminded us that no one is saved alone. What ties us to one another is what we commonly call solidarity. Solidarity is more than acts of generosity, important as they are, it is the call to embrace the reality that we are bound by bonds of reciprocity.*

12. REFERENCES

1. Wu, F., Zhao, S., Yu, B., Chen, Y.M., Wang, W., Song, Z.G., et al. 2020. A new coronavirus associated with human respiratory disease in China. *Nature*, 579(7798), 265–9.
2. Naming the coronavirus disease (COVID-19) and the virus that causes it. 2020. Geneva: WHO. Available at: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-(covid-2019)-and-the-virus-that-causes-it#:~:text=Human%20disease%20preparedness%20and%20response,virus%20on%2011%20February%202020 (Accessed on 9 October 2021).
3. Corman, V.M., Landt, O., Kaiser, M., Molenkamp, R., Meijer, A., Chu, D.K., et al. 2020. Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR. *Euro Surveill*, 25(3), 2000045.
4. World Health Organization (WHO) Emergency Committee. 2020. Statement on the second meeting of the International Health Regulations (2005) Emergency Committee regarding the outbreak of novel coronavirus (2019-nCoV). Geneva: WHO. Available at: https://www.who.int/news-room/detail/30-01-2020-statement-on-the-second-meeting-of-the-international-health-regulations-(2005)-emergency-committee-regarding-the-outbreak-of-novel-coronavirus-(2019-ncov) (Accessed on 12 August 2021).
5. Wu, Z., McGoogan, J. 2020. Characteristics of and important lessons from the coronavirus disease 2019 (COVID19) outbreak in China. *JAMA*, 323(13), 1239–42.
6. Our World in Data. 2020.12. European CDC. Coronavirus pandemic data explorer. Situation update worldwide. Available at: https://ourworldindata.org/coronavirus-data-explorer?zoomToSelection=true&time=2020-03-01.latest&country=~OWID_WRL&region=World&casesMetric=true&interval=smoothed&perCapita=true&smoothing=7&pickerMetric=total_cases&pickerSort=desc (Accessed on 12 August 2021).
7. Dong, E., Du, H., Gardner, L. 2020. An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect Dis*, 20(5), 533–4.
8. Neil Ferguson. 2020. UK Government COVID-19 Advisor. Delay in lockdown UK's 'biggest' pandemic mistake. Available at: https://www.theguardian.com/world/2020/jun/10/uk-coronavirus-lockdown-20000-lives-boris-johnson-neil-ferguson (Accessed on 9 October 2021).
9. V'kovski, P., Kratzel, A., Steiner, S., Stalder, H., Thiel, V. 2021. Coronavirus biology and replication: implications for SARS-CoV-2. *Nat Rev Microbiol*, 19(3), 155–70.
10. Miranda, C., Silva, V., Igrejas, G., Poeta, P. 2021. Genomic evolution of the human and animal coronavirus diseases. *Mol Biol Rep*, 12 Aug, 1–9.
11. Cohen, J. 2021. The hunt for SARS-CoV-2's ancestors heats up. *Science*, 373(6559), 1076–7.
12. Sanyal, S. 2020. How SARS-CoV-2 (COVID-19) spreads within infected hosts—what we know so far. *Emerg Top Life Sci*, 4(4), 371–8.
13. Buitrago-Garcia, D., Egli-Gany, D., Counotte, M.J., Hossmann, S., Imeri, H., et al. (2020) Occurrence and transmission potential of asymptomatic and presymptomatic SARS-CoV-2 infections: a living systematic review and meta-analysis. *PLoS Med*, 17(9), e1003346.
14. World Health Organization. Considerations in adjusting public health measures in the context of COVID-19. Interim guidance, 16 April 2020. WHO reference number: WHO/2019-nCoV/Adjusting_PH_measures/2020.1 WHO test positivity rate. Available at: https://www.who.int/publications/i/item/considerations-in-adjusting-public-health-and-social-measures-in-the-context-of-covid-19-interim-guidance (Accessed on 9 October 2021).
15. Rocklöv, J., Sjödin, H., Wilder-Smith, A. 2020. COVID-19 outbreak on the Diamond Princess cruise ship: estimating the epidemic potential and effectiveness of public health countermeasures. *J Travel Med*, 27(3), taaa030.

16. Malone, J.D. 2020. USS Theodore Roosevelt, COVID-19, and ships: lessons learned. *JAMA Network Open*, 3(10), e2022095.
17. Patolia, S.K., Mosenifar, Z. 2021. What is the R naught of coronavirus diseases 2019 (COVID-19)? 12 Jan 2021. Medscape.
18. Anderson RM, May RM. 1979. Population biology of infectious diseases: part I. *Nature*, 280(5721), 361–7.
19. Bettencourt, L.M., Ribeiro, R.M. 2008. Real time Bayesian estimation of the epidemic potential of emerging infectious diseases. *PLoS One*, 3(5), e2185.
20. O'Driscoll, M., Harry, C., Donnelly, C.A., Cori, A., Dorigatti, I. 2021. A comparative analysis of statistical methods to estimate the reproduction number in emerging epidemics, with implications for the current coronavirus disease 2019 (COVID-19) pandemic. *CID*, 73(1), e215–23.
21. Senthuraman, N., Jeremiah, S.S., Ryo, A. 2020. Interpreting diagnostic tests for SARS-Cov-2. *JAMA*, 323(22):2249–51.
22. Abdool Karim, S.S. 2020. The South African response to the pandemic. *N Eng J Med*, 382, e95.
23. Mukhrjee, S. 2021. Why does the pandemic seem to be hitting some countries harder than others? *New Yorker* 2021, 22 Feb. Available at: https://www.newyorker.com/magazine/2021/03/01/why-does-the-pandemic-seem-to-be-hitting-some-countries-harder-than-others (Accessed on 9 October 2021).
24. Nuclear Threat Initiative. 2019. Global Health Security Index. Baltimore, MD: Johns Hopkins Center for Health Security & The Economist Intelligence Unit. Available at: https://www.ghsindex.org/wp-content/uploads/2019/10/2019-Global-Health-Security-Index.pdf (Accessed on 28 September 2020).
25. Baum, F., Freeman, T., Musolino, C., Abramovitz, M., De Ceukelaire, W., Flavel, J., et al. 2021. Explaining covid-19 performance: what factors might predict national responses? *BMJ*, 372, n91.
26. International Labour Organization (ILO). 2021. COVID-19 and the World of Work: Country Policy Responses Geneva: ILO. Available at: https://www.ilo.org/global/topics/coronavirus/regional-country/country-responses/lang--en/index.htm (Accessed on 9 October 2021).
27. World Bank (2020). *The COVID-19 Pandemic: Shocks to Education and Policy Responses*. Washington DC: World Bank.
28. Razavi, A., Erondu, N.A., Okereke, E. 2020. The Global Health Security Index: what value does it add? *BMJ Global Health*, 5, e002477.
29. Scheepers, E., Lakhani, E., Armstrong, K. 2020. Making a community action net (work): organising in the times of COVID-19. Open Global Rights. Available at: https://www.openglobalrights.org/organising-in-the-times-of-COVID-19/ (Accessed on 9 October 2021).
30. UN Inter-agency Standing Committee. 2020. Public health and social measures for Covid-19 preparedness and response in low capacity and humanitarian settings. Interim Guidance. Version 1, May 2020. Available at: https://www.who.int/publications/m/item/public-health-and-social-measures-for-covid-19-preparedness-and-response-in-low-capacity-and-humanitarian-settings (Accessed on 9 October 2021).
31. Powdthavee, N., Riyanto, Y.E., Wong, E.C.L., Yeo, J.X.W., Chan, Q.Y. 2021. When face masks signal social identity: explaining the deep face-mask divide during the COVID-19 pandemic. *PLoS One*, 16(6), e0253195.
32. Verschuur, J., Koks, E.E., Hall, J.W. 2021. Global economic impacts of COVID-19 lockdown measures stand out in high-frequency shipping data. *PLoS One*, 16(4), e0248818.
33. Boza-Kiss, B., Pachauri, S., Zimm, C. Deprivations and inequities in cities viewed through a pandemic lens. *Front Sustain Cities*, 19 March 2021. Available at:

https://doi.org/10.3389/frsc.2021.645914 (Accessed on 9 October 2021).
34. Baxter C, Abdool Karim Q, Abdool Karim SS. 2021. Identifying SARS-CoV-2 infections in South Africa: balancing public health imperatives with saving lives. *Biochem Biophys Res Commun*, 538, 221–5.
35. Tsai, P.H., Lai, W.Y., Lin, Y.Y., Luo, Y.H., Lin, Y.T., Chen, H.K., et al. 2021. Clinical manifestation and disease progression in COVID-19 infection. *J Chin Med Assoc*, 84(1), 3–8.
36. World Health Organization. 2021. Clinical features and prognostic factors of COVID-19 in people living with HIV hospitalized with suspected or confirmed SARS-CoV-2 infection. July 2021. Available at: https://www.who.int/publications/i/item/WHO-2019-nCoV-Clinical-HIV-2021.1 (Accessed on 9 October 2021).
37. Callard, F., Perego, E. 2021. How and why patients made long COVID. *Soc Sci Med*, 268, 113426.
38. Feng, Z., Yu, Q., Yao, S., Luo, L., Zhou, W., Mao, X., et al. 2020. Early prediction of disease progression in COVID-19 pneumonia patients with chest CT and clinical characteristics. *Nat Commun*, 11(1), 1–9.
39. Garcia, L.F. 2020. Immune response, Inflammation, and the clinical spectrum of COVID-19. *Front Immunol*, 16 June. Available at: https://doi.org/10.3389/fimmu.2020.01441 (Accessed on 9 October 2021).
40. Dalan, R., Bornstein, S.R., El-Armouche, A., Rodionov, R.N., Markov, A., Wielockx, B., et al. 2020. The ACE-2 in COVID-19: foe or friend? *Horm Metab Res*, 52(05), 257–63.
41. Zhang, J., Wang, X., Jia, X., Li, J., Hu, K., Chen, G., et al. 2020. Risk factors for disease severity, unimprovement, and mortality in COVID-19 patients in Wuhan, China. *Clin Microbiol Infect*, 26(6), 767–72.
42. Katzmarzyk, P.T., Salbaum, J.M., Heymsfield, S.B. 2020. Obesity, noncommunicable diseases, and COVID-19: a perfect storm. *Am J Hum Biol*, 32(5), e23484.
43. Iyengar, K., Bahl, S., Vaishya, R., and Vaish, A. 2020. Challenges and solutions in meeting up the urgent requirement of ventilators for COVID-19 patients. *Diabetes Metab Syndr*, 14(4), 499–501.
44. White, D.B., Lo, B. 2020. A framework for rationing ventilators and critical care beds during the COVID-19 pandemic. *JAMA*, 323(18), 1773–4.
45. Fifi, J.T., Mocco, J. 2020. COVID-19 related stroke in young individuals. *Lancet Neurol*, 19(9), 713–15.
46. Biswas, S., Thakur, V., Kaur, P., Khan, A., Kulshrestha, S., Kumar, P. 2021. Blood clots in COVID-19 patients: simplifying the curious mystery. *Med Hypotheses*, 146, 110371.
47. Blomberg, B., Mohn, K.G., Brokstad, K.A., Zhou, F., Linchausen, D.W., Hansen B.A., et al. 2021. Long COVID in a prospective cohort of home-isolated patients. *Nat Med*, 27(9), 1607–13.
48. Patel, A.B., Verma A. 2020. Nasal ACE2 levels and COVID-19 in children. *JAMA*, 323(23), 2386–7.
49. Consiglio, C.R., Cotugno, N., Sardh, F., Pou, C., Amodio, D., Rodriguez, L., et al. 2020. The immunology of multisystem inflammatory syndrome in children with COVID-19. *Cell*, 183(4), 968–81.
50. World Health Organization. 2021. Therapeutics and COVID-19: Living guideline. Available at: https://www.who.int/publications/i/item/WHO-2019-nCoV-therapeutics-2021.3 (Accessed on 9 October 2021).
51. National Institutes of Health (NIH). 2021. Coronavirus disease 2019 (COVID-19) treatment guidelines (updated 4 August 2021). Available at: https://www.covid19treatmentguidelines.nih.gov/ (Accessed on 12 August 2021).
52. Maciorowski, D., Idrissi, S.Z.E., Gupta, Y., Medernach, B.J., Burns, M.B., Becker, D.P., et al. 2020. A review of the preclinical and clinical efficacy of remdesivir, hydroxychloroquine, and lopinavir-ritonavir treatments against COVID-19. *SLAS DISCOVERY: Advancing the Science of Drug Discovery*, 25(10), 1108–22.

53. WHO Solidarity Trial Consortium, Pan, H., Peto, R., Henao-Restrepo, A.M., Preziosi, M.P., et al. 2021. Repurposed antiviral drugs for COVID-19—interim WHO solidarity trial results. *N Engl J Med*, 384(6), 497–511.
54. Beigel J.H., Tomashek, K.M., Dodd, L.E., Mehta, A.K., Zingman, B.S., Kalil, A.C., et al. 2020. Remdesivir for the treatment of COVID-19—final report. *N Engl J Med*, 383, 1813–26.
55. Shen, C., Wang, Z., Zhao, F., et al. 2020. Treatment of 5 critically ill patients with COVID-19 with convalescent plasma. *JAMA*, 323(16), 1582–9.
56. Marovich, M., Mascola, J.R., Cohen, M.S. 2020. Monoclonal antibodies for prevention and treatment of COVID-19. *JAMA*, 324(2), 131–2.
57. Horby, P.W., Mafham, M., Peto, L., Campbell, M., Pessoa-Amorim, G., Spata, E., et al. 2021. Casirivimab and imdevimab in patients admitted to hospital with COVID-19 (RECOVERY): a randomised, controlled, open-label, platform trial. medRxiv January 2021. doi: 10.1101/2021.06.15.21258542. Available at: https://www.medrxiv.org/content/10.1101/2021.06.15.21258542v1 (Accessed on 9 October 2021).
58. Abani, O., Abbas, A., Abbas, F., Abbas, M., Abbasi, S., Abbass, H., et al. 2021. Tocilizumab in patients admitted to hospital with COVID-19 (RECOVERY): a randomised, controlled, open-label, platform trial. *Lancet*, 397(10285), 1637–45.
59. OECD. 2021. COVID-19: A focus on the global economy. Paris: OECD. Available at: https://www.oecd.org/coronavirus/en/themes/global-economy (Accessed on 13 August 2021).
60. Pak, A., Adegboye, O.A., Adekunle, A.I., Rahman, K.M., McBryde, E.S., Eisen, D.P. 2020. Economic consequences of the COVID-19 outbreak: the need for epidemic preparedness. *Front Public Health*, 8, 241.
61. Congressional Research Service. 2021. *Global Economic Effects of COVID-19*. Washington, DC: Congressional Research Service.
62. Stiglitz, J. 2020. Conquering the great divide. Finance and Development 2020 fall issue. Available at: https://www.imf.org/external/pubs/ft/fandd/2020/09/COVID19-and-global-inequality-joseph-stiglitz.htm (Accessed on 9 October 2021).
63. Noy, I., Dean, N., Ferrarini, B., Park, D. 2020. The economic risk of COVID-19 in developing countries: where is it highest? In: S. Djankov and U. Panizza (eds). *COVID-19 in Developing Economies* (pp. 38–52). Geneva: Centre for Economic Policy Research.
64. McKinsey and Company. 2021. The coronavirus effect on global economic sentiment. Survey: McKinsey and Company (updated 28 July 2021). Available at: https://www.mckinsey.com/business-functions/strategy-and-corporate-finance/our-insights/the-coronavirus-effect-on-global-economic-sentiment (Accessed on 13 August 2021).
65. Yeyati, E.L., Filippini, F. 2021. Pandemic divergence: the social and economic costs of COVID-19. Available at: https://voxeu.org/article/social-and-economic-costs-covid-19 (Accessed on 16 August 2021).
66. Pickard, J., Giles, C., Parker, G. 2021. Rishi Sunak delivers spend now, tax later budget to Kickstart UK Economy. *Financial Times*. Available at: https://www.ft.com/content/da66ce9a-6dfc-4a3a-bde7-d4f4faed6c4a (Accessed on 13 August 2021).
67. International Monetary Fund. n.d. Policy responses to COVID-19. Available at: https://www.imf.org/en/Topics/imf-and-covid19/Policy-Responses-to-COVID-19 (Accessed on 13 August 2021).
68. Patrick, A. 2021. COVID-19 lockdowns a 'mass sacrificial event'. *Australian Financial Review* 16 July. Sydney. Available at: https://www.afr.com/policy/economy/covid-19-lockdowns-a-mass-sacrificial-event-20210715-p58a6n (Accessed on 9 October 2021).
69. Independent Panel for Pandemic Preparedness and Response. 2021. COVID-19: Make it the last pandemic. Available at: https://theindependentpanel.org/mainreport/ (Accessed on 12 August 2021).

70. Lassale, C., Gaye, B., Hamer, M., Gale, C.R., Batty, G.D. 2020. Ethnic disparities in hospitalisation for COVID-19 in England: the role of socioeconomic factors, mental health, and inflammatory and pro-inflammatory factors in a community-based cohort study. *Brain, Behavior, and Immunity*, 88, 44–9.
71. Santos, V.S., Souza Araújo, A.A., de Oliveira, J.R., Quintans-Júnior, L.J., Martins-Filho, P.R. 2021. COVID-19 mortality among Indigenous people in Brazil: a nationwide register-based study. *J Public Health*, 43(2), e250–1.
72. National Aboriginal Community Controlled Health Organisation (NACCHO). 2021. National Management Plan for Aboriginal and Torres Strait Islander Peoples. Canberra: NACCHO. Available at: https://www.naccho.org.au/aboriginal-health-alerts-coronavirus-covid-19/ (Accessed on 13 August 2021).
73. World Health Organization. 2021. Refugees and migrants in times of COVID-19: mapping trends of public health and migration policies and practices. Geneva: WHO. Available at: https://www.who.int/publications-detail-redirect/9789240028906 (Accessed on 13 August 2021.
74. International Labour Organization (ILO). 2021. COVID-19 and the world of work: Country Policy Responses. Geneva: ILO. Available at: https://www.ilo.org/global/topics/coronavirus/regional-country/country-responses/lang--en/index.htm (Accessed on 9 October 2021).
75. Pokora, R., Kutschbach, S., Weigl, M., Braun, D., Epple, A., Lorenz, E., et al. 2021. Investigation of superspreading COVID-19 outbreak events in meat and poultry processing plants in Germany: a cross-sectional study. *PLoS One*, 16(6), e0242456.
76. Pachirat, T. 2021. *Every Twelve Seconds: Industrialized Slaughter and the Politics of Sight*. New Haven, CT: Yale University Press.
77. Kinniburgh, C. 2020. COVID-19: How the meat industry became a global health liability. France 24. Available at: https://www.france24.com/en/20200524-covid-19-how-the-meat-industry-became-a-global-health-liability (Accessed on 7 October 2020).
78. Human Rights Watch. 2005. Blood, sweat, and fear: workers' rights in US meat and poultry plants. Available at: https://www.hrw.org/report/2005/01/24/blood-sweat-and-fear/workers-rights-us-meat-and-poultry-plants (Accessed on 7 October 2020).
79. Meng G, Williamson H. 2020. Germany to protect meatpacking workers better: US Should emulate Germany's workers-first response: human rights watch. Available at: https://www.hrw.org/news/2020/05/29/germany-protect-meatpacking-workers-better (Accessed on 7 October 2020).
80. Waltz, L. 2018. *Hog Wild: The Battle for Workers' Rights at the World's Largest Slaughterhouse*. Iowa City, IA: University of Iowa Press.
81. Boseley, M. 2020. Cedar Meats cluster: why abattoir workers are on the coronavirus frontline. *The Guardian*. Available at: https://www.theguardian.com/australia-news/2020/may/09/cedar-meats-cluster-why-abattoir-workers-are-on-the-coronavirus-frontline (Accessed on 7 October 2020).
82. Thompson, R. 2020. Are slaughterhouses turning into coronavirus factories? Euronews (updated 24 June 2020). Available at: https://www.euronews.com/2020/06/24/are-slaughterhouses-turning-into-coronavirus-factories (Accessed on 7 October 2020).
83. Jin, J.M., Bai, P., He, W., Wu, F., Liu, X.F., Han, D.M., et al. 2020. Gender differences in patients with COVID-19: focus on severity and mortality. *Frontiers in Public Health*, 8, 152.
84. Hawkes, S., Tanaka, S., Pantazis, A., Gautam, A., Kiwuwa-Muyingo, S., Buse, K., et al. 2021. Recorded but not revealed: exploring the relationship between sex and gender, country income level, and COVID-19. *Lancet Global Health*, 9(6), e751–2.
85. Pinchoff, J., Austrian, K., Rajshekhar, N., Abuya, T., Kangwana, B., Ochako, R., et al. 2021. Gendered economic, social and health effects of the COVID-19 pandemic and mitigation policies in Kenya: evidence from a prospective cohort survey in Nairobi informal settlements. *BMJ Open*, 11(3), e042749.

86. Hamadani, J.D., Hasan, M.I., Baldi, A.J., Hossain, S.J., Shiraji, S., Bhuiyan, M.S.A., et al. 2020. Immediate impact of stay-at-home orders to control COVID-19 transmission on socioeconomic conditions, food insecurity, mental health, and intimate partner violence in Bangladeshi women and their families: an interrupted time series. *Lancet Global Health*, 8(11), e1380–9.
87. UN Women. 2020. COVID-19 and its economic toll on women: the story behind the numbers. Available at: https://www.unwomen.org/en/news/stories/2020/9/feature-covid-19-economic-impacts-on-women (Accessed on 2 August 2020).
88. Azcona, G., Bhatt, A., Encarnacion, J., Plazaola-Castaño, J., Seck, P., Staab, S., et al. 2020. From insights to action: gender equality in the wake of COVID-19. United Nations Entity for Gender Equality and the Empowerment of Women (UN Women). Available at: https://www.unwomen.org/en/digital-library/publications/2020/09/gender-equality-in-the-wake-of-covid-19 (Accessed on 2 August 2020).
89. Rao, B., Tewari, S. 2020. Distress among health workers in COVID-19 fight: Article-14; 2020 [updated 9 June 2020; cited 7 October 2020]. Available at: https://www.article-14.com/post/anger-distress-among-india-s-frontline-workers-in-fight-against-covid-19 (Accessed on 7 October 2020).
90. Lotta, G., Wenham, C., Nunes, J., Pimenta, D.N. 2020. Community health workers reveal COVID-19 disaster in Brazil. *Lancet*, 396(10248), 365–6.
91. Nepomnyashchiy, L., Dahn, B., Saykpah, R., Raghavan, M. 2020. COVID-19: Africa needs unprecedented attention to strengthen community health systems. *Lancet*, 396(10245), 150–2.
92. Ballard, M., Bancroft, E., Nesbit, J., Johnson, A., Holeman, I., Foth, J., et al. 2020. Prioritising the role of community health workers in the COVID-19 response. *BMJ Glob Health*, 5, e002550.
93. UNICEF data hub. COVID-19 and children. Available at: https://data.unicef.org/covid-19-and-children (Accessed on 7 October 2020).
94. World Economic Forum. 2020. Coronavirus has exposed the digital divides like never before. Available at: https://www.weforum.org/agenda/2020/04/coronavirus-covid-19-pandemic-digital-divide-internet-data-broadband-mobbile/ (Accessed on 7 October 2020).
95. Mpungose, C.B. 2020. Emergent transition from face-to-face learning to online learning in a South African university in the context of the coronavirus pandemic. *Humanit Soc Sci Commun*, 7, 113.
96. Han, J., Meyer, B.D., Sullivan, J.X. 2020. Income and poverty in the COVID-19 pandemic. BPEA Conference drafts, 25 June 2020. Brookings papers on economic activity. Available at: https://www.brookings.edu/wp-content/uploads/2020/06/Han-et-al-conference-draft.pdf (Accessed on 7 October 2020).
97. International Labour Organization. 2020. COVID-19 and the world of work. Global impact and policy recommendations. Available at: https://ilo.org/global/topics/coronavirus/impacts-and-responses/lang--eng/index.htm (Accessed on 7 October 2020).
98. Abdool Karim, Q., Abdool Karim, S.S. 2020. COVID-19 affects HIV and tuberculosis care. *Science*, 369(6502), 366–8.
99. UNAIDS. 2021. Global commitments, local action: after 40 years of AIDS, charting a course to end the pandemic. Geneva, Switzerland: Joint United Nations Programme on HIV/AIDS (UNAIDS). Available at: https://www.unaids.org/sites/default/files/media_asset/global-commitments-local-action_en.pdf (Accessed on 7 July 2021).
100. de Oliveira, T., Kharsany, A.B., Gräf, T., Cawood, C., Khanyile, D., Grobler, A., et al. 2017. Transmission networks and risk of HIV infection in KwaZulu-Natal, South Africa: a community-wide phylogenetic study. *Lancet HIV*, 4(1), e41–e50.
101. Pillay, S., Giandhari, J., Tegally, H., Wilkinson, E., Chimukangara, B., Lessells, R., et al. 2020. Whole genome sequencing of SARS-CoV-2: adapting illumina protocols for quick and accurate outbreak investigation during a pandemic. *Genes*, 11(8), 949.

102. Tegally, H., Wilkinson, E., Giovanetti, M., Iranzadeh, A., Fonseca, V., Giandhari, J., et al. 2021. Detection of a SARS-CoV-2 variant of concern in South Africa. *Nature*, 592, 438–43.
103. Nachega, J., Sam-Agudu, N.A., Mellors, J.W., Zumla, A., Mofenson, L.M. 2021. Scaling up COVID-19 vaccination in Africa—lessons from the HIV pandemic. *N Eng J Med*, 385, 196–8.
104. Madhi, S., Gray, G.E., Ismail, N., Izu, A., Mendelson, M., Cassim, N., et al. 2020. COVID-19 lockdowns in low- and middle-income countries: success against COVID-19 at the price of greater costs. *SAMJ*, 110, 724–6.
105. Dorward, J., Khubone, T., Gate, K., Ngobese, H., Sookrajh, Y., Mkhize, S., et al. 2021. The impact of the COVID-19 lockdown on HIV care in 65 South African primary care clinics: an interrupted time series analysis. *Lancet HIV*, 8(3), e158–65.
106. UNAIDS. 2014. Fast track: ending the HIV epidemic by 2030. Geneva: Joint United Nations Programme on HIV/AIDS (UNAIDS). Available at: http://www.unaids.org/en/resources/documents/2014/JC2686_WAD2014report (Accessed on 7 July 2021).
107. Hoffman R, Moyo C, Balakasi KT, et al. 2021. Multimonth dispensing of up to 6 months of antiretroviral therapy in Malawi and Zambia (INTERVAL): a cluster-randomised, non-blinded, non-inferiority trial. *Lancet Glob Health*, 9(5), e628–38.
108. Davies, M.A. 2020. HIV and risk of COVID-19 death: a population cohort study from the Western Cape Province, South Africa. *medRxiv*, 3 July. doi: https://doi.org/10.1101/2020.07.02.20145185.
109. Millett, G.A., Jones, A.T., Benkeser, D., Baral, S., Mercer, L., Beyrer, C., et al. 2020. Assessing differential impacts of COVID-19 on black communities. *Ann Epidemiol*, 47, 37–44.
110. Bhaskaran, K., Rentsch, C.T., MacKenna, B., Schultze, A., et al. 2021. HIV infection and COVID-19 death: a population-based cohort analysis of UK primary care data and linked national death registrations within the OpenSAFELY platform. *Lancet HIV*, 8, e24–32.
111. Boulle, A., Davies, M.A., Hussey, H., Ismail, M., Morden, E., Vundle, Z., et al. 2021. Risk factors for COVID-19 death in a population cohort study from the Western Cape Province, South Africa. *Clin Infect Dis*, 73(7), e2005–15.
112. Jassat, W., Cohen, C., Tempia, S., Masha, M., Goldstein, S., Kufa, T., et al. 2021. Risk factors for COVID-19-related in-hospital mortality in a high HIV and tuberculosis prevalence setting in South Africa: a cohort study. *Lancet HIV*, 8(9), e554–67.
113. Hadi, Y.B., Naqvi, S.F.Z., Kupec, J.T., Sarwari, A.R. 2020. Characteristics and outcomes of COVID-19 in patients with HIV: a multicentre research network study. *AIDS*, 34(13), F3–8.
114. Karim, F., Moosa, M.Y., Gosnell, B., Sandile, C., Giandhari, J., Pillay, S., et al. 2021. Persistent SARS-CoV-2 infection and intra-host evolution in association with advanced HIV infection. medRxiv, 1 Jan. doi: https://doi.org/10.1101/2021.06.03.21258228
115. Spinelli, M.A., Lynch, K.L., Yun, C., Glidden, D.V., Peluso, M.J., Henrich, T.J., et al. 2021. SARS-CoV-2 seroprevalence, and IgG concentration and pseudovirus neutralising antibody titres after infection, compared by HIV status: a matched case-control observational study. *Lancet HIV*, 8(6), e334–41.
116. Collier, D.A., Ferreira, I.A., Kotagiri, P., Datir, R., Lim, E., Touizer, E., et al. 2021. Age-related immune response heterogeneity to SARS-CoV-2 vaccine BNT162b2. *Nature*, 596(7872), 417–22.
117. Frater, J., Ewer, K., Ogbe, A., Pace, M., Adele, S., Adland, E., et al. 2021. Safety and immunogenicity of the ChAdOx1 nCoV-19 (AZD1222) vaccine against SARS-CoV-2 in HIV infection: a single-arm substudy of a phase 2/3 clinical trial. *Lancet HIV*, 8, e474–85.
118. Madhi, S., Koen, A.l., Izu, A., Fairlie, L., Cutland, C.L., Baillie, V., et al. 2021. Safety and immunogenicity of the ChAdOx1 nCoV-19 (AZD1222) vaccine against SARS-CoV-2 in people living with and without HIV in South Africa: an interim analysis of a randomised, double-blind, placebo-controlled, phase 1B/2A trial. *Lancet HIV*, 8, e568–80.

119. Claeson, M., Hanson, S. 2021. COVID-19 and the Swedish enigma. *Lancet*, 397(10271), 259–61.
120. Hall, V.J., Foulkes, S., Charlett, A., Atti, A., Monk, E.J.M., Simmons, R., et al. 2021. SARS-CoV-2 infection rates of antibody-positive compared with antibody-negative health-care workers in England: a large, multicentre, prospective cohort study (SIREN). *Lancet*, 397, 1459–69.
121. Creech, C.B., Walker, S.C., Samuels, R.J. 2021. SARS-CoV-2 vaccines. *JAMA*, 325(13), 1318–20.
122. Forni, G., Mantovani, A., & on behalf of the COVID-19 Commission of Accademia Nazionale dei Lincei, Rome. 2021. COVID-19 vaccines: where we stand and challenges ahead. *Cell Death Differ*, 28, 626–39.
123. Baden LR, El Sahly HM, Essink B, Kotloff, K., Frey, S., Novak, R., et al. 2020. Efficacy and safety of the mRNA-1273 SARS-CoV-2 vaccine. *N Eng J Med*, 384, 403–16.
124. Polack, F.P., Thomas, S.J., Kitchin, N., Absalon, J., Gurtman, A., Lockhart, S., et al. 2020. Safety and efficacy of the BNT162b2 mRNA COVID-19 vaccine. *N Eng J Med*, 383, 2603–15.
125. Voysey, M., Clemens, S.A.C., Madhi, S.A., Weckx, L.Y., Folegatti, P.M., Aley, P.K., et al. 2021. Safety and efficacy of the ChAdOx1 nCoV-19 vaccine (AZD1222) against SARS-CoV-2: an interim analysis of four randomised controlled trials in Brazil, South Africa, and the UK. *Lancet*, 397, 99–111.
126. Fauci, A.S. 2021. The story behind COVID vaccines. *Science*, 372, 6538.
127. Callway, E. 2020. The race for coronavirus vaccines: a graphical guide. *Nature*. Available at: https://www.nature.com/articles/d41586-020-01221-y (Accessed on 7 October 2020).
128. Bian, L., Gao, F., Zhang, J., He, Q., Mao, Q., Xu, M., Liang, Z. 2021. Effects of SARS-CoV-2 variants on vaccine efficacy and response strategies. *Expert Rev Vaccines*, 20(4), 365–73.
129. Johnson & Johnson. 2021. Press release: Johnson & Johnson announces single-shot Janssen COVID-19 vaccine candidate met primary endpoints in interim analysis of its Phase 3 ENSEMBLE Trial. Available at: https://www.jnj.com/johnson-johnson-announces-single-shot-janssen-covid-19-vaccine-candidate-met-primary-endpoints-in-interim-analysis-of-its-phase-3-ensemble-trial (Accessed on 9 February 2021).
130. Logunov, D.Y., Dolzhikova, I.V., Shcheblyakov, D.V., Tukhvatulin, A.I., Zubkova, O.V., Dzharullaeva, A.S., et al. 2021. Safety and efficacy of an rAd26 and rAd5 vector-based heterologous prime-boost COVID-19 vaccine: an interim analysis of a randomised controlled phase 3 trial in Russia. *Lancet*, 397(10275), 671–81. Erratum in: *Lancet* (2021), 397(10275), 670.
131. Palacios, R., Patiño, E.G., de Oliveira Piorelli, R., Conde, M.T.R.P., Batista, A.P., Zeng, G., et al. 2020. Double-blind, randomized, placebo-controlled phase iii clinical trial to evaluate the efficacy and safety of treating healthcare professionals with the adsorbed COVID-19 (inactivated) vaccine manufactured by Sinovac–PROFISCOV: a structured summary of a study protocol for a randomised controlled trial. *Trials*, 21(1), 853.
132. Nature Editorial. 2020. The COVID-19 vaccine challenges that lie ahead. *Nature*, 587, 522.
133. Katz, I.T., Weintraub, R., Bekker, L.-G., Brandt, A. 2021. From vaccine nationalism to vaccine equity—finding a path forward. *N Engl J Med*, 384, 1281–3.
134. Scobie, H.M., Johnson, A.G., Suthar, A.B., Severson, R., Alden, N.B., Balter, S., et al. Monitoring incidence of COVID-19 cases, hospitalizations, and deaths, by vaccination status—13 US jurisdictions, April 4–July 17 2021. *MMWR Morb Mortal Wkly Rep*, 70(37), 1284–90.
135. BBC News.2021. Covid: More cases but fewer deaths in third wave. Available at: https://www.bbc.com/news/uk-wales-57896047 (Accessed on 13 October 2021).
136. Prunas, O., Warren, J.L., Crawford, F.W., Gazit, S., Patalon, T., Weinberger, D.M., et al. Vaccination with BNT162b2 reduces transmission of SARS-CoV-2 to household contacts in Israel medRxiv 2021.07.13.21260393; doi: https://doi.org/10.1101/2021.07.13.21260393
137. Milman, O., Yelin, I., Aharony, N., Katz, R., Herzel, E., Ben-Tov, A., et al. 2021. Community-level evidence for SARS-CoV-2 vaccine protection of unvaccinated individuals. *Nat Med*, 27,

1367–9.
138. Dror, A.A., Eisenbach, N., Taiber, S., Morozov, N.G., Mizrachi, M., Zigron, A., et al. 2020. Vaccine hesitancy: the next challenge in the fight against COVID-19. *Eur J Epidemiol*, 35, 775–9.
139. Loomba, S., de Figueiredo, A., Piatek, S.J., de Graaf, K., Larson, H.J. 2021. Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. *Nat Hum Behav*, 5, 337–48.
140. Machingaidze, S., Wiysonge, C.S. 2021. Understanding COVID-19 vaccine hesitancy. *Nat Med* 27, 1338–9.
141. Joseph, S., Dore, G.J. 2021. Vaccine apartheid: a human rights analysis of COVID-19 vaccine inequity (30 June 2021). Available at: https://ssrn.com/abstract=3876848 or http://dx.doi.org/10.2139/ssrn.3876848 (Accessed on 7 October 2020).
142. Loembé, M.M., Nkengasong, J.N. 2021. COVID-19 vaccine access in Africa: Global distribution, vaccine platforms, and challenges ahead. *Immunity*, 54(7), 1353–62.
143. Mathieu, E., Ritchie, H., Ortiz-Ospina, E., Roser, M., Hasell, J., Appel, C., et al. 2021. A global database of COVID-19 vaccinations. *Nat Hum Behav*, 5, 947–53.
144. Gonsalves, G., Yamey, G. 2021. The COVID-19 vaccine patent waiver: a crucial step towards a 'people's vaccine'. *BMJ*, 373, n1249.
145. Nkengasong, J.N., Ndembi, N., Tshangela, A., Raji, T. 2020. COVID-19 vaccines: how to ensure Africa has access. *Nature*, 586, 197–9.
146. Robbins, R., Mueller, B. 2021. COVID vaccines produced in Africa are being exported to Europe. *New York Times*. 17 August 2021. Available at: https://www.nytimes.com/2021/08/16/business/johnson-johnson-vaccine-africa-exported-europe.html (Accessed on 7 October 2020).
147. Slavov, S.N., Patané, J.S.L., Bezerra, R.D.S., Giovanetti, M., Fonseca, V., Martins, A.J., et al. 2021. Genomic monitoring unveil the early detection of the SARS-CoV-2 B.1.351 (beta) variant (20H/501Y.V2) in Brazil. *J Med Virol*, 10.1002/jmv.27190.
148. Tegally, H., Wilkinson, E., Giovanetti, M., Iranzadeh, A., Fonseca, V., Giandhari, J., et al. 2021. Detection of a SARS-CoV-2 variant of concern in South Africa. *Nature* 592, 438–43.
149. Ramírez, J.D., Muñoz, M., Patiño, L.H., Ballesteros, N., Paniz-Mondolfi, A. 2021. Will the emergent SARS-CoV2 B.1.1.7 lineage affect molecular diagnosis of COVID-19? *J Med Virol*, 93, 2566–8.
150. Konings, F., Perkins, M.D., Kuhn, J.H., Pallen, M.J., Alm, E.J., Archer, B.N., et al. 2021. SARS-CoV-2 variants of interest and concern naming scheme conducive for global discourse. *Nat Microbiol*, 6, 821–3.
151. Chakraborty, C., Bhattacharya, M., Sharma, A.R. 2021. Present variants of concern and variants of interest of severe acute respiratory syndrome coronavirus 2: Their significant mutations in S-glycoprotein, infectivity, re-infectivity, immune escape and vaccines activity. *Rev Med Virol*, 27 Jun, e2270.
152. Fenwick, C., Turelli, P., Pellaton, C., Farina, A., Campos, J., Raclot, C., et al. 2021. A high-throughput cell- and virus-free assay shows reduced neutralization of SARS-CoV-2 variants by COVID-19 convalescent plasma. *Sci Transl Med*, 13(605), eabi8452.
153. Vogels, C.B.F., Breban, M.I., Ott, I.M., Alpert, T., Petrone, M.E., Watkins, A.E., et al. 2021. Multiplex qPCR discriminates variants of concern to enhance global surveillance of SARS-CoV-2. *PLoS Biol*, 19(5), e3001236.
154. Gómez, C.E., Perdiguero, B., Esteban, M. 2021. Emerging SARS-CoV-2 variants and impact in global vaccination programs against SARS-CoV-2/COVID-19. *Vaccines*, 9, 243.
155. Bernal, L.J., Andrews, N., Gower, C., Gallagher, E., Simmons, R., Thelwall, S., et al. 2021. Effectiveness of COVID-19 vaccines against the B.1.617.2 (Delta) Variant. *N Engl J Med*, 385(7), 585–94.

156. Griffin S. 2021. COVID-19: Fully vaccinated people can carry as much Delta virus as unvaccinated people, data indicate. *BMJ*, 374, n2074.
157. Burioni, R., Topol, E.J. 2021. Has SARS-CoV-2 reached peak fitness?. *Nat Med*, 27, 1323–4.
158. Crotty, S. 2021. Hybrid immunity: COVID-19 vaccine responses provide insights into how the immune system perceives threats. *Science*, 372(6549), 1392–3.

SECTION 1

The development of the discipline of public health

1.1 The scope and concerns of public health

1.2 Roger Detels and Chorh Chuan Tan The history and development of public health in developed countries
Simon Szreter

1.3 The history and development of public health in low- and middle-income countries
Than Sein

1.4 Public health priorities in countries undergoing economic transition: the middle-income countries

1.5 Stephen Tollman and Jessica Price Globalization
Kelley Lee

1.1

The scope and concerns of public health

Roger Detels and Chorh Chuan Tan

Introduction to the scope and concerns of public health

There have been many definitions and elaborations of public health. The definition offered by the Acheson Report (Acheson 1988, p. 1) has been widely accepted:

> Public health is the science and art of preventing disease, prolonging life, and promoting health through the organized efforts of society.

This definition underscores the broad scope of public health and the fact that public health is the result of all of society's efforts viewed as a whole, rather than that of single individuals.

In 2003, Roger Detels defined the goal of public health as:

> The biologic, physical, and mental well-being of all members of society regardless of gender, wealth, ethnicity, sexual orientation, country, or political views. (cited in Detels 2009, p. 3)

This definition or goal emphasizes equity and the range of public health interests as encompassing not just the physical and biological, but also the mental well-being of society., This is in keeping with the spirit of the 1948 United Nations declaration that health is a basic human right, and the declaration of the United Nation's Millennium Development Goals. These have now been superseded by the declaration of Sustainable Development Goals (SDGs) in 2012. The 17 SDG goals focus on a broad range of topics including inequalities, poverty, climate change, and other areas essential for achieving optimal health of the public. The SDGs underscore public health as being concerned with more than mere elimination of disease, and place public health issues as a fundamental component of development.

To achieve the World Health Organization (WHO) goal of ‘health for all’, it is essential to bring to bear many diverse disciplines to the attainment of optimal health, including the physical, biological, political, and social sciences. The field of public health has adapted and applied these disciplines for elimination and control of disease, and promotion of health.

Functions of public health

Public health is concerned with the process of mobilizing local, state/provincial, national, and international resources to assure the conditions in which all people can be healthy (Detels and Breslow 2002). To successfully implement this process and to make health for all achievable, public health must perform the functions listed in Box 1.1.1.

Box 1.1.1 Functions of public health

1. Prevent disease and injuries and their negative outcomes
2. Promote healthy lifestyles and good health habits
3. Identify, measure, monitor, and anticipate community health needs (e.g. surveillance)
4. Investigate and diagnose health problems, including microbial, societal, and chemical hazards
5. Formulate, promote, and enforce essential health policies
6. Organize and ensure high-quality, cost-effective public health and healthcare services
7. Reduce health disparities and ensure access to healthcare for all
8. Promote and protect a healthy environment
9. Disseminate health information and mobilize communities to take appropriate action
10. Promote health
11. Plan and prepare for natural and man-made disasters
12. Reduce the impact of interpersonal violence and aggressive war
13. Conduct research and evaluate health-promoting/disease-preventing intervention strategies
14. Develop new methodologies for research and evaluation
15. Train and ensure a competent public health workforce
16. Continually evaluate the effectiveness of existing public health programmes

Source: data from the Office of the Director, National Public Health Performance Standards Program, *10 Essential Public Health Services*, Centers for Disease Control, 1994, available from https://www.cdc.gov/od/ocphp/nphpsp/EssentialPHServices.htm.

Public health *identifies*, *measures*, and *monitors* health needs and trends at the community, national, and global levels through surveillance of disease

and risk factor (e.g. smoking) trends. Analysis of these trends and the existence of a functioning health information system provide the essential information for predicting and anticipating future community health needs.

In order to ensure the health of the population, it is necessary to *formulate*, *promote*, and *enforce* sound health policies to prevent and control disease, and to reduce the prevalence of factors impairing the health of the community. These include policies requiring reporting of highly transmissible diseases and health threats to the community and control of environmental threats through regulation of environmental hazards (e.g. water and air quality standards and smoking). It is important to recognize that influencing politics and promoting sound health policies is an essential function of public health at the local, national, and global levels.

There are limited resources that can be devoted to public health and the assurance of high-quality health services. Thus, an essential function of public health is to effectively *prioritize*, *plan*, *manage*, *coordinate*, and *administer cost-effective health services*, and to ensure their availability to all segments of society. In every society, there are *health inequalities* that limit the ability of some members to achieve their maximum ability to function. Although these disparities primarily affect the poor, minority, rural, and remote populations and the vulnerable, they also impact on society as a whole, particularly in regard to infectious and/or transmissible diseases. Thus, there is not only an ethical imperative to reduce health disparities, but also a pragmatic imperative.

Technological advances and increasing commerce have done much to improve quality of life, but these advances have often come at a high cost to the environment. In many cities of both the developed and developing world, the poor quality of air—contaminated by industry and commerce—has affected the respiratory health of the population and has threatened to change the climate, with disastrous consequences locally and globally (Fig. 1.1.1). We have only one world. If we do not take care of it, we will ultimately have difficulty living in it. Through education of the public, formulation of sound regulations, and influencing policy, public health can contribute much to the protection and monitoring of the environment to *ensure that it is conducive for the population to live healthily*.

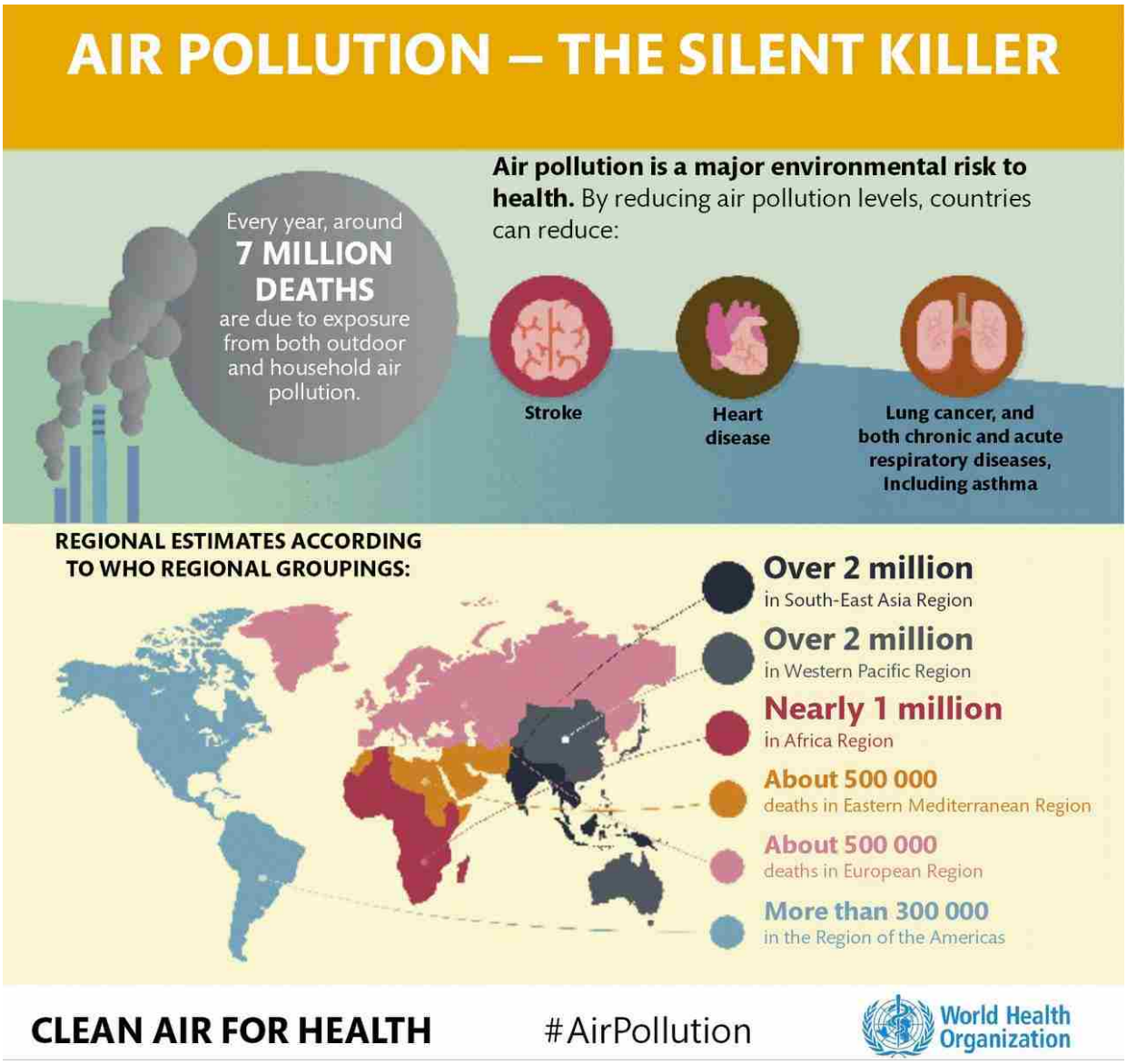


Fig. 1.1.1 Air pollution—the silent killer.
From http://www.euro.who.int/en/health-topics/environment-and-health/air-quality/news/news/2018/5/over-half-a-million-premature-deaths-annually-in-the-european-region-attributable-to-household-and-ambient-air-pollution/infographic-air-pollution-the-silent-killer

Behavioural patterns account for about 40% of all deaths in the United States, while social circumstances and environmental factors account for 15% and 5%, respectively (Schroeder 2007). To ensure that each individual in the population functions to his or her maximum capacity, public health needs to *educate the public, promote adoption of behaviours associated with good health outcomes*, and *stimulate the community* to take appropriate

actions to ensure the optimal conditions for the health of the public. This is vital since the majority of public health problems are linked to human behaviour and lifestyles. Ultimately, public health cannot succeed without the support and active involvement of the community.

We cannot predict and rarely can we prevent the occurrence of natural and man-made disasters, but we can prepare for them to ensure that the resulting damage is minimized. Thus, *disaster preparedness* is an essential component of public health, whether the disaster is an epidemic such as influenza or the occurrence of typhoons and other natural disasters.

Unfortunately, in the modern world, interpersonal *violence* and *war* have become common. In some segments of society (particularly among adolescent and young adult minority males), violence and suicide have become leading causes of death and productive years of life lost. Public health cannot ignore the fact that violence and wars are major factors dramatically reducing the quality of life for millions.

Many of the advances in public health have become possible through *research*. Research will continue to be essential for identifying and anticipating health problems and the optimal strategies for addressing them. Strategies that seem very logical may, in fact, not succeed for a variety of unforeseen reasons. Therefore, public health systems and programmes cannot be assumed to function cost-effectively without continuous monitoring and evaluation. Thus, it is essential that new public health strategies undergo rigorous evaluation before being scaled up, and once scaled up, are periodically reviewed to ensure their continuing effectiveness in diverse groups and populations.

Over the last century, the quality of research has been enhanced by the *development of new methodologies*, particularly in the fields of epidemiology, biostatistics, social, and laboratory sciences. Rapid advances in computational hardware, software, and techniques, including artificial intelligence, have greatly increased our ability to analyse massive amounts of data that incorporate thousands of variables, and to use multiple strategies to aid in the interpretation of data. It will be a major challenge to keep pace with the explosive growth in the volume and complexity of data being generated, driven by a range of factors from the pervasive use of the internet, social media, and mobile phones, to the masses of data from population studies (e.g. census), molecular biology, and whole genome sequencing studies. To glean valuable insights pertinent to public health from these huge

data sets will require new approaches, strategies, and methodologies, such as further development of machine (computer) learning techniques for finding associations with public health outcomes of relevance and discovering patterns such as geographic distributions of relevant variables within geographic subunits that can be targeted for implementation of specific interventions. The increasing use of robotics and artificial intelligence will have a huge impact on public health in the coming decades. Increasingly humans in the workforce will be replaced by robots, especially for performing dangerous and repetitive jobs. These technologic advances are likely to affect low level workers disproportionately. Thus, there will be an increasing demand for training in science, technology, engineering, and mathematics (STEM disciplines) which are essential for designing and programming of robots and other technology advances. It is essential that public health continues to use leading-edge technologies to develop more sophisticated research strategies to address current and future public health issues.

A major problem in public health has been translating research advances into effective health practice and policy in a timely manner. A recent area of research, *implementation science*, attempts to delineate barriers to and factors that facilitate rapid translation of scientific advances into improvements in health practice and development of more effective policies promoting health.

The quality of public health is dependent on the competence and vision of the public health *workforce*. Thus, it is an essential function of public health to *ensure the continuing availability of well-trained and competent professionals* at all levels, including leaders with the vision essential to ensure the continued well-being of society and the implementation of innovative, effective public health measures.

Finally, a common thread that runs through all these functions is the necessity for much greater international collaboration in data sharing, policy formulation and implementation, and management of specific public health issues. With globalization, the rapid flow of information, people, goods, and services across national boundaries means that many public health concerns and issues are interconnected in complex ways. Effective solutions will therefore often depend on joint action between different countries. In addition, now more than ever, there are many opportunities for major public health attainments to be made through cooperative action at the global level,

as has been demonstrated in the past by milestones such as the eradication of smallpox. The media can play an important role in educating the public and in facilitating public health interventions, but often must be educated about public health strategies and goals.

Contemporary health issues

Underlying the bulk of the public health problems of the world is the issue of poverty. More than half of the world's population (more than 3 billion, with one-third being children) live below the internationally defined poverty line of less than US $2.50 per day, with 1.3 billion on less than US $1.25 per day (World Bank 2012). Although the majority of the world's poor live in developing countries, there are many poor living in the wealthiest countries of the world, underscoring the disparity of wealth between the poor and the rich in all countries (United Nations Development Programme 2014). In the United States, 40.6 million Americans were living below the official poverty level in 2016. The proportion was highest among African-Americans and Hispanic-Americans (23.2%). Unfortunately, the disparity between the rich and the poor is increasing within countries (US Census Bureau 2016). Poverty causes a cascade of problems leading to poor health (Table. 1.1.1). Worldwide, 844 million people do not have access to clean drinking water, and at least 2 billion use a drinking water source contaminated with faeces, which causes diarrhoea, cholera, dysentery, typhoid, and polio (WHO Drinking Water). It is incumbent on public health to work to reduce the impact of these disparities to ensure that all members of the global society share in a healthy quality of life.

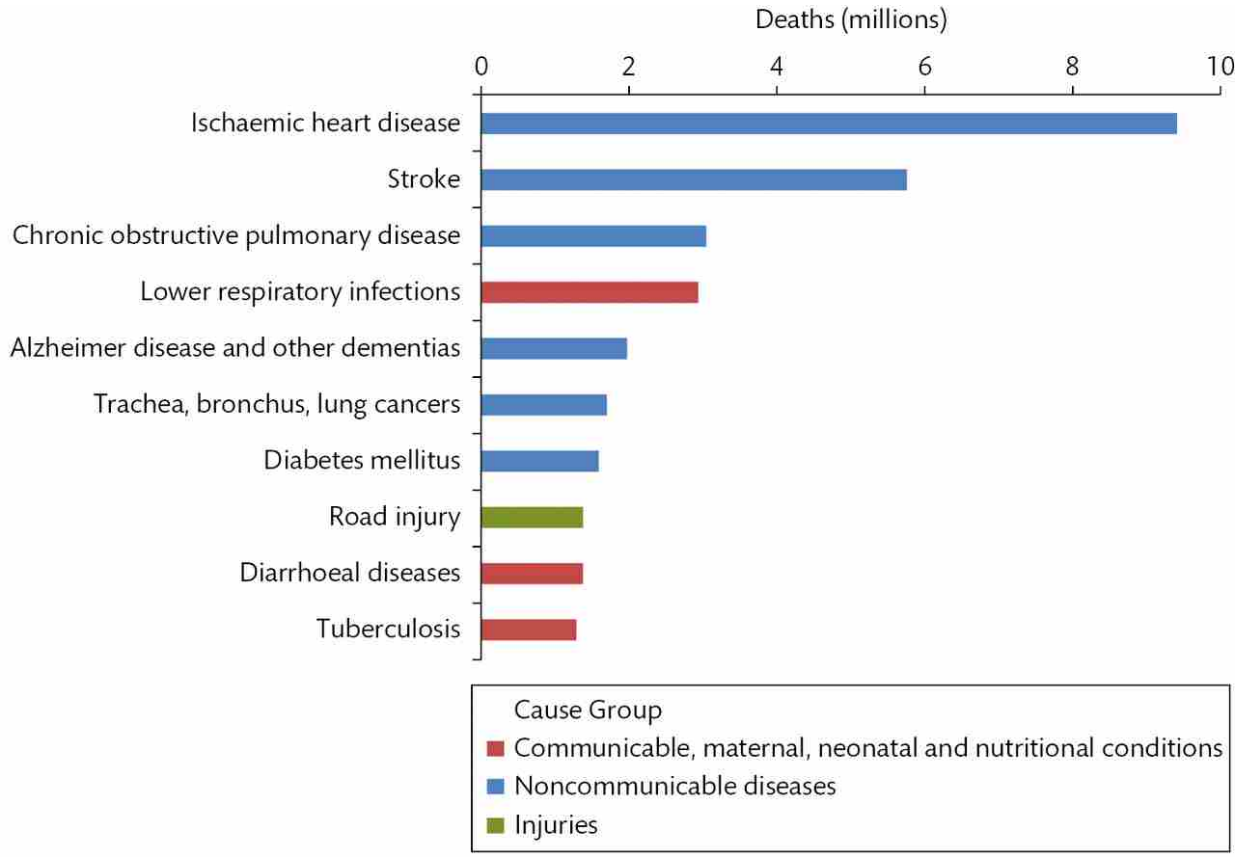
The population of the world is changing. In 2017 the global population was over 7 billion people. By 2050 it is estimated to be more than 9 billion. Population growth will be flat in many areas of the world, but in the African countries the population will continue to increase changing the relative distribution of population across the globe. Meeting these changes will be a challenge for assuring the health of the global population.

Table 1.1.1 Deaths caused by poverty-related diseases

% of deaths caused by/in	High mortality low-income countries	Low mortality low-income countries	High-income countries
Infectious and parasitic disease	34.1	24.8	2.1
Respiratory infections	9.9	8.0	3.2
Perinatal and maternal conditions	8.4	6.8	0.4
Nutritional deficiencies	1.3	1.1	0.0
Tropical diseases	0.5	0.3	0.0
Total ‘poverty-related’ diseases	54.1	40.7	6.2

Source: data from Stevens P. Diseases of poverty and the 10/90 Gap, International Policy Network, 2004.

The twentieth century witnessed the transition of major disease burdens, defined by death from infectious and/or communicable diseases, to non-communicable diseases (NCDs). In 1900, the leading causes of death in the United States and other developed countries were reported to be pneumonia and influenza. By the beginning of the twenty-first century, diseases of the heart and other chronic diseases were the leading causes of death, and pneumonia and influenza were no longer among the top ten causes of death and primarily affected the elderly (Fig. 1.1.2).



Source: Global Health Estimates 2016: Deaths by Cause, Age, Sex, by Country and by Region, 2000–2016. Geneva, World Health Organization; 2018.

Fig. 1.1.2 Top 10 global causes of deaths, 2016 (WHO Global Health Estimates).
Reproduced with permission from World Health organization. (2018) The top 10 causes of death. Copyright (c) 2017 WHO. Available at http://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

The prevalence of chronic diseases will continue to increase further stressing the healthcare system. It also underscores the need for cities, which will accommodate an ever increasing proportion of the population, to innovate and adapt in order to facilitate health promotion, e.g. through providing conducive walking paths and stairs, and supporting those with reduced function and mobility or who have disabilities.

Reductions in communicable diseases were not primarily due to development of better treatments, although vaccines and antibiotics played an important role in the second half of the twentieth century. Instead, public efforts to reduce crowding and improve housing, enhancing nutrition, and providing clean water and safe disposal of wastes were key to reducing communicable diseases.

By 1980, many leading public health figures felt that infectious diseases had been eliminated as a primary concern for public health; however, the discovery and expanding pandemic of acquired immunodeficiency syndrome (AIDS) caused by the human immunodeficiency virus (HIV) in the early 1980s, and subsequently, outbreaks of the severe acute respiratory syndrome (SARS), Ebola, and Zika in the early twenty-first century underscored the fallacy of these declarations as do the persisting high rates of infectious diseases, particularly in Africa.

Although communicable diseases persist as a major public health concern, globally, NCDs have become the leading health challenge, accounting for 40 million or 71.4% of all deaths in 2016 (Fig. 1.1.3). Even in poor developing countries, NCDs are a dominant and growing concern. In 2015, 30.7 million NCD deaths occurred in low- and middle-income countries (World Health Organization 2017b). The age-standardized death rate due to NCDs among males in low- and middle-income countries was 65% higher, and among females, 85% higher than for men and women in high-income countries. This figure is particularly disturbing because low- and middle-income countries have far fewer resources and capacity to address the epidemic of NCDs. Communicable diseases, however, still accounted for almost 30% of the burden of disease worldwide, but caused a majority of deaths only in Africa. The majority of communicable diseases are now preventable through vaccines, improved sanitation, behavioural interventions, and better standards of living, although new diseases emerge that will require public health action.

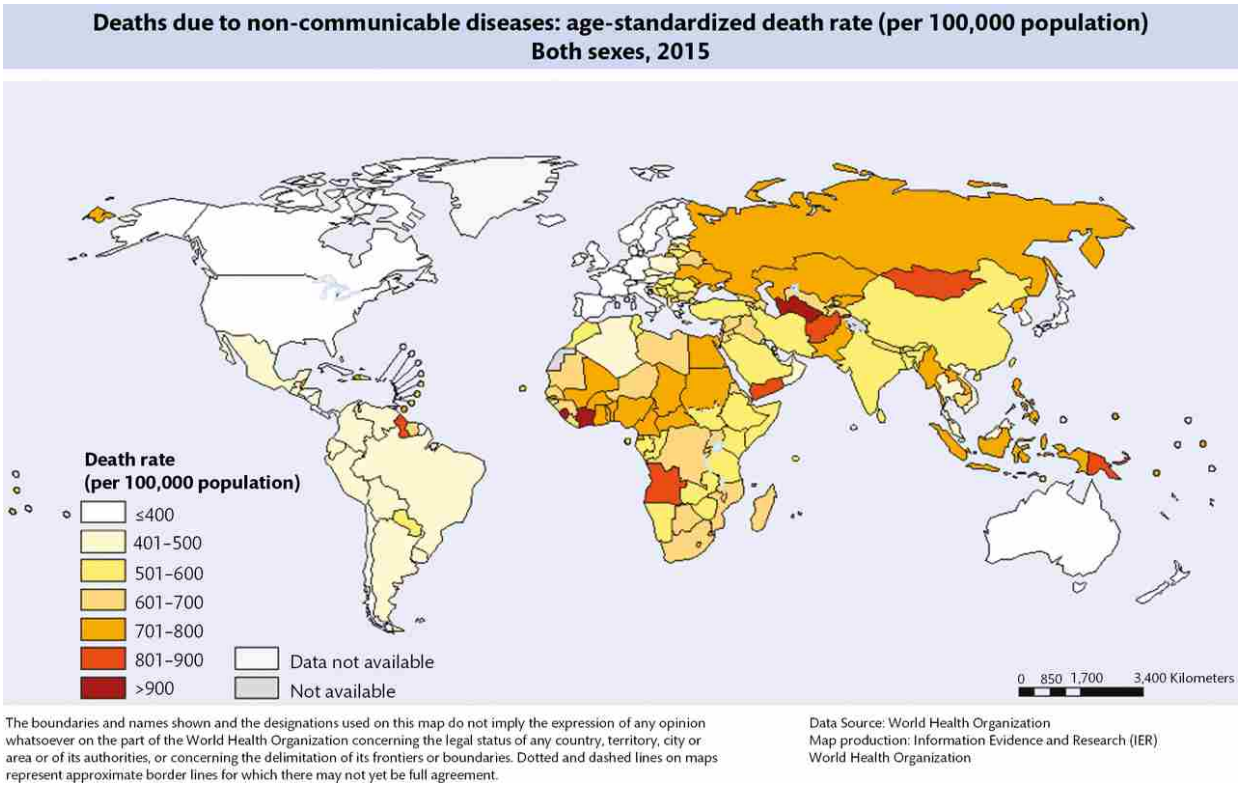


Fig. 1.1.3 Deaths due to noncommunicable diseases, 2015.

Reproduced with permission from World Health organization. (2017) Deaths due to noncommunicable diseases: age standardized death rate (per 100 000 population). Copyright (c) 2017 WHO. Available at http://gamapserver.who.int/mapLibrary/Files/Maps/Global_NCD_deaths_2015.png

Compounding the global shift to NCDs is the rapidly rising age of populations in many countries, due to increased longevity (71.4 years globally) and dramatically decreased birth rates (United Nations 2015a). Population growth is already below replacement in many countries, both developed and developing. This demographic shift to an ageing population has widespread and profound implications (Fig. 1.1.4). It will increase the burden of chronic disease in these countries, place increasing demands on healthcare and social support services, and strain public financing systems. This will occur as the proportion of the population in the productive ages decreases, which in turn will impact economic growth. The low- and middle-income countries in particular will be affected by the ageing of their populations.

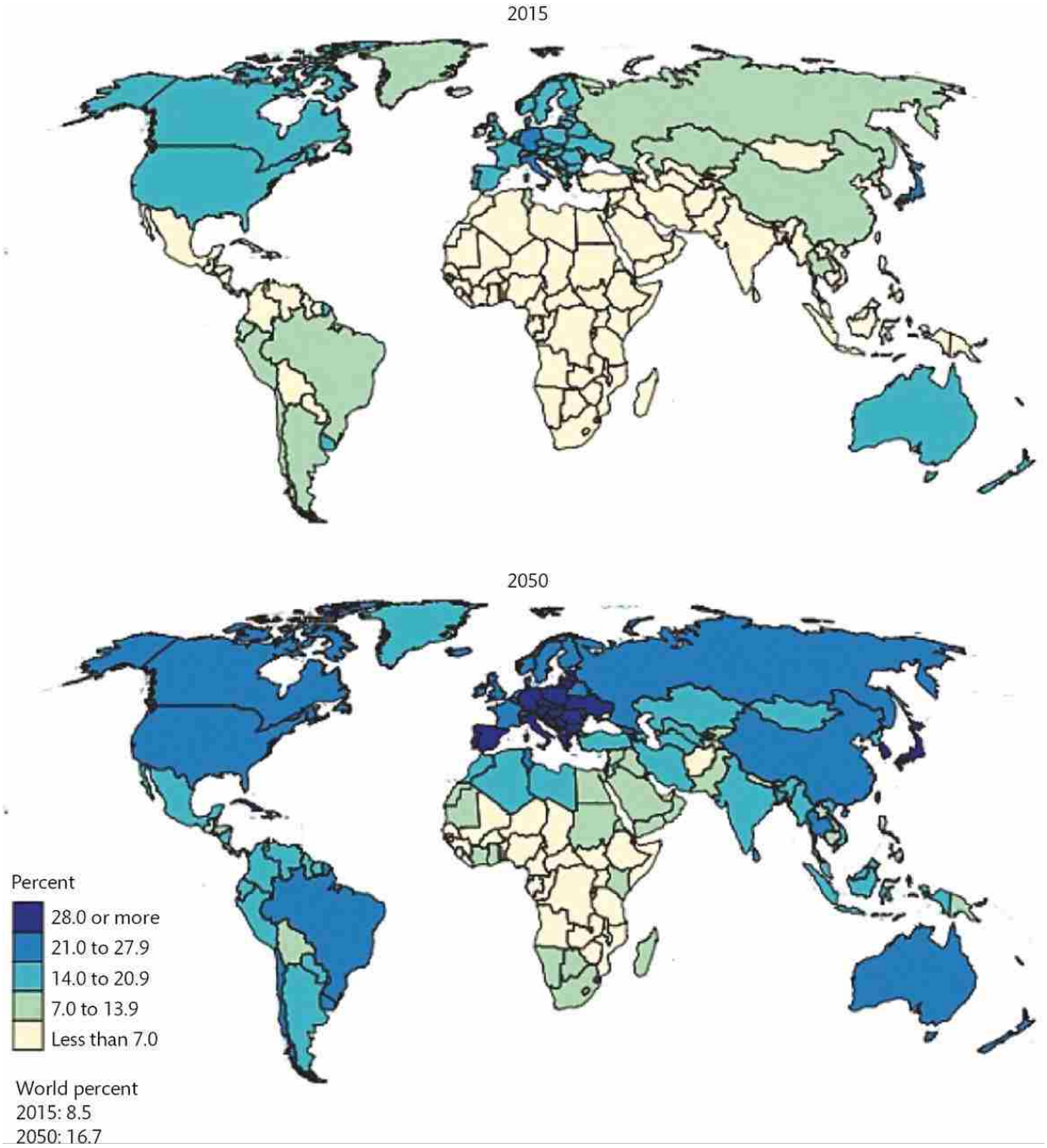


Fig. 1.1.4 Percentages of population aged 65 and over in 2015 and 2050.
Source: data from US Census Bureau, 2013, 2014a, 2014b, International Data Base, U.S. Population Estimates and U.S. Population Projections.

Measuring disease occurrence

An essential step in defining health is to identify appropriate methods for measuring it. Traditionally, public health has defined disease in terms of mortality rates because they are relatively easy to obtain and death is indisputable. The use of mortality rates, however, places the greatest emphasis on diseases that end life, and tends to ignore those which compromise function and quality of life without causing death. Thus, the problems of mental illnesses, accidents, and disabling conditions are seriously underestimated if one uses only mortality to define them.

Two other strategies to measure health that evolved in the last half of the twentieth century have been 'years of productive life lost' (YPLL) (Lopez et al. 2007) and 'disability-adjusted life years' (DALYs) (Murray and Lopez 1997). The former emphasizes those diseases that reduce the productive lifespan (currently arbitrarily defined as 75 years), whereas the latter emphasizes those diseases that compromise function and also includes a measure of premature mortality. Using either of these alternatives to define health results in very different orderings of diseases and/or health problems as public health priorities.

Using death to identify disease priorities, the leading cause is NCDs, which account for 70% of diseases worldwide (Fig. 1.1.3). Among the chronic diseases, cardiovascular diseases account for half of the deaths. The proportion, however, varies markedly by regions of the world and level of affluence of the countries. Communicable diseases remain the major cause of death only in Africa, although they account for a significant proportion of deaths in Southeast Asia and the eastern Mediterranean. The majority of victims of these communicable diseases are infants and children under 5 years old. The persistence of communicable diseases in these areas represents a continuing major public health challenge.

DALYs and YPLL may be considered better measures of the quality of life and functioning capacity of a country than mortality (Figs. 1.1.5a, 1.1.5b, and 1.1.6). Using DALYs to establish global disease priorities emphasizes communicable diseases and injuries, which tend to disproportionately affect the young, and reduces the relative importance of cardiovascular diseases and other chronic diseases that primarily affect the elderly. The WHO has projected that the ranking of total DALYs for neuropsychiatric disorders, injuries, and non-communicable and/or chronic diseases will increase by 2020, whereas the ranking for communicable diseases will decline. Communicable diseases, which currently account for

39% of global DALYs, are expected to decline to 30% by 2030 (Mathers and Loncar 2006).

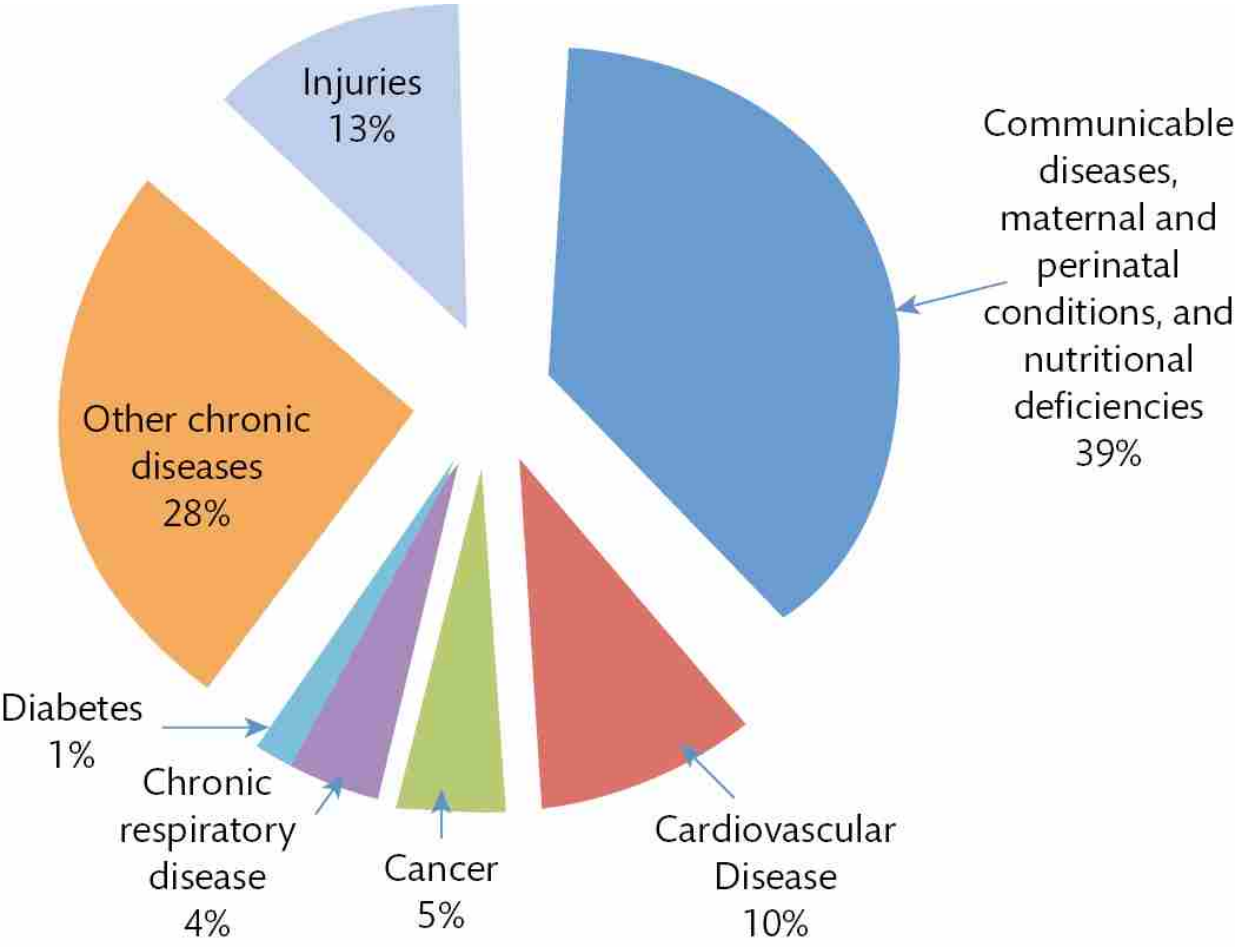


Fig. 1.1.5 Global burden: disability-adjusted life year (DALY).
From https://ourworldindata.org/health-meta under the CC BY 4.0 license.

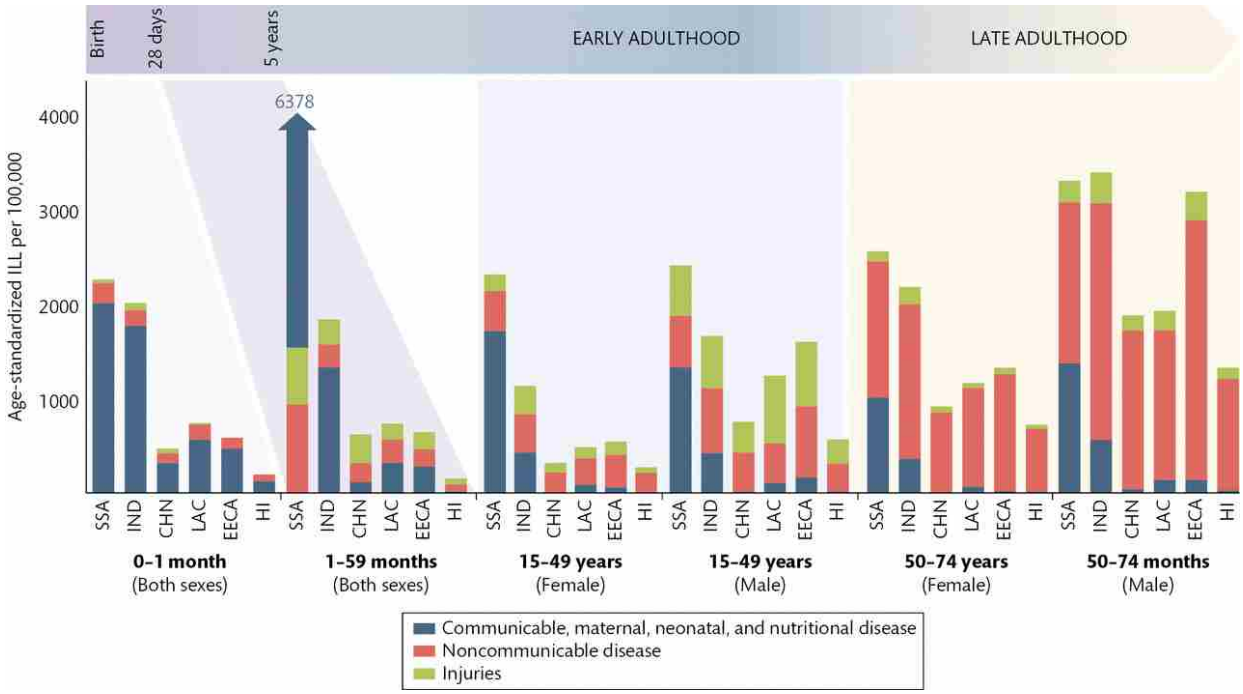


Fig. 1.1.6 YLL (potential years of life lost): estimator of the impact of premature death by world regions.
From https://libguides.acomedu.org/GlobalHealth.

According to projections by the WHO, while lower respiratory infections and diarrhoea remain the most prevalent communicable diseases, infections such as HIV, tuberculosis, and malaria will rise in terms of YPLL per 1,000 population by 2030, even as other communicable diseases will yield to intervention efforts and account for progressively fewer YPLL (World Health Organization n.d.). The YPLL per 1,000 population due to NCDs that primarily affect older people, however, is projected to remain constant, perhaps reflecting the optimism regarding the development of strategies for earlier diagnosis, better health habits, and better drugs to sustain life in patients with these conditions.

In the 1990s syndemicity introduced the concept that diseases occur in specific contexts. These include such factors as the environment (physical and social/psychologic) in which disease occur and the concept that there are interactions between diseases which affect vulnerability and severity. This theory emphasizes that disease occurs as the result of a constellation of factors, only some of which are biologic (see Fig 1.1.7). This theory, thus, suggests the need for a broader approach to public health promotion and disease intervention (Mendenhall 2017).

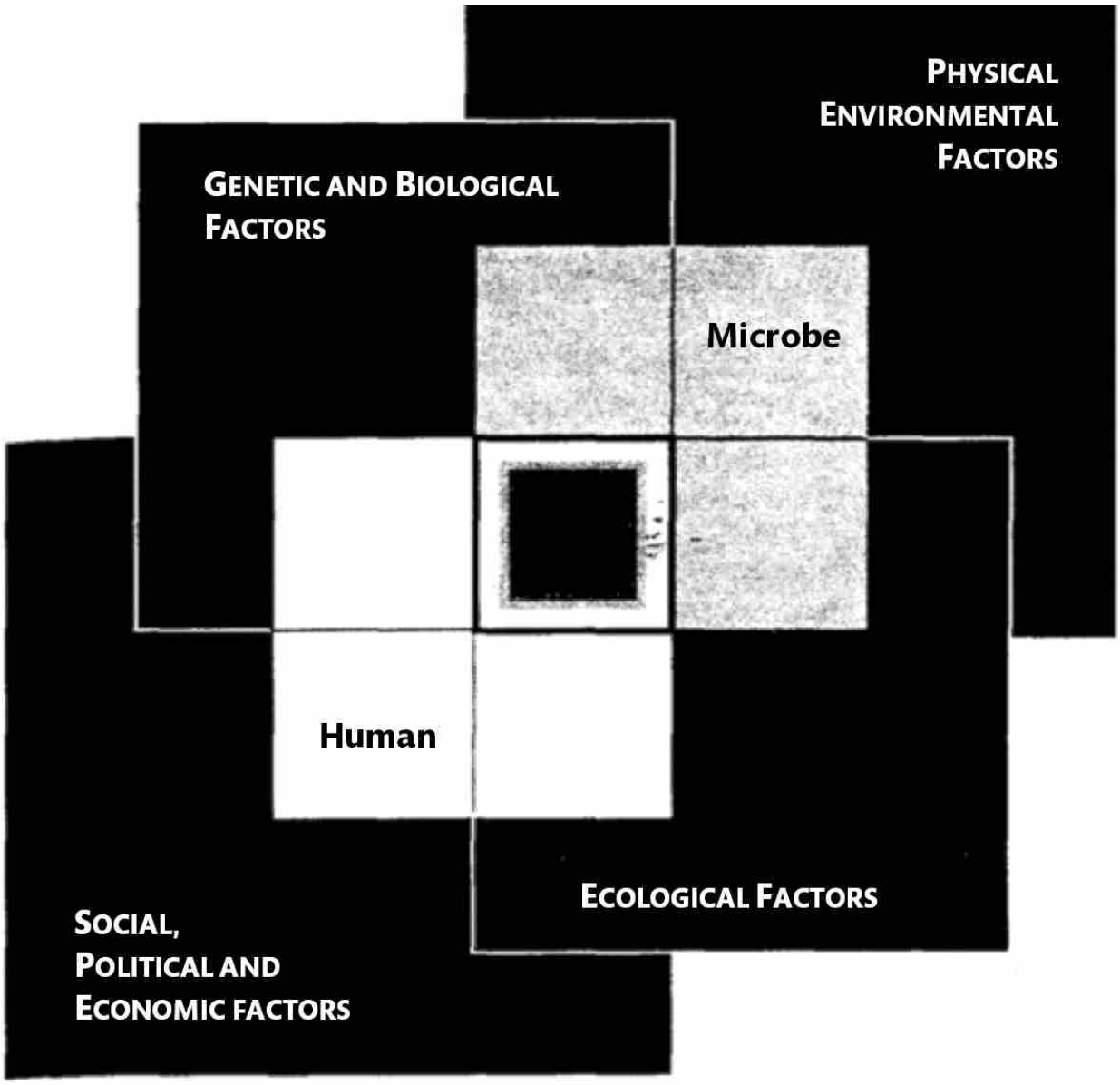


Fig. 1.1.7 The convergence model.
Reprinted from Smolinski, S.M., Hamburg, M.A., Lederberg, J. (editors) (2003). *Microbial Threats to Health: Emergence, Detection, and Response, Institute of Medicine, Board on Global Health, Committee on Emerging Microbial Threats to Health in the 21st Century*, Figure ES-1, p. 5, with permission from The National Academies Press.

Communicable diseases

Many new vaccines against infectious agents have been and are being developed, and many have become more affordable (Qadri et al. 2016). The WHO's regional offices working with individual countries have conducted intensive immunization programmes against the major preventable infectious diseases of childhood, but there are significant barriers to complete coverage, including poverty, geographic obstacles, low levels of education affecting willingness to accept vaccination, logistical problems, civil unrest and wars, corruption, and mistrust of governments. Poverty, weak governments, and misuse of funds have also prevented control of disease vectors that play a key role in diseases such as malaria and dengue, provision of clean water, and safe disposal of human waste, all of which are essential for control of communicable diseases.

Another major factor in the rapid spread of communicable diseases has been the rapid growth in transportation. It is now possible for an individual with a communicable disease to circumnavigate the globe while still infectious but asymptomatic. Thus, cases of SARS were reported throughout Southeast Asia and as far as Canada within weeks of the recognition of the first cases in Hong Kong (Lee 2003). Similarly, due to the extensive global food supply chains, food-borne infections can spread rapidly within and across countries.

Another source of communicable diseases is the continuing emergence of new infectious agents, many of them adapting to humans from animal sources. Box 1.1.2 lists new disease outbreaks from 1994 to 2019, including newly drug-resistant variants of new diseases and emerging infections such as MERS, Zika, and SARS-CoV2 occurring worldwide. Changes in food production, crowding of animals, mixing of live animal species in 'wet markets' (selling live animals for food) in Asia and elsewhere, and the introduction of hormones and antibiotics into animal feed have all contributed to the emergence of these new diseases.

In early January 2020 a new agent, SARS-CoV2, was recognized. The virus rapidly spread globally from China to cause a pandemic which, in the first 4 months since its identification, caused over 3.25 million cases and over 226,000 deaths severely impacting the global economy and the way society functions (Wu and McGoogan 2020; Singhal 2020). As the seventh edition of the *Oxford Textbook of Global Public Health* goes to press, many key characteristics of the virus are unknown making predictions difficult,

although it is clear that the pandemic will have a major impact on the future of the planet (Chapter 8.13).

Box 1.1.2 Newly identified infectious diseases and pathogens

Year	Disease or pathogen
2019	SARS-CoV2
2018	Acute flaccid myelitis
2012	Zika virus
2012	MERS-CoV (don’t kiss camels!)
2009	H1N1
2004	Avian influenza (human cases)
2003	SARS
1999	Nipah virus
1997	H5N1 (avian influenza A virus)
1996	New variant of Creutzfeldt–Jacob disease; Australian bat lyssavirus
1995	Human herpesvirus 8 (Kaposi’s sarcoma virus)
1994	Savia virus; Hendra virus

Source: data from workshop presentation by David Heymann, World Health Organization, 1999.

In addition to the diseases listed in Box 1.1.2, antibiotic-resistant strains of known agents have emerged rapidly due, in part, to the widespread inappropriate use of antibiotics (see also Box 1.1.3). Thus, resistant strains of gonorrhoea, *Staphylococcus*, tuberculosis, and malaria have become major problems, with the latter two being among the most important infectious disease problems globally. Development of drug-resistant malaria has been compounded by the emergence of vectors resistant to the commonly used chemical insecticides. The frightening potential for rapid and dangerous spread of antibiotic resistance through genetic elements that are transmissible between different bacterial species was highlighted by

reports of the speed and extent to which New Delhi metallo-beta-lactamase (NDM-1), which encodes multiple-antibiotic resistance, spread to many different countries (Moellering 2010).

More than one billion people, one-seventh of the world's population, suffer from one or more tropical disease, including Buruli ulcer, Chagas' disease, cholera, dengue, dracunculiasis, trypanosomiasis, leishmaniasis, leprosy, lymphatic filariasis, onchocerciasis, schistosomiasis, helminthiasis, and trachoma (World Health Organization 2015). The functional ability of those afflicted is severely compromised, in turn affecting the economic competitiveness of the poorest countries, which suffer the greatest burden of these tropical diseases. However, major strides have been achieved in reducing the burden of diseases such as leprosy, guinea worm disease, and lymphatic filariasis. Continuing efforts are needed to further reduce the impact of these and other tropical diseases.

Box 1.1.3 Factors contributing to the emergence or re-emergence of infectious diseases

1. Human ‘demographic change’ by which persons begin to live in previously uninhabited remote areas of the world and are exposed to new environmental sources of infectious agents, insects, and animals
2. People living in close proximity to domestic animals and poor animal husbandry in many parts of the developing world, leading to zoonotic infections
3. Breakdown of sanitary and other public health measures in overcrowded cities and in situations of civil unrest and war
4. Economic development and changes in the use of land, including deforestation, reforestation, and rapid urbanization
5. Climate changes causing changes in the geographic range of agents and vectors
6. Changing human behaviours, such as increased use of child-care facilities, sexual and drug-use behaviours, and patterns of outdoor recreation
7. Social inequality
8. International travel and commerce that quickly transport people and goods vast distances
9. Changes in food processing and handling, including foods prepared from many different animals and transported great distances
10. Evolution of pathogenic infectious agents by which they may infect new hosts, produce toxins, or adapt by responding to changes in the host immunity (e.g. influenza, HIV)
11. Development of resistance of infectious agents such as *Mycobacterium tuberculosis* and *Neisseria gonorrhoeae* to chemoprophylactic or chemotherapeutic drugs
12. Resistance of the vectors of vector-borne infectious diseases to pesticides
13. Immunosuppression of persons due to new medical treatments or new diseases that result in vulnerability to infectious diseases caused by agents not usually pathogenic in healthy persons (e.g. leukaemia patients, AIDS)
14. Deterioration in surveillance systems for infectious diseases, including laboratory support, to detect new or emerging disease problems at an early stage
15. Illiteracy limiting knowledge and implementation of prevention strategies
16. Lack of political will—corruption, competing priorities
17. Biowarfare/bioterrorism—an unfortunate potential source of new or emerging disease threats (e.g. anthrax)
18. War, civil unrest—creating refugees, food and housing shortages, increased density of living, etc.
19. Famine
20. Declining immunization rates (e.g. measles, mumps)

We now recognize that we will continue to see new human pathogens emerge in the future, and need to be prepared to contain them. Unless the world realizes the consequences of not protecting the environment in which we live and acts upon it, newly emerging diseases will continue to plague us.

Non-communicable diseases

With increasing control of communicable diseases and increasing lifespan, NCDs have emerged as the major global health problem in both developed

and developing countries. Even in developing countries, NCDs have assumed greater importance. The prevalence of type 2 diabetes in rural India is 7.3% (Anjana et al. 2017). Cardiovascular diseases have become a major cause of death in China. During 2000–2008, the incidence of stroke in low- and middle-income countries exceeded that in high-income countries by 20% (Feigin et al. 2009).

The causes of NCDs are many and complex. Although the immediate causes are factors such as elevated blood pressure, increased blood glucose, abnormal lipids and fat deposition, and diabetes, the underlying causes are behavioural and social. These behavioural factors include unhealthy diets that substitute pre-packaged and fast foods high in fats for a balanced diet, physical inactivity, and, especially, tobacco use; these in turn are the products of social change, including globalization, urbanization, economic improvement, and ageing. The WHO estimated that globally, insufficient physical activity contributes to 3.2 million deaths and 69 million DALYs annually (Barreto 2013), and that obesity contributes to 2.8 million deaths and 35.8 DALYs globally (World Health Organization 2017b). Some NCDs have been associated with infectious disease agents. For example, *Chlamydia pneumoniae* has been implicated in the development of atherosclerosis (Kuo and Campbell 2000), hepatitis C is a leading cause of hepatocellular (liver) cancer, and human papilloma virus (HPV) is the major cause of cervical cancer. Recently, an effective vaccine against HPV was developed that protects against cervical cancer, but it is expensive, must be administered before sexual activity begins (i.e. early adolescence) and requires multiple injections.

Fig. 1.1.8 shows the global distribution of deaths from NCDs, demonstrating the higher rates of death from NCDs in developing countries, especially in Africa. Another aspect of NCDs is the increasing survival of treated individuals who would not have survived as long previously. Many of them are left with disabilities that require modified environments to allow a reasonable quality of life and to realize their full potential in order to contribute to society. Most NCDs can be reduced by a combination of healthy behaviours, including not smoking, moderate alcohol use, and exercise (Breslow and Breslow 1993). Many developed countries have been promoting healthy lifestyles, but there is a need for greater emphasis and development of these programmes in developing countries, where the major global burden of chronic diseases exists.

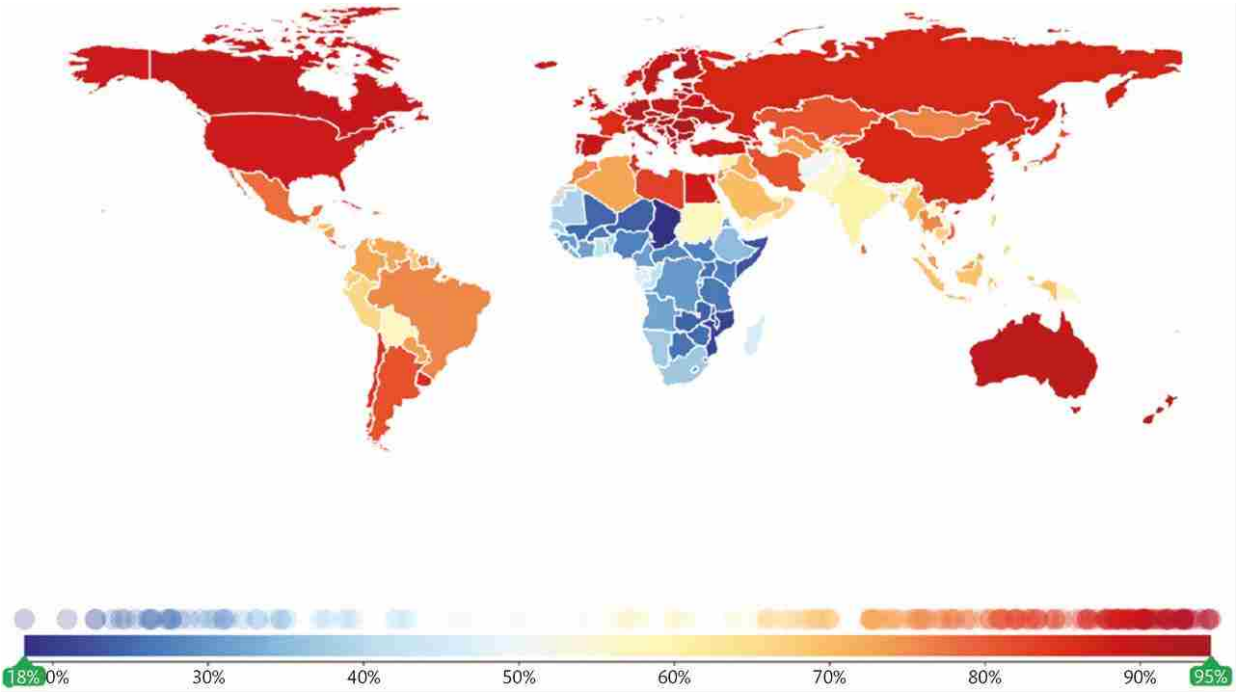


Fig. 1.1.8 Percentage of deaths from non-communicable diseases by country, 2013.
Reprinted from Mokdad A. (2016) "Global non-communicable disease prevention: Building on success by addressing an emerging health need in developing countries" *The Journal of Health Specialties* 4(20: 92–104), with permission from Wolters Kluwer: Medknow.

Mental illness

Public health professionals have only relatively recently recognized the importance of addressing the mental health needs of society on a global scale, partly due to the difficulties in defining it. It is now estimated that 1.1 billion people suffer from mental health problems, and that psychiatric disorders account for 15.5% of the global burden of disease (Collins et al. 2011) (Table 1.1.2). Mortality rates seriously underestimate the burden of mental health on society and the true extent of mental illness is probably greater.

Table 1.1.2 Global percentages of mental disorders

Disorder	Share of global population with disorder (2017) [difference across countries]	Number of people with the disorder (2017)	Share of males:females with disorder (2017)
Any mental health disorder	10.70%	792 million	9.3% males
			11.9% females
Depression	3.40%	264 million	2.7% males
	[2–6%]		4.1% females
Anxiety disorders	3.80%	284 million	2.8% males
	[2.5–7%]		4.7% females
Bipolar disorder	0.60%	46 million	0.55% males
	[0.3–1.2%]		0.65% females
Eating disorders (clinical anorexia & bulimia)	0.20%	16 million	0.13% males
	[0.1–1%]		0.29% females
Schizophrenia	0.30%	20 million	0.26% males
	[0.2–0.4%]		0.25% females
Any mental or substance use disorder	13%	970 million	12.6% males
	[11–18%]		13.3% females
Alcohol use disorder	1.40%	107 million	2% males
	[0.5–5%]		0.8% females
Drug-use disorder (excluding alcohol)	0.90%	71 million	1.3% males
	[0.4–3.5%]		0.6% females

Reprinted from Hannah Ritchie and Max Roser (2020)—'Mental Health'. Published online at OurWorldInData.org. Retrieved from: 'https://ourworldindata.org/mental-health' [Online Resource] under the CC BY 4.0 license.

Updates of global provisions for treatment of mental illness are still significantly below what is necessary to adequately address the problem.

The global median mental health expenditure per capita is US$ 2.5. Based on the WHO Global Health Expenditure database, the global median of domestic general government health expenditure per capita in 2015 was US$ 141, thus making government mental health expenditure less than 2% of global median of government health expenditure, with large variations between world regions. Levels of public expenditure on mental health are very meagre in low and middle-income countries, and more than 80% of these funds go to mental hospitals. Seventeen per cent of WHO Member States reported that persons pay mostly or entirely out of their own pockets for mental health services and 18% pay for psychotropic medicines themselves. The highest rates for out-of-own-pocket expenditures on mental health were noted in African and Southeast Asian regions. Globally, the median number of mental health workers is 9 per 100,000 population, but there is extreme variation (from below 1 in low-income countries to 72 in high-income countries) (World Health Organization 2017a). Of those with mental illness in developing countries, 76–85% do not have access to appropriate care. Although 87% of the world's governments offer some mental health services at the primary care level, 30% of them have no relevant programme and 28% have no budget specifically identified for mental health. Mental illness robs society of a significant number of potentially productive persons. With the diminishing proportion of productive people of working age and the increasing proportion of elderly dependents, it is important to assist those who are not productive because of mental illness to become healthy, productive members of society.

Population projections

Although the rate of growth of the world's population has slowed in the latter half of the twentieth century, the world's population, currently (2018) over 7.6 billion people, is still estimated to grow to 9.8 billion by 2050 (World Health Organization 2017b). An increasing proportion of the population will be mostly among the elderly and the old elderly (those over 80 years of age).

The well-being of society is dependent on the ratio of those who produce to those who are dependent. Improved technology and strategies will be required to increase worker productivity. The majority of the population growth in the coming decades will be among the old and old elderly, not

through increasing birth rates, and will result in a diminishing proportion of producers and an increasing proportion of dependents. It is expected, however, that more of the elderly will need to continue to be economically productive for longer periods of time. In 2017, the proportion of the world's population that was 60 years and older was close to 1 billion (13% of the global population and doubled since 1980), and is growing at approximately 3% each year, with the total expected to double to 2 billion by 2050 (World Population Prospects: the 2017 Revision). East Asia and the Pacific have the greatest increases in ageing populations. This will be further exacerbated because the majority of the oldest elderly will be single women who traditionally have more limited resources and lower levels of education, particularly in developing countries. The productivity and efficiency of those who produce must increase if we are to sustain and improve the quality of life for all.

The occurrence of disease in old age is directly correlated with unhealthy behaviours developed in early life. Unfortunately, concurrent with population growth, there has been a worldwide epidemic of obesity and decreased physical activity, which has increased the proportion of elderly who suffer from chronic debilitating diseases in both the developed and developing world. Thus, unless efforts to promote healthy lifestyles are successful, not only will there be an increase in the proportion of elderly, there will also be an increasing proportion of them who will require assistive care, placing a further economic and social burden on families and society.

As the population grows, there is increasing pressure to provide food, water, and other necessities to maintain a high quality of life. Shifts in dietary preferences in developing countries towards greater meat consumption also put additional strains on food production. Fertile farmlands are increasingly being converted to residential, commercial, and industrial use. Thus, more people will effectively need to be supported on less arable land. Food security will hence be a key issue for the future and this will be affected by a multitude of factors, most notably, insufficient water resources and whether there will be major increases in agricultural productivity. It is not clear how large a population planet earth can sustain.

Other public health issues

Nutrition

Appropriate nutrition is essential for health. In many developing countries (as well as among the poor and homeless in developed countries), undernutrition is a problem. Beyond access to nutritious and safe food, many of the poor have little knowledge about what constitutes a healthy diet, compounding the difficulties. At the same time, in developed and many rapidly developing countries such as China, overnutrition and obesity are a major problem underlying the current epidemic of diabetes. Industrial sectors have been more effective in promoting calorie-dense, salt-rich, and unhealthy foods than the efforts of public health professionals to champion healthy food choices consumed in reasonable quantities. This has particularly been the case for the young, and has contributed to rising childhood obesity in many countries, including the more affluent in developing countries.

Oral health

Good dental health is essential for maintaining adequate nutrition and a good quality of life. Worldwide, however, 60–90% of school children and nearly 100% of adults have dental cavities. About 30% of adults aged 65–74 years have lost all of their teeth. It was also estimated that globally, there was an average of 1.6 decayed, missing, or filled teeth among children aged 12 years old (FDI World Dental Federation, 2015). These high rates of dental problems reflect poor dental hygiene and preventive care (Pine and Harris 2007). Unfortunately, many people believe that dental care is an expendable luxury, and that visits to dentists are only necessary when there is a problem. Oral cancers are the eighth most common cause of cancer globally and the third most common in Asia. The estimated incidence, mortality and 5-year survival due to lip and, oral cavity cancer in world is 300,373 (2.1%) 145,328 (1.8%) and 702,149 (2.2%), respectively (Gupta et al. 2016). Approximately 90% of all oral cancers are squamous cell carcinomas, and the 5-year survival rate of oral cancer is estimated to be 50% globally (FDI World Dental Federation, 2015). Worldwide, approximately 450,000 new cases are found each year. Death rates are high, due to the disease usually being found in late stages (Oral Cancer Foundation). Poor dental hygiene is an important risk factor, together with smoking, as well as the habit of betel nut chewing, which is common in parts of Asia. Clearly the public health

message regarding the importance of good dental hygiene, regular brushing of teeth, and regular dental check-ups is not reaching the majority of the people.

Injuries

Injuries and violence cause about 5.8 million deaths annually (about 10% of all global deaths), nearly 1.7 times the number of fatalities that result from HIV/AIDS, tuberculosis, and malaria combined, of which about one-third are due to interpersonal violence, one-quarter to traffic accidents, and 15% to suicide. Tens of millions of people suffer non-fatal injuries that require treatment. Deaths due to injuries are almost three times greater in developing than in developed countries. Furthermore, they occur more commonly among younger persons and children; for those aged 5–44 years, injuries are the third leading cause of death (World Health Organization 2014a). Injuries can be broadly categorized into the following groups: motor vehicle accidents, suicide, homicide, and unintentional injuries, including occupational injuries and falls. Motor vehicle accidents account for the largest proportion of deaths due to injury (23%; over 3,400 deaths each day), and tens of millions of people are injured or disabled every year (World Health Organization 2014b). Falls, particularly among the elderly, are also a major cause of DALYs.

Unintentional injuries are largely preventable through community and governmental intervention. Thus, improved roads, separation of different modes of transportation, enactment, and enforcement of seat belt and helmet laws, implementation, and enforcement of safe practices at work, and improved designs of houses, automobiles, ladders, and other equipment and tools have all been shown to significantly reduce injuries and deaths due to accidents.

Homicide, violence, and suicide

Homicide, violence, and suicide represent a growing problem, particularly among the young. Homicide and suicide are among the leading causes of death globally. In some minority groups in the United States, homicide and violence are the leading cause of death of youth, followed by suicide. Globally, approximately 30% of women have experienced physical and/or

sexual violence from their intimate partners (World Health Organization 2014d). In China, suicide remains the leading cause of death among women in rural areas. Globally, the WHO predicts that homicide and suicide will account for an increasing proportion of deaths.

Vulnerable populations

Public health has always been concerned with the health and well-being of vulnerable groups who require special attention. The definition of a vulnerable population varies by time, situation, and culture, but the common characteristic across all vulnerable groups is their special susceptibility to poor health and quality of life. The list of vulnerable groups includes the poor, minorities, women, children, the elderly, the handicapped, the illiterate, orphans and street children, immigrants, rural-to-urban migrants, refugees and displaced people, the homeless, and the mentally ill. In certain situations, other groups may be considered vulnerable. For example, in the face of epidemics such as HIV/AIDS, one should also consider adolescents to be a vulnerable group. Often vulnerable individuals live at the margins of society and have difficulty accomplishing the basic functions of living and accessing healthcare. Thus, they require assistance. In many societies, particularly in developing countries, the family acts as the safety net for these groups, but if the family itself is vulnerable or dysfunctional, this safety net is compromised. Societies with sufficient resources have developed social support programmes that assist the vulnerable, but these programmes seldom cover the full range of vulnerable groups, and may not adequately support those whom they target. Universal access to healthcare is one component of assisting the vulnerable, but presently, even in wealthy developed countries such as the United States, healthcare is not available to all, and strategies to fund universal healthcare are difficult to implement.

In almost every country, developed and developing, there are homeless people, many of whom suffer from multiple problems, including mental illness. The ability to function adequately and achieve good health among many vulnerable groups, including the homeless, mentally ill, alcoholics, and drug addicts, is adversely impacted by additional factors such as poverty, prejudice, and stigmatization by society. The US is experiencing an epidemic of overuse of opioid drugs, due to overprescribing. Thus, programmes to assist the vulnerable need to also encourage society to take

supportive action in order to be optimally effective. This is a particular challenge in respect to persons with handicaps. Many developed countries have provisions for persons with handicaps, but in poorer countries, those with handicaps face substantial difficulties to function in society, and many do not survive.

Over the last decade our understanding of the diversity of society has increased as we have recognized, for example, that gender encompasses more than the dichotomy of male and female and is actually a spectrum of sexual orientations and behaviours. Traditionally society has either ignored or ostracized groups that deviate from the 'norm'. In order to achieve true equity it is incumbent on public health to champion diversity and equality as we recognize more diverse subgroups in our society.

In designing programmes for vulnerable groups, a further complication is the fact that the specific problems and needs of each of these groups differ, and they thus require public health actions which are tailored to their requirements. For some of these groups, such as mothers and children and the handicapped, there are well-established programmes, although coverage is far from complete and the quality of these programmes varies widely. For others, such as illiterate persons and migrants, there are fewer established programmes. If we are to meet the public health goal of 'health for all', we need to identify and assist the vulnerable groups within societies to achieve their maximum possible health and function.

The environment

> Environmental health comprises those aspects of human health, including quality of life, that are determined by physical, chemical, biological, social, and psychosocial processes in the environment.
>
> *Draft definition developed at a WHO consultation in Sofia, Bulgaria, 1993* (WHO Regional Office for Europe 1994)

Currently, one of the major problems that the world faces is deterioration of the environment caused by the increasing numbers of people and the accumulation of wastes produced by them, their vehicles, and the industries they support. Thus, the quality of the air that we breathe has declined, especially in developing countries where rapid economic growth has been achieved at the expense of the environment (Table. 1.1.3). The most polluted cities of the world are concentrated in developing countries, which often

have the least capacity and political will to reduce pollutant levels. Pollution of the world's oceans, which receive massive amounts of biological and chemical wastes daily, affects not only the quality of the water, but also the ability of the ocean to sustain marine life, an important source of food.

The number of known chemicals globally exceeds 14 million, of which over 60,000 are commonly used. All of these ultimately end up in the environment. They are the result of the huge proliferation of industry, technology, and automobiles in the twentieth century. The full health effects of many of these chemicals are still unclear and difficult to establish. Twenty-three per cent (23%) of all deaths globally are attributable to the environment (approximately 12.6 million deaths annually) (WHO Environmental Impacts on Health). As the twenty-first century progresses, the number of pollutants will continue to increase. Exposure to ambient $PM_{2.5}$ (particles with diameters of less than 2.5 microns) was ranked fifth of all global risk factors in 2014. It is estimated that it caused 4.2 deaths and 103 million DALYs globally (4.2%) in 2015, with attributable deaths increasing from 3.5 million in 1990 to 4.2 million in 2015 (Cohen et al. 2017).

Problems of the environment occur at the personal level (at home and the workplace), the community level (e.g. air and water pollution), and globally (e.g. global warming, hazardous and radioactive waste). Although these problems may be viewed separately, they are in fact all global issues affecting both local and remote populations. Thus, slash-and-burn agricultural practices in Indonesia result in periodic and sometimes severe air pollution problems in neighbouring Singapore and Malaysia. Industrial pollutants released in the industrial states of northeastern United States cause acid rain, which adversely affects crops and people in the midwestern United States and southern Canada. Pollution of rivers upstream can adversely affect communities and countries downstream, as happened in 2005 when nitrobenzene was released into the Songhua River in Heilongjiang, China, contaminating drinking water downriver in both China and Siberia, Russia.

Air pollution

The rapid increase in automobiles and industry has caused widespread air pollution in most urban areas of the world, with the worst occurring in the developing countries that have rapidly industrialized at the expense of their environment (Table. 1.1.3). The main pollutants are ozone, particulate matter, carbon monoxide, nitrogen oxides, sulphur dioxide, and lead. Now,

in the early part of the twenty-first century, many of these countries are realizing the need to protect the environment. Unfortunately, reversal of decades of pollution is far more difficult and costly than prevention.

Table 1.1.3 Global estimates of comparative occupational accidents and work-related illnesses by world region and sector categories

Region	Fatality rates of each economic sector					
	Agriculture		Industry		Service	
	2010	2014	2010	2014	2010	2014
HIGH	7.8	no change	3.8	no change	1.5	no change
AFRO	18.9	no change	21.1	no change	17.7	no change
AMRO	9.3	8.7↓	9.5	11.2↑	6.0	5.7↓
EMRO	13.0	no change	14.9	no change	12.3	no change
EURO	15.7	17.0↑	10.3	13.4↑	5.5	3.5↓
SEARO	24.0	27.5↑	9.7	9.9↑	5.1	4.4↓
WPRO	24.0	27.5↑	9.7	9.9↑	5.1	4.4↓

• High-income countries (HIGH) • Low- and middle-income countries of the African Region (AFRO) • Low- and middle-income countries of the Americas (AMRO) • Low- and middle-income countries of the Eastern Mediterranean Region (EMRO) • Low- and middle-income countries of the European Region (EURO) • Low- and middle-income countries of the Southeast Asia Region (SEARO) • Low- and middle-income countries of the Western Pacific Region (WPRO).

Agriculture: farming, fishing, and forestry.

Industry: mining, manufacturing, energy production, and construction services.

From
http://www.icohweb.org/site/images/news/pdf/Report%20Global%20Estimates%20of%20Occupational%20Accidents%20and%20Work-related%20Illnesses%202017%20rev1.pdf

The harmful effects of air pollution extend beyond the environment. Many members of society, including asthmatics and persons with chronic respiratory disease, are vulnerable to even relatively low levels of pollutants. Levels of pollutants observed in many developing countries, especially in China and India, are considerably higher than in developed countries. A

study carried out in China in 2015 (Levin 2015) estimated that 1.6 million deaths per year in China are attributable to air pollution, and cancer rates have increased dramatically (as much as fourfold in industrial provinces), with an estimated 4 million cases in 2015 that resulted in nearly 3 million deaths (Carney 2016). However, the true cost of uncontrolled industrialization and pollution in these countries is not fully known.

Indoor pollution is a particular problem in developing countries, where cooking is typically fuelled by coal or charcoal fires in poorly ventilated houses. However, it is also a problem in wealthier countries in which harmful chemicals are used for cleaning and household construction.

Water pollution

The main types of water pollutants are mercury, nitrates, phosphorous, faecal coliform, and bacterial pollution. Those who live in developed countries take the provision of safe drinking water for granted, but 40% of the world's population does not have access to clean drinking water, a basic necessity of life. As the world population expands, production of waste increases, and the problem of protecting water supplies also rises. Approximately 60% of the world does not have adequate facilities for waste disposal. Even in leading cities in developed countries, pollution of the water supply can occur. The increased rate of upper respiratory infections and gastrointestinal disorders among surfers and others using the ocean for recreational purposes has been well documented. Beaches in most urban areas are frequently closed when the sewage disposal systems that drain into the vicinity become overwhelmed. Acid rain from industrialization has caused acidification of lakes, making them inhospitable for fish and other marine life, thus compromising the food supply. Recently there has been discussion about whether the benefits of omega-3 fatty acids found in fish outweigh the risk of mercury poisoning among those who eat large quantities of fish. Ensuring a safe adequate water supply for people in both developed and developing countries must remain a public health priority.

Other pollutants

The most common types of soil pollutants are heavy metals such as cadmium, chromium, copper, zinc, mercury, pesticides or herbicides, organic chemicals, oils and tars, explosive or toxic gases, combustible or radioactive

materials, biologically active compounds, and asbestos; these pollutants can enter the soil through poor agricultural practices, industrial runoff, mining, landfill leakage, littering, or the improper or illegal dumping of household or industrial waste materials. As the population of the world rapidly increases and technology produces new substances and processes, not only the amounts of pollutants, but also increased varieties of pollutants. As new substances are developed, it would be ideal if their use was not permitted until plans and provisions are developed and implemented for their safe disposal. However, this is often not the case in practice.

Biodegradable pollutants have a limited lifespan in the environment, but there is increasing production of non-biodegradable substances such as plastics, which are now ubiquitous, and hazardous materials such as radioactive wastes that persist for generations. The problem of discarding these materials safely has become a major public health issue. In some cases, developed countries are paying developing countries to accept their hazardous waste products. This strategy does not solve the problem, but shifts it to those countries that have fewer resources with which to deal with the challenge, and which are increasingly reluctant to accept wealthy countries' garbage.

In the past decade, nanoparticles (1 nanometre is 1 billionth of a metre) have been increasingly used in the production of foods, drugs, cosmetics, and other products used by humans. Particles of this small size become reactive in the body and, according to reports (Li et al. 2009), can cause serious damage to lung, liver, and brain cells. Given this problem, it is important that the use of nanoparticles in products intended for human use be regulated.

Climate change

One of the most serious, long-term challenges of the twenty-first century is global warming due to the release of carbon dioxide and other 'greenhouse gases'. Climate changes affect the social and environmental determinants of health, including safe homes, clean air, safe drinking water, sufficient food, and safe homes. Climate change may affect health through multiple ways, such as increased frequency and intensity of heat waves, reduction in cold-related deaths, increased natural disasters such as floods and droughts, changes in the distribution of vector-borne diseases, and effects on the risk of disasters and malnutrition. There is growing evidence that the

consequential climate change will be associated with increased public health risks such as heat-related disorders, infectious diseases, respiratory and allergic disorders, malnutrition, mental health problems, and violence (Patz et al. 2005).

A major review by WHO to assess the public health impact of anthropogenic climate change to date, estimated that from the mid-1970s, climate change might already have contributed to 150,000 deaths and about 5 million DALYs per year. These are largely related to increased incidence of conditions such as malnutrition, diarrhoea, and malaria (World Health Organization 2014c). Mitigating the health impact of global warming and climate change at the regional level is a major public health issue for this century, which unfortunately has been politicized.

Rescuing the environment

To prevent further degradation of the environment and to tackle the threat of global climate change, strong political will is required of the countries of the world. For example, while the United States is one of the world's major producers of carbon dioxide and other greenhouse gases, in 2017, President Trump refused to endorse the Paris Accord on Climate Change, an international treaty aimed at stabilizing global greenhouse gases at levels that would avoid dangerous climate change.

From a public health perspective, it is unrealistic to expect that the risks from environmental pollution and hazardous waste can be reduced to zero. Instead, the concept of 'acceptable risk' will continue to be a part of the process. Determining the level of acceptable risk will probably be arrived at through an interplay of the scientific data and evidence, with policy and political considerations. Public health professionals and researchers must play a strong role in these determinations.

Occupational health

> Occupational diseases are different from other diseases, not biologically, but socially.
>
> Henry Sigerist, 1958 (Sigerist 1958–1961)

The International Labour Organization's Health and Safety Programme estimated that each year globally there are over 2 million work-related

deaths, over 300 million suffering from work-related diseases, and approximately 3.7 million fatal and non-fatal work-related injuries. In some developing countries, child labour is still the norm. Twelve million serious injuries occurred among young workers in 2011. This affects more people than those who have myocardial infarcts (heart attacks), strokes, or newly diagnosed malignancies annually. A significant proportion of these deaths and injuries are preventable by improving safety in the workplace. However, safeguarding the health of the worker often receives less priority than the need to produce goods cheaply, especially in developing countries, despite the increase in workplace fatalities in low- and middle-income countries in four of the six WHO regions (Hamalainen et al. 2017) (Table 1.1.3).

The nature of the workplace is constantly changing, with increasing proportions of workers being involved in service industries rather than in manufacturing. Over the last decade particularly, production of goods has shifted rapidly to developing countries, where labour costs are lower. The manufacturing industries in these countries are often subject to fewer and less comprehensive safety regulations, and in some cases, are associated with low salaries and few or low healthcare benefit provisions for workers, especially for migrant workers from rural areas. Increasingly, women are entering the workforce and must manage both work and family. Larger numbers of workers are being employed on an informal part-time basis. While this reduces labour costs for industry, these informal part-time workers do not usually receive work-related benefits such as retirement and health insurance. They now represent 50% or more of the workforce globally. This segment of the workforce is particularly vulnerable to injury and limited access to healthcare.

As noted earlier, the population is ageing, and the proportion of the population that is economically active is diminishing. In response to this change, the age of eligibility for social security benefits in the United States is increasing, and mandatory retirement is being phased out. It is now projected that the proportion of workers over the age of 60 will increase to 20% in Japan and 10% in the United States by 2030 (Population Projections 2000). The United Nations report on World Population Ageing 2015 noted that among individuals aged 65 years or over, 30% of men and 15% of women were active in the labour force (United Nations 2015b). The needs of older workers are different and greater than those of younger workers. Thus, the changing nature of the workforce will require corresponding changes in

work safety regulations and health benefits to ensure a healthy, productive workforce.

Provision of and access to healthcare

Access to preventive and curative care is a requirement for health in every society, whether rich or poor. Access to affordable healthcare has long been a problem for the poor and for rural residents, especially in developing countries. Even in the United States, full access to healthcare remains a problem. Health insurance is prohibitively expensive and beyond the reach of many in the middle classes, unless it is subsidized by employers or government. Increasingly, employers are attempting to free themselves from the cost of health insurance for their employees through a variety of strategies. The elderly also have problems with healthcare; because healthcare utilization and costs tend to increase with age, insurance companies have been less willing to cover the elderly, and many governments, even in developed countries, do not provide adequate support for the elderly. The Affordable Care Act enacted under President Obama extended coverage significantly, but President Trump and the Republican Congress have undermined it, causing 13 million Americans to lose their coverage. In developing countries, the rural poor are particularly at risk. Few health professionals are willing to work in rural areas, and the cost of providing care in less populated areas, even in wealthier countries, is greater than in urban areas. Innovative strategies are needed to promote universal coverage and ensure that the rural poor and elderly have access to reasonable healthcare.

Bioterrorism and war

The history of use of biological weapons in war extends back for hundreds of years. In the Middle Ages, corpses of plague victims were catapulted into castles under siege. In more modern times, anthrax was used to contaminate the US postal system, resulting in several deaths. The WHO and public health agencies of individual countries have developed plans to quickly diagnose and control bioterrorist incidents. However, these threats to the

health of the public will continue until the underlying causes of terrorism and bioterrorism are addressed.

Few actions can have the magnitude of negative impact on the health of the public that war has. Men, women, and children are killed, children are forced to serve as soldiers, homes are destroyed, major segments of the population become displaced refugees, women are raped, and the social and/or economic fabric of the countries involved is destroyed. Recovery usually takes years to decades. The outside world, particularly those countries adjacent to warring nations, must cope with the huge influx of displaced persons, and action needs to be implemented to help those still in the country suffering from the impact of the war. The consequences of war are so severe and wide-ranging that it is imperative that better ways be reached to resolve international conflicts. The resources ploughed into armed conflict could be better deployed on humanitarian and public health support.

Ethics in public health

Although ethical issues are implicit in the delivery of public health, it was only after the Second World War and the recognition that ‘scientific experiments’ in Nazi Germany violated human rights that an emphasis was placed on recognizing the ethics of public health actions, particularly research. The Declaration of Helsinki (World Medical Association 2002), the Belmont Report (US National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research 1979), and the Council for International Organizations of Medical Science (2002) have promulgated ethical guidelines for research and the establishment of institutional review boards worldwide to ensure that medical and/or public health research is conducted ethically and does not violate human rights. However, there are inherent ethical conflicts in many public health actions. For example, some consider that the human rights of ‘typhoid Mary’, a typhoid carrier who insisted on working as a cook in the early twentieth century, were violated when she was incarcerated to prevent her from continuing to prepare food that initiated typhoid epidemics. Protecting the human rights of a man to refuse testing for HIV may result in him unknowingly infecting his wife, yet-to-be-born children, and other sexual contacts. By protecting his human rights, the human rights of his wife or partner and future family will be violated, underscoring the complexity of human rights issues.

Implementing public health programmes and research often results in ethical conflicts and the need to balance the good of society against potential harm to the individual. It is imperative to inform society, particularly those who will be involved in the public health programmes or research, about the nature of the ethical conflicts inherent in action. For example, a trial evaluating the effectiveness of prophylactic treatment to prevent HIV infection in sex workers in Cambodia was stopped by the prime minister, who felt that the prevention trial exploited Cambodian sex workers. A more intense effort on the part of the researchers to inform the public and politicians about the nature of the study and the potential benefit to sex workers, not only in Cambodia, but globally, might have averted this unfortunate outcome.

Public health interventions

One important task of public health professionals is to raise the level of anxiety of the public about public health problems to the level at which they will be willing to take an appropriate action. In this, an appropriate balance must be struck, since raising the level of anxiety too little will result in inadequate or no action, while raising it too high may promote a fatalistic attitude and, as in the case of the HIV/AIDS epidemic, may promote stigmatization and isolation of affected individuals, seriously complicating the task of intervention. The difficulty for the public health professional is creating the level of anxiety that results in the required action while minimizing unintended consequences.

Public health interventions can be divided into four categories: social/biological/environmental, behavioural, political, and structural. The public health professional must use strategies in all four categories to achieve the maximum health of the public.

Social, biological, and/or environmental interventions

The strategy that has had the greatest impact on improving the health of the public has been an improved standard of living, including provision of clean water and safe disposal of wastes. Unfortunately, these interventions have not reached much of the world where crowding, unsafe and insufficient

water, accumulation of wastes, inadequate housing, and lack of economic development persist.

The most cost-effective biological intervention strategy is immunization, in part because it requires minimal behavioural change and usually involves only a single series of actions. The WHO has taken the lead in promoting vaccine coverage worldwide through its Expanded Programme on Immunization. The appropriate use of vaccines has reduced the majority of childhood infections in developed countries, and significantly reduced them in most developing countries. Smallpox, a major infectious disease problem until the latter half of the twentieth century, has now been eliminated. The WHO has now targeted polio, but more challenges, such as hepatitis, tuberculosis, and measles, remain. However, it is important to realize that development and production of a vaccine is only the first step. An effective vaccine against smallpox was available for over 150 years before smallpox was eliminated. The key was the strategy of vaccine coverage, ‘search and contain’, together with adequate political will and funding that permitted global elimination of that disease. Thus, the strategy for utilizing the vaccine is perhaps equally important as the efficacy of the vaccine itself. Public education and engagement must continue to be a major priority. Recently, attacks from the media, particularly social media, have caused acceptance of vaccines to decline, reducing the level of ‘herd immunity’ required to prevent epidemics (Karlamangla 2018).

Another biological strategy is to eliminate the vectors of disease, the major approach currently in use for the control of dengue, arboviral diseases, and many of the parasitic diseases. However, overzealous use of pesticides can also create problems. For example, dichloro-diphenyl-trichloroethane (DDT), used widely in the twentieth century as an insecticide, still contaminates the food supply, creating other health problems, including the risk of malignancy.

Treatment can also be considered a biological intervention strategy. To confront tuberculosis, one of the major infectious diseases of the twenty-first century, directly observed treatment short course (DOTS) has been successfully implemented in countries where the disease persists, and reduces transmission and development of multidrug-resistant tuberculosis (TB). Treatment of sexually transmitted infections and contacts to reduce the reservoir of infectious people is a major strategy for control of transmission,

but has yet to prove effective in stopping the current epidemic. ‘Treatment as prevention’ has become a mantra of HIV/AIDS control.

Behavioural interventions

Most public health interventions depend ultimately on behaviour change, whether it is personal or community behaviour. At the personal or individual level, promotion of good health habits, including avoidance of smoking and excessive alcohol use and other dependency disorders are important interventions that have a major impact on health. At the community level, confronting attitudes towards acceptable sexual behaviour and persons with dependency disorders and stigmatizing diseases are key to establishing community ‘norms’ that promote a healthy lifestyle and include all segments of society. However, modifying individual behaviour and community norms is difficult, and it is even more difficult to ensure persistence of the modified behaviour. However, the majority of the public health interventions will not be successful unless they are embraced and sustained by the community at the individual, local, national, and international levels, and accompanied by appropriate regulations and policies. The success of the antismoking campaigns in the United States and Britain and population control in China (the one-child policy) affirm that it is possible to change community norms.

Many theories identifying strategies to modify behaviour have been proposed. One of the most interesting is the Popular Opinion Leader model (Kelly 2004), which utilizes the natural leaders found in any social group as agents of change. In the United States, this strategy has been demonstrated to change behaviour in groups of men who have sex with men, and has been evaluated in other populations worldwide (Rotheram-Borus et al. 2011). Ultimately, if public health interventions are to be effective, they must be adopted by the community. The media can play an important role in facilitating these interventions and assuring that they become community norms.

Political interventions

Public health *is* politics. Any process that involves obtaining the support of the public will involve politics and differing points of view. For example, the campaign to stop smoking was strongly opposed by the tobacco industry,

which spent millions of dollars trying to counter the many reports on the adverse health effects of smoking. Countering the efforts of the tobacco industry required obtaining the political support of the public in order to pass laws and regulations limiting smoking, placing health warnings on cigarette packages, and raising taxes on cigarettes. Many needed regulations transcend boundaries and require international cooperation (e.g. regulation of greenhouse gases).

If we are to succeed in safeguarding the oceans, inland waters, and the air we breathe, it will be through political process requiring global coordination and joint action. This process has already begun in many developed countries, which have passed strong laws regulating emissions from automobiles and factories, although they are constantly under attack. Now this process must be expanded to the developing countries, where the worst pollution is currently occurring.

One of the most urgent issues before the public today is the battle over emission of 'greenhouse gases', which are causing a rise in temperatures globally. This temperature rise will adversely affect the quality of life of our children, grandchildren, and their grandchildren. Unfortunately there is still political action in the US to not take the necessary steps to arrest or reverse this detrimental warming trend.

It is important that the political process puts in place measures to improve the health of the public, based on sound scientific evidence. Pushing agendas not based on sound scientific evidence will undermine the credibility of public health professionals and our ability to accomplish our legitimate goals. Obtaining this evidence is not always easy. For example, accumulating evidence on the long term (induction period of years to decades) impact of adverse exposures is not easily established, and often requires extrapolation from data from the impact of acute high-dose exposures to lower long-term doses. This often requires relying on models that are difficult for the public to understand, and are often subject to debate, even within the scientific community.

Structural interventions

An important end-point of the political process is the passage of laws and regulations. This action, if implemented, can have a very significant impact on the improvement of the health of the public. The passage and

enforcement of helmet laws for motorcycles in Indonesia, the US, and Vietnam reduced the incidence of associated brain injuries and deaths. The incidence of lung cancer and heart disease among men has been significantly reduced, partly due to the laws regulating smoking in public spaces and the high taxes imposed on cigarettes. In 2013, China began very strong antipollution measures, particularly to stop use of coal, and the air quality in China has now greatly improved (Economist 2018). Many of the current public health problems of the world, particularly those involving protection of the environment, can best be addressed through structural changes requiring passage and implementation of laws and regulations. To accomplish this will require changing the attitudes and behaviour of the public and ensuring that public health regulations and laws are enforced.

Private support of public health

Private support has played an important role in the development of public health, especially in the twentieth century. The Rockefeller Foundation supported the first school of public health in the United States at Johns Hopkins University; set up the International Health Commission in 1913; established the China Medical Board in 1914, which established the first public health university in China, the Peking Union Medical College, in 1921; and has continued to contribute to global health since its founding in 1913 (Brown 1979; Berman 1983). Other foundations, including the Gates, Ford, Carnegie, and Robert Wood Johnson Foundations, have made similar significant contributions to public health.

Private support of public health has been implemented through three strategies: establishment of charitable foundations by industry; development of international, national, and local non-governmental organizations (NGOs); and direct contributions by industry. Each makes and can continue to make a significant contribution to the health of the public.

Foundations have contributed enormously to the advancement of public health, but most identify their own priorities for funding. Usually they provide support for important public health needs, but foundations and public health leaders do not always agree on what the most important priorities are. Massive infusions of money into public health by organizations such as the Gates Foundation, which makes contributions to fight HIV, malaria, and tuberculosis, can have a significant positive impact,

but they also tend to influence public health priorities. Some argue that developing strong public health infrastructures in developing countries will have a much greater impact on improving health than focusing funds on specific health issues (Garrett 2007).

NGOs tend to focus on specific health problems (e.g. American Cancer Society), specific health issues such as refugee health or medical care for the underserved (e.g. Doctors without Borders), and specific populations (e.g. drug users and sex workers). Often they can be more effective in reaching vulnerable populations and addressing specific health problems and issues because they are closer to the problem than health professionals who must handle a broad range of concerns. Public health programmes can increase their cost-effectiveness by cooperating with NGOs in addressing specific issues, health problems, and populations.

Industry is often viewed as part of the problem. Public health needs to convince industry to temper its profit motive to incorporate responsible citizenship at all levels. Industry is frequently a significant contributor to public health problems (e.g. air and water pollution). On the other hand, economic development can lead to an improved economic situation that reduces poverty and benefits all of society. Industry, particularly advertising, has clearly demonstrated that they are better at creating demand and influencing lifestyles than public health professionals. Thus, it behooves public health organizations to learn from industry and to work with industry to develop and implement healthy economic growth, while safeguarding the environment and benefiting the public.

Private support greatly benefited public health in the twentieth century. The challenge for the twenty-first century is for public health and private support to agree on the most effective use of private funds for achieving the greatest public health advances.

Social activism

Action by groups within society has led to improvements in the health of the public. Social activism by workers and others in the nineteenth and early twentieth century in England and the United States resulted in improved working conditions for workers in a variety of industries and to development of the field of occupational health. More recently, social activism by social groups adversely affected by HIV/AIDS led to the policy to provide access

to treatment for all with HIV infection and the need to recognize the human rights of marginalized groups in both developed and developing countries. Demands for better health conditions led by society itself are most likely to result in positive changes.

The future of public health

Public health does not lack challenges requiring solutions. Poverty is the major cause of poor health globally, yet income disparities in most countries of the world continue to grow. Developing countries must continue to cope with infectious diseases while confronting the epidemic of NCDs, further compounded by the threat of emerging diseases such as new variants of influenza. Rapid communications and transportation greatly increase the likelihood that local problems will quickly become global problems. This underscores the urgent need for much better international cooperation. An increasing proportion of the world's population will live longer. We have been successful at adding 'years to life', but chronic diseases such as Alzheimer's have reduced the quality of some of the years of life added. We must now concentrate on adding 'life to years', helping older people to continue to be healthy and productive.

We cannot afford to continue to ignore the quality of the environment. Continuing contamination of the air and water will not only cause and/or exacerbate chronic and infectious diseases, but will also compromise global food production. The world's population is still growing and, together with increasing urbanization, will further exacerbate the problem of environmental degradation.

Despite the economic and health advances of the past century, disparities between the rich and the poor in many countries are widening. This gap needs to be narrowed, not at the expense of those who are better-off, but by improving the economic situation and health of the poor and disadvantaged. The rising cost of healthcare will make closing the gap in access to healthcare even more challenging.

Injuries and violence are robbing an increasing number of people of their ability to function and to enjoy a reasonable quality of life. Injuries can be easily prevented through a variety of preventive strategies, including better design of the workplace and tools, as well as behavioural and structural approaches.

Violence and war present a particularly great challenge, and will require new strategies not hitherto widely used in public health. Public health must contribute to strategies to resolve differences between countries by promoting cross-national and international cooperation in confronting global health problems, and contribute to strategies to implement successful conflict resolution.

It is clear that the pandemic of SARS-CoV2, which has caused many deaths globally, has severely impacted the global economy and the ways in which society functions. As of this writing it is difficult to predict what those changes will be, but we must be prepared to deal with them.

Public health must convince people and provide the environment that allows them to adopt healthy lifestyles. The major strategies to combat the current epidemic of NCDs are regular exercise, a healthy diet, and development of good health habits.

Much is known about what needs to be done to significantly reduce the incidence of NCDs such as cardiovascular diseases, stroke, and cancer, but much more effective ways are needed to effect the necessary changes in personal and community behaviour, and to promote healthy lifestyles. The growing prominence of data analytics, artificial intelligence and technology raises the promise of more effective identification of at-risk groups in the general and patient populations, and more targeted and effective interventions. However, these need to be well integrated into broader public health goals and strategies.

Tremendous strides have been made to improve the health of the public, but the challenge to do better remains. In subsequent chapters, public health experts discuss the challenges and potential solutions in greater detail.

REFERENCES

Acheson, E.D. (1988). On the state of the public health. [The fourth Duncan lecture]. *Public Health*, 102(5), 431–37.

Anjana, R.M., et al. (2017). Prevalence of diabetes and prediabetes in 15 states of India: results from the ICMR-INDIAB population-based cross-sectional study. *The Lancet Diabetes Endocrinology*, 5(8), 585–96.

Barreto, P.D.S. (2013). Why are we failing to promote physical activity globally? *Bulletin of WHO*, **91**, 390–390A.

Berman, E.H. (1983). *The Ideology of Philanthropy: The Influence of the Carnegie, Ford, and Rockefeller Foundations on American Foreign Policy*. New York: University of New York Press.

Breslow, L. and Breslow, N. (1993). Health practices and disability: some evidence from Alameda County. *Preventive Medicine*, **22**, 86–95.
Brown, E.R. (1979). *Rockefeller Medicine Men: Medicine and Capitalism in America*. Berkeley, CA: University of California Press.
Carney M (2016). China's cancer rates exploding, more than 4 million people diagnosed, study says. [Online] Available at: http://www.abc.net.au/news/2016-03-24/chinas-cancer-rates-exploding-study-says/7272266
Cohen, A.J., et al. (2017). Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the Global Burden of Diseases Study 2015. *The Lancet*, 389(10082), 1907–18.
Collins, P.Y., et al. (2011). Grand challenges in global mental health. *Nature*, **475**, 27–30.
Detels, R. (2009). The scope and concerns of public health. In: Detels, R., Beaglehole, R., Lansang, M.A., and Gulliford, M. (eds.) *Oxford Textbook of Public Health* (Vol. 1, 5th ed.), pp. 3–19. Oxford: Oxford University Press.
Detels, R. and Breslow, L. (2002). Current scope and concerns in public health. In: Detels, R., McEwen, J., Beaglehole, R., and Tanaka, H. (eds.) *Oxford Textbook of Public Health* (Vol. 1, 4th ed.), pp. 3–20. Oxford: Oxford University Press.
The Economist (2018). Editorial. How China cut its air pollution. *The Economist* [Online]. Available at: https://www.economist.com/the-economist-explains/2018/01/25/how-china-cut-its-air-pollution
FDI World Dental Federation (2015). *Oral Health Worldwide*. [Online] Available at: https://www.fdiworlddental.org/sites/default/files/media/documents/2015_wohd-whitepaper-oral_health_worldwide.pdf
Feigin, V.L., et al. (2009). Worldwide stroke incidence and early case fatality reported in 56 population-based studies: a systematic review. *The Lancet Neurology*, 8(4), 355–69.
Garrett, L. (2007). The challenge of public health. *Foreign Affairs*, **86**, 14–38.
Gupta, N., et al. (2016). Changing trends in oral cancer—a global scenario. *Nepalese Journal of Epidemiology*, **31**, 6(4):613–19.
Hamalainen, P., Takala, J., and Kiat, T.B. (2017). Global estimates of occupational accidents and work-related illnesses 2017. *Ministry of Social Affairs and Health*, Finland.
International Labour Organization. International. Labour Standards on Occupational Safety and Health. [Online] Available at: https://www.ilo.org/global/standards/subjects-covered-by-international-labour-standards/occupational-safety-and-health/lang--en/index.htm
Kelly, J.A. (2004). Popular opinion leaders and HIV prevention peer education: resolving discrepant findings, and implications for the development of effective community programmes. *AIDS Care*, 16(2), 139–50.
Karlamangla S. Pushback against immunization laws leaves some California schools vulnerable to outbreaks. *Los Angeles Times*, July 2018. [Online] Available at: http://www.latimes.com/local/lanow/la-me-ln-sears-vaccines-fight-20180713-story.html
Kuo, C.C. and Campbell, L.A. (2000). Detection of *Chlamydia pneumoniae* in arterial tissues. *Journal of Infectious Diseases*, **181**, S432–36.
Lee, S.H. (2003). The SARS epidemic in Hong Kong. *Journal of Epidemiology and Community Health*, 57(9), 652–54.
Levin D. (2015). Study links polluted air in China to 1.6 million deaths a year. *New York Times*. [Online] Available at: https://www.nytimes.com/2015/08/14/world/asia/study-links-polluted-air-in-china-to-1-6-million-deaths-a-year.html
Li, C., et al. (2009). PAMAM nonparticles promote acute lung injury by inducing autophagic cell death through the Akt-TSC2-m TOR signaling pathway. *Journal of Molecular and Cell Biology*, 1(1), 37–45.
Lopez, A., Mathers, C., Ezzati, M., Jamison, D., and Murray, C. (2007). Global and regional burden of disease and risk factors, 2001: a systematic analysis of population health data. *The*

Lancet, **367**, 1747–57.
Mathers C.D. and Loncar D. (2006). Projections of global mortality and burden of disease from 2002 to 2030. *PLoS Medicine*, 3(11), e442.
Mendenhall, E. (2017). Syndemics: a new path for global health research. *The Lancet*, 389(10072), 889–91.
Moellering, R.C. Jr (2010). NDM-1—a cause for worldwide concern. *The New England Journal of Medicine*, **363**, 2377–9.
Murray, C.J.L. and Lopez A.D. (1997). Alternative projections of mortality and disability by cause 1990–2020: Global Burden of Disease Study. *The Lancet*, 349(9064), 1498–504.
Oral Cancer Foundation. [Online] Available at: https://oralcancerfoundation.org/research/
Patz, J.A., Campbell-Lendrum, D., Holloway, T., and Foley, J.A. (2005). Impact of regional climate change on human health. *Nature*, **438**, 310–17.
Pine, C. and Harris, R. (2007). *Community Oral Health*. London: Quintessence Publishing.
Population Projections (2000). Population projections. *Health Affairs*, **3**, 191–203.
Qadri, F., et al. (2016). Efficacy of a single-dose, inactivated oral cholera vaccine in Bangladesh. *The New England Journal of Medicine*, 374(18), 1723–32.
Rotheram-Borus, M.J., et al.; the NIMH Collaborative HIV/STD Prevention Trial Group. (2011). Reductions in sexually transmitted infections associated with popular opinion leaders in China in a randomised controlled trial. *Sexually Transmitted Infections*, **87**, 337–43.
Schroeder, S. A. (2007). Shattuck lecture: we can do better—improving the health of the American people. *The New England Journal of Medicine*, **357**, 1221–8.
Sigerist, H.E. (1958–1961). *A History of Medicine*. Oxford: Oxford University Press.
Singhal, T. (2020). A review of coronavirus disease 2019 (COVID-19). *Indian Journal of Pediatrics, 87*(4), 281–6.
United Nations (2015a). *World Population Prospects*. [Online] Available at: https://esa.un.org/unpd/wpp/publications/files/key_findings_wpp_2015.pdf
United Nations (2015b). *World Population Ageing 2015*. [Online] Available at: https://www.un.org/en/development/desa/population/publications/pdf/ageing/WPA2015_Report.pdf
United Nations Development Programme (2014). *Human Development Report 2014*. [Online] Available at: http://www.undp.org/content/undp/en/home/librarypage/hdr/2014-human-development-report.html
US Census Bureau (2016). *International Database*. [Online] Available at: https://www.census.gov/programs-surveys/international-programs/about/idb.html
US National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research (1979). *The Belmont Report: Ethical Principles and Guidelines for Protection of Human Subjects of Biomedical and Behavioral Research*. Washington, DC: US Department of Health & Human Services.
World Bank (2012). *World Bank Sees Progress Against Extreme Poverty, But Flags Vulnerabilities*. [Online] Available at: http://www.worldbank.org/en/news/press-release/2012/02/29/world-bank-sees-progress-against-extreme-poverty-but-flags-vulnerabilities
World Health Organization (WHO) (2015). *Control of Neglected Tropical Diseases. Investing to Overcome the Global Impact of Neglected Tropical Diseases. Third WHO Report on Neglected Tropical Diseases*. [Online] Available at: https://www.who.int/neglected_diseases/9789241564861/en/
World Health Organization (WHO). *Drinking Water*. [Online] Available at http://www.who.int/news-room/fact-sheets/detail/drinking-water
World Health Organization (WHO) (n.d.). *Environmental Impacts on Health*. [Online] Available at: http://www.who.int/phe/PHE-prevention-diseases-infographic-EN-1-1200px.gif?ua=1
World Health Organization (WHO) (n.d.). *Global Health Estimates*. [Online] Available at: http://www.who.int/healthinfo/global_burden_disease/en/index.html

World Health Organization (WHO) (2014a). *Injuries and Violence: The Facts*. [Online] Available at http://www.who.int/violence_injury_prevention/key_facts/VIP_key_facts.pdf?ua=1
World Health Organization (WHO) (2014b). *Road Traffic Injuries.* [Online] Available at: http://www.who.int/violence_injury_prevention/road_traffic/en/
World Health Organization (WHO) (2014c). Fifth Assessment Report Climate Change 2014: Impacts, Adaptation and Vulnerability. Geneva: WHO.
World Health Organization (WHO) (2014d). *Violence Against Women*. [Online] Available at: http://www.who.int/en/news-room/fact-sheets/detail/violence-against-women
World Health Organization (WHO) (2017a). *Mental Health Atlas*. Geneva: WHO.
World Health Organization (2017b). *Global Health Observatory Data*, 2017. [Online] Available at: http://www.who.int/features/factfiles/obesity/en/
World Medical Association (2002). *Declaration of Helsinki* (5th rev.). Washington, DC: World Medical Association.
World Population Prospects (2017). [Online] Available at: https://population.un.org/wpp/
Wu, Z.Y. and McGoogan, J. (2020). Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China: summary of a report of 72 314 cases from the Chinese Center for Disease Control and Prevention. *Journal of the American Medical Association*, 323(13), 1239–42.

1.2

The history and development of public health in developed countries

Simon Szreter

Introduction: prioritizing the public's health and engendering economic development

It has been conventional to locate the origins of public health in early efforts to combat epidemics and to regulate the sanitary environment accompanying urban life and to trace its history in the gradual evolution of such measures in relation to politics, administrative practices, public laws, and medical science's changing aetiology (Rosen 1993; Porter 1994). Such an historical account provides important insights and understanding but it is also delimited in one significant sense. Public health is cast in a responsive role in relation to the processes of economic development. However, the nature of the relationship between public health and global economic development can appear very differently when it is viewed over the long-term, encompassing the whole process of modern economic transformation from the singular matrix of its origins in the early modern society and economy of England, c.1600–1800.

This chapter will show that state policies to promote the social order, security, and health of the population in fact pre-date and crucially underpinned the process through which the modern world's economic transformation originated—in early modern England. Furthermore, the relationship has since remained an interdependent and dialectical one. Economic development has required healthier and more productive citizens to sustain both the supply and the demand sides of the growing economy. Furthermore, as Amartya Sen has recently had to remind the discipline of economics, the purpose of economic growth is to enhance human

capabilities and functioning, not simply to amass a global stock of capital, regardless of welfare and distributional issues (Sen 1999).

The historic origins of public health policies in ensuring food security and preventing epidemics

Public health is a term which refers to a complex body of scientific expertise, policies, laws, and behavioural norms which have evolved in relation to a changing knowledge base about the prevention of harm and the promotion of health among human communities. It is also a term which refers to the observable state of relative health in those communities. Wise statecraft in the past dictated that a secure leader or government ruling over a robust polity aimed for the populace to be at least as vigorous and economically productive as those of rival states. The most elementary public health policy pursued by leaders throughout history has therefore always been the search for food security.

Given the vagaries of climate and weather, food security has posed a fundamental problem of potential disorder for leaders of settled human groups of any significant size. Threats of dearth and famine are literally Old Testament biblical themes. The importance of the relationship between sound statecraft and ensuring the public's health through food security is encapsulated in Genesis 41 in the story of Joseph's interpretation of the Pharaoh's dream of the seven fat years to be followed by seven lean years and the institution of state granaries in Egypt to preserve the people and the ruler. Ancient Chinese doctrine of an emperor's 'Mandate of Heaven' enshrined the legitimacy of popular uprising against a ruler who could not protect his people from natural disaster, including from famines. Most ancient civilizations which were sufficiently successful to have endured across many generations and to have supported the elaborate cultural productions which we now marvel at in museums, such as that of the Egyptians or the Chinese, the Incas, or the Romans, also solved this basic public health problem of food security through state-controlled granaries.

Inadequate food is the most elementary of public health threats, but sufficient and clean water is not far behind. Most historic settlements could only prosper and expand where water supplies were adequate to their needs. However, the Romans (and quite independently in South America, e.g. the

Moche-Chimú and the Incas) famously developed impressive hydraulic technological solutions which allowed settlements and cities to grow to sizes way beyond their immediately accessible water supplies. Their irrigation channels, aqueducts, public fountains, and bathing facilities were publicly funded engineering feats supplying dense populations with secure and clean water supplies, separated from the equally necessary sanitary waste systems.

Two other crucially important measures in the modern public health armamentarium which also have their genesis in much earlier practices—from the late fourteenth century onwards in Ragusa (Dubrovnik) first and then across Italy—are, firstly, the policy of physical segregation of those believed to be carrying a life-threatening communicable disease. This practice is generally known as quarantine, signalling its historic association in Europe with efforts to ward off plague outbreaks (40 days free from disease was found to be the required period to clear infected ships). It is closely related to the later practice of isolation of the infected, which lay behind the creation of separate fever wards and hospitals. Secondly, the administrative practice of counting burials, as an early warning system to monitor signs of epidemic disease outbreaks within urban populations. Also first observed in Italian city states, the most developed of these were London's weekly Bills of Mortality, which publicized the data collated from its parishes sporadically in the sixteenth century and then continuously from the early seventeenth century onwards, when they also began to include information on attested causes of death. This provided the basis for the beginnings of epidemiology in the analysis pioneered by John Graunt in his *Natural and Political Observations* of 1662 (although information on age at death was lacking until the eighteenth century).[1]

Thus, most of the key principles and policies which inform the preventive practices of 'modern' public health—and which cause it to have an integral relationship with the capacity of populations to attain high densities of settlement and associated sophisticated levels of commercial and cultural development—can be found repeatedly in history long before the advent of the most recent era of global industrialization (on earlier globalizations in history, see Hopkins 2002).

The emergence of the world's first universalist social security system in England c.1530–1830: facilitator of labour mobility,

economic growth, and the public's health

Between approximately 1600 and 1850 English society experienced a transformative rise in the material productivity of its people and economy that was without historical precedent in its speed and global reach; and it is from the proliferation of this process that all of today's developed economies and countries have emerged. Economic historians will no doubt continue to debate till the end of time the relative importance of all the different factors, both within British society and through its relationships within an expanding European, trans-Atlantic, and also global set of trading networks, which have contributed to explaining how this extraordinary process unfolded during these two centuries.[2]

However, one crucially important part of the story of this initiating episode of modern, industrial economic development relates intimately and directly to the history of a peculiarly English variant of the long-standing preoccupation of sovereign rulers with preserving and promoting the health of their subject populations. The strategy for doing this took an important, entirely innovative turn in England during the course of the sixteenth and seventeenth centuries, with extraordinary, novel implications for the productivity of the economy.

During the fateful decade of the 1530s Henry VIII broke with the Catholic Papacy in Rome, proclaiming himself head of the Church of England. One of the prime functions of the mediaeval church all across Europe had been to dispense alms to the needy, fulfilling its Christian vocation to offer mercy and charity. As the new head of the Church in England the ultimate responsibility to provide alms to the indigent poor now rested with the Crown. The Tudor state therefore needed to find an alternative, practical means to fulfil its Christian obligations to support the poor without bankrupting the Crown's limited resources, having sold off much of the rent-generating Catholic church's estates, which previously funded this activity. A sequence of experiments during the rest of the sixteenth century by Henry and his successors finally culminated in a system for placing this practical responsibility primarily on the approximately 10,000 individual parishes of the new Church of England, using the funds generated by a progressive tax rate levied on the value of land occupied by individuals within the parish. This was the famous Elizabethan or 'Old' Poor Law, enacted by statutes in 1598 and 1601. Buttressed also by a reforming 1601 Charities Act, these

Acts established an absolute right to ‘relief’—a universal social security entitlement—for every Anglican subject of the Crown in all circumstances. Parish officials (usually unpaid) had the duty, overseen by local Justices of the Peace—answerable ultimately for the fair administration of the system to the Crown—to ensure that nobody starved due to lack of work, being orphaned, old age, or illness.[3]

The parish-administered Poor Laws became the all-purpose social security and antifamine system for England and Wales, preserving the public’s health throughout the ensuing two centuries. It proved so effective it enabled the entire population to avoid completely the fundamental insecurity of famine and starvation-related disease deaths throughout this period, achieving a unique level of protection from the perennial threat which still afflicted the rural populations of the continent. The key difference in England and Wales was that every rural parish had its Poor Law fund and officers (Solar 1995). This necessitated an associated scale of public expenditure on the rural poor which French observers in the eighteenth century found literally incredible (Smith 2011). Whereas on the continent of Europe episodes of dearth resulted in raised death rates as hungry, workless, diseased rural refugees tried to make for urban centres with their grain stores, in England and Wales cash relief to purchase food was doled out to all in need in their own parishes. Unlike in the rest of Europe, after the 1620s there is no further evidence that high food prices caused peaks in death rates across England (Galloway 1988; Healey 2014). As Paul Slack has pointed out this approach to preserving social order by stabilizing the populace in times of potential disruption was also followed through in relation to the great epidemic threat of plague, where Tudor practices of isolating infected families by shutting them up in their houses, while continuing to supply them with food, were consolidated in the 1604 Plague Act (Slack 1985).

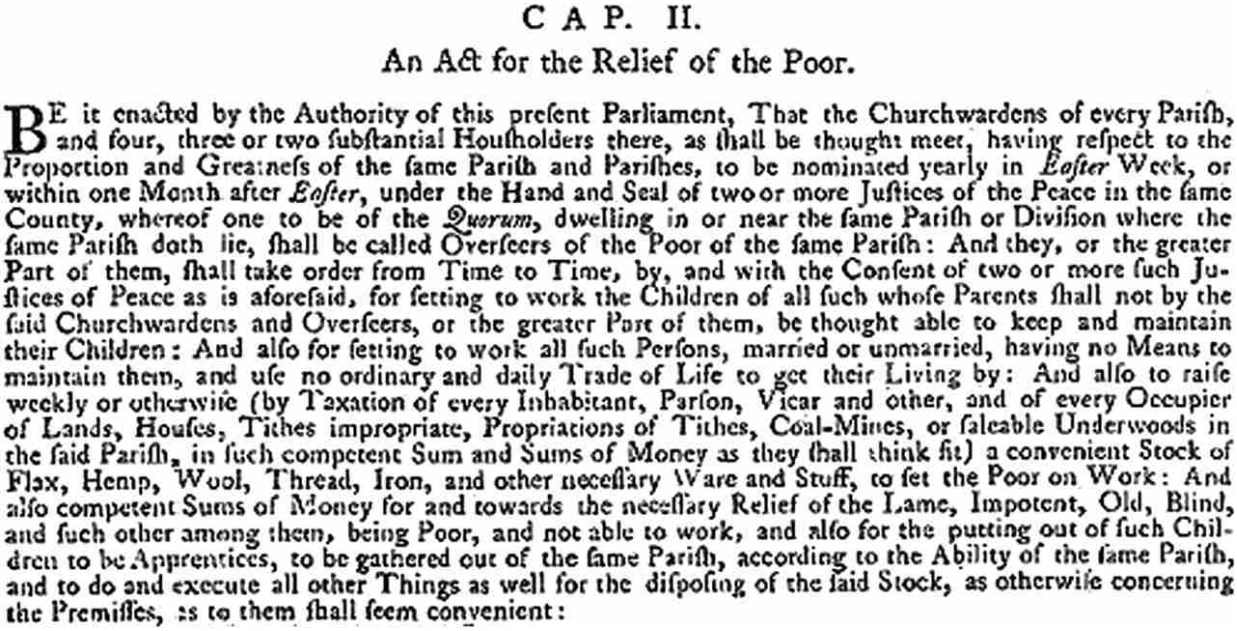
Quite unlike previous, successful antifamine strategies, as for instance that of the state granary system of the Qing dynasty in China at this very same time (Wong 2011), the English parish Poor Law did so without interrupting the normal commercial functioning of the market economy. Even when food prices were high, rather than interdicting the market price, the poor received adequate financial support to maintain them; and producers continued to make their profits.[4] The continuity of markets gave greater confidence to landlords and farmers making major investments to promote higher productivity. This raised the overall productivity of British agriculture,

enabling it successfully to feed a population of increasingly landless, urban workers which grew to proportions of the national total never previously sustained (Wrigley 1987, ch.7).

The absolute commitment, given in black-letter statute law, to maintain at parish expense all subjects of the Crown, including not only the unemployed but also the ill, orphans, and the old, was a most extraordinary mandate laid on the collective funds of each community to maintain the health of all, including a duty of care for the vulnerable—both young and old. This was an incipient free healthcare system, a comprehensive, if embryonic, public health policy. It did indeed lead to the largest proportion of the regular expenses of many parishes being consumed in supporting the incomes of the elderly in pensions, and in the building of almshouses to accommodate them and payments for doctors' fees in attending them (Williams 2011: 161–2). The most comprehensive recent study of healthcare under the Poor Law in seven provincial counties has found 15% of Poor Law funds spent on medical care throughout the period 1750–1834 (King 2018). London's largest parishes, charged under the Poor Law with caring for the vulnerable, were operating large infirmaries giving free treatment to patients throughout the eighteenth century (Siena 2004; Reinarz and Schwarz 2013). The link between illness and liability to support the poor was sufficiently obvious that leading figures across England were founding voluntary subscription hospitals at a rate of one every two years during the period 1720–75 (Slack 1990: 19, 50–1). The provincial city of Chester, for instance, was where the physician, Dr John Haygarth, trained under the great William Cullen at Edinburgh, innovated in 1783 the practice of segregated fever wards, by having a top-floor extension constructed for the Chester Infirmary; he was also an active promoter of smallpox inoculation in the region and conducted a pioneering census of the city to estimate more accurately the prevalence of different diseases. Enlightenment principles of systematic observation were leading physicians towards preventive conceptions for collective promotion of the public's health (Risse 1992).

It has been estimated that Poor Law transfer payments amounted to approximately 1% of gross national product (GNP) by 1696 and this had further risen to 2% by the 1780s (Slack 1990, Table 1). The Poor Law's primary virtue, in terms of promoting economic development, was that by underwriting security for the older generation, it released a historically novel degree of voluntary labour mobility among the younger generation. During

the initial, gradual stages of commercial expansion in the world's first industrializing economy the most important factor of production was flexible labour power, free to migrate to wherever nodes of growing labour demand were located. Between 1600 and 1800, the single statistic which consistently marked-out English society as exceptional from the rest of Europe and the world was the rapidity with which its rising generations were consistently moving, decade by decade, to urban places of manufacturing and commerce, leaving behind their ageing, landless parents secure in the knowledge that their parishes would support them (Wrigley 2004: Fig. 3.8). Parish pensions provided the elderly with an alternative security to the small subsistence patrimonies handed down from one impoverished generation to the next, as in much of the rest of Europe, including France, where the immiseration of *morcellement* is widely considered one of the underlying causes of the 1789 Revolution. Throughout history honouring and looking after elders had gone with a general need to maintain close geographic proximity; the parish Poor Law of England released the younger generation from this constraint to pursue their best options for income-earning. Furthermore, the absolute entitlement to community support muted the political consequences of the grievances of the rural poor at the productivity-raising enclosure of commons lands and the buying-up of smaller farms by larger, more commercial enterprises.



CAP. II.

An Act for the Relief of the Poor.

BE it enacted by the Authority of this preſent Parliament, That the Churchwardens of every Pariſh, and four, three or two ſubſtantial Houſholders there, as ſhall be thought meet, having reſpect to the Proportion and Greatneſs of the ſame Pariſh and Pariſhes, to be nominated yearly in *Eaſter* Week, or within one Month after *Eaſter*, under the Hand and Seal of two or more Juſtices of the Peace in the ſame County, whereof one to be of the *Quorum*, dwelling in or near the ſame Pariſh or Diviſion where the ſame Pariſh doth lie, ſhall be called Overſeers of the Poor of the ſame Pariſh: And they, or the greater Part of them, ſhall take order from Time to Time, by, and with the Conſent of two or more ſuch Juſtices of Peace as is aforeſaid, for ſetting to work the Children of all ſuch whoſe Parents ſhall not by the ſaid Churchwardens and Overſeers, or the greater Part of them, be thought able to keep and maintain their Children: And alſo for ſetting to work all ſuch Perſons, married or unmarried, having no Means to maintain them, and uſe no ordinary and daily Trade of Life to get their Living by: And alſo to raiſe weekly or otherwiſe (by Taxation of every Inhabitant, Parſon, Vicar and other, and of every Occupier of Lands, Houſes, Tithes impropriate, Propriations of Tithes, Coal-Mines, or ſaleable Underwoods in the ſaid Pariſh, in ſuch competent Sum and Sums of Money as they ſhall think fit) a convenient Stock of Flax, Hemp, Wool, Thread, Iron, and other neceſſary Ware and Stuff, to ſet the Poor on Work: And alſo competent Sums of Money for and towards the neceſſary Relief of the Lame, Impotent, Old, Blind, and ſuch other among them, being Poor, and not able to work, and alſo for the putting out of ſuch Children to be Apprentices, to be gathered out of the ſame Pariſh, according to the Ability of the ſame Pariſh, and to do and execute all other Things as well for the diſpoſing of the ſaid Stock, as otherwiſe concerning the Premiſſes, as to them ſhall ſeem convenient:

Fig. 1.2.1 The Poor Law of England.
Source: © Peter Higginbotham (workhouse.org.uk).

The birth of the modern Public Health movement: minimizing the costs of disease prevention

Although the Elizabethan Poor Law was the midwife to England's labour mobility and, thereby, its economic dynamism, during the second half of the eighteenth century its costs were rising as the industrializing economy disrupted and drew workers from ever larger proportions of its rural hinterland. The British economy's globally expansionary growing pains were also involving the nation increasingly in foreign wars during this same period and so the Poor Laws were picking up the indirect domestic social costs of migration, overseas trade, and war, rising 8-fold in cost c.1750–1815. It was the search by the governing class for containment of the rising costs of managing these disruptions, within a continuing commitment to universal social security, which was to dictate the shape of the world's first public health administration during the formative period c.1834–1914.

This directly resulted in the first ever national Public Health Act passed in 1848, the same year in which the continent of Europe was convulsed in republican rebellions violently put down by reactionary monarchical regimes. By contrast, Britain's 1848 reconfirmed a further stage in the

evolution of the exceptionally devolved governance system the Tudor Crown had first instituted to deal with the public policy issues of social order and social security. Although important scientific initiatives and policy ideas, relevant to the overall unfolding history of public health, can be identified as occurring in several other countries, too, such as in Austro-Hungary, France, Germany, and Sweden, it was primarily in Britain, the most economically dynamic, rapidly urbanizing, and wealthy state, that a public health administration—gradually and contestedly—first came into existence as a nationwide institution over the half-century following the 1848 Act.

Influential advocates for Poor Law reform included Adam Smith and Thomas Malthus, though their detailed arguments have been demonstrated by modern historians to have been unevidenced and incorrect.[5] Strongest support for reform in Parliament came from those seeing themselves as paying disproportionately for the Poor Law, who were principally the tax-rated landowners and farmers of the nation, not the urban industrialists (Mandler 1987). Eventually in 1834 a radically reformed New Poor Law was enacted. Jeremy Bentham's amanuensis, Edwin Chadwick, deployed his master's individualist and behaviourist pain-pleasure principle of 'scientific' government to redesign deterrence into the nation's social security system.

The new ideology of free markets and individual pursuit of self-interest deemed the collectively funded parish Poor Law an inefficient creator of dependency and pauperism and so insisted on cut-backs and deterrence. The state was attempting to withdraw systematically from intervention in society and economy, following the logic of political economy and rationalist utilitarianism replacing mercantilism and Christian mercy as guides to public policy. Although the state would continue to honour the right to relief of all subjects of the Crown, this would now only be available to those entering a workhouse to earn their keep by hard labour, with the sexes kept segregated to ensure the idle poor had no inducement to reproduce further at the expense of the community. Henceforth outdoor relief (cash doles handed to the unemployed in their homes) were supposed to cease, except for the old and the sick who genuinely could not work.

Globally expanding free trade, promoted by an increasingly dominant Royal Navy able to enforce British trading wishes even as far away as China, also brought with it more intensive intercontinental exchange of diseases, signified for Europe in the sequence of cholera pandemics, 1829–49, 1852–9, 1863–79, 1881–96, 1899–1923, each originating in India, the

prize colony in Britain's expanding commercial empire. The successive epidemics, reached Britain itself in 1831–2, 1847–8, 1853–4, and 1866–7, provoking fear and panic, even among the property-owning section of society, who increasingly sought refuge in the suburbs for their families, away from the filth and poverty of the urban centres where their wealth was manufactured.

The public health tool of information collection and dissemination, used for centuries to warn of such epidemics, was developed further, initially in post-Napoleonic France to promote public awareness of the scale of truly shocking quantities of non-epidemic, premature mortality afflicting the poorest districts of Paris, increasingly revealed from 1829 onwards in the pages of the new journal, *Annales d'Hygiène Publique et de Médecine Légale* (Coleman 1982; La Berge 1992). The pioneering French studies then inspired new official analyses beginning a decade later in Britain. William Farr had studied in 1829–30 in the Paris of the pioneering investigative hygienists such as Louis-René Villermé and Alexandre Parent-Duchâtelet and the great exponent of the numerical method, Pierre Louis, and this influenced his own statistical approach to medicine. On appointment to the new post of Medical Superintendent of Statistics at the General Register Office in 1839, Farr began to analyse the whole nation's newly instituted civil death registers to conduct a classic public health information campaign bombarding the British public and its newly freed-up provincial and national press with evidence of the egregious scale of preventable deaths occurring in the industrial cities. He innovated graphic survival tables which aimed to shock, in demonstrating that half of all infants failed to reach their fifth birthday in Liverpool and Manchester; and he went on to devise league tables of death rates to goad city's elected representatives to take action (Eyler 1979; Szreter 2005, ch.8).

From quite distinct premises Edwin Chadwick was himself coming to parallel conclusions about the costs of disease. He was vexed to find that his deterrent workhouses were not, after all containing Poor Law costs. His investigations in East London convinced him that this was because disease caused incapacity to work and hence dependence of the sick on the Poor Law (Hamlin 1997). As Hamlin points out, there were other medical leaders at this time, notably W.P. Alison in Edinburgh, who championed the more radical argument that poverty—inadequate pay and insecure employment–caused the disease revealed by Farr's statistics. The great German medical

scientist, Rudolf Virchow, having concluded from his pioneering studies of typhus among starving Polish peasants in Silesia in 1847–8 that it was due to the exploitative conditions imposed by the German aristocratic landowners, also backed socially radical policies to address workers' wages and living conditions in his new weekly journal, *Die Medicinische Reform*. However, the flawed Smithian wages fund theory prevailed at this time in Britain and increasingly across Europe, with its claim that any rise in one set of workers' wages could only be at the expense of others. It was endorsed also by the French hygienists, who accepted the Revolution's principles regarding the sanctity of private property (against monarchical despotism) and the necessity of a market economy, championed in France by J.-B. Say (Ramsey 1994, p. 57). The wages fund theory was not to be abandoned by the Victorian high priest of economic orthodoxy, J.S. Mill, until 1869. As Hamlin has shown, Chadwick's Public Health Act did not question economic orthodoxy and was devised as a pragmatic, sanitary engineering solution to specific problems he had diagnosed with the practical operation of his masterly New Poor Law.

In medical terms Chadwick's 'Sanitary Idea' was premised on the obsolescent miasmatist theory—that bad smells given off by decaying organic matter—caused disease. The notion that cleanliness was close to godliness was, however, a conventional homily of the period, appealing to MPs' common sense and to evangelical religionists (Smith 2007). The proposal to cleanse the city with an inflow of constant pressure, clean water, and an outflow of the used water through a self-scouring arterial sewers system (with the product sold to farmers as night-soil to help defray the cost) passed parliamentary scrutiny and became law. It would engineer the sanitized city, remove disease, and thereby minimize poverty and claims on the Poor Law, without in any way impugning the laws of the free market in labour.

Chadwick's sanitary city was still going to be extremely expensive to build. Westminster envisaged that most of the money would come from local taxation—rates levied on urban property values. Proud 'traditions' of liberty, sanctity of private property, and local self-government were all now mobilized against the will of Edwin Chadwick. As James Hanley's important recent study has shown, when it came to implementing the spirit of the 1848 Public Health Act or its watered-down successor, the 1858 Local Government Act, the legal battles that ensued between private landowners

and urban and other corporations, were many and protracted (Hanley 2016). The net effect was to delay by as much as a whole generation—until the 1870s—the implementation in many of Britain's fast-growing towns and cities the kind of ambitious public health scheme envisaged in Chadwick's blueprint. While urban water supplies improved somewhat after 1848, especially in textile towns where millowners wanted increased supply because of their industrial needs, it was often the sewering dimension which was neglected. Partial sewering in the better-off districts was initially left to discharge into inland rivers until this polluting practice was legally addressed through mandating treatment and filtration plants from the 1880s (Hassan 1985; Luckin 1986; Szreter and Mooney 1998).

The nineteenth-century public health movement in other developed countries

However, for all the gradualness and delay, most historians who have studied the distinctive evolution of public health history in other, comparable major countries also undergoing rapidly urbanizing economic growth, tend to portray the British as relatively hyperactive and, overall, endowed with a more effective set of institutions. In his meticulous and wide-ranging comparative examinations of England and nineteenth-century Germany, E.P. Hennock finds a relative disinterest in Germany in its appalling urban death rates and notes that the more radical German medical reformers, like Virchow, had been politically incapacitated and excluded from power by prematurely tying their flag to the failed 1848 republican revolutionary movement (Hennock 1984; 2000). By contrast Chadwick's 'Sanitary Idea' was politically brilliant, as Hamlin has put it (Hamlin 2015: 27), because it was ideologically safe. When, later in the 1880s, Bismarck initiated innovative compulsory accident and health insurance, the spirt of 1848 was long past and his primary motive was to promote productivity of the German industrial economy (Hennock 2007: 94–6).

Hygiène publique in France had been the standout scientific and intellectual leader in the wake of the French Revolutionaries' inclusion of health within their Declaration of the Rights of Man. However, the Revolution's commitment also to the sanctity of private property and an acceptance of J.-B. Say's liberal political economy became the over-riding,

safe consensus of post-Napoleonic French policy, surviving all subsequent regime changes through to 1945 (Ramsey 1994: 50, 57, 70, 82–3). There was suspicion and rivalry almost everywhere between medical professionals selling their therapeutic services in the market-place and preventive public medicine, which would seemingly undercut their market. In France this exerted an exceptionally strong influence impeding the emergence of a public health medical administration. Despite the scientific achievements of bacteriology and the Institut Pasteur in France, the power of the incumbent private practitioners ensured that even when finally a national Public Health act arrived in 1902 it did not produce a salaried public health profession—by contrast with the UK's hundreds of trained MOHs all across the country by this time. Ramsey concludes that the French medical profession has been the most privately dominated of any modern society, apart from USA's (Ramsey 1994: 99).

Indeed, in addition to USA, none of the Anglophone, developed countries beyond Europe—Canada, Australia New Zealand—followed the British lead during the nineteenth century, either in terms of a universal social security system or in creating a national cadre of medically trained public health officials with enforceable powers (Bryder 1994; Fee 1994). The thrust of the foregoing analysis is that in fact the latter absence followed from the former. Like USA, these were all frontier societies of opportunity and immigrant settlers, who bridled at paternalist authorities and associated taxes necessary for collective provision, preferring to take their chances and retain the profits of their own individualist enterprise. This meant that in the various states of Australia, although public health acts were passed, the same factors which had militated against their effectiveness in mid-Victorian Britain were even more of an impediment: local authority jurisdiction countermanding central executive initiatives in the name of upholding private property rights; disinclination to self-tax; and consequently inadequately remunerated health officials and lack of legal enforcement agencies. For instance, though Victoria passed a Public Health Act in 1854, 'marvellous Melbourne' was still a deadly city lacking a sewers system until 1897; New South Wales was unable to pass its own first Public Health Act until the year after it finally and reluctantly introduced its first Income and Land Tax Acts in 1895 (Lewis 2003: 70–90).

In the USA, despite its large number of urban centres, public health medicine was not integrated into urban government as in Britain. Although

the foundation of the APHA (American Public Health Association) dates from 1872, it was initially short of medically trained members and was relatively easy to tar with the brush of being unscientific, paternalist do-gooders when the majority of American medical practitioners subscribed to conventional individualist, free market values. New York city, having commenced the nineteenth century with a colonial inheritance of a municipally organized food supply, deregulated its public food markets in 1843 in response to commercial private interests and promptly unleashed rising inequalities in food access and quality to compound the inequalities of unregulated housing conditions, from which landlords benefitted while immigrants suffered sanitation and crowding diseases (Baics 2016). Although major cities like Chicago and New York did subsequently invest—out of sheer necessity—in municipal water and sanitation projects, nationally there was no co-ordination; not even an investigative Public Health Service until 1912. There was also no universal death registration system until 1926, the essential foundation for intelligent public health work.

Sweden was the principal other European state which did, unlike France and Germany, institute effective public health policies already in the nineteenth century. Sweden was also an early modern unified state with a Protestant established church of loyal local priests, as in England. The Swedish Crown had first become concerned to take measures to boost its population after absorbing in 1748 the disappointing results of its then-new parish population registers, the *Tabellverket*, showing a population of just two million. The following decades saw royal decrees that parishes create poorhouses and over 20 locally financed and state-subsidized hospitals were built so that illness no longer robbed the Crown of productive subjects (Johanisson 1994: 166–8). This pro-populationist public health policy was informed by mercantilism, the dominant economic orthodoxy in this era. This held that successful sovereigns accumulated the limited stock of the world's gold and silver (necessary to employ armies) through the capacity of a large and productive population to sell abroad more products than were bought-in. The mercantilist Swedish public health efforts culminated in the 1815 smallpox legislation and in the same year a unique nationally funded system to care for patients with venereal diseases—many of whom were young men of prime productive and military age, of course (Sköld 1996; Lundberg 1999).

An ambitious state public health policy could be justified on mercantilist grounds and the most comprehensive blueprint for this was the *Medicinische Polizey*, famously laid out in minute detail across nine volumes, 1779–1827, by the eminent German physician, Johann Peter Frank (Rosen 1993: 137–43). When he had begun his publications in the 1770s he had been addressing Europe's enlightened despots, such as Joseph II of Austro-Hungary. This was at a time when for instance, Austro-Hungary maintained a 1,000-mile militarized cordon sanitaire, the *Militärgrenze*, against plague incursion from the Ottoman empire (Flinn 1981, ch.4). However, neither the enlightened despots of Prussia or Russia, nor even Joseph II of Austro-Hungary actually implemented Frank's autocratic and expensive masterplan. It is generally considered that the Swedish state's policies represent the furthest practical elaboration of such mercantilist-inspired state public health measures (Porter 1994: Introduction; Sundin and Willner 2007). As with England, Sweden's historical social-inclusionary commitment to promote the population's health also fed through into early adoption of collectively funded preventive sanitary investment. Sweden's 1874 Public Health Act was strongly influenced by British experience but was passed sufficiently early in Sweden's trajectory of industrialization that it successfully prevented Sweden's urbanizing population from suffering the same epidemiological consequences as those of Britain, France, Germany, and the USA (Nelson and Rogers 1994).

Thus, among the leading industrializing countries during the nineteenth century it was mainly in Britain that important scientific advances led also to practical public health interventions, though these only really began to be extensively funded from the 1870s, when a provincial movement to promote municipal sanitation finally took-off, led by Joseph Chamberlain's mayorship in Birmingham, 1873–6 (Hennock 1973; Szreter 2005, chs.7, 9). The French had pioneered comparative epidemiology in the 1820s but it was in Britain that comparative death rates subsequently became a motivator for a nationwide movement towards the adoption of sanitary measures. Scientific leadership in the application of germ theory to identify the pathogens of communicable diseases clearly occurred first in France (The Institut Pasteur) and in Germany (Koch and the Prussian Institute for Infectious Diseases), yet it was in Britain that an extensive public system for the notification of infectious diseases was developed before the end of the century. Why was this?

Here, once again, the influence of the unusual, universal social security system of England and Wales, even though it was now in its reformed, deterrent variant of the post-1834 New Poor Law, still continued to provide a crucial influence and explanatory context. We have already seen that it was the encompassing Poor Law statutory assumptions of universalist responsibility for the poor which repeatedly led the responsible authorities into innovative forms of healthcare provision, with the building of parish almshouses followed in the eighteenth century by workhouse infirmaries and subscription hospitals; and that the need to constrain costs had led Chadwick to argue the case for sanitary investment to prevent diseases. Exactly the same logic of Poor Law cost containment, while still adhering to principles of universal provision, drove the precocious investment in disease notification in the last decades of the nineteenth century.

The moralizing aim of the New Poor Law was to end outdoor relief and drive all able-bodied workers into the labour market; however, the law recognized that the sick should instead be paid for in their own homes, not in the workhouse, as their incapacity to work was not due to idleness and in any case workhouse infirmaries did not want the risks of infectious patients. Chadwick had tried to reduce the costs of illness by sanitation in 1848. But incapacitating diseases were due to more than unclean water. Two decades later Sir John Simon, the nation's chief medical officer, was able to demonstrate statistically that infected individuals left in the community to spread their afflictions were driving up Poor Law costs and that it would be far more efficient to make a range of such diseases compulsorily 'notifiable' and to then remove the sufferers to isolation facilities (Mooney 2015: 11–12, chs. 1, 3). Mooney points out it was this administrative efficiency reasoning (and the moral and economic ideologies standing behind it), as much as microbial science, which resulted in compulsory notification for various lethal diseases with identifiable causes, such as scarlet fever, diphtheria, enteric fever (typhoid) and smallpox, beginning in London in 1889—following experiments in certain other cities—and then throughout the country with the Infectious Disease (Notification) Act, 1899. In England and Wales unavoidable fiscal responsibilities, dating back to 1601, were laid on local authorities for their sick and poor; and it was crucially this economic and politically generated framework of accountability which enabled bacteriological science to be turned into a practical, funded regime. Doctors, though they were not even state officials, were legally compelled to notify

infectious disease cases; and cities had to invest in freely provided isolation hospital facilities to make the system work as efficiently as possible while meeting its universalist commitment.

Public health in the twentieth century

Sustained global economic growth in the twentieth and twenty-first centuries has, of course, contributed a sequence of important scientific developments, with a range of potential implications for the enhancement of the public's health, such as antitoxins (diphtheria was the first in the 1890s), chemotherapies (salvarsan and neo-salvarsan to treat syphilis was the first in the 1910s), antibiotics (penicillin from the 1940s) and vaccinations (enabling smallpox to be globally eradicated in the 1970s). At the same time economic growth has contributed major new epidemiological challenges, notably the tobacco and obesity epidemics due to insufficiently regulated marketing of profitable, addictive products, HIV-AIDs, and most recently the COVID-19 pandemic, global warming, and various pollutions of land, water and air, due both to industrial practices and to affluence; and also a range of mental health-related conditions such as anorexia and depression. None of these—neither the challenging new afflictions, nor the scientific breakthroughs to counter them—can be simply attributed to the unmediated effects of economic growth on the public's health. Ideology, culture, and politics, interacting with contingent major events such as the World Wars, the Great Depression, and the Cold War, have in all cases been involved as crucial socially and politically determining influences over the exact way in which the public's health and economic growth have co-evolved differently in each distinct national polity.

Most southern and eastern European states, such as Spain, Italy, Greece, Yugoslavia, or Poland, for instance—and also those populations in the Far East, such as Japan, Singapore, and Korea, which became developed countries in terms of per capita income levels in the course of the twentieth century, did so after many of the scientific breakthroughs had been made to deal with communicable diseases and the sanitary problems of dense, urban living. In some cases, such as pre-war Japan or Franco's Spain, these advances were driven by authoritarian regimes, others were more liberal and democratic, such as postwar Japan and Italy. Provided governments promoted universal access to the necessary primary care and public health

services, their populations could achieve better health outcomes while their economies were still growing from a relatively low per capita income level, as the famous Preston curves first demonstrated (Preston 1975). The interwar period also saw the emergence of ambitious public international bodies, notably the League of Nations Health Organization and US philanthropics, such as the Rockefeller Foundation, seeking to promote public health initiatives internationally (Weindling 1995). This pattern has since expanded further in the postwar era to include the World Health Organization (WHO), World Bank, the Gates Foundations, and many other western governmental departments such UK's DfID (Packard 2016; Cueto et al. 2019). However, since Preston's publication of 1975 new global problems of growing inequality and antibiotic resistance and, most especially, pollutants and climate change are proving to be enormous health challenges for all developed and developing societies, alike. Indeed, many species are now threatened, not only poor humans. It is proving extremely challenging for national governments and powerful corporations to set aside their short-term interests to engage in a new level of global civic and political commitment to effective action, with the UN Climate Summit of 2019 the latest in a series of only partially successful attempts to bring this about. The Montreal protocols of 1987 (to address the human threat to stratospheric ozone) still remains the only such UN treaty ratified by all the world's states.

Although democracies have been strongly associated with the rise of welfare states among the economically most advanced societies since the Second World War, it is recognized that these welfare states have taken on many diverse forms influenced by their historical inheritances (Esping-Andersen 1990; Castles 2010). Germany had pioneered state social insurance in the 1880s, whereas in USA there was no nationwide Social Security Act until half a century later in the depths of the Great Depression in 1933. The Second World War, as elsewhere, further empowered public health in Roosevelt's New Deal, as the poor health of many GI recruits was revealed. However, the subsequent Cold War fears of Soviet rivalry enabled the commercial, insurance and medical vested interests once again to reassert a dominant role for private medicine in postwar USA at a time when most European developed countries were endorsing universal compulsory insurance models of healthcare for their new, fully, democratic electorates (Fee 1994: 241, 246, 249, 254). Unlike in McCarthyite USA, these were predominantly seen in western Europe, in the long tradition established by

Bismarck's pioneering 'state socialism', not as 'red', but as the best antidote to communism: social democratic methods to ameliorate working-class insecurity and grievances in a free market economy.

The UK went further than most. The principles of 1601 re-emerged, once those of 1834 were abandoned. Beveridge's welfare state was once again unconditionally universalist, supporting in principle all the vulnerable—old, young and unemployed—and was to be funded by national progressive taxation. In fact, as early as 1917 the UK government had endorsed a free national public health service, but only for venereal diseases (STIs), thereby adopting Sweden's precocious policy of a century earlier, in the interests of protecting its conscripts (Eyler 1987). From 1948 onwards the NHS meant that in principle the entire population was entitled to receive the best preventive and therapeutic medical care the nation could afford for all diseases, not just for syphilis and gonorrhoea. For instance, BCG vaccination against tuberculosis, which had been adopted in Sweden since the 1920s, now became a public health policy protecting the whole British nation (Bryder 1988: 139, 253–4).

With such universalist welfare and public health systems, national average life expectancy among most developed nations now rose above 60 for the first time in history in the immediate postwar decades as most of the sanitary and infectious diseases were tamed. Expectation of life at birth has since risen even above 80 in many advanced nations, though notably not in USA where a system of universal healthcare insurance was only introduced in 2010 and remains politically contested. The stalling of population health improvements in USA both before and since President Obama's 2010 Patient Protection and Affordable Care Act, a stalling also seen to a lesser extent in UK under the Conservative government's decision to impose an austerity policy of cuts on public services rather than additional taxes on the financial sector (despite the uncontested culpability of the latter for the financial crash of 2007–8), has drawn attention to the importance for public health of trends in socio-economic inequality. These are the two developed societies in which economic inequality and its justifying ideology of economic liberalism have had most impact on public policy during the period since c.1980, with regimes in power extolling the virtues of low-taxation and negative liberty (in Isiah Berlin's sense of individual- and therefore corporate—freedom from government regulation and intervention).

As the foregoing has shown, public health does not progress simply because of scientific medical advances or directly because of increases in average per capita income. Economic growth does not necessarily translate into public health. Ideologies which favour private, individualist wealth accumulation over more universalist, more egalitarian distribution militate against providing the funding necessary for effective public health policies. These require collectively funded measures both to prevent those with less private resources from exposure to harm in their environment, and to make available relatively equally the health-enhancing advances of the bio-medical sciences. They attempt to redress what are now termed the social determinants of health (Marmot 2008). Alternative, individualist approaches focus on behaviours and 'lifestyle risks', appearing to presume that the major causes of ill-health in developed societies today are amenable to a significant degree to personal control, even though these notably include cancers, heart disease, obesity, mental afflictions, and dementias.

Conclusions

While it is possible to produce an account of public health history in terms of a sequence of scientific advances, it is more important to examine carefully when, where, and how such advances were adopted into public health practice. Due to the complex administrative and substantial financial resources required to translate medical science into enhancements in the public's general state of health—most particularly in the highly disruptive conditions of living provoked by continuous economic growth—major conflicts of interests, ideologies, and ethics are all involved. This has meant that public health has tended to be promoted most effectively in societies subscribing to collectivist principles of welfare provision. England and Sweden were two Protestant confessional states initially influenced by mercantilist and humanist ideologies, which were historically innovative in adopting universalist principles of policy in relation to social security and health. This endowed both countries with a profoundly pro-public health commitment which has endured through different forms across the centuries. Other developed countries, both the other monarchies of Europe and the individualist, settler societies of 'neo-Europe', have only acquired such collectivist and universalist dispositions towards their citizens' health claims

much later in their histories and in the case of USA incompletely so even until the twenty-first century.

The history of public health in developed countries across the whole era of modern economic history thus reveals several major features, of which all current practitioners need to be fully aware. Firstly, the continuing long-term productivity of all modern national economies, as each century in world history has passed, has always required an ever more well-educated and generally healthy populace, not only to justify, but also both to produce and to consume an increasingly diverse and complex range of goods and services. Secondly, both the public's health and the institutional practice of public health has always been and necessarily always will be interrelated to the prevailing, wider social security system (or lack of it) in the country in which it functions. Thirdly, that this combined social security and public health matrix has enormous generative significance for the potential and the patterns of economic growth achievable in any national context. Public health, efficiently managed, is far more than a cost-burden on the economy; it is an integral, productive constituent of the character of both citizens' and the economy's future growth potential (Keating and Hertzman 1999). Finally, the public health apparatus extant in any national polity is intimately related to its political and fiscal sensitivities. These, in turn, are subject to a process of historical conflict and negotiation over the balance to be struck between Isiah Berlin's (1969) concepts of positive and negative freedoms. To what extent, within a democratic polity respectful of the rule of law, do its citizens place their trust (and taxes) in collectively provided social, environmental, and health services, as against reliance on their own private resources to purchase commercially provided variants of such services? (Nathanson 2007)

With its universalist and progressive taxation principles, the Elizabethan Poor Law statutes are, in effect, jointly the world's first social security and Public Health Act because of their coverage of the sick, the 'lame, impotent, old, blind' as well as those 'not able to work' (ER 43, cap.2 first paragraph). This extraordinary innovation may still provide a fruitful resource of ideas and administrative practices to return to, when considering the relationship between statecraft, economic development, and public health in the future of the world's developed societies.

REFERENCES

Ahearn, B. (2014). The British Industrial revolution in a European Mirror. In: Floud, R., et al. (eds.) *The Cambridge Economic History of Modern Britain, Vol. I 1700–1870*. Cambridge: Cambridge University Press.

Baics, G. (2016). *Feeding Gotham. The Political Economy and Geography of Food in New York, 1790–1860*. Princeton, NJ: Princeton University Press.

Bryder, L. (1988). *Below the Magic Mountain A Social History of Tuberculosis in Twentieth-Century Britain*. Oxford: Oxford University Press.

Bryder, L. (1994). A new world? Two hundred years of public health in Australia and New Zealand. In: Porter, D. (ed.) *The History of Public Health and the Modern State*. Amsterdam: Rodolpi.

Campbell, R.H. and Skinner, A.S. (1976). General introduction to *Adam Smith An Inquiry into the Nature and Causes of the Wealth of Nations, Vol I*. Oxford: Clarendon Press.

Cassel, J. (1994). Public health in Canada. In: Porter, D. (ed.) *The History of Public Health and the Modern State*. Amsterdam: Rodolpi.

Castles, F. et al. (eds.) (2010). *The Oxford Handbook of the Welfare State*. Oxford: Oxford University Press.

Cipolla, C. (1978). The 'Bills of Mortality' of Florence. *Population Studies*, 32(3), 543–8.

Coleman, W. (1982). *Death is a Social Disease: Public Health and Political Economy in Early Industrial France*. Madison, WI: University of Wisconsin Press.

Cueto, M., Brown, T.M., and Fee, E. (2019). *The World Health Organization: A History*. Cambridge: Cambridge University Press.

Eyler, J.M. (1979). *Victorian Social Medicine: The Ideas and Methods of William Farr*. Baltimore, MA: Johns Hopkins University Press.

Eyler, J.M. (1987). *Sir Arthur Newsholme and State Medicine, 1885–1935*. Cambridge: Cambridge University Press.

Esping-Andersen, G. (1990). *The Three Worlds of Welfare Capitalism*. Cambridge: Policy Press.

Fee, E. (1994). Public health and the state: the United States. In: Porter, D. (ed.) *The History of Public Health and the Modern State*. Amsterdam: Rodolpi.

Flinn, M.W. (1981). *The European Demographic System, 1500–1820*. Baltimore, MD: The John Hopkins University Press.

Galloway, P. (1988). Basic patterns in annual variations in fertility, nuptiality, mortality, and prices in pre-industrial Europe. *Population Studies*, **42**, 275–303.

Hamlin, C. (1997). *Public Health and Social Justice in the Age of Chadwick*. Cambridge: Cambridge University Press.

Hamlin, C. (2015). The history and development of public health in developed countries. In: Detels, R., Gulliford, M., Karim, Q.A, and Tan, CC. (eds.) *The Oxford Textbook of Global Public Health* (6th ed.). Oxford: Oxford University Press.

Hanley, J. (2016). *Healthy Boundaries: Property, Law, and Public Health in England and Wales, 1815–1872*. Rochester, NY: Rochester U.P.

Hassan, J.A. (1985). The growth and impact of the British water industry in the nineteenth century. *Economic History Review*, **38**, 531–47.

Healey, J. (2014). *The First Century of Welfare. Poverty and Poor Relief in Lancashire 1620–1730*. Woodbridge: Boydell Press.

Hennock, E.P. (1973). *Fit and Proper Persons: Ideal and Reality in Nineteenth-Century Urban Government*. London: Edward Arnold.

Hennock, E.P. (1984). The creation of an urban local government system in England and Wales. In: Naunin, H. (ed.) *Stadteordnungen des 19. Jahrunderts*. Koln: Bohlau Verlag.

Hennock, E.P. (2000). The urban sanitary movement in England and Germany, 1838–1914: a comparison. *Continuity and Change*, **15**, 269–96.
Hennock, E.P. (2007). *The Origin of the Welfare State in England and Germany, 1850–1914*. Cambridge: Cambridge University Press.
Hopkins, A.G. (ed.) (2002). *Globalization in World History*. London: Pimlico.
Ishkanian, A. and Szreter, S. (eds.) (2012). *The Big Society Debate. A New Agenda for Social Welfare?* Cheltenham: Edward Elgar.
Johanisson, K. (1994). The people's health: public health policies in Sweden. In: Porter, D. (ed.) *The History of Public Health and the Modern State*. Amsterdam: Rodolpi.
Keating, D.P. and Hertzman, C. (eds.) (1999). *Developmental Health and the Wealth of Nations. Social, Biological and Educational Dynamics*. New York: Guilford Press.
King, S, (2018), *Sickness, Medical Welfare and the English poor 1750–1834*. Manchester: Manchester University Press.
Kreager, P. (1988). New light on Graunt. *Population Studies*, 42(1), 129–40.
La Berge, A.F. (1992). *Mission and Method. The Early Nineteenth-Century French Public Health Movement*. Cambridge: Cambridge University Press.
Lewis, M. (2003). *The People's Health. Vol. I: Public Health in Australia, 1788–1950*. London: Praeger.
Luckin, B. (1986). *Pollution and Control. A Social History of the Thames in the Nineteenth Century*. Bristol: Adam Hilger.
Lundberg, A. (1999). *Care and Coercion: Medical Knowledge, Social Policy and Patients with Venereal Disease in Sweden 1785–1903*. Report no. 14 from the Demographic Data Base. Umeå: Umeå University.
Mandler, P. (1987). The making of the new poor law redivivus. *Past and Present*, **117**, 132.
Marmot, M. (2008). *Closing the Gap in a Generation: Health Equity Through Action on the Social Determinants of Health*. WHO Commission on Social Determinants of Health—final report. Geneva: WHO.
Mokyr, J. (1990). *The Lever of Riches*. Oxford: Oxford University Press.
Mooney, G. (2015). *Intrusive Interventions. Public Health, Domestic Space, and Infectious Diseases Surveillance in England, 1840–1914*. Rochester: Rochester University Press.
Nathanson, C.A. (2007). *Disease Prevention as Social Change: The State, Society, and Public Health in the United States, France, Great Britain, and Canada*. New York: Russell Sage Foundation.
Nelson, M.C. and Rogers, J. (1994). Gleaning up the cities: application of the first comprehensive public health law in Sweden. *Scandinavian Journal of History*, **19**, 17–39.
Rosen, G. (1993). *A History of Public Health*. Expanded edition [orig. 1958]. Baltimore, MD: Johns Hopkins University Press.
Packard, R.M. (2016). *A History of Global Health: Interventions into the Lives of Other Peoples*. Baltimore, MD: Johns Hopkins University Press.
Porter, D. (1994). Introduction. In: Porter, D. (ed.) *The History of Public Health and the Modern State*. Amsterdam: Rodolpi.
Preston, S.H. (1975). The changing relation between mortality and level of economic development. *Population Studies*, 29(2), 231–48.
Ramsey, M. (1994). Public health in France. In: Porter, D. (ed.) *The History of Public Health and the Modern State*. Amsterdam: Rodolpi.
Reinarz, J. and Schwarz, L. (eds.) (2013). *Medicine and the Workhouse*. Rochester, NY: Rochester University Press.
Risse, G. (1992). Medicine in the Age of Enlightenment. In: Wear, A. (ed.) *Medicine and Society*. Cambridge: Cambridge University Press.
Sen, A. (1999). *Development as Freedom*. Oxford: Oxford University Press.

Siena, K. (2004). *Venereal Disease, Hospitals and the Urban Poor: London's 'Foul Wards,' 1600–1800*. Rochester, NY: Rochester U.P.
Sköld, P. (1996). *The Two Faces of Smallpox. A Disease and its Prevention in Eighteenth and Nineteenth-Century Sweden*. Report no. 2 from the Demographic Data Base. Umeå, Sweden: Umeå University.
Slack, P. (1985). *The Impact of Plague in Tudor and Stuart England*. London: Routledge & Kegan Paul.
Slack, P. (1990). *The English Poor Law 1531–1782*. Basingstoke: Macmillan.
Smith, R.M. (2011). Social security as a development institution?. In: Bayly, C.A., et al. (eds.) *History, Historians and Development Policy*. Manchester: Manchester University Press.
Smith, V. (2007). *Clean. A History of Personal Hygiene and Purity*. Oxford: Oxford University Press.
Solar, P.M. (1995). Poor relief and English economic development before the industrial revolution. *Economic History Review*, **48**, 1–22.
Sundin, J., and Willner, S. (2007). *Social Change and Health in Sweden: 250 Years of Politics and Practice*. Stockholm: Swedish National Institute of Public Health.
Szreter, S. (2005). *Health and Wealth. Studies in History and Policy*. Rochester, NY: Rochester University Press.
Szreter, S. and Mooney, G. (1998). Urbanisation, mortality and the standard of living debate. *Economic History Review*, **51**, 84–112.
Weindling, P. (ed.) (1995). *International Health Organisations and Movements, 1918–1939*. Cambridge: Cambridge University Press.
Williams, S. (2011). *Poverty, Gender and Life-cycle Under the English Poor Law, 1760–1834*. Woodbridge: Boydell Press.
Wong, R.B. (2011). Historical lessons about contemporary social welfare: Chinese puzzles and global challenges. In: Bayly, C.A., et al. (eds.) *History, Historians and Development Policy*. Manchester: Manchester University Press.
Wrigley, E.A. (1987). *People, Cities and Wealth*. Oxford: Basil Blackwell.
Wrigley, E.A. (1999). Corn and crisis: Malthus on the high price of provisions. *Population and Development Review*, **25**, 121–8.
Wrigley, E.A. (2004). British population during the 'long' eighteenth century, 1680–1840. In: Floud, R., and Johnson, P. (eds.) *The Cambridge Economic History of Modern Britain Volume I —1700–1860*. Cambridge: Cambridge University Press.
Wrigley, E.A. and Smith, R. (2020). Malthus and the Poor Law. *The Historical Journal*, **63**, 33–62.

---

[1] As Cipolla carefully pointed out, although the governments of several Italian city states began collating burials in 'Books of the Deceased' from as early as the fourteenth century, they did not seek to make the information public in 'Weekly Bills' circulated to the population. Since the Italian data subsequently 'remained buried … in the cities' archives' it did not contribute to the early development of demography, which occurred when John Graunt set in train a scientific public discourse based on his systematic analysis of the London Bills (Cipolla 1978: 543; Kreager 1988).

[2] For a review of the range of such interpretations, see Ahearn (2014) and Mokyr (1990: 3–6) on Solovian, Smithian, Boserupian, and Schumpeterian causes of growth.

[3] See Slack (1990: 57) on the reasons for the failure of a similar, effective parish scheme to develop in the course of the seventeenth century in Scotland (Ireland was not part of the United Kingdom until 1801). See Ishkanian and Szreter (2012: 16–17, 42–3) on the relation between the Poor Law and the Charity statute.

[4] As Wrigley (1999) notes, this was a positive effect of the Poor Laws, favourably pointed out by Malthus himself despite his generally critical disposition towards them.

[5] Smith's critique was empirically flawed because he failed properly to acknowledge the significance for labour mobility of the flexibility which certification and other widespread practices permitted (Campbell and Skinner 1976: 53–4), while Malthusian fears that generous outdoor relief encouraged the unemployed to marry too young and have too many children were also unevidenced and have since been proven to be incorrect: Wrigley and Smith (2020).

1.3

The history and development of public health in low- and middle-income countries

Than Sein

Introduction to the history and development of public health in low- and middle-income countries

Historically, public health interventions are those actions to promote and protect people's health, and are chiefly undertaken by the governments. The notion of public, within the term public health, encompasses actions for health development by the people themselves individually and collectively. Public health interventions not only deal with the control and management of both communicable and non-communicable diseases, but also address the promotion, prevention, or reduction of risks and root causes of these diseases. Fundamental characteristics of public health actions lie primarily on the social and other determinants of health that fall outside the domain of the government's health sector—and also beyond the individual's action. The majority of people in low- and middle-income countries are still living in poverty with very low health status. According to the United Nations (UN) classification of countries by their development as measured by per-capita gross national income (GNI) in 2017: countries that have per-capita GNI below US$1,005 are designated as low-income countries (32 in total); those which have per-capita GNI between US$1,006 and US$3,955 are lower-middle-income countries (38 in total); and those between US$3,956 and US$12,235 as upper-middle-income countries (42 in total). Countries with per-capita GNI of more than US$12,235 are categorized as high-income countries (UN 2018a).

This chapter describes the history and development of public health in 112 countries, classified as low-, lower- and upper-middle-income countries, also referred to as low- and middle-income countries (LMICs). The population of

LMICs (excluding those of Brazil, China, and India) is around three billion (about half of the world's population); thus, the advancement of health development in LMICs has a major impact on the health of the world. The present chapter provides an updated review of the history and development of public health in LMICs, covering the key factors and events of the last decade.

Historical development

The majority of LMICs came under the colonial territories during the eighteenth and nineteenth centuries, and some have not yet been independent until near the end of the twentieth century. The initial foundation for public health was formally laid down through the intensification of international and local public health actions by the colonialists in their territories, including international health diplomacy among themselves from the eighteenth century onwards. The United States and the European countries applied protective legislative measures and preventive actions through several series of international sanitary conventions. Soon after the League of Nations and its health organization were founded, the consciousness on public health prevention and control at the international level became more realized. The World Health Organization (WHO), the first truly international and intergovernmental agency, was founded in 1946. Its two main functions were to direct and coordinate international health work and cooperate with the Member States and partners in international health development. After WHO, several international agencies had come up under the UN to deal with health, social and economic development, such as UN Development Program (UNDP), UN Children's Fund (UNICEF), UN Population Fund (UNFPA), UN High Commissioner for Refugees (UNHCR), UN AIDS, etc. The original mandates for establishing these agencies have improved and changed over time. By the late 1990s, several international philanthropies and international civil societies had come out as major financing and development partners.

Control and eliminating old scourges

While effective prevention and control measures for old scourges like smallpox, cholera, plague, and yellow fever, that are major cripplers and

killers over the last 500 years, are available in high-income countries for centuries, a substantial portion of the world's population living in LMICs have not been able to gain access to these measures. The control of smallpox in LMICs after the invention of smallpox vaccination in the early nineteenth century in Europe was hindered by several factors, such as: impurity and variable potency of the vaccine; insufficient quantities of heat-stable vaccine; poor vaccination technique; poor coverage by the health workers due to lack of appropriate human resources; lack of commitment by the colonial administration; and lack of confidence in vaccination among the local populace. The continued global advocacy and provision of technical, material (heat-stable freeze-dried vaccines), financial, and human resources support from WHO, UNDP, and many donor countries had led to successful containment of the disease by the mid-1960s. Combined strategies of routine vaccination of children and mopping up mass vaccination, coupled with intensive case-detection, contact tracing, isolation of cases, and vaccination of individuals in the vicinity who could have been exposed to the infection, had managed to control smallpox epidemics and cut the transmission. After the last naturally acquired human smallpox case in Somalia in October 1977, in May 1980 the world was declared free from natural transmission of the smallpox virus. An international arrangement was made to create an emergency stockpile of over 60 million doses of the first- and second-generation smallpox vaccine in safe and secure places, to be used for the effective control of any smallpox outbreak that might occur (WHO 2011a; WHO 2016a). Since smallpox has been used historically as a notorious biological warfare agent, and it is easier than ever to create and spread such a deadly pathogen, the world has been warned that a bioterrorist attack that could wipe out 30 million people is becoming increasingly likely to happen (The Telegraph 2017). The 1980 eradication of smallpox by natural transmission stands out as the most successful public health intervention achieved globally in the twentieth century.

Another devastating disease, cholera, occurred worldwide for many centuries due to its severity and quick and easy spread across borders. Although it was identified as a globally notifiable disease under the old version of the International Health Regulations (IHR), report on its occurrence was no more required under the new version of IHR except in a public health emergency. Cholera remains a significant global public health problem, since 38 countries reported a total of 132,121 cases including 2,420

deaths, resulting in an overall case fatality ratio (CFR) of 1.8% in 2016. The Democratic Republic of the Congo, Haiti, Kenya, the United Republic of Tanzania, and Yemen accounted for 80% of all cases (WHO 2017a). Local armed conflicts, mass urbanization, climate change, continuing unsanitary practices (e.g. open defecation), and growing competition for water resources may, in fact, contribute to the high disease burden. Effective emergency outbreak control with good personal hygienic practices and safe water supply and sanitation remain the best preventive and control measure. Properly administered rehydration therapy and oral vaccination at the time of outbreaks could save thousands of lives. Around 20 countries affected by cholera pledged to eliminate the disease by 2030 and reduce the deaths by 90% in 2030 (WHO 2017b).

Human plague, a bacterial infection mainly affecting small mammals and transmitted by fleas but infecting humans, is often considered a problem of the ancient disease. Between 1 January 2010 and 31 December 2015, more than 3,200 cases of human plague were reported, resulting in 584 deaths, mainly occurring in 11 countries of Africa, Asia, and America. A recent outbreak of human plague with a higher frequency of pneumatic form in Africa has illustrated the breakdown of health systems due to social and political crises. The occurrence of human plague clearly demonstrates the need for strengthening surveillance on rat falls (deaths of rats around human habitats), early diagnosis and prompt treatment of human plague cases, and improved preventive and control measures in endemic countries. All these require an effective public health system (WHO 2016b). Human plague is still a global challenge as a scourge of humanity in the twenty-first century.

Another old scourge, yellow fever (YF), has been almost eliminated in several parts of Africa and South America with an effective vaccination programme, together with vector control and widespread adoption of personal protection measures. Since there is no effective specific treatment, immunization is the only means of protection against YF. The disease remains endemic in many LMICs, with an estimated 200,000 cases worldwide each year due to declining population immunity to infection, deforestation, urbanization, population movement, and climate change. While YF spreads in historical times from the original sites in Africa to the Americas, it has never established itself in Asia, where suitable conditions for the vectors and hosts exist (WHO 2017c). The reasons for this are not well understood; hence, it is difficult to assess the probability of YF

establishing itself in Asia in the near future. It could have serious consequences, given the large and dense immune-naive rural and urban population in Asia. The ongoing circulation of wild YF virus, coupled with low vaccination coverage and the presence of vector mosquitoes around the world, means that YF remains an important threat for international spread in the twenty-first century.

Elimination of childhood vaccine-preventable diseases

After increasing availability and accessibility of improved vaccines for major infectious diseases affecting infants and young children, all LMICs have commenced national immunization programmes in mid-1980, under the sponsorship of the global Expanded Program of Immunization (EPI). This global EPI initially planned to control a few childhood vaccine-preventable diseases with the ultimate aim of eliminating or eradicating them. With the continuous technical, material, and financial inputs by the Global Alliance for Vaccines and Immunization (later known as the GAVI Alliance) since early 2000, immunization coverage has been expanded both by population and antigens. Globally over 116.5 million children (85%) received a third dose of DPT (diphtheria, pertussis, and tetanus) (DPT3) in 2016, compared to 24.2 million in 1980. Measles vaccine coverage was around 70% in Africa, whereas it was over and around 90% in Europe, Asia, and the Pacific. While bacille Calmette–Guérin (BCG) and poliomyelitis vaccination remained over 80% in all WHO regions, the coverage for third doses of hepatitis B (HepB3) and *H. Influenza* (Hib3) remained below 80% in Africa (WHO 2017d). This situation highlights the need to address limited national and community resources, competing health priorities, poor health systems management, and inadequate monitoring and supervision. Despite these challenges, many public health professionals have considered the global EPI initiative a silent revolution in public health. Figure 1.3.1 shows the immunization coverage of major childhood vaccine-preventable diseases by WHO regions in 2016.

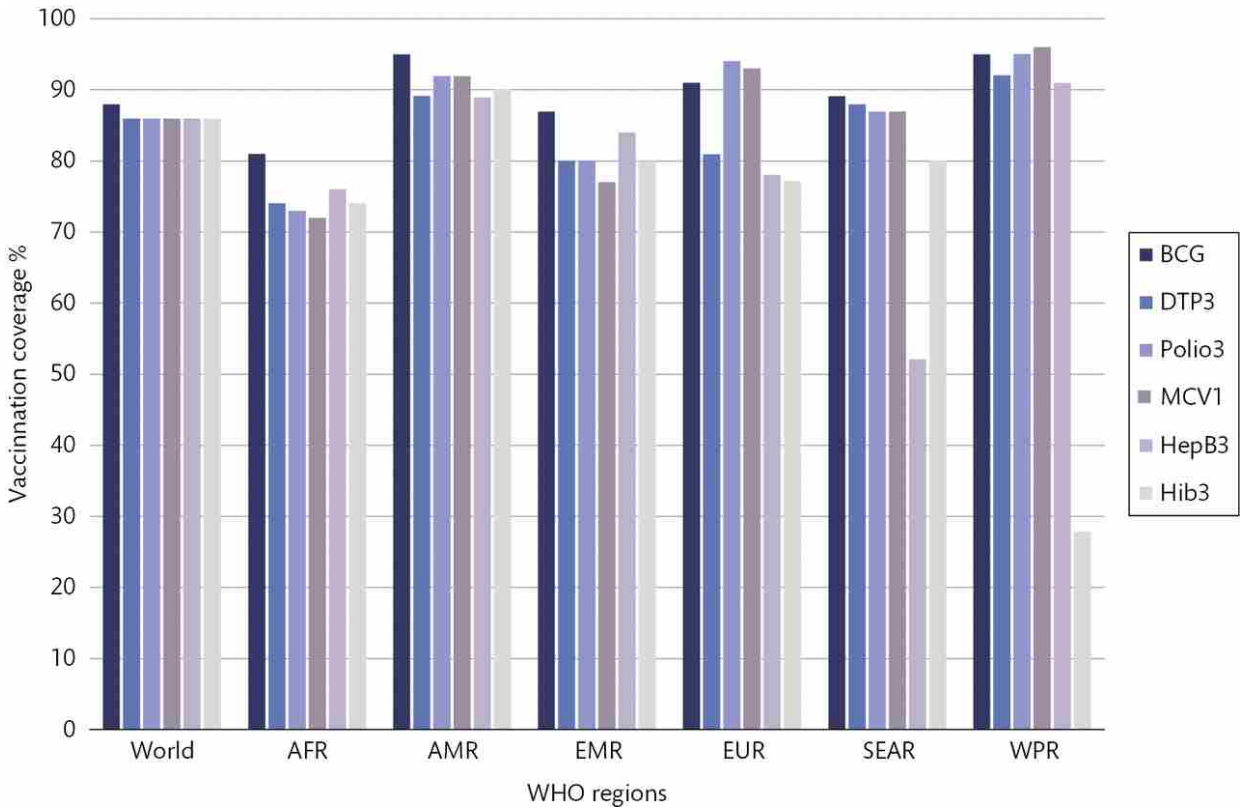


Fig. 1.3.1 Vaccination coverage of major vaccine-preventable diseases by vaccines and WHO Regions, 2016.

Abbreviations: BCG = bacille Calmette–Guérin; DTP3 = 3 doses of diphtheria-tetanus-pertussis vaccine; Polio3 = 3 doses of polio vaccine; MCV1 = 1 dose of measles-containing vaccine; HepB3 = 3 doses of hepatitis B vaccine; Hib3 = 3 doses of haemophilus influenza type B vaccine. WHO Regions: AFR = WHO African Region; AMR = WHO American Region; EMR = WHO Eastern Mediterranean Region; EUR = WHO European Region; SEAR = WHO South-East Asia Region; and WPR = WHO Western Pacific Region.

Source: Data from World Health Organization, *Global routine vaccination coverage, 2016*, WER 92(46), 701–16.

Among the vaccine-preventable diseases, the goal of eliminating poliomyelitis, an acute potentially paralysing disease of children, was launched in 1988 under the Global Eradication of Poliomyelitis programme. In the beginning, the wild poliovirus was circulating in more than 125 countries on five continents, disabling more than 1,000 children every day. With concerted efforts and additional support from the GAVI Alliance, the global coverage of polio immunization over the next two decades had reached around 80% in all WHO Regions. WHO Region of the Americas (AMR) was certified polio-free in 1994, followed by WHO Western Pacific Region (WPR) in 2000, and then European Region (EUR) in 2002. After

several rounds of vaccination against polio with designated national immunization days (NIDs), many countries were able to announce the interruption of indigenous polio transmission (WHO 2012). By 2014, the WHO South-East Asia Region (SEAR) was certified polio-free. This achievement of the majority of WHO regions marks a significant leap forward in global eradication, with 80% of the world's population now living in certified polio-free places. More than 16 million people are able to walk today who would otherwise have been paralysed. By 2018, only three countries, viz. Afghanistan and Pakistan in Asia, and Nigeria in Africa, had endemic infections due to indigenous transmission of wild polioviruses. These countries had accounted for 90% of all new cases of poliomyelitis. Failure of the global polio eradication effort, while seemingly unthinkable, would result in the return of polio infection worldwide and constitute the most expensive public health failure in history. If and when the world successfully eradicates wild poliovirus, this would represent a major milestone of public health success in the twenty-first century. Figure 1.3.2 shows the trends for the coverage of poliomyelitis immunization (three doses) for infants by WHO Regions, in 1980, and 1990 to 2016.

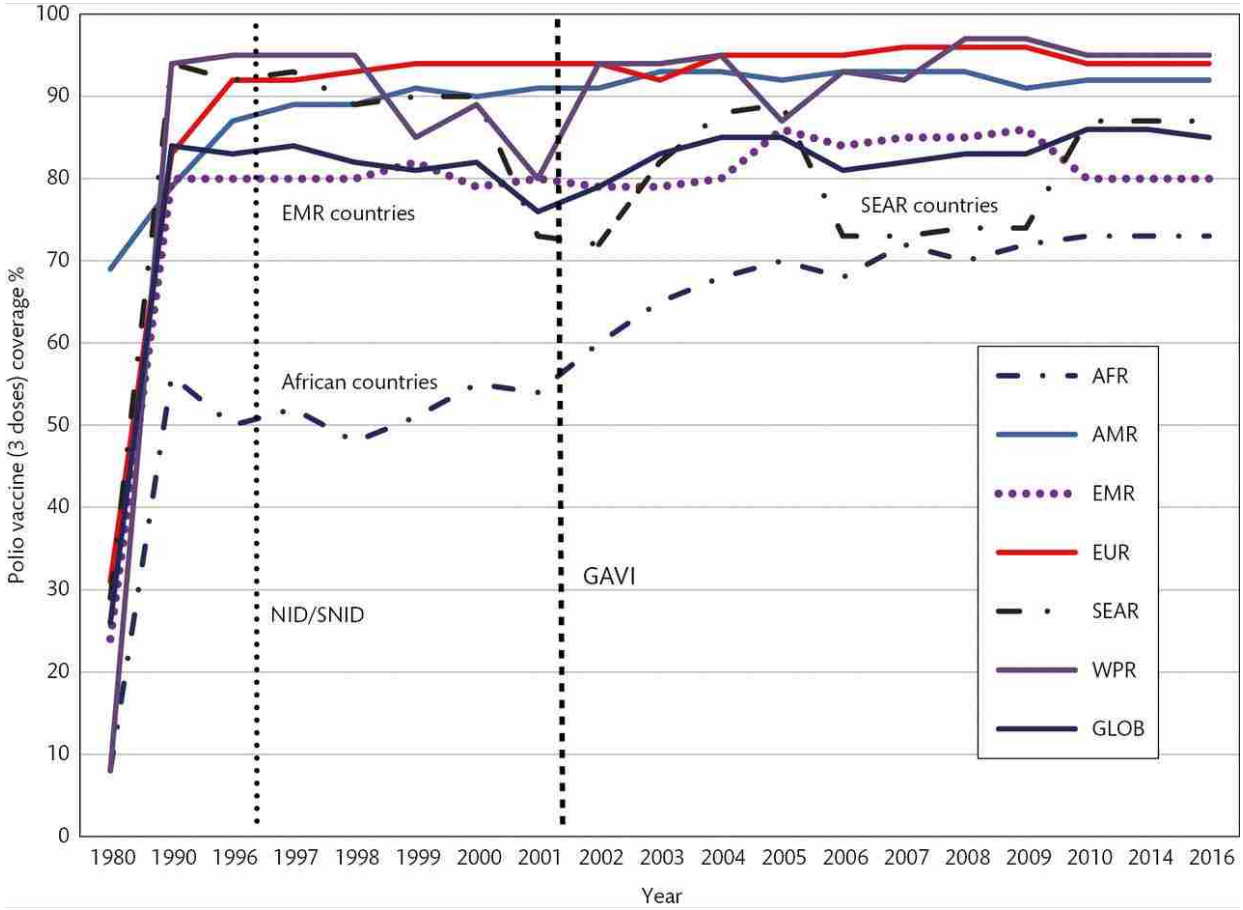


Fig. 1.3.2 Poliomyelitis immunization (three doses) coverage for <1 year of age by WHO regions, 1980 and 1990–2016.

Abbreviations: WHO Regions: AFR = WHO African Region; AMR = WHO American Region; EMR = WHO Eastern Mediterranean Region; EUR = WHO European Region; SEAR = WHO South-East Asia Region; and WPR = WHO Western Pacific Region

Source: Data from WHO Global Health Observatory Data Repository, Copyright World Health Organization 2017.

Measles is an easily transmitted viral infection responsible for around 10% of deaths from all causes among under-five children globally, and could easily be preventable by immunization. Measles control was included as part of the routine EPI programme in LMICs since mid-1980. By 2000, 72% of the world's children received at least a single dose of measles vaccine (MCV1) (against 16% in 1980). The entire American continent had eliminated measles infection (i.e. no indigenous case for more than 12 months) by 2002. This successful experience had prompted other continents to launch similar measles elimination campaigns. From 2000 to 2016, MCV1 coverage had increased globally from 72% to 85%. The number of

measles cases reported annually worldwide decreased from over 854,000 cases in 2000 to 132,000 in 2016. For the same period, the number of measles deaths declined from 550,100 in 2000 to 89,780 in 2016 (WHO 2017e). While it may not be possible for many LMICs to eliminate measles in the near future, maintaining the high level of coverage of measles immunization of all infants (more than 85%) is nevertheless highly beneficial in reducing sickness, lifelong disability, and deaths at early ages.

High coverage of immunization with tetanus toxoid (TT) to all women of childbearing age and high coverage of DPT vaccine for infants and with boosters among school children, are the simplest and most cost-effective ways to reduce tetanus infection in women and neonates, as well as deaths due to neonatal tetanus (NNT). Efforts have been made since 1989 to eliminate the disease (i.e. reducing the level of incidence of one NNT case per 1,000 live births in all districts of the country). However, by 2016, around 25 countries of the world, all from LMICs, remained tetanus endemic.

Immunization against *hepatitis B* (HepB), another highly prevalent viral infection of the liver that could lead to cirrhosis and cancer, was introduced as part of a routine EPI programme in 183 countries since early-1990. Approximately 257 million people live with hepatitis B virus infection (defined as hepatitis B surface antigen-positive). In 2015, hepatitis B resulted in 887,000 deaths, mostly from complications (including cirrhosis and hepatocellular carcinoma). For the last 25 years, almost all LMICs have used both plasma-derived and recombinant HepB vaccines, achieving over 80–90% coverage globally in 2016 (WHO 2017f). If the continuous vaccine supply can be ensured and the high coverage maintained, HepB infection could be under control and even eliminated within the next two decades.

A potent vaccine against another viral infection caused by *Haemophilus influenzae* type b or Hib has been available since mid-1990, being introduced under routine immunization programmes in 189 countries. Global coverage with three doses of Hib vaccination was estimated at around 52%, while some regions have better coverage of over 90% (WHO 2006a).

Control of neglected tropical diseases

A group of infectious diseases that cause illness and death of thousands of people living in the tropical zones worldwide have been neglected for

various reasons for complete prevention and control. Leprosy, the most neglected tropical disease (NTD), has been put under control by accelerated efforts in the last 25 years with the introduction of multidrug therapy (MDT). By 2001, 107 of 122 endemic countries were able to achieve the global target of leprosy elimination (i.e. reducing the prevalence to a level below one case per 10,000 people). Globally, the number of new leprosy cases reported from 143 countries had come down from a peak of over 775,000 cases in 2001 to 214,783 in 2016. Over 90% of the new cases during the last decade were represented from 22 priority countries in Asia, the Pacific, and Africa (WHO 2017g). If the current trend in disease reduction in endemic countries is well sustained with full support from the Global Alliance, the incidence and total burden of leprosy could be further reduced in the near future. Leprosy elimination could be another public health landmark during the twenty-first century.

Trachoma is a chronic infection of the eyes by the bacterium *Chlamydia trachomatis*, mainly affecting poor people and children living in 41 tropical countries of Asia and Africa, especially in areas where unhygienic personal practices still prevail and inadequate access to safe water supply, and is responsible for the blindness or visual impairment of about 1.9 million people. Under the national integrated trachoma control programme, many endemic countries implemented the WHO-advocated SAFE strategy (surgery, antibiotics, facial cleansing, and environmental improvement). In 2016, 190.2 million people still lived in trachoma-endemic areas and are at risk of trachoma blindness. Global-level antibiotic coverage was 44.8%, a considerable increase compared to the 29.6% achieved in 2015. As of 1 July 2017, 10 countries (Cambodia, China, Gambia, Ghana, Islamic Republic of Iran, Lao People’s Democratic Republic, Mexico, Morocco, Myanmar, and Oman) had reported achieving elimination goal, which signifies a major milestone in the campaign to eliminate trachoma (WHO 2017h).

The global effort for eradicating yaws, another disabling NTD, has been launched in many LMICs around the world in the last five decades through mass vaccination programmes and effective mass treatment. However, sporadic epidemics and scattered foci of infection persist in some parts of Latin America, the Pacific, and South and South-East Asia. Extensive surveillance and case finding, use of newer antibiotics, and follow-up of treated cases and contacts became important strategies for the eradication of yaws in the remaining endemic countries.

Visceral leishmaniosis or kala-azar, another NTD, has been put under control by using an oral medication, *Miltefosine*. Until newer medicines and appropriate vaccine(s) are available at affordable prices, and made accessible to the mass population in the endemic countries, the misery and suffering of thousands of people from visceral leishmaniosis would continue.

Onchocerciasis, also known as river blindness, is another parasitic infection still endemic in 37 countries in Africa and Latin America. A million people are visually impaired, and over 300,000 people are blind due to this disease. The Global Onchocerciasis Control Program launched in 1974 had tried to reduce severe pathological manifestations through community-based mass management with a microfilarial drug (*ivermectin*) combined with vector control. Although the disease is a preventable cause of blindness and its elimination goal is achievable, there is a great deal of work to be done by the health systems in endemic countries (Diawara et al. 2009).

Dracunculiasis is a vector-borne guinea worm infestation, but an eradicable disease, as diagnosis is easy and unambiguous, and control interventions are simple, cost-effective, and relatively easy to implement. The disease has limited geographical distribution and transmission is seasonal. Many endemic countries in Asia and Africa with strong political commitment from governments were able to achieve the global eradication target set in 2004 by reducing the annual incidence from 3.5 million in 1986 to less than 1,100 cases in 2011 (WHO 2011b). As of 31 December 2015, 196 countries, territories, and areas have been certified free of dracunculiasis transmission. By 2016, the disease was confined only to three endemic countries (Sudan, Chad, and Ethiopia), where less than 30 cases were reported (WHO 2018). The challenge for eradication efforts in these countries is the emergence of another host, i.e. domestic dogs (Molyneux and Sankara 2017). If concerted efforts continued, eradication of dracunculiasis could be achieved in the near future without any use of vaccine and medicine.

Lymphatic filariasis (LF) is also a disabling NTD caused by blood-borne infection of microscopic thread-like parasitic worms—*microfilariae*. The global programme to eliminate LF through mass drug administration (MDA) was launched in 2000. With 556 million people in 39 of 73 endemic countries treated during 2015, the cumulative total of MDA now exceeds 6.2 billion. A total of 18 countries have halted MDA nationwide and six of these (Cambodia, Cook Islands, Maldives, Niue, Sri Lanka, and Vanuatu) were

acknowledged by WHO as having eliminated LF as a public health problem. Nonetheless, 2738 implementation units across 54 countries, with approximately 946 million people, are still considered to require MDA to achieve elimination targets (WHO 2017i).

Controlling non-communicable diseases

While communicable diseases continue to be major public health problems, chronic non-communicable diseases (NCD) are emerging in LMICs as public health challenge for the 21st century. In 2015, an estimated 40 million deaths occurred due to NCD, accounting for 70% of the overall total of 56 million deaths. The majority of such deaths were caused by the four main NCD, namely: cardiovascular disease—17.7 million deaths (accounting for 45% of all NCD deaths); cancer—8.8 million deaths (22%); chronic respiratory disease—3.9 million deaths (10%); and diabetes—1.6 million deaths (4%) (WHO 2017j). Many large-scale, well-designed population-based control programmes had shown that lowering the common risk factors like tobacco and alcohol use, high sugar or salt including unhealthy diets, can reduce morbidity and deaths from NCDs. The dual approach of early screening and mitigating risk factors, and promoting population-wide preventive measures like policy and legislation is an appropriate strategy for LMICs.

Diabetes mellitus, either insulin-dependent (Type 1) or non-insulin-dependent (Type 2), has affected around 200 million people of above 20 years of age worldwide. Gaps in early diagnosis and treatment of diabetes in LMICs are the biggest challenges, and health systems are not well-prepared for narrowing these gaps. The number of patients with diabetes may increase to 370 million by 2030, if proper promotive, preventive, and control strategies are not in place. With early screening and diagnosis followed by effective management with diet, exercise, and essential medicines, diabetes, and its complications can be potentially controlled.

Cancer is a well-known NCD and a leading cause of death worldwide, with more than 70% of all cancer deaths occurring in LMIC. Liver cancer affects approximately 5.5 million people, predominantly caused by viral hepatitis B and C. Stomach cancer caused by *Helicobacter pylori* infection could be easily prevented by early prevention and control of infection with appropriate medicines. The global burden of cancer in women has recently

received much attention, but there are few comprehensive reviews of the burden and policy approaches to reduce it. Early detection with proper screening and improved breast cancer treatment has reduced mortality by 45% in women above 50 years of age. Cancer of the uterine cervix, which is prevalent among women of childbearing age, is mainly due to chronic infection with sexually transmitted *human papillomavirus* (HPV). In LMICs, cervical cancer is the leading cause of cancer death in women, and 91% of global estimated HPV-related cancer deaths are due to cervical cancer. HPV vaccines are very effective at preventing the disease, and the protection lasts for at least 5 years (WHO 2009). Pap smear screening and early treatment also significantly reduced mortality from cervical cancer.

There are examples of proven public health interventions with a particular focus on the primary prevention of risk factors. Many of these measures are extremely cost-effective, especially in the case of tobacco control and vaccination against cancer-causing diseases. LMICs must prioritize efforts to reduce known cancer risk factors and make prevention and treatment accessible to all. Early diagnosis, effective treatment, and palliative care are also required. Expanding population-based cancer registries and research in LMICs are needed to set cancer control priorities and determine the most effective interventions.

Environmental health promotion

For centuries, the promotion of environmental health, including public housing, public water supply, personal hygiene, and sanitation, has been recognized as key elements for public health development. Despite legislative measures supplemented by public health education campaigns, public subsidy support, and private donation, the progress in environmental health (specifically safe water supply and sanitation) in LMICs has not been satisfactory. According to WHO, the population using improved drinking-water sources in LMICs had increased from 70% in 1990 to 87% by 2010, while the population using improved sanitation in the same group of countries in 2010 was still less than 50% (WHO 2017j). About 2.4 billion people around the world, mostly in Asia and Africa, were still lacking access to improved sanitation facilities, mainly due to low priority given to the sector, inadequate financial resources, and lack of sustainable public water

supply and sanitation services, especially the existence of unimproved sanitation and open defecation, and poor personal hygienic practices.

The situation is exacerbated by inadequate provision of public water supply and sanitation facilities, even in public places like wet markets, parks, movie theatres, schools, hospitals, health centres, offices, and religious places. Providing access to low-cost, adequate, and simple, improved drinking-water sources and improved facilities for sanitary disposal of excrement, and promoting sound hygienic behaviours among the general population in LMICs is of critical importance in reducing the burden of communicable diseases. Given the current economic growth prospect, LMICs have the challenge of getting additional financial support from external agencies, in order to achieve the UNSDG for environmental health improvement (WHO/ UNICEF 2017).

The UN Sustainable Development Goals (SDGs) on health has a target that by 2030, there will be substantially reducing the number of deaths and illnesses from hazardous chemicals and air, water, and soil pollution and contamination. Around 4.5 million deaths occur every year as a result of household exposure to smoke from dirty cook stoves and fuels, and 3 million deaths every year as a result of exposure to ambient (outdoor) air pollution. Most air-pollution deaths are premature deaths, largely as a result of increased mortality from stroke, heart disease, chronic obstructive pulmonary disease, lung cancer, and acute respiratory infections. People living in LMICs disproportionately experience the burden of outdoor air pollution, with 87% of the 3 million premature deaths (WHO 2016c). Policies and investments supporting cleaner transport, energy-efficient housing, power generation, industry, and better municipal waste management, agricultural waste incineration, forest fires, and certain agro-forestry activities (e.g. charcoal production) would reduce key sources of urban outdoor air pollution.

Climate change and its impact

Climate change is now affecting every country on every continent. It is disrupting national economies and affecting lives, costing people, communities, and countries dearly today and even more tomorrow. People are experiencing the significant impacts of climate change, which include changing weather patterns, rising sea levels, and more extreme weather

events. The greenhouse gas emissions from human activities are driving climate change and continue to rise. Without action, the world's average surface temperature is projected to rise over the twenty-first century. It is likely to surpass 3ºC this century—with some areas of the world expected to warm even more. Implementation of the Paris Agreement adopted in 2015 provides a roadmap for climate actions that will reduce emissions and build climate resilience (UN 2018b).

New generation of health systems reforms

For effective control of both communicable and NCDs and their risks, there is a need to have viable and efficient health systems. Historically, many LMICs tried to establish their healthcare facilities and public health measures, based on the models of their colonial masters. In general, these health systems were publicly funded, except in some African countries where basic healthcare relied upon health facilities and staff supported by non-governmental organizations. Several medical schools and paramedical training institutes, public health training and research institutions and healthcare facilities were later added with technical and financial assistance from their colonial rulers.

Health systems development in the twenty-first century

Until mid-1970, many LMICs had weak health infrastructure for maternal and child health (MCH) care and communicable disease control, leading to high mortality and morbidity from maternal and childhood illnesses. Immunization coverage for children was very low, and essential MCH care had to rely on poorly trained human resources at hospitals and hospital-based clinics, located mostly in urban areas. It was later realized that the vertical approach of opening MCH centres and deploying MCH workers alone did not achieve the purpose of expanding MCH care. In 1980–1990, many LMICs started adopting comprehensive population policies and family planning services as part of MCH care to address both demographic challenges and maternal health problems. Even though simple and effective family planning techniques were available and within reach globally, only a fraction of women in LMICs could have access to these services.

Another critical challenge is making essential medicines affordable and accessible for needy patients. While many newer food supplements, medicines, and vaccines have been made available to the vast population, the control of substandard, spurious, falsely labelled, falsified, or counterfeit medical products has become a big challenge. LMICs must continually strengthen their food and drug regulatory agencies to ensure effective implementation, especially control of these counterfeit, substandard, spurious medicines. With the expansion of the private sector and health insurance, ensuring access to, and—in particular—affordability of essential medicines and vaccines become more vital.

Healthcare systems utilizing varieties of health professionals and allied health workers, hospitals and modern allopathic medicine intended to benefit the whole population barely existed a century ago in many LMICs. To date, major responsibility for healthcare provision in LMICs is through the publicly financed health facilities in conjunction with private healthcare providers. The development of infrastructure for public health, mainly for prevention and control of communicable disease, solely depended upon the support and expansion of vertical disease control campaigns such as vaccine-preventable diseases, HIV/AIDS, malaria, tuberculosis, or reproductive health, supported with international funds.

There are glaring contrasts between high-income countries and LMICs as well as among LMICs themselves in achieving a higher level of health status. The average life expectancy at birth in 2015 in low-income countries was around 62 years, and that of lower-middle-income countries was around 66 years, while the data for upper-middle and high-income countries ranged from 74 to 80 years (WHO 2017j). Many low-income countries still have an under-five mortality rate of above 120 per 1000 live births.

At the turn of the twenty-first century, while some countries still concentrated on vertical campaigns for expanding personal healthcare interventions like MCH, family planning, growth monitoring, breastfeeding, and immunization, as well as vertical disease control campaigns for malaria, HIV/AIDS, TB, etc., other LMICs tried to implement comprehensive health programmes including health legislative and health finance reforms. There was inadequate coordination and collaboration among specific health intervention initiatives, like immunization or specific disease campaigns or capital investment for hospital care on the one hand, and the development of basic health systems infrastructure on another. This dichotomy situation

resulted in the inadequate involvement of communities in their health development efforts and the slow pace of integrating vertical disease control campaigns into the basic health services. It was compounded by weak planning and management of health development, especially at the operational levels at districts and below, and the imbalance and inappropriateness of some human resources for health.

Some LMICs acknowledged that their health development programmes utilizing the conventional approach of merely expanding basic health services were inadequate. The public sector could not bear the cost of expanding basic healthcare services to the entire population in the face of existing resource constraints. Thus, many countries sought partnerships with private healthcare providers, worked with community-based civil society organizations, and deployed a large number of community-based health workers, who were volunteers and trained for short periods to serve as a third workforce of human resources for health (Sein 2006; WHO 2007). The increasing use of community-based volunteers in health development had proved successful for quickly expanding essential healthcare coverage in many countries. With the involvement of private healthcare providers and community-based health workers, many essential public health interventions were enhanced, especially in disease prevention and control, health promotion, maternal and child healthcare, information gathering and surveillance, treatment of minor ailments, and environmental health promotion.

Health policy, planning, and financing

The centrally directed health planning framework applying country health planning (CHP) or national planning process for health sector development, supported by WHO and UN agencies during 1970–1990, had moved many LMICs to a higher level of health attainment. It was highly successful when the health development programmes based on selective healthcare interventions, like immunization, malaria, leprosy, HIV/AIDS, or tuberculosis control, or selective maternal and childcare and family planning, growth monitoring, were promoted on the national scale through centrally controlled, externally funded programmes. In the late 1990s, the development partners moved their financing aid from planning multiple

vertical health development projects, to integrated sectoral planning of comprehensive health development programmes.

Around 2005, the UN introduced the principle of delivering as one (DaO), basically covering four areas—one leader, one budget, one programme, and one office, especially to integrate multiple UN-supported national programmes under one umbrella. The UN also worked with multiple development partners to establish the multi-donor trust fund (MDTF) mechanism by pooling multiple donor-supported funds into one integrated plan, and getting more partners to support a single national strategy in selected LMICs (UNDG 2010). A group of development partners, led by WHO and the World Bank that share a common interest in improving health services and health outcomes for LMICs, in 2007 launched another global initiative called the International Health Partnership plus related initiatives (IHP+). An interagency working group of IHP+ had developed a joint assessment tool for national health plans and strategies (JANs) and associated guidelines in 2009. Since 2010, the JANs tool was applied in several LMICs as part of the national health policy development and planning process (WHO 2015b). The national health planning processes now cover a comprehensive, balanced, and evidence-based assessment of the country's health and health system challenges. Many LMICs tried to strengthen their institutional capacity to harmonize and align donor programmes with their national policies, strategies, priorities, and plans.

LMICs, while struggling to improve their health systems, increasingly realized the value of financially protecting their own people in the fairest way possible. The key challenge of reform in health financing in LMICs, particularly the collection of revenue and pooling and use of financial resources, is to find a good balance between affordability and fairness to the people while providing appropriate incentives to healthcare providers. This is in keeping with the ideal for universal coverage, which means equity of access to key promotive, preventive, curative, and rehabilitative health interventions for all at an affordable cost (WHO IBRD/WB 2017).

In 2015, the world had spent US$ 7.3 trillion on health, close to 10% of the global gross domestic product (GDP). In fact, high-income countries with only 16% of the world's people account for 80% of global health spending. Conversely, 76% of the world's people living in middle-income countries accounted for less than 20% of global health spending, and the low-income countries with more than half a billion people accounted for less

than 1% of the world's health spending. In 2015, the global average health spending per capita was US$ 1,011, which conceals a very large difference between the highest and lowest spending countries, ranging from over US$ 9,000 to less than US$ 20 per capita. People spend more on health through out-of-pocket payment (OOP) in constant absolute terms for all income groups. Today, there are one billion people living in countries where out-of-pocket spending is 50% or more (WHO 2017k).

The way healthcare is financed varies considerably across countries. Middle-income and high-income countries tend to have a higher share of health spending funded from compulsory prepaid sources, such as government budgets (from various types of taxes) and/or social health insurance contributions. Many LMICs still rely heavily on external assistance to health expenditure, estimated at about 33% on average. Middle-income countries with high reliance on external resources are almost all exclusively small states. There is evidence of the fungibility of external funds, particularly in low-income countries. In addition, there is evidence of non-additionality of development assistance for health, where development assistance is spent in the health sector, but the recipient government re-allocates its own resources to fund other priorities (WHO 2017k). The proliferation of bilateral and multilateral donors, multiple global initiatives, and international non-governmental agencies (NGOs) also generates the potential to weaken health systems rather than strengthen them, allowing countries to avoid issues like long-term sustainability.

Strengthening health workforce for 2030

One key factor for success in public health and health system development is having a critical mass of competent health workforce. As human resources consume as much as 60–70% of the public health expenditure, they need to be fully developed and optimally utilized. The World Health Report 2006 had provided a global analysis and identified effective strategies to strengthen the development of appropriate human resources for health (WHO 2006b). There is an anticipated mismatch between demand, need, and supply. In addition, there is an acceleration of domestic and international migration of all types of the health workforce. Furthermore, the challenge of deploying and retaining health professionals in rural, remote, and underserved areas is truly global. The major reasons for the health workforce

shortage, while varying greatly among LMICs, include inadequate numbers of human resources due to under-investment in basic undergraduate and graduate education and training; loss of existing workforce due to emigration to other countries; mismanagement and coordination; and inefficient planning. Newer models of deployment, health workers' education, and management are in need (WHO 2016d).

Some of the impediments to the development of a relevant and adequately sized public health workforce are lack of incentives and career structure for public health professionals, public health training being seen as an extension of the biomedical model of clinical training, and the inclusion of many training courses and core curriculum which are either obsolete or irrelevant to the public health needs of the countries. Several networks, both formal and informal, among public health education institutions in Asia and Africa have been established to collaborate and support exchange programmes for faculty members and students, sharing of curriculum and training materials, and organizing joint research studies.

Health research and development

While science and technology research and development and their application have advanced rapidly in high-income countries, these disciplines have remained rudimentary in LMICs. Many LMICs must rely on the results of research and development on pharmaceuticals and vaccines from high-income countries, and in fact, many breakthroughs have been contributed by the experiences gained and the studies carried out in LMIC. In many LMICs, there are very few health research and development institutions as well as poor career prospects for researchers.

Joint collaborative studies between LMICs, especially upper-middle-income countries, and the high-income countries, are carried out for the development and use of medicines and vaccines for infectious diseases, promotion of human health and reproduction specifically on contraception and other fertility control measures, strengthening of health systems, promotion and protection of environmental health, and control of NCDs, and development of essential healthcare technologies. Considerable progress has been made in capability strengthening for promoting health systems research, which has further paved the way for developing effective national health research systems.

In recent years, the development of health research has been promoted through international exchanges of experience and expertise among the networks such as INCLEN Trust, Asia-Pacific Health Research System Network and African Health Research System Network. A major issue faced by LMICs today is that while scientific advances and technological development of the high-income countries provide them with vast opportunities for their overall economic development, many people in LMICs are still suffering from preventable NCDs, neglected tropical diseases, and under-nutrition.

Strive for social determinants of health

Many LMICs are particularly susceptible to the negative impact of socially determined health inequities due to heavy exposure to health risks, social exclusion, and other broader determinants associated with health risks, including low-income, unsafe living conditions, poor work environment, unsafe sex, and changing consumption patterns of food, water, and information. LMICs considered the need to build up their national capacities to address advocacy, intersectoral action, evidence gathering of health inequity, and social determinants across sectors.

Reducing tobacco and alcohol use

Tobacco is the most common risk factor for many NCDs. In 2015, more than 1.1 billion people smoked tobacco, with far more males than females currently engaging in this behaviour. Tobacco-related illnesses and deaths had increased globally during the past decades with seven million deaths every year, and if no proper control measures are instituted, the death toll may go up to 10 million in the next decade. In LMICs, the increasing use of betel-quid (betel leave combining betel-nut, slake-lime, and tobacco) has become a major issue, since it causes oral, laryngeal, and gastrointestinal cancer, and cardiovascular diseases. After adoption in May 2003 of the WHO Framework Convention on Tobacco Control (WHO FCTC), LMICs have either adopted or amended their tobacco control legislation, conforming to the provisions of the WHO FCTC. By 2017, 180 parties to WHO FCTC had ratified the Convention, and adopted tobacco control measures called MPOWER (M for monitoring tobacco use, P for protecting people from

tobacco smoke, O for offer help to quit, W for warning about dangers of tobacco, E for enforcing ban on tobacco advertising, promotion and sponsorship, and R for raising tax for tobacco), the six best-buy strategies promoted by WHO and adopted by all parties to WHO FCTC. Around 4.77 billion people are covered by at least one best practice policy intervention at the national level, a substantial increase from nearly two billion people (WHO 2017l). There is also an urgent need to make effective control for e-cigarettes, water pipes, and heated tobacco products, as they became newer forms of tobacco marketed to LMICs. It is critical that all countries act urgently to protect their people more effectively with evidence-based tobacco control policies at the most comprehensive level.

Worldwide consumption of alcohol has increased from 6 litres of pure alcohol consumed per person aged 15 or older, in 2008, as compared to 6.4 litres of pure alcohol per person aged 15 years or older, with considerable variation between WHO regions, in 2016. A large portion of this—28.6% or 1.76 litres per person—was homemade, illegally produced or sold outside normal government controls (WHO 2017j). Harmful use of alcohol is a major factor contributing to premature deaths and avoidable disease worldwide. A wide range of policy options and interventions for reducing public health problems caused by the harmful use of alcohol was advocated for all countries.

Preventing injuries

Injuries, caused intentionally and non-intentionally, have emerged as a major public health problem in the twenty-first century. Globally, road traffic injuries claim more than 1.2 million lives each year and greatly impact health and development. The rapidly rising number of motor vehicles (48% of the world's vehicles are registered in LMICs), and motorcycles (adding more than a million a day on the roads) has led to an equally rapid increase in traffic-related injuries and deaths. LMICs have a higher road-traffic-related death rate of over 20 per 100,000 population, and accounted for over 90% of the world's fatalities.

They are the leading cause of death among young people aged between 15 and 29 years, and cost governments approximately 3–5% of GDP. The rise in global road traffic deaths has been largely driven by the escalating death toll on roads in LMICs—particularly in emerging economies where urbanization

and motorization accompany rapid economic growth. In many LMICs, necessary infrastructural development, policy changes, and levels of enforcement have not kept pace with increasing vehicle use. Introduction and upgrading of safe public transport, improvement of road signs and pedestrian crossings, compliance with traffic legislation like wearing helmets and seatbelts, and avoiding speeding and drink-driving, are important measures for reducing road traffic injuries (WHO 2015). In addition to deaths on the roads, up to 50 million people incur non-fatal injuries each year as a result of road traffic crashes, while there are additional indirect health consequences that are associated with this growing epidemic. The UN SDG3 target aims to halve the number of global deaths and injuries from road traffic crashes by 2030, while SDG11 relates to providing access to sustainable transport systems for all, improving road safety, and expanding public transport (WHO 2016d).

Nutrition promotion

Despite the significant steps taken towards promoting nutrition and reducing associated health burdens over recent decades, the global community is grappling with multiple burdens of malnutrition. Global Nutrition Report (2017) showed that 88% of countries face a serious burden of either two or three forms of malnutrition (i.e. childhood stunting, anaemia in women of reproductive age, and/or overweight in adult women). It also indicated that progress on global nutrition targets is slow or moving backwards. Around 2 billion people lack essential micronutrients like iron and vitamin A in their daily diet, 155 million children have stunted growth, 52 million children are wasted, while 41 million children are obese. 88% of 193 UN member countries have faced a serious burden of either two or three forms of malnutrition (Development Initiatives 2017). There is an opportunity to identify—and to take—double-duty actions at once which tackle more than one form of malnutrition. These will increase the effectiveness and efficiency of the investment of time, energy, and resources to improve nutrition. Likewise, triple duty actions which tackle malnutrition and other development challenges could yield multiple benefits across the UN SDGs.

Social medicine and public health

The initial idea of social medicine or the social dimensions of public health had emerged in the early twentieth century. The concept of social medicine was developed as a new public health discipline in the late 1940s. At that time, the leaders of both clinical medicine and public health questioned the perceived polarization of curative medicine and preventive medicine. The situation was exacerbated by continued specialization and the lack of cross-disciplinary interaction (Ko Ko 2011).

The social and behavioural aspects that influenced illness and well-being are recognized again, and revisited during the dawn of the twenty-first century. Many social and personal behavioural interventions are introduced as part of health promotion and disease prevention. Many universities and medical schools converted their departments, faculties, or schools of hygiene and public health into those of preventive and social medicine from the 1960s. Instead of integrating curative and preventive medicine in the teaching of social medicine, educational institutions equated public health and hygiene with preventive and social medicine, and conventional public health teaching still has a primary focus on the prevention and control of infectious diseases and hygiene. The general population is now becoming increasingly aware of the social and economic determinants impacting health. As a result, national debates at both public and political levels on health linkages with social, environmental, economic, and political factors are needed. There are more discussions at the global level to give a wider political and social dimension to international public health.

The history of public health development has shown the need for increased cooperation between high-income and LMICs to promote the total health and well-being of all world citizens. Over the past 50 years, many international, intergovernmental, bilateral, and multilateral development agencies and organizations, as well as professional and technical alliances, have emerged out of necessity to deal with international health developmental issues both in policy and programme terms. Partnerships could be established to sustain and support health system development so that equitable health outcomes are achieved and the needs of the people are met. More recently, debates have intensified around global public health development and the global governance of health (Kickbusch and Gleicher 2012).

Whither to new public health

Now, entering the twenty-first century, many LMICs move towards another era of public health development with new thoughts from a narrow view of vertical disease control interventions to a wider perspective of multisectoral interventions beyond the health sector. Besides the usual public health measures undertaken by the governments, an increasing number of NGOs and the private sector agencies, both at the local and global level, were involving in public health development. In addition to external financing and technical support by traditional international agencies like WHO and UN, bilateral aid agencies from high-income countries and multilateral financial institutions like the IBRD, the World Bank, and Regional Development Banks, there are a myriad of multinational foundations, international funds, and financing groups like the GAVI Alliance and the Global Fund, as well as private international foundations like Bill and Melinda Gates Foundation or Bloomberg Philanthropies, etc., that have transformed international health support and governance on national and international health cooperation.

Globalization of public health

The UN had launched the Joint UN Program on HIV/AIDS (UNAIDS) in January 1996, with the UN-wide mission to lead, strengthen and support an expanded response to HIV/AIDS. Another UN initiative was the adoption of UN Millennium Development Goals (UN MDG), with targets for 2015, representing the world's commitment to deal with global poverty and health in several dimensions (UN 2011). After reviewing the progress of UN MDG, the UN Summit in 2015 adopted a development agenda titled—Transforming our world: the 2030 Agenda for Sustainable Development, which contained the UN SDG to be achieved by 2030. The SDG positions health as a key feature of human development in a more integrated manner than was the case for the MDG. Certain important health issues that were not addressed in the MDG now have targets in the SDG, such as mental health, substance abuse, health impact from hazardous chemicals, water, and soil pollution and contamination and the effective implementation of the WHO FCTC (UN 2018c).

By the start of the twenty-first century, several series of analyses of the impact of rapid globalization on public health are available in many health and development literature. They all revealed that the global trade, finance, information and people flows, governance, and security could contribute to

improving the health of people everywhere. The threats to and opportunities in public health, including cross-country control of infectious diseases and international governance, have been analysed extensively within the context of globalization (Fidler 2004). An updated version of IHR is a guide for all countries to prevent, protect against and control, and provide a public health response to the international spread of a disease or diseases, in various ways that are commensurate with and restricted to public health risks, and that avoid unnecessary interference with international traffic and trade. The UN SDG outline the need to support countries to detect and manage health risks, build resilient health systems, and focus on protecting the most vulnerable populations.

Addressing healthy ageing

Many LMICs have experienced an expansion of the proportion of people above 60 years of age, from 1.2 billion in 2015 to 2 billion in 2050. They have started paying attention to dealing with older people's health by adopting appropriate comprehensive national policies and related legislation. They have focused on creating closer linkages between social welfare and health, providing care at home and in the community, promoting traditional family ties, optimizing the use of existing healthcare delivery systems and setting up elderly care homes. Access to essential healthcare for older people is still inadequate in many LMICs, since knowledge and skill among health and social workers for specific needs of older people is minimal. The economic, social, and health status of the fast-growing elderly population in LMICs poses a great challenge to all sectors.

Looking ahead

LMICs have undertaken a series of health reforms by improving the content of essential health packages and introducing a new way of financing. Many health financing models either at national or local levels were developed, including reforms in expanding social health insurance. Another global trend in health development is the increasing role of the private sector, both for-profit and non-profit. It is not a simple solution that can fit within a country's existing socioeconomic, political, and health situation. The extent of the

national health plans ensures that wider reforms would address the gaps in healthcare and create a pro-poor health system.

A multiplicity of new actors in national and international health arena had redefined the boundaries of the health sector. Groups of individuals, united to a particular cause, such as patient groups or civil societies both at national and international levels, have become major players for policy and programme development in public health by creating powerful lobbies and raising public awareness. A new paradigm in health development, in countries that must rely heavily on external aid for health in earlier years and even now, has led to a complex relationship among traditional and new players in international health and health-related development, for effective planning and use for resources, and the need for delineation and harmonization of responsibilities.

Conclusion

The history and development of public health in LMICs shows that important breakthroughs in public health interventions have led to great improvements in economic development. Health determines economic productivity and prosperity, and the physical and emotional well-being of the people. The mindset of the population afflicted with a high infant mortality rate usually lacks the secure knowledge of its children's longevity, witnesses higher fertility rates, and experiences the quality-quantity trade-off in child-rearing. Ever since the health-for-all movement was initiated over three decades ago, health, equity, and social justice remain the main themes of social and health policy. It is essential for all public health professionals to sustain these values, especially those in LMICs and the international community.

REFERENCES

Development Initiatives (2017). *Global Nutrition Report 2017: Nourishing the SDGs*. Bristol, UK: Development Initiatives.

Diawara, L., Traoré, M.O., Badji, A., et al. (2009). Feasibility of onchocerciasis elimination with ivermectin treatment in endemic foci in Africa: first evidence from studies in Mali and Senegal. *PLoS Neglected Tropical Diseases*, 3(7), e497.

Fidler, D. (2004). *SARS, Governance, and the Globalization of Diseases*. London, UK: Palgrave Macmillan.

Kickbusch, I. and Gleicher, D. (2012). Governance for health in the 21st century. Geneva, Switzerland: WHO EURO. Available at: http://www.euro.who.int/en/publications/abstracts/governance-for-health-in-the-21st-century

Ko Ko, U. (2011). *Wither To—Preventive and Social Medicine?* Paper presented at the 57th Myanmar Medical Conference, 22 January 2011. Yangon, Myanmar: Myanmar Medical Association.

Molyneux, D. and Sankara, D.P. (2017). Guinea worm eradication: progress and challenges—should we beware of the dog? *PLoS Neglected Tropical Diseases*, 11(4), e0005495.

Sein, T. (2006). Health volunteers: third workforce for health-for-all movement. *Regional Health Forum*, 10(1), 38–48.

The Telegraph (2017). Bill Gates: terrorists could wipe out 30 million people by weaponising a disease such as smallpox. *The Telegraph, Science*, 19 April 2017. Available at: https://www.telegraph.co.uk/science/2017/04/19/bill-gates-terrorists-could-wipe-30-million-people-weaponising/

United Nations (2011). *Millennium Development Goals Report 2011*. New York: UN.

United Nations (2018a). *World Economic Situation and Prospects 2018*, Table E, 144. New York: UN.

United Nations (2018b). *SDG and Climate Change, Climate Change Affects Everyone*. New York: UN. Available at: http://www.un.org/sustainabledevelopment/climatechange/

United Nations (2018c). *Sustainable Development Agenda, Sustainable Development Goals*. New York: UN. Available at: http://www.un.org/sustainabledevelopment/development-agenda/#22a1d196b9763 ced3

United Nations Development Group (2010). *UNDG Guidance Note on Establishing Multi-Donor Trust Funds*. New York, NY: UN.

World Health Organization (2006a). WHO position paper on haemophilus influenzae type b (Hib) conjugate vaccines. *Weekly Epidemiology Record*, 81(47), 445–52.

World Health Organization (2006b). *World Health Report 2006: Health Workforce*. Geneva, Switzerland: WHO.

World Health Organization (2007). *Community Health Workers: What Do We Know About Them?* Geneva, Switzerland: WHO. Available at: http://www.who.int/hrh/documents/community_health_workers.pdf

World Health Organization (2009). Human papilloma virus vaccines: WHO position paper, *Weekly Epidemiological Records*, 84(15), 118–31.

World Health Organization (2011a). *Smallpox Eradication: Destruction of Variola Virus Stocks* (WHA A64/17), Geneva, Switzerland: WHO.

World Health Organization (2011b). Dracunculiasis eradication: global surveillance summary, 2011. *Weekly Epidemiology Records*, 86(20), 189–204, May 2011, Geneva.

World Health Organization (2012). *Poliomyelitis: Intensification of the Global Polio Eradication Initiative* (WHA Resolution A65/20). Geneva, Switzerland: WHO.

World Health Organization (2015a). *Global Status Report on Road Safety 2015*, Geneva

World Health Organization (2015b). *Joint Assessment of National Health Strategies and Plans. Combined Joint Assessment Tool and Guidelines Draft, Version 2, 2014*. Geneva, Switzerland: WHO. Available at: http://www.who.int/workforcealliance/knowledge/toolkit/24/en/

World Health Organization (2016a). *WHO Advisory Committee on Variola Virus Research: Report of the Eighteenth Meeting*. November 2016. Geneva, Switzerland: WHO.

World Health Organization (2016b). Plague around the world, 2010–2015. *Weekly Epidemiological Records*, 91(8), 89–104.

World Health Organization (2016c). *Ambient Air Quality and Health, Fact Sheet*, September, 2016. Geneva, Switzerland: WHO. Available at: http://www.who.int/mediacentre/factsheets/fs313/en/

World Health Organization (2016d). *Global Strategy on Human Resources for Health: Workforce 2030*. Geneva, Switzerland: WHO.

World Health Organization (2016e). *World Health Statistics 2016: Monitoring Health for the SDG*. Geneva, Switzerland: WHO.
World Health Organization (2017a). Cholera, 2016, *Weekly Epidemiological Records*, 92(38):521–36
World Health Organization (2017b). *Ending Cholera: A Global Roadmap to 2030, Global Taskforce on Cholera Control*. Geneva, Switzerland: WHO.
World Health Organization (2017c). Eliminate yellow fever epidemics: a global strategy 2017–2026. *Weekly Epidemiological Records*, **92** (16), 193–204.
World Health Organization (2017d). Global routine vaccination coverage, 2016. *Weekly Epidemiological Records*, 92(46), 701–16.
World Health Organization (2017e). Progress towards regional measles elimination-worldwide, 2010–2016. *Weekly Epidemiological Records*, 92(43), 649–60.
World Health Organization (2017f). *Hepatitis B Fact Sheet*, July 2017. Geneva, Switzerland: WHO. Available at: https://www.who.int/news-room/fact-sheets/detail/hepatitis-b
World Health Organization (2017g). Global leprosy update, 2016: accelerating reduction of disease burden. *Weekly Epidemiology Records*, 92(35), 501–19.
World Health Organization (2017h). *Trachoma, Fact Sheet*, July 2017, WHO, Geneva. Available at: http://www.who.int/mediacentre/factsheets/fs382/en/
World Health Organization (2017i). The 25th meeting of the International Task Force for Disease Eradication, held at the Carter Center. Atlanta, GA, USA, on 8 November 2016. *Weekly Epidemiological Records*, 93(9/10).
World Health Organization (2017j). *World Health Statistics 2017: monitoring health for the SDGs, Sustainable Development Goals*. Geneva, Switzerland: WHO. Licence: CC BY-NC-SA 3.0 IGO.
World Health Organization (2017k). *New Perspectives on Global Health Spending for Universal Health Coverage*. Geneva, Switzerland: WHO.
World Health Organization (2017l). *WHO Report on the Global Tobacco Epidemic, 2017: Monitoring Tobacco Use and Prevention Policies*. Geneva, Switzerland: WHO.
World Health Organization (2018). Monthly report on dracunculiasis cases, January–November 2016. *Weekly Epidemiological Records*, 93(3), 31.
World Health Organization and United Nations Children's Fund (2017). *Progress on Drinking Water, Sanitation and Hygiene: 2017 Update and SDG Baselines*. Geneva, Switzerland: WHO and UNICEF.
World Health Organization, International Bank for Reconstruction and Development/The World Bank (2017). *Tracking Universal Health Coverage: 2017 Global Monitoring Report*. Licence: CC BY-NC-SA 3.0 IGO. Geneva, Switzerland: WHO and The World Bank.

1.4

Public health priorities in countries undergoing economic transition

The middle-income countries

Stephen Tollman and Jessica Price

Introduction

Sustainable Development Goal (SDG) 3 highlights the range of health challenges which continue to reduce quality of life, extending the horizon of global health well beyond maternal and child health, HIV, and malaria—which dominated the Millennium Development Goals—to include non-communicable diseases, mental health, injuries, and exposure-related illnesses. SDG 3 further addresses the goal of universal health coverage (UHC) (World Health Organization n.d., b). However, if SDG 3 targets are to be met, the complex health challenges accompanying social and economic transitions in middle-income countries (MICs) must be confronted.

MICs are diverse in geography, population size, politics, and economic activity. The World Bank defines lower MICs as countries with a gross national income (GNI) per capita of $1,006 to $3,955 (currently 47 countries), and upper middle-income economies as those with a GNI per capita between $3,956 and $12,235 as of 2018 (60 countries) (The World Bank in Middle Income Countries 2018).

MICs are of special interest to public health researchers, practitioners, and policy makers first, because of the sheer number of people living in them—5 of the world's 7 billion; second, because despite rapid economic growth, many MICs remain highly unequal—reflected by a high GINI index and widespread persistent poverty; and third, because prolonged economic and social transitions drive a series of health transitions. These in turn put a premium on adaptive health systems able to respond to changing health needs (The World Bank in Middle Income Countries 2018).

Economic transition refers to structural, organizational, and institutional changes in national economies that affect their growth paths, policy context, and private sector evolution. Changes to the job market include growing numbers of skilled jobs and a rise in women's participation in the formal economy (with a move from unpaid home and care work to paid formal employment), increased labour-related migration, the establishment of new microeconomies, and rising expectations of standards of living, government services and social protection (International Monetary Fund 2000). Economic transitions are broadly separated into those required to boost countries from low- to middle-income (industrialization, greater privatization, movement to market-based economies); and transition from middle-income to high-income (characterized by a move from industrial to knowledge-based economies). Countries which transition successfully to middle-income, but fail to transition to high-income nations are said to be caught in a 'middle-income trap'. This describes the slowing of economic growth in MICs that are unable to develop new growth models and institutions required to drive knowledge-based economies (Gill and Kharas 2007). The 'middle-income-trap' is an important reminder that despite being described as linear processes, economic transitions are typically prolonged, heterogenous and non-linear with some countries even regressing to a lower income status—as in Argentina which was classified as high-income in 2014 and 2018, but as upper middle-income 2015–2017 and again in 2019 (World Bank 2019).

The upstream changes associated with economic transition lead to far-reaching behavioural transitions which feed through to health, risk and disease profiles—outcomes directly influenced by greater urbanization, changes in diet and physical activity, as well as educational and infrastructure development. Economic and social transitions have important downstream health system effects: management of diseases improve as numbers of health professionals increase with the upskilling of the population, and increased spending on healthcare as gross domestic product (GDP) grows. Positive effects on the social determinants of health include better female education, improvements in sewage and hygiene, and introduction of social protection policies—accompanied by declining maternal and child mortality, and infectious disease. However, offsetting impacts on health arise from overcrowding, air pollution, physical inactivity, alcohol and tobacco consumption, and high fat, sugar and salt diets often linked with hyperaccessibility to 'fast foods' (Miranda et al. 2019). This tends to manifest in growing

burdens of non-communicable diseases: diabetes, cardiovascular and respiratory diseases, and cancers. These changes in disease burden have been termed health transitions (Beaglehole and Bonita 2004).

In this chapter we start by reviewing the relationship between the economy and health, looking at changes in health outcome as income increases, demographic transitions related to economic growth, and the impact of urbanization on health outcomes. We then review major health challenges facing MICs, including the triple burden of disease, multimorbidity, mental health across the life course and the impacts of migration, climate change, and antimicrobial resistance. Finally, we discuss health system challenges and the importance of building resilient and 'learning' health systems with the agility to adapt to meet MICs' evolving health needs. We discuss strategies to achieve universal health coverage in MICs and the importance of intersectoral collaboration in promoting health. The key messages are outlined in Box 1.4.1.

Box 1.4.1 Key messages

1. MICs are countries undergoing social and economic transitions, which can be prolonged and non-linear.
2. Economic and social transitions have profound effects on the health burdens and related health behaviours in MICs. This is compounded by greater vulnerability to health impacts from climate change, migration, and antimicrobial resistance.
3. MICs are best served by resilient and 'learning' health systems, able to respond with agility to the changing health needs of the population.
4. Policies designed to enable universal health coverage and intersectoral collaboration can position MICs to meet the complex health needs of their populations.

Health and the economy

Income and health

Growth in per capita GDP is associated with improved health outcomes at a national level. This association was described by Samuel Preston in 1975, who showed that life expectancy at birth was higher in wealthier nations, but that the relationship flattens at higher incomes (Preston 1975). As shown in Figure 1.4.1 using 2012 data, in countries with a GDP per capita of up to ~$4,000 small increases in gross domestic product (GPD) are associated with large increases in life expectancy. At higher levels of GDP per capita the association weakens (Biciunaite 2014). Rising GDP per capita is also associated, in aggregate, with declining neonatal, infant and under-5 mortality (Pritchett and Summers 1993). As income rises and under-5 mortality falls, neonatal deaths typically account for a greater proportion of all under-5 deaths.

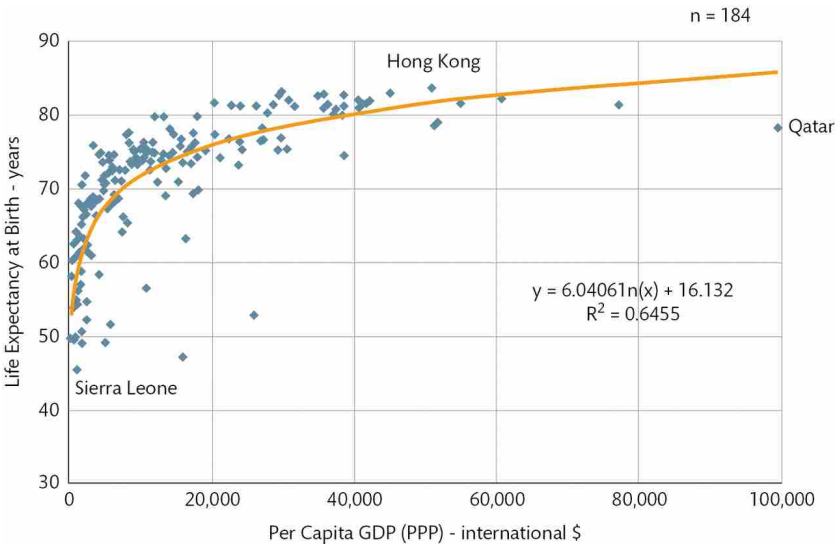


Fig. 1.4.1 Preston curve 2012: change in life expectancy at birth by per capita GDP (from Euromonitor International). Note: GDP based on purchasing power parity (PPP) and takes account of differences in pricing between countries.
Reprinted from Biciunaite A (2015) 'Economic Growth and Life Expectancy—Do Wealthier Countries Live Longer?' Euromonitor International
©Euromonitor.

While the relationship between income and health outcomes holds across countries, national averages can mask significant differences in outcomes between the wealthy and the poor within the same country. This is especially problematic in MICs, many of which remain highly unequal. A recent review found that in India, a child in the poorest socioeconomic quintile is over 2.5 times more likely to die as one from the wealthiest quintile, with a difference in the under-5 mortality rate of over 30/1,000 live births. The ratio of under-5 mortality in the poorest quintile compared to the wealthiest in Brazil exceeded 2.5, and is above in South Africa (Chao et al. 2018). Table 1.4.1 illustrates differences in infant and child mortality by socioeconomic quintile for selected MICs where Demographic and Health Survey (DHS) data is available (World Health Organization, n.d., a; Hosseinpoor et al. 2016).

Table 1.4.1 Infant and under-5 mortality by socioeconomic quintile for selected MICs

Country	Year	Infant mortality (per 1,000 live births)					Under-5 mortality (per 1,000 live births)					GINI index (year)
		Quintiles					Quintiles					
		1	2	3	4	5	1	2	3	4	5	
South Africa	1998	61.4	51.2	35.7	34.3	16.7	87.5	70.4	48.7	39.7	21.1	63.0 (2014)
Namibia	2013	50.9	45.0	41.7	37.6	22.6	67.1	68.4	64.9	56.2	30.1	59.1 (2015)
Zambia	2013	54.7	49.0	46.6	47.8	39.0	99.1	82.8	78.4	71.9	57.7	57.1 (2015)
Lesotho	2014	62.1	62.4	73.5	93.8	50.6	76.3	87.0	104.4	120.6	69.2	54.2 (2010)
Brazil	1996	83.6	46.2	31.2	23.9	27.8	99.0	55.3	37.7	26.1	32.9	53.3 (2017)
Colombia	2015	21.9	21.2	10.3	13.9	5.4	27.3	23.6	11.6	16.6	6.9	49.7 (2017)
Cameroon	2011	90.1	71.2	66.2	58.3	51.3	183.9	143.2	120.3	88.6	72.4	46.6 (2016)
Bolivia	2008	88.7	64.4	48.1	37.1	26.3	115.7	85.0	66.4	44.7	31.4	44.0 (2017)
Ghana	2014	54.9	44.6	39.4	47.8	51.1	90.9	73.0	60.7	55.0	64.0	43.5 (2016)
Nigeria	2013	91.1	93.9	70.5	64.3	47.7	187.5	186.9	126.4	100	72.2	43.0 (2009)
Angola	2015	62.2	63.7	47.0	42.4	22.8	101.1	102.4	72.3	65.2	36.3	42.7 (2008)
Turkey	2003	62.2	43.6	30.9	24.9	17.0	80.0	50.5	37.3	29.7	22.4	41.9 (2016)
Indonesia	2012	52.0	34.1	33.0	27.8	16.9	69.7	42.2	39.0	34.5	21.8	38.1 (2017)
Myanmar	2015	77.7	75.0	52.4	35.0	21.3	97.9	88.6	66.7	42.7	25.0	38.1 (2015)
India	2005	82.0	73.4	66.0	51.2	34.2	116.5	97.5	81.3	60.8	39.2	35.7 (2011)
Vietnam	2002	39.5	28.0	19.2	15.4	13.9	53.0	38.1	22.8	21.8	15.3	35.3 (2016)
Jordan	2012	25.4	14.7	14.9	19.7	8.3	27.6	18.4	19.9	20.7	11.2	33.7 (2010)
Pakistan	2012	90.0	97.2	84.4	74.4	43.9	117.6	115.3	97.3	83.6	48.5	33.5 (2015)
Bangladesh	2014	51.8	50.0	42.7	43.2	28.6	61.6	60.3	55.2	51.1	37.1	32.4 (2016)

* GINI index indicates inequality of income with zero expressing perfect equality and 100% expressing maximal inequality.

Source: data from WHO Health Equity Monitor (*Global Health Observatory data repository: Health Equity Monitor Database*. World Health Organization. Available at: http://apps.who.int/gho/data/node.main.nHE-1540?lang=en)

Income and demography

In essence, demographic transition can be understood as a change in population dynamics from high mortality and birth rates (fertility), to low mortality and fertility as socioeconomic development (notably female education) occurs. The demographic transition results in changes to population age structure, with implications for labour capacity, child and older adult dependence, and disease and disability burdens (Coale 1989).

As a group, MICs span multiple stages of demographic transition (Table 1.4.2): those at an earlier stage of transition experience a growing proportion of young people in the population; while at later stages, fertility declines and population size stabilizes. Concurrent declines in mortality and fertility can produce a 'demographic dividend' with greater productivity of adults in the labour force resulting from fewer child dependents (so a decline in the dependency ratio) (United Nations Population Fund, n.d.). A greater proportion of economic resources can therefore be invested to speed development. However, the benefits of a demographic dividend are not guaranteed: in addition to education and health systems to support more productive working populations, the demographic dividend requires a buoyant economy that can withstand shocks. In contrast, in countries with high youth unemployment, economic growth remains slow even as the dependency ratio falls (Table 1.4.2). In addition, where economic benefit is realized, this can be short-lived: as the population ages greater care burdens are place on the state and on working adults to care for older generations. Growth in the population 65 years and older results in a rise in the old-age dependency ratio and is generally associated with a slowing of per capita income growth (United Nations Population Fund n.d.).

Although characterized as a linear transition with sequential stages, the reality of demographic transitions in MICs is generally more complex. In southern and east Africa, the HIV epidemic resulted in rising mortality in the 1990s and early 2000s despite a continued decline in fertility. AIDS-related mortality mostly affected adults age 30–44 (i.e. the working generation) resulting in significant burdens placed on older persons to take over the roles of primary caregiver and breadwinner (with households increasingly dependent on old-age pensions as sources of income) (Hosegood et al. 2007). Although demographic transitions tend to accompany economic transition, it is possible for demographic transitions to continue despite economic stagnation.

Table 1.4.2 provides trend data for selected MICs from 1997 and 2017, demonstrating the heterogeneity of demographic transitions across countries.

Table 1.4.2 Population demographics for nine middle-income countries

Country	Life expectancy at birth		Under-5 mortality (per 1000 live births)		Fertility rate (average births per woman)		Population 0–14 years (% total population)		Population 15–65 years (% total population)		Population over 65 years (% total population)		Dependency ratio		GDP growth (annual %)		Unemplo (% of tot labour fo
	1997	2017	1997	2017	1997	2017	1997	2017	1997	2017	1997	2017	1997	2017	1997	2017	1997
Brazil	69.06	75.46	42.2	14.8	2.48	1.74	31.67	21.68	63.38	69.71	4.95	8.61	57.78	43.45	3.40	1.06	8.16
Russian Federation	66.7	72.12	21.2	7.6	1.22	1.76	20.13	17.63	67.51	68.1	12.36	14.26	48.13	46.83	1.40	1.63	11.81
India	61.23	69.17	102.4	38.7	3.51	2.24	35.97	27.48	59.86	66.54	4.18	5.98	67.07	50.29	4.05	7.17	2.14[a]
China	70.43	76.47	43.5	9.2	1.61	1.68	26	17.94	67.57	71.72	6.43	10.35	47.99	39.45	9.23	6.76	3.90
South Africa	59.49	63.54	65.7	35.3	2.92	2.43	36.06	29.15	59.54	65.63	4.41	5.22	67.97	52.37	2.60	1.42	25.00[c]
Mexico	73.27	74.95	31.2	13.4	2.88	2.16	35.62	26.94	59.49	66.02	4.9	7.04	68.11	51.47	6.85	2.07	4.06
Indonesia	64.95	71.28	60.3	25.9	2.58	2.34	32.5	26.92	63.14	67.4	4.37	5.68	58.39	48.37	4.70	5.07	4.68
Nigeria	45.92	53.95	199.6	122.1	6.19	5.46	44.05	44.01	53.08	53.24	2.87	2.75	88.39	87.83	2.94	0.81	3.94[d]
Turkey	68.19	77.16	47.4	11.2	2.66	2.08	32.02	25	62.41	66.74	5.57	8.26	60.23	49.84	7.58	7.44	6.84

[a] 1994 data;

[b] 2018 data;

[c] 1998 data;

[d] 1986 data.

[*] Unemployment refers to share of the labour force without work but seeking employment. Definitions of labour force and unemployment differ by country.

Data taken from the World Bank, accessible at www.data.worldbank.org

Urbanization and health

Urbanization is arguably the most ubiquitous and powerful social transition with mixed effects on personal and population health. On one hand, given the concentration of services, urbanization can support marked health gains through improved access to education, healthcare, and employment. However, when urbanization occurs in a rapid, unplanned, and uncontrolled manner—as seen in the creation of urban slums—urbanization itself can become a risk factor for poor health outcomes (Patil 2014).

Urban slums are overcrowded with a lack of proper hygiene and sanitation systems, informal housing, and unreliable water and energy supply. While some slums have a quasi-permanent character, others have high turnover of residents (Ezeh et al. 2017). As a result of poor living conditions, slum-dwellers are at risk of infectious diseases such as TB, acute respiratory infections, diarrhoea, and vector-borne diseases including malaria (Elsey et al. 2019). Conditions of poverty including a lack of food put children of the urban poor at increased risk of undernutrition and stunting (Patil 2014).

Behaviour changes influenced by urbanization include dietary preferences, tobacco, and alcohol consumption and physical activity—contribute at population level to a rising incidence of non-communicable disease and risk including obesity, hypertension, diabetes, and cardiovascular disease. Urbanization has also been shown to negatively affect mental health—with increasing stress and social isolation, and higher levels of violence and injury (Patil 2014). Changes in household structure—tending to more households of smaller size occur in rural and urban areas (Wittenberg et al. 2017) and exacerbate the loss of social support (Patil 2014).

Increasing pace of migration—whether unidirectional or circular—of sizeable numbers to cities increases demands on transport and leads to more (often less roadworthy) vehicles. Furthermore, vehicles in low- and middle-income countries are typically more polluting than those in high-income countries (HICs). Combined with rapid industrialization, this contributes to worsening air pollution[1] which in turn may exacerbate chronic respiratory conditions and contribute to cardiovascular events and cancers. In 2018, air pollution was assessed the fifth leading risk factor for death globally, and the third and fourth leading risk factor for mortality and disability-adjusted life years (DALYs) combined in India and China respectively (Institute for Health Metrics and Evaluation (IHME) 2018; Institute for Health Metrics and Evaluation (IHME) 2018b; Health Effects Institute 2019).

Finally, urbanization is associated with localized rises in temperature relative to surrounding areas—known as the 'Urban Heat Island' effect. This results from increased heat production, reduced surface evaporation, changing thermal characteristics and rising air pollution (Grover and Singh 2015). The negative impacts of rising temperatures, including heat stroke, cardiovascular and respiratory mortality, are discussed in the following section on climate change and health.

[1] Air pollution is measured using particulate size >25 µm (PM2.5). Over the past 30 years, PM2.5 concentrations have increased by ~25% across South Asian cities. Furthermore, growth of urban populations means more people exposed to harmful levels of air pollution (Miranda et al. 2019).

Health challenges in MICs

Triple burden of disease

The triple burden of disease describes the concurrent, often interacting health burdens of (1) communicable, maternal, neonatal, and nutritional disease (CMNN); (2) a rising burden of non-communicable disease; and (3) substantial morbidity and mortality from injuries.

As seen in Figure 1.4.2, non-communicable diseases (NCDs) are now the leading cause of death and account for the largest proportion of DALYs globally and in MICs (Institute for Health Metrics and Evaluation (IHME) 2018c). This reflects changes in the social determinants, risk factors, and behaviours associated with the economic and social transitions occurring in MICs, as well as the impact of demographic transitions—with growing proportions of the world population over 65 years (Table 1.4.2). According to figures from the Global Burden of Disease study, stroke, ischaemic heart disease, chronic obstructive pulmonary disease (COPD), and lung cancer were leading causes of DALYs in China in 2017. This marks a significant change from 1990 when lower respiratory infections and neonatal disorders were first, followed by stroke and COPD (Zhou et al. 2019). In India, leading causes of death in 2017 were ischaemic heart disease, COPD, and stroke (IHME 2018); and in Mexico ischaemic heart disease, chronic kidney disease, and diabetes remained the leading causes of death from 2007 to 2017 (IHME 2018d).

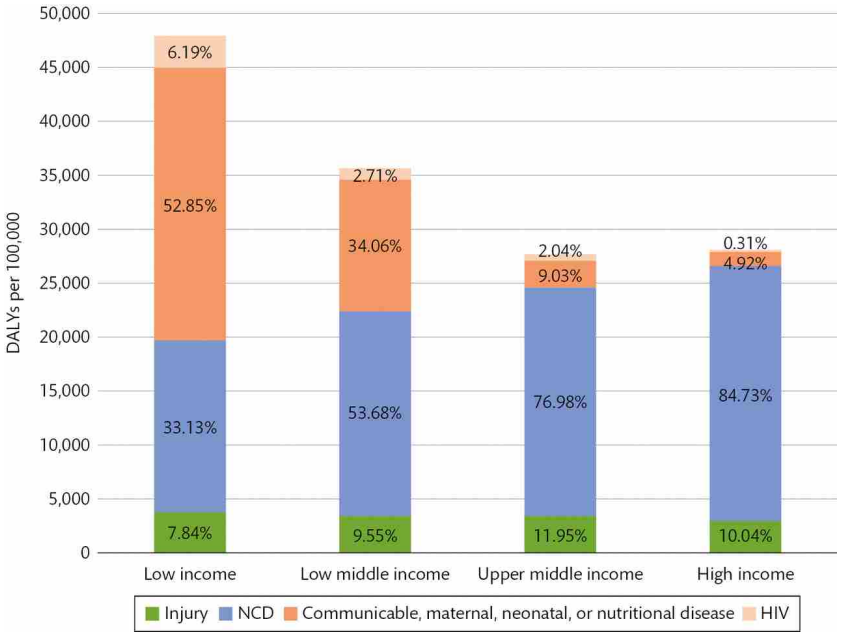


Fig. 1.4.2 Causes of DALYs across World Bank Income groups, 2017.
Data from Global Burden of Disease study 2017 (IHME 2018c).

However, MICs continue to face a substantial burden from infectious disease, maternal, neonatal, and nutrition-related conditions (Fig. 1.4.2) (Bawah et al. 2016; IHME 2018c). This despite improvements in hygiene and sanitation, declining fertility rates, greater female education, and targeted interventions resulting in increased vaccination rates, skilled birth attendants at delivery and the use of artemisinin-based treatment for malaria among others. Furthermore, there is growing recognition of the interaction between infectious and NCDs, especially cancers (e.g. human immunodeficiency virus and Kaposi's sarcoma as well as lymphomas, human papilloma virus and cervical cancer, and chronic hepatitis B infection and liver cancer).

Finally, injuries (from road traffic accidents, interpersonal violence, homicides, and self-harm) continue to account for approximately 10% of the global burden of disease and the same is true in MICs (Fig. 1.4.2). However, the impact of DALYs from injuries are disproportionately felt in MICs (and low-income countries) where car ownership is rising rapidly, preventive strategies are less developed, traffic and alcohol regulations are poorly enforced and health systems are less prepared to manage injury emergencies (Gosselin et al. 2009).

South Africa—along with many countries in southern and eastern Africa—is described as having a quadruple burden of disease (Groenewald et al. 2017) so recognizing the impact of HIV and TB. HIV/AIDS remains the leading cause of death in South Africa in 2017, as well as the leading cause of premature death and disability (IHME 2018f).

The complexity of disease burdens facing MICs is also reflected by the concurrence of risk factors that promote both communicable and non-communicable disease. This is best highlighted by the coexisting epidemics of underweight and overweight populations within the same country (Manyanga et al. 2014; Ford et al. 2017). Furthermore, both epidemics are typically experienced by the poorest socioeconomic quintiles, with wealthier quintiles having better health outcomes across all risk factor and disease groupings. In addition to changing nutritional patterns—driven in part by the expansion of multinationals into MICs and the hyperaccessibility of 'fast foods' (Miranda et al. 2019)—increased consumption of tobacco and alcohol, high blood pressure, raised glucose and air pollution are among the leading risk factors contribution to DALY's across the BRICS nations and many other MICs (IHME 2018; IHME 2018a, 2018e, 2018b, 2018f).

Multimorbidity

A consequence of these complex, sometimes interacting infectious, non-communicable and injury-related disease burdens, is the growing challenge of multimorbidity—referring to the presence of two or more chronic conditions in the same individual (Academy of Medical Sciences 2018). Increasingly individuals in MICs have coexisting conditions which manifest at younger ages. Thus multimorbidity is not only due to population ageing, but also

because of the premature growth in prevalence of NCDs, and persistence of communicable disease including HIV and TB (Academy of Medical Sciences 2018). See Box 1.4.2.

Box 1.4.2 Multimorbidity

Multimorbidity refers to the coexistence of two or more chronic conditions, each one of which is either:

- A physical non-communicable disease of long duration, such as a cardiovascular disease or cancer.
- A mental health condition of long duration, such as a mood disorder or dementia.
- An infectious disease of long duration, such as HIV/AIDS or hepatitis C.

The World Health Organization (WHO) Study on global AGEing and adult health (SAGE)—which considered non-communicable conditions in China, Ghana, India, Mexico, Russian Federation and South Africa—found that NCD-related multimorbidity affected 19% of the adult population (compared to 17% with only one NCD). Furthermore, 17.43% of 18–49 year olds already had 1 NCD with 9.27% having two or more NCDs. This burden only grows as the population ages: a third of 50–64 year olds had two or more NCDs (Sum et al. 2019).

Though recognition of the impact of multimorbidity on treatment requirements, patient outcomes, and health system demands is growing, there remains limited evidence about the most common clusters of chronic conditions and the relevance of the different combinations. In attempting to group coexisting conditions, those which are similar in their origin and/or treatment requirements have been termed concordant multimorbidities (e.g. ischaemic heart disease and stroke are both related to hypertension), while those which differ in their origin and treatment are classified as discordant multimorbidities (e.g. ischaemic heart disease and depression) (Academy of Medical Sciences 2018). Contrasting patterns of multimorbidity have also been described, a European pattern reflecting combinations of two or more NCDs (that may be similar to China, South Asia, and South America), whereas multimorbidity patterns including HIV/AIDS or TB and non-communicable disease is common in African settings. Successful antiretroviral treatment programmes and subsequent rise in life expectancy of people living with HIV means that multimorbidity including HIV/AIDS will become more prevalent (Bawah et al. 2016).

There is evidence to suggest that the different chronic disease dyad and triad combinations have different health system implications; for example, patients with multimorbidity including depression having fewer outpatient visits but more hospitalizations and generally reporting a lower quality of life (hypothesized to result from limited use of primary care services and delayed presentation to more specialized care leading to hospital admission (Sum et al. 2019)). Work from South Africa suggests that HIV positive patients on antiretroviral therapy (ART) have better outcomes across a range of coexisting conditions including better blood pressure control and lower blood glucose. The ‘ART advantage’ is thought to arise through increased healthcare utilization in people living with HIV/AIDS (Manne-Goehler et al. 2017; M’yambo Magodoro et al. 2019).

However, the health system implications of multimorbidity extend beyond utilization rates: these patients contribute to an excessive workload and cost (beyond simply the sum of treating the individual conditions). Vertically designed programmes tend to be specialized and may miss opportunities to manage concurrent conditions—though further data is needed to assess the contrasting impacts of vertical, horizontal, and diagonal health system designs (Academy of Medical Sciences 2018).

The impact of ART on life expectancy is resulting in rising prevalence of HIV/AIDS in ageing populations also at risk of other chronic NCDs (Bawah et al. 2016; Chang et al. 2019). The implications of managing patients’ concurrent HIV and NCD treatment needs demand greater consideration, particularly in southern and eastern Africa which account for almost 60% of all people taking ARVs globally (Chang et al. 2019).

Finally, despite evidence that multimorbidity and associated risk affects both adolescent and ageing populations, most health systems are poorly configured to meet the needs of these key groups. Notably, in adolescents, emphasis on preventive strategies and behaviour change interventions will be required (e.g. promoting healthier lifestyles and reducing obesity which will pay dividends later in life) (Patton et al. 2016).

Mental health across the life course

The burden from mental, neurological and substance use disorders continues to grow globally, in part reflecting demographic and epidemiological transitions as well as greater awareness. Mental health across the life course is a particular concern for MICs as known social determinants of mental illness and substance use disorders include

inequality, poverty, childhood adversity, displacement, and violence (Patel et al. 2018). As such, mental health status is an important pathway in the cycle of intergenerational poverty (Fig. 1.4.3) (Leon and Walt 2000; Patel et al. 2018).

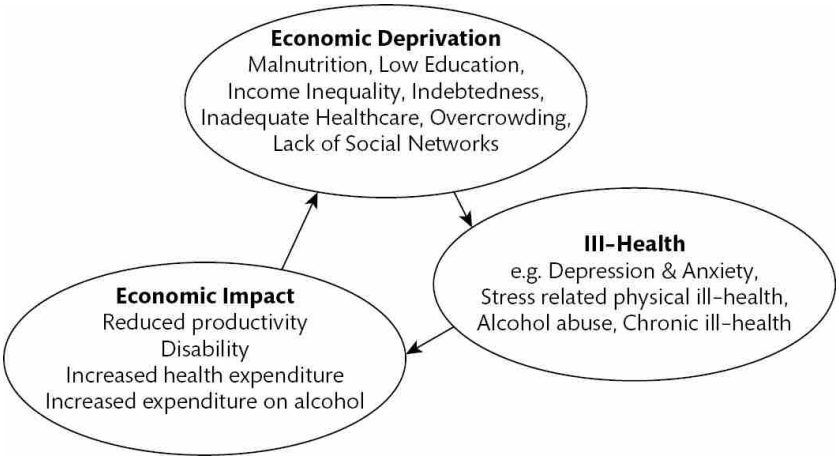


Fig. 1.4.3 The cycle of poverty and mental illness. From poverty, inequality, and mental health in developing countries (2000).
Reprinted from Leon, D. A. and Walt, G. (2000) 'Poverty, inequality, and mental health in developing countries', in Leon, D. and Walt, G. (eds) *Poverty, Inequality, and Health*. Oxford University Press, with permission from Oxford University Press.

A life course approach highlights critical periods which increase the risk for mental ill-health and require multilevel preventive and treatment strategies (Patel et al. 2018). Similarly, the concept of allostatic load (McEwen and Seeman 1999) recognizes the impact of cumulative psychosocial stresses which in turn lead to physiological 'wear and tear' contributing to negative health outcomes, including poor mental health.

Childhood experiences of early life stresses associated with low socioeconomic status and lack of stimulation in infancy are shown to increase the risk of mental health problems in later life. The presentation of mental disorders in adolescents is partly explained by the centrality of peer opinions which tend to replace those of parents and family and can be influenced by complex patterns of brain and cognitive development (Patton et al. 2016). This is a challenge in MICs where health systems seldom meet the health needs of adolescents despite growing adolescent and young adult populations. In older adults, mental health outcomes often relate to earlier life experiences, and a lack of formal education increases the chances of cognitive decline. Furthermore, cognitive impairment is associated with depression over the life course and so societies more exposed to risk factors for depression may experience increased burdens of dementia as populations age, a concern for MICs (Patel et al. 2018).

In addition, significant treatment and care gaps exist and surveys from China and India suggest that over 80% of people with a mental or substance use disorder did not seek treatment (Wei et al. 2013; Murthy 2017). Lack of access to care partly reflects underspending on mental health, with MICs spending approximately 2% of health budgets on mental health despite mental illness accounting for almost 30% of DALYs. Furthermore, most of this is spent on tertiary centres and inpatient institutions with little investment in community-based and primary care interventions (Patel et al. 2018).

Quality of services also tends to be poor: the World Mental Health Surveys report that only 11.4% of patients living with major depressive disorder in upper MICs receive minimally adequate treatment, and this drops to 3.7% in lower MICs (Thornicroft et al. 2017). Regarding substance use disorders, only 4.3% of patients in upper MICs and 1% of patients in lower MICs receive minimally adequate treatment (Degenhardt et al. 2017).

Migration

Most global migration occurs within LMICs (Abubakar et al. 2018). There were an estimated 258 million international (cross-border) migrants in 2017 (UN Department of Economic and Social Affairs Population Division 2017). However, internal migration—referring to movement within national borders—accounts for far more movement globally with over 740 million internal migrants in 2009 (four times the number of international migrants in the same year) (UNDP 2009). Internal labour migration may be from rural to urban areas (accounting

for much of urban growth in MICs), or from rural to rural areas based on work opportunities in agriculture or mining (with the proportion of female migrants growing substantially in recent decades. Importantly, evidence is strong that labour migrants, whether international or internal, generally contribute positively to a host country's economy. There are also consistently more internally displaced people in LMICs (due to conflict or natural disasters) than refugees or asylum seekers (being forced international migrants) (Abubakar et al. 2018). While it remains challenging to assess the full range of migrant health needs due to limitations in data quality, migration should be recognized as a core determinant of health in MICs.

The UCL-*Lancet* Commission on Migration and Health (2018) highlighted that migration is rarely a single journey (Abubakar et al. 2018)—with patterns involving variable distances within and across borders, temporarily or permanently, and often multiple times. Migration may involve several phases (Fig. 1.4.4) (Abubakar et al. 2018), each of which poses particular health risks relevant to planning health systems in MICs. Health of individuals pre-departure may be poor; the mode of transport can pose risks of injury or exposure, as can environmental and pathogen exposures on the journey. In the new country or region of residence, migrants face different infection and occupational exposures. And when returning to their origin home, the health risks may increase—some having lost immunity to local pathogens or having returned home following injury or disability without being able to access the necessary services (Abubakar et al. 2018).

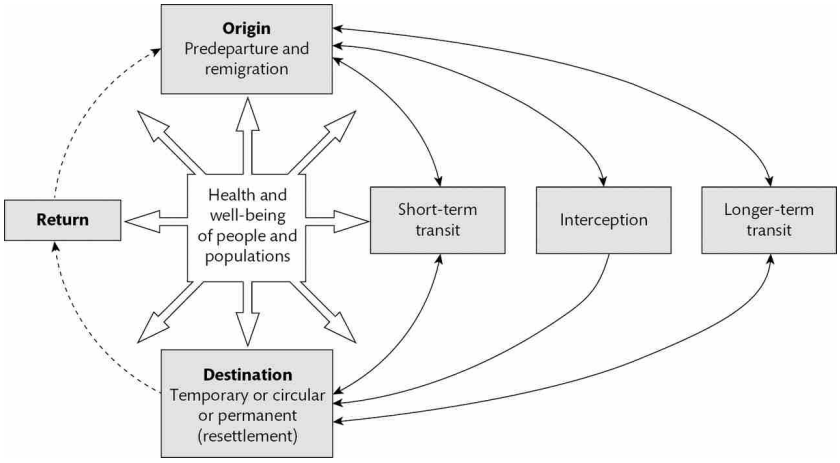


Fig. 1.4.4 The migration cycle.
Reprinted from Abubakar A et al. (2018) 'The UCL-*Lancet* Commission on Migration and Health: the health of a world on the move'. *The Lancet* 392(10164):2606–54, with permission from Elsevier.

Mortality rates among migrants have been studied in various settings. While a 'healthy migrant effect'—whereby health among migrants tends to be better than expected with lower risk for disease—the theory overlooks the diversity among migrant groups, and that the health benefits of migration may reduce over time. Data on mortality among migrants in MICs is limited however work in sub-Saharan Africa concluded that internal migrants often display higher mortality than long-term residents. This health disadvantage arose due to both infectious and non-communicable disease, and was likely linked to poor social integration and access to care, and inadequate living conditions (Abubakar et al. 2018; Ginsburg et al. 2018).

Evidence on the relationship between migration and mental health is mixed: international migrants, particularly asylum seekers and refugees, tend to have higher rates of depression, anxiety, and post-traumatic stress disorder. This often reflects differences in the precipitating risk factors such as experience of violence and traumatic events (Patel et al. 2018). The economic and psychosocial stresses of migration and urbanization may place migrants at greater risk of poor mental health outcomes (Ginsburg et al. 2018). Finally, migration appears linked to mental health outcomes in children: both in children who migrate, where research suggests they are more likely to have depression, anxiety, aggression, or hyperactivity, and among children left behind. A meta-analysis of studies on left-behind children suggested they are at increased risk of depression (RR 1.52, 95%CI 1.27–1.82), anxiety (RR 1.85, 1.36–2.53), substance use (RR 1.24, 1.00–1.52) and suicidal ideation (RR 1.70, 1.28–2.26) (Patel et al. 2018).

In South Africa, internal ‘circular’ labour migration—whereby workers move to town but maintain regular contact with their rural home—was an important factor in propagating the HIV epidemic: migration led to an increased risk of HIV infection, though acquisition and transmission could be bidirectional, e.g. from migrants to their rural partners, and from rural residents to their migrant partners (Lurie et al. 2003). As migrant labourers developed AIDS (or other serious chronic condition), becoming too sick to work, they chose to ‘return home to die’ (Welaga et al. 2009), with significant attendant care burdens and health-related costs borne by their households during their final stage of illness.

As a principle, consistent with the goal of UHC, health systems need to incorporate the competencies required to address the specific health needs of the migrant population, whether citizens or not. Furthermore, health systems need to strengthen their capability to provide services to highly mobile groups unlikely to access care from the same place consistently over time.

Climate change and health

MICs include countries heavily affected by climate change through altered temperature and rainfall patterns coupled with extreme weather events resulting in widespread health, social, and economic impacts. MICs are at risk of climate-related health effects for three reasons: (1) geography viz. a higher concentration of heatwaves, adverse weather events such as droughts and floods, and loss of labour capacity especially in areas close to the equator; (2) national systems lack the resilience needed to adapt to persisting weather changes, respond to disasters, or diversify food sources—with single crop markets at particular risk; (3) less developed regulatory frameworks to limit key drivers of climate change such as fossil fuels, and provide incentives for clean energy.

Climate change has both direct and indirect health impacts. Direct effects include exposure to heat waves and rising temperatures, as well as weather-related disasters. Indirect effects are those mediated through changes in ecosystems, land use, air pollution and water quality (Fig. 1.4.5).

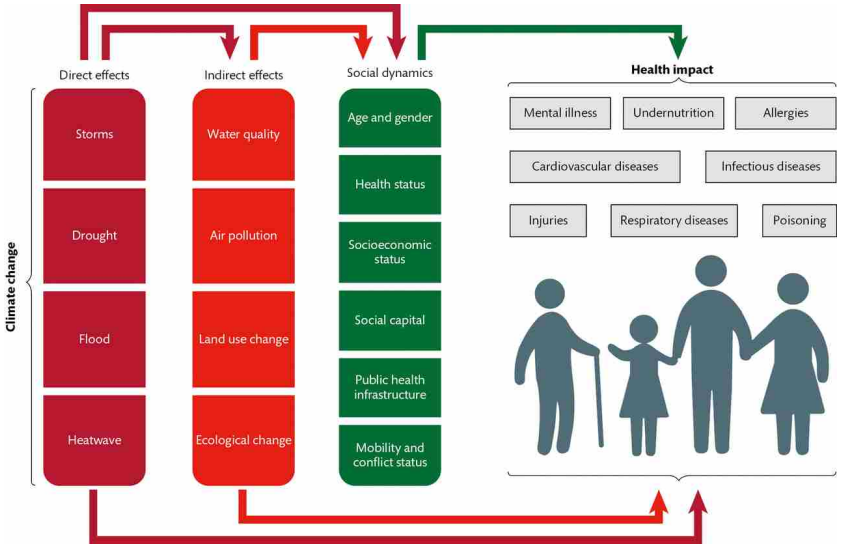


Fig. 1.4.5 The direct and indirect effects of climate change on health and wellbeing. From *Lancet* Commission on Health and climate change: policy responses to protect public health.

Reprinted from Watts, N., *et al.* (2015) ‘Health and climate change: policy responses to protect public health’, *The Lancet*. 386(10006), pp. 1861–914, with permission from Elsevier.

Direct health effects

Increased temperature and heatwaves can exacerbate chronic conditions, lead to new cardiovascular and renal conditions, and provoke dehydration—to which older persons are especially prone. Furthermore, when combined with urbanization and air pollution, not only is the temperature rise amplified, but acute and chronic respiratory disease tends to worsen (Watts et al. 2015).

The impact of heat exposure is also evident in the fall in outdoor labour productivity (some 5.3% decline worldwide 2000–2016—with the greatest reductions in countries worst affected by rising temperatures, many of which are MICs). The decline in outdoor labour productivity is paralleled by both a rise in health burden and reduction in economic output—of particular concern for MICs where much of the labour force continues to work in industry (Watts et al. 2015, 2018).

Concern around climate-sensitive diseases is growing: greater exposure to UV rays resulting from damage to the ozone layer is leading to rising incidence of malignant melanoma (Watts et al. 2018); and climate-related weather disasters result in direct morbidity and mortality, and can cause mental health disorders linked to the trauma of the extreme weather events (Watts et al. 2015).

Indirect health effects

Most climate-related health risks are mediated through ecological and social processes. Food insecurity is among the greatest impacts of climate change and arises through multiple pathways including crop failure, reduced food production, extreme weather events, increased food-borne and other infectious diseases, and civil unrest. While severe food insecurity and hunger are associated with undernutrition, wasting, and stunting in children, mild to moderate food insecurity has been associated with obesity in vulnerable populations (Swinburn et al. 2019), a paradox of concern in the context of high inequality and rising NCDs in MICs.

Diseases related to water quality (e.g. cholera) and vector spread (including malaria and dengue fever) are impacted by rising temperatures and changes in rainfall patterns. Warmer temperatures promote cholera bacterial growth, while droughts and extreme weather events can limit access to safe drinking water. In areas previously protected by lower temperatures, climate change is affecting the range and reproductive rate of malaria-carrying mosquitoes, even at higher altitudes, with growing risk of new malaria epidemics (Watts et al. 2015). This increase in vector capacity also contributes to the climbing incidence of dengue fever with an estimated 58.4 million cases in 2013 (95%CI 23.6 million –121.9 million) accounting for some 10 000 deaths and 1.14 million DALYs. Much of this burden is in the Asia-Pacific, Latin American and Caribbean regions—mostly MICs. The *A aegypti* and *A albopictus* mosquitoes which transmit dengue are also vectors for emerging and re-emerging arboviruses including Yellow Fever, Chikungunya and Zika virus (Watts et al. 2018).

Antimicrobial resistance

MICs face a paradox when it comes to antimicrobial use: on one hand, infections continue to account for a significant burden and are the leading cause of death in some countries, notably in children under-5. An estimated 445, 000 pneumonia-related child deaths could be averted through universal provision of antibiotics (Laxminarayan et al. 2016). On the other hand, antimicrobial resistance (AMR) is a growing problem, driven largely by increasing and inappropriate use of antibiotics (Miller-Petrie et al. 2018). With regard to neonatal mortality alone, resistant sepsis infections kill approximately 214, 000 neonates each year primarily in India, Pakistan, Nigeria, the Democratic Republic of Congo, and China (four of these five countries are MICs) (Miller-Petrie et al. 2018).

Tackling antimicrobial resistance in MICS is challenging for several reasons:

Increasing consumption of antibiotics: although per capita antibiotic use remains lower in MICs than HICs, MICs saw the highest growth rates in antibiotic consumption 2000–2010 and have higher drug resistance indices (Klein et al. 2019). In part this is due to rising income and greater access to antibiotics, however, it also reflects supply-side problems including: lack of enforcement of prescribing regulations, high rates of over-the-counter purchasing, and a lack of incentives for rational prescribing, all of which lead to inappropriate antibiotic use. Substandard and falsified medications are also a concern; while this is common in India, the WHO estimates that it affects 7% of antibiotics globally (Miller-Petrie et al. 2018; World Health Organization 2018b). Lastly, changing patterns of antibiotic use have resulted in increased consumption of broad-spectrum antibiotics including cephalosporins, broad-spectrum penicillins, and fluoroquinolones, as well as increased use of last-resort antibiotic classes (carbapenems and polymixins) (Laxminarayan et al. 2016).
Antimicrobial use in animals: two-thirds of all antibiotic use, including antibiotics to treat infections in humans, are used in animals at subtherapeutic levels to promote growth and prevent disease, and at therapeutic levels to treat infection in livestock (Laxminarayan et al. 2016). Estimated use of antibiotics in food production is expected to virtually double (98% increase) in BRICS nations (Brazil, Russia, India, China, South Africa) by 2030, including a projected 312% increase in India alone (Laxminarayan et al. 2016). Widespread use of antibiotics in farming creates strong selection pressure, encouraging the development of resistant infections (Miller-Petrie et al. 2018) which will have impact human and animal health (encompassed in the concept of One Health). However, few MICs regulate or monitor the use of antimicrobials in animals—or have adequate capacity to do so (Hu and O'Neill 2019).
Lack of antimicrobial resistance surveillance programmes: Accurate data on AMR is required to effectively design programmes to prevent the spread of infection and tackle inappropriate prescribing practices. However, few MICs have managed to establish routine antimicrobial surveillance systems. And where these do exist, data is usually publicized in annual reports of resistance profiles, making it difficult to respond rapidly to changes in resistance and outbreaks of resistant organisms (Perez and Villegas 2015). Furthermore, no AMR surveillance programmes in MICs include data on antimicrobial use and antibiotic resistance in agriculture and livestock (World Health Organization 2014). A global AMR surveillance system (GLASS) is being developed by the WHO to support the global action plan on antimicrobial resistance, and aims to improve surveillance capacity and data quality particularly in LMICs (World Health Organization 2018a).
Lack of data from primary care settings: Where data on AMR is available—whether from surveillance programmes or specific research studies—this tends to focus on resistance in hospital settings and is extrapolated to estimate the community prevalence of AMR (World Health Organization 2018b). However, the impact of antibiotics used in farming and the environment, substandard and falsified medications, and self-medication with antibiotics are more likely to affect community-acquired infections. Most patients with bacterial infections sensitive to first line antibiotics will improve with treatment at primary care level and not present to higher levels of care. As a result, extrapolating trends in AMR from hospital level data could over- or underestimate AMR in the community, and resistant strains may vary between hospital and community settings.

The consequences of AMR can be more severe in MICs (and low-income countries (LICs)) than HICs. Mortality from resistant infections may be higher—due to a lack of access to newer and more effective drugs, compounded by poorer social determinants of health (World Health Organization 2018b). Where second- and third-line drugs are available they tend to be more expensive: the World Bank estimates that AMR could push 28 million people into poverty by 2050 (World Bank 2017). Furthermore, the growth of fluoroquinolone resistance is of particular concern as fluoroquinolones are used in the treatment of multidrug resistant TB which remains a priority in many MICs (Hu and O'Neill 2019).

Finally, strategies to limit AMR include preventing infection with regular hand washing and infection control in health facilities (Miller-Petrie et al. 2018), practices which tend to be less well implemented in MICs compared to HICs. Similarly, vaccinations can protect against common infections, reducing the need for antibiotics (Miller-Petrie et al. 2018). However as MICs graduate from GAVI, the Vaccine Alliance, they will need to take on higher vaccine costs—as of December 2020 only India, Nigeria and Indonesia will remain eligible for assistance out of the nine BRICS and MINT (Mexico, Indonesia, Nigeria, Turkey) nations (Hu and O'Neill 2019).

Health systems in MICs

Given the complex and heterogeneous health challenges facing MICs as a consequence of the evolving, dynamic transitions underway, it is critical that health systems in these countries adapt to respond to societies' changing

health needs. This cannot be achieved by simply improving the efficiency of current approaches; rather it requires an all-round rethinking of health systems in these settings, taking cognisance of inevitable resource constraints. We suggest the following tenets underpin future-oriented health systems in MICs—that they

be *resilient and learning health systems*
strive to provide *UHC*
focus on the *quality of health services* in addition to their availability
take into account the contributions possible from *intersectoral collaboration.*

Resilient and learning health systems

'Resilience' describes the degree to which health systems can absorb shocks and adapt to provide the required services (Kruk et al. 2015; Massuda et al. 2018). Although the phrase gained traction following the 2014–2016 Ebola crisis in West Africa, there is growing recognition of the importance of everyday resilience in resource-constrained health systems which experience repeated shocks (such as funding shortages, unreliable staffing, changing policies and guidelines, and growing administrative burdens) (Gilson et al. 2017; Massuda et al. 2018).

Resilient health systems can be characterized by five elements: (1) awareness of available assets and vulnerabilities (human, physical, informational); (2) capacity to address a diversity of health needs across communities (measured in part by the degree to which universal health coverage has been achieved); (3) self-regulation with the capability to rapidly identify and contain new health threats while continuing to provide other essential health services; (4) integration of health services with other sectors to maximize the innovation in solutions to health challenges (which requires clear communication and information sharing); (5) adaptation to meet new health needs, both in response to emergencies and slower changes in evolving disease burdens (Kruk et al. 2015).

A learning health system is also one which has as a central tenet, a focus on routinely collected data as a basis for continuous improvement integrated into the everyday functioning of services (English et al. 2016)—in this way, learning health systems can ore rapidly adjust their practices to adopt new evidence and improve the quality of patient care.

Both resilient and learning health systems depend on accurate and up-to-date data upon which to base decisions, evaluate progress, and adapt in response to available evidence. To-date, many MICs lack comprehensive civil registration and vital statistics (CRVS) systems. Investment in routine information systems is necessary as MICs endeavour to strengthen their health systems. Doing so as part of initial health system strengthening can prove far more cost-effective than retro-fitting data collection systems later (English et al. 2016).

Universal Health Coverage

In 2005, all WHO member states committed to working to achieve UHC, i.e. that all people are able to access high-quality health services when needed (including preventive, promotive, curative, rehabilitative and palliative services) without risk of financial hardship (WHA58.33 Sustainable health financing, universal coverage, and social health insurance 2005). UHC is a particularly powerful and unifying principle for health systems in MICs: systems designed around the provision of essential care to all regardless of citizenship, nationality, age, or gender should be far better placed to address the complex disease burdens and provide integrated care for patients with multimorbidity, often at young adult ages. UHC is also a critical element in strategies to better provide migrant health services by lowering barriers to care related to entitlement to use services, where and which services, and essential information sharing between providers when patients routinely do not attend the same clinic.

While few MICs can claim to have achieved UHC, some have made substantial strides as detailed in the World Bank's *Going Universal* report (Cotlear et al. 2015). Although approaches vary, key features include increased health spending, pooling of funds, decentralization of care management, effective private–public partnerships, and expansion of comprehensive primary care services (Cotlear et al. 2015; Castro et al. 2019). Box 1.4.3 details elements of the Brazilian effort to achieve UHC.

Box 1.4.3 Achieving UHC in Brazil

The basis of UHC in Brazil is the Sistema Unico de Saúde (SUS)—a unified health system established in 1990 following the enshrining of health as universal human right in the Brazilian constitution of 1988. SUS decentralized health programme financing and management, sharing power between

federal and municipal structures. All citizens are entitled to services provided by SUS, however the purchasing of private insurance by higher income groups means that SUS services are largely utilized by poorer and uninsured members of the population.

As explained by Castro and colleagues (2019), SUS changed the model of healthcare delivery, expanding comprehensive primary care services and investing in the Family Health Strategy (a community-level health delivery platform) which increased service provision from some 7 million people (4% of the population) in 1998, to 130 million people (62% of the population) in 2018. In addition, mental health services, emergency care services, and specialist outpatient services were developed. Private–public partnerships were established to boost access to specialist care whereby the public sector funds service and facilities, but management and staff are provided by the private sector (Castro et al. 2019).

The SUS is credited with significant improvements in health outcomes including rising life expectancy at birth, a fall in under-5 mortality from 64/1,000 live births in 1990 to 15/1,000 in 2015, and increased immunization coverage. SUS enabled Brazil to meet the rapidly changing health needs of its population, driven by migration from rural to urban areas, re-emerging infectious disease due to expanded human occupation of regions of the Amazon, rapid demographic and epidemiological transitions resulting in an increasing burden from non-communicable diseases (Castro et al. 2019).

However, Brazil also serves as a warning to MICs of the fragility of health gains, and the ease of reversal when political leadership changes and investment in health systems decline. Introduction of austerity measures in 2016, including capped increases in health spending, and changes to reproductive and adolescent health services which threaten to limit access to care (given fear of promoting abortion), failure to explicitly include the lesbian, gay, bisexual, transgender, and questioning (LGBTQ+) community in ministerial mandates (claiming instead that 'diversity policies have threatened the Brazilian family'), education policies which prevent teachers from discussing gender, identity, and sexuality, amendments to gun laws which will increase their availability, and possible reductions in tobacco taxation all threaten to reverse—at least in part—the hard won health gains of the past 30 years (Castro et al. 2019).

Although the principle of UHC is increasingly accepted, financing remains more controversial. The move towards UHC will see countries increase total health expenditure even in scenarios where there are significant efficiency gains. Ability to afford this increase is largely driven by economic growth and so increased available national income. However, in addition to increased spending, UHC requires pooled spending with criteria for contributions distinct from criteria for use (which is simply health need). Different strategies to pool spending are available (Schuster et al. 2018): some countries use general taxation; however, this can be challenging in countries with high unemployment and small tax bases (which may already be heavily taxed). Some countries have used compulsory payroll contributions (Hong 2009), while others have targeted public health insurance to the most vulnerable groups who do not meet criteria for employment-based schemes (such as the Seguro Popular programme in Mexico (Frenk et al. 2006)).

Countries also need to resolve what health services to fund (often referred to as a 'minimum benefits package'). This too will vary depending on what countries can afford. However, the Disease Control Priorities Project (3rd ed.) recommended standardized essential care packages that all countries should strive to provide based on value for money, disease burden and feasibility of implementation (Watkins et al. 2017).

There is ongoing debate on the extent to which migrants gain access to the various models of UHC emerging. Thailand is among the best-known examples of migrant-inclusive policies: migrants with work permits are fully covered by the Thai social security scheme which is funded through mandatory payroll taxes including contributions from the employer, employee, and government. However, the majority of migrants in Thailand are undocumented and so ineligible for social health insurance. Therefore, in 2001, Thailand established a migrant health insurance scheme which migrants (documented or not) could join through a voluntary annual prepayment premium. The scheme was later extended to cover migrants' dependents. Furthermore, Thailand invested in innovations to create migrant-friendly health services, including use of volunteer community workers to assist with health education, mobile clinics, and bilingual signposts. These efforts have served to increase access care for migrants, improve diagnosis and treatment, and reduce financial risk from large out-of-pocket payments, though work remains to increase the proportion of migrants enrolled (Tangcharoensathien et al. 2017).

Quality healthcare services

The quality of care is as important as the extent of care provided: the Lancet Global Health Commission on High Quality Health Systems in the SDG Era found that improving quality of care could save an estimated 8 million lives in LMICs each year. Furthermore, promoting quality care can be expected to increase patient trust and improve patients' experience of engaging the healthcare system (Kruk et al. 2018). However, health systems in MICs remain underresourced relative to the need and so innovative strategies are required to deliver quality care at an affordable cost. Task shifting and digital technologies are strategies increasingly used to extend high-quality healthcare in MICs.

Task shifting

Task shifting, an approach that is widespread in global health, refers to the process of redistributing tasks to less specialized healthcare workers (World Health Organization 2007) and was successfully employed at scale in response to the HIV epidemic in sub-Saharan Africa: counselling and testing 'shifted' from nurses to lay counsellors or community health workers (CHWs); treatment initiation and monitoring transferred from doctors to

nurses as did diagnosis and management of opportunistic infections; and management of advanced disease shifted from infectious disease specialists to non-specialist physicians (World Health Organization 2007).

Task shifting has been employed in prevention and management of cardiovascular disease (Ogedegbe et al. 2014) and to improve maternal and child health outcomes, with CHWs tasked to provide health education to pregnant women and new mothers, identify high-risk pregnancies for referral to medical centres, conduct newborn checks, treat newborn infections and refer for further care where necessary, and monitor growth and vaccinations among others (Balsari et al. 2017). In global surgery, obstetrics, and anaesthetics, associate clinicians are increasingly tasked with operations and anaesthesia in place of doctors (Federspiel et al. 2018).

While multiple trials have demonstrated the success of task shifting in expanding access to care without sacrificing the quality of care, there is less evidence to show which factors are critical to maintaining quality of care when such interventions are scaled-up. However, supervision, clear guidelines, adequate initial training, and regular retraining all appear important elements of successful programmes (Mason et al. 2006; World Health Organization 2008).

Digital health technologies

Digital technologies are increasingly included in studies that: address coverage gaps, increase communication with patients, improve diagnosis and treatment, improve data management, and assist in financial transactions (and limit fraud and abuse) within the healthcare system. Technologies include mHealth (mobile health) and computer-based information applications. With rapidly rising access to mobile phones and the internet, delivering health interventions using mobile technologies is increasingly feasible (Lewis et al. 2012).

There is now evidence to demonstrate the value of mHealth interventions across LMICs. SMS reminders have been shown to improve treatment adherence and care seeking as seen with regards to ARVs and malaria diagnostics (Liu and Modrek 2016; Amankwaa et al. 2018); and educational interventions such as MomConnect have been effective at delivering targeted health promotion messages to pregnant women (Seebregtsi et al. 2016). Technology has also enabled task shifting with mobile or tablet-based applications designed to assist health workers in diagnosis and treatment in the field or in primary care (Lewis et al. 2012). Virtual consultations reduce the need for in-person visits to clinics and so increase access to care where geography, transport and financial barriers might exist (Lewis et al. 2012). Furthermore, virtual consultations may be preferred by patients who experience stigma and resist accessing care for fear of being ostracized. In contrast, the use of information system interventions, including electronic health records and data management systems is less well established in MICs.

Despite successful pilot studies, few programmes have been scaled successfully (exceptions including MomConnect in South Africa (Seebregtsi et al. 2016) and the mDiabetes project in Senegal (World Health Organization 2017)). Scaling digital interventions often requires additional investment; electricity and internet access are inconsistent and mobile data is expensive; and there remain cultural and generational barriers to widespread adoption of technological interventions (Lewis et al. 2012; Labrique et al. 2018). Furthermore, despite the imperative for data that underpin learning health systems, many MICs are not making use of available technologies to assist in data collection and management (Lewis et al. 2012).

Intersectoral collaboration

The impact of social and economic transitions on disease burdens and outcomes highlights the multidimensional nature of health and points to the value of intersectoral collaboration in tackling major public health challenges. We have noted the importance of functioning CRVS systems and regular health outcome reporting which often falls outside the remit of health departments; improvements in housing, water, and sanitation are critical to reducing the burden of communicable disease. Intersectoral collaboration is required to tackle persistent inequality in many MICs: targeted social welfare policies seek to support the large ‘tail’ of the population still to emerge from poverty. These tend to be high health burden populations, and policies which redistribute wealth across socioeconomic quintiles assist in lessening the burden on the next generation. The South African non-contributory old-age pension and child support grants are exemplars of pro-poor policies where positive externalities have included health gains beyond the direct grant recipient (Aguero et al. 2006; Case and Menendez 2007).

Conditional cash transfers have been used to address the burdens of poverty and incentivize health behaviours in high-risk populations: in Indonesia, conditional cash transfers improved vaccination rates by up to 30% (Kusuma et al. 2017). while the Oportunidades CCT programme in Mexico and the Bolsa Família programme in Brazil have been associated with improved child health, growth, and development, and increased care seeking among other positive health and education outcomes (Fernald et al. 2008; Shei et al. 2014).

Finally, national taxation policies and regulation of nutrient content in foods are powerful tools for addressing public health burdens. Excise taxes on tobacco, alcohol, and sugar-sweetened beverages are effective strategies to reduce consumption with an ensuing decline in risk, morbidity, and mortality from related diseases (including cardiovascular disease, cancers, interpersonal violence) (Task Force on Fiscal Policy for Health 2019). Mandated salt targets in processed foods are expected to further reduce death rates from heart disease (Webster et al. 2017). Such policies demonstrate the value in governments becoming more adept at nudging their populations to make healthy choices, and suggest that similar approaches can be applied to other products. See Box 1.4.4 for a discussion of tobacco.

Box 1.4.4 The power of tobacco, alcohol, and sugar-sweetened beverage excise taxes

Excise taxes on tobacco have been used effectively in MICs including Brazil, South Africa, Ukraine, Turkey, and the Philippines (Task Force on Fiscal Policy for Health 2019), though there remains room for further increases to meet the WHO recommendation of a 70% excise tax on tobacco (World Health Organization 2010). Figure 1.4.6 shows changes in per capita cigarette sales as prices rose in Brazil: over 2006–2011, the real price of cigarettes rose by 34% per pack, with a corresponding fall of 19% in per capita cigarette consumption. In 2012–2016, the real price of cigarette packs rose by a further 33%, corresponding to an overall decline of almost 50% in cigarette consumption per adult (Task Force on Fiscal Policy for Health 2019).

Russia continues to have one of the highest alcohol consumption rates in the world. However in 2004, laws banning alcohol advertising, restricting availability, and taxing alcoholic beverages were introduced and resulted in a decline in alcohol consumption by a third by 2014 (Task Force on Fiscal Policy for Health 2019). In Mexico the introduction of a 10% sugar-sweetened beverage tax is credited with reducing consumption of sugary drinks by 7.6% over the first 2 years of the tax, with the largest reductions noted among the poorest socioeconomic groups (Colchero et al. 2017). Further projections suggest the tax will prevent 189,300 incident cases of type 2 diabetes, 20,400 strokes and heart attacks and 18,900 deaths in adults 35 years and older between 2016 and 2026 (Sánchez-Romero et al. 2016). A similar policy (Health Promotion Levy) was introduced in South Africa in 2018 and is projected to prevent 8,000 diabetes-related premature deaths over the coming 20 years (Saxena et al. 2019).

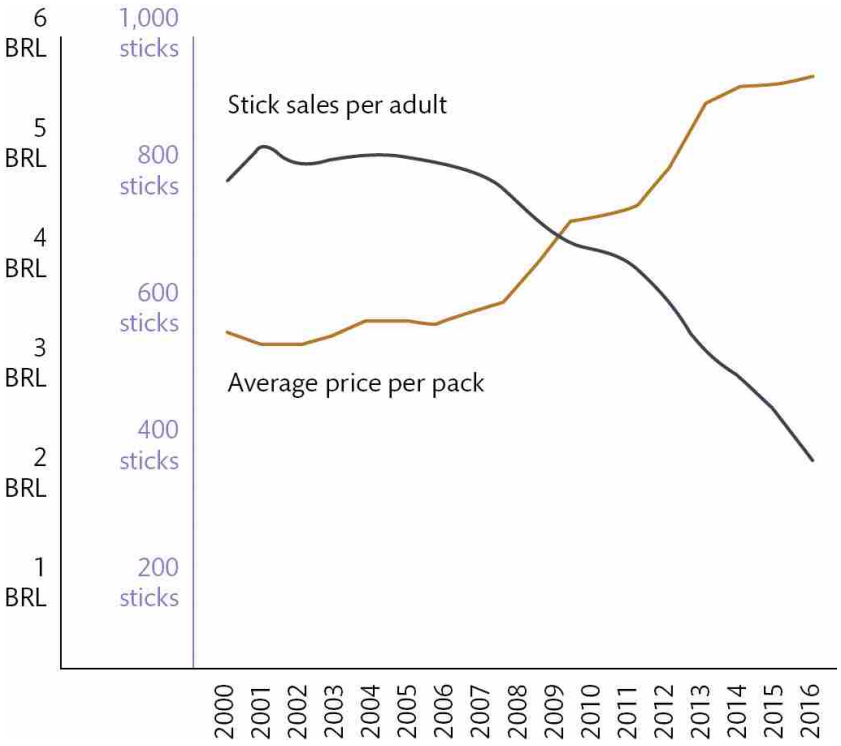


Fig. 1.4.6 Cigarette Prices and Per Capita Cigarette Sales, Brazil, 2000–2016.
Reprinted from Health Taxes to Save Lives (2019) Task Force on Fiscal Policy for Health (2019) 'Health Taxes to Save Lives: Employing Effective Excise Taxes on Tobacco, Alcohol, and Sugary Beverages'. New York: Bloomberg Philanthropies.

Discussion

MICs are diverse in geography, politics, and population size—but all are undergoing social, economic and health transitions which necessitate adaptations of their health system to meet changing needs across the life course. Investing in resilient and learning health systems based on the principle of UHC should enable MICs to deliver care in a context of heterogeneity of outcomes related to persistent inequality in MICs, a triple (or quadruple) disease burden, growing prevalence of multimorbidity, mental and substance use disorders, and the health impacts arising from migration, climate change, and antimicrobial resistance. There is every likelihood that failure to adapt current health system functioning will lead to a divergence between health spending and outcomes—with increased spending failing to generate increased impact. Moreover, the nature of health profiles and burdens evolving in transitioning countries also requires interventions beyond the health sector, highlighting the value of collaboration across sectors and reinforcing a 'health in all policies' approach.

While much of the chapter is focused on the health transitions that accompany economic growth, there are of course economic effects from changing health status: with improved health comes greater productivity and reduced treatment costs (and conversely poor health negatively impacts economic growth). Investment in effective preventive and promotive strategies can be extremely cost-effective across the life course and translate with time into enhanced education and livelihoods. However there is a tension between prioritizing spending on prevention of future disease in the face of prevalent burdens requiring immediate attention. Furthermore, many preventive strategies require investment beyond the health sector—to alter more upstream social, structural, and environmental determinants of health. UHC and intersectoral collaboration are key to managing these competing priorities in order to meet the health needs of transitioning populations given the resources available.

Achieving the necessary health system reforms across MICs will require sustained investment in research and cognisance of the complexity of cultural and contextual 'fit'. We have highlighted a number of areas where further research is needed: accurate data on risk, morbidity, and mortality patterns is critical to designing and monitoring interventions. Better understanding of multimorbidity is urgent, with particular focus on adolescent and young adult populations, both poorly served by existing health infrastructure. The full impact of migration, climate change, and antimicrobial resistance on primary healthcare and agriculture are research agendas which warrant ongoing attention. Research is also necessary to evaluate interventions and generate the data to drive learning health systems—including questions of cultural acceptability and contextual fit which will influence the effects of interventions, and strategies to scale-up. Implementation scientists can advise on integrating routine data into evaluation of health systems performance, and how to effectively convert evidence into practice and policy.

Costs are a challenge facing all countries and a particular challenge for MICs. Health costs will grow as life expectancy increases and disease burdens evolve, and intervention strategies will likely differ for communicable, non-communicable and injury-related burdens—with the implication that parallel investments are required. The suggestion that MICs strive to achieve UHC places further demands on national health services—and an imperative to set priorities. The reality, however, is that MICs will be unable to cover all interventions for all conditions for all people. This is true of even the largest economies (such as China, India, and Brazil) which remain vulnerable to the pressures inherent in the global financial system including trade wars, depressions, and recessions. Many MICs require external support and investment—from bilateral donor countries, philanthropies, and the International Monetary Fund (IMF), World Bank, and UN system—to meet immediate costs and support investment in systems strengthening and health promotion which yield health benefits in the future.

The leadership demands posed by the health challenges facing MICs in transition are substantial with few tried and tested leadership strategies that guarantee improved health outcomes. The challenges highlighted in this chapter impose leadership requirements across all levels of the health sector and spill over into other sectors (including infrastructure, water, social development, housing, education, environmental policy, city planning, legislation, and policing). Successful leaders will be able to influence the political agenda towards achieving UHC, while balancing the challenges of competing interests, and using evidence to buttress policy decisions.

Good policy, though necessary, is not sufficient. Governance—being the capability to strategically implement policy, achieve buy-in, monitor compliance, and hold parties to account—is critical. Although many MICs need to build institutions and strengthen their capacity to monitor public health legislation, successes such as the enforcement of tobacco legislation in Brazil and South Africa can serve as motivation to further develop governance structures.

Conclusion

MICs are countries undergoing complex, often rapid economic and social transitions. Such transitions are typically prolonged and non-linear and pose major challenges to the conventional organization of health systems. Innovative

policies should focus on building resilient and learning health systems better able to deliver UHC, capitalize on intersectoral collaboration, and make use of task shifting and digital health technologies to deliver high-quality care. Investment in further research—to generate new data and better translate evidence into policy and programmes—and enhanced leadership and governance will support MICs in meeting their populations' changing health and development needs.

REFERENCES

Abubakar, I., et al. (2018). The UCL-*Lancet* Commission on Migration and Health: the health of a world on the move. *The Lancet*, 392(10164), 2606–54.

Aguero, J.M., Carter, M.R., and Woolard, I. (2006). *The Impact of Unconditional Cash Transfers on Nutrition: The South African Child Support Grant. Southern Africa Labour and Development Research Unit Working Paper Number 06/08*. Cape Town [Online]. Available at: https://www.saldru.uct.ac.za

Amankwaa, I., et al. (2018). Effectiveness of short message services and voice call interventions for antiretroviral therapy adherence and other outcomes: a systematic review and meta-analysis. *PLoS One*, 13(9), e0204091.

Balsari, S., et al. (2017). *Task Shifting in Healthcare: Reframing the AYUSH Debate*. Cambridge, MA. [Online] Available at: https://cdn2.sph.harvard.edu/wp-content/uploads/sites/114/2017/12/Task-Shifting-in-Indian-Healthcare-Jan-2017.pdf

Bawah, A., et al. (2016). The evolving demographic and health transition in four low- and middle-income countries: evidence from four sites in the INDEPTH network of longitudinal health and demographic surveillance systems. *PloS One*, 11(6), e0157281.

Beaglehole, R. and Bonita, R. (2004). *Public Health at the Crossroads: Achievements and Prospects*. Cambridge: Cambridge University Press. [Online] Available at: https://www.cambridge.org/vu/academic/subjects/medicine/epidemiology-public-health-and-medical-statistics/public-health-crossroads-achievements-and-prospects-2nd-edition?format=PB&isbn=9780521540476

Biciunaite, A. (2014). *Does Economic Growth Raise Life Expectancy—Do Wealthier Countries Live Longer? Euromonitor International*. [Online] Available at: https://blog.euromonitor.com/economic-growth-and-life-expectancy-do-wealthier-countries-live-longer/

Case, A. and Menendez, A. (2007). Does money empower the elderly? Evidence from the Agincourt demographic surveillance site, South Africa. *Scandinavian Journal of Public Health*, **69**, 157–64.

Castro, M.C., et al. (2019). Brazil's unified health system: the first 30 years and prospects for the future. *The Lancet*, 394(10195), 345–56.

Chang, AY., et al. (2019). Chronic multimorbidity among older adults in rural South Africa. *BMJ Global Health*, 4(4), e001386.

Chao, F., et al. (2018). National and regional under-5 mortality rate by economic status for low-income and middle-income countries: a systematic assessment. *The Lancet*, 6(5), e535–47.

Coale, A.J. (1989). Demographic transition. In: Eatwell, J., Millgate, M., and Newman, P. (eds.) *Social Economics*. London: Palgrave Macmillan UK, pp. 16–23.

Colchero, M.A., et al. (2017). In Mexico, evidence of sustained consumer response two years after implementing a sugar-sweetened beverage tax. *Health Affairs*, 36(3), 564–71.

Cotlear, D., et al. (2015). *Going Universal: How 24 Developing Countries are Implementing Universal Health Coverage from the Bottom Up*. Washington, DC: The World Bank.

Degenhardt, L., et al. (2017). Estimating treatment coverage for people with substance use disorders: an analysis of data from the World Mental Health Surveys. *World Psychiatry*, 16(3), 299–307.

Elsey, H., et al. (2019). Rethinking health systems in the context of urbanisation: challenges from four rapidly urbanising low-income and middle-income countries. *BMJ Global Health*, 4(3), e001501.

English, M., et al. (2016). Building learning health systems to accelerate research and improve outcomes of clinical care in low- and middle-income countries. *PLOS Medicine*, 13(4), e1001991.

Ezeh, A., et al. (2017). The history, geography, and sociology of slums and the health problems of people who live in slums. *The Lancet*, 389(10068), 547–58.

Federspiel, F., et al. (2018). Global surgical, obstetric, and anesthetic task shifting: a systematic literature review. *Surgery*, 164(3), 553–8.

Fernald, L.C., Gertler, P.J., and Neufeld, L.M. (2008). Role of cash in conditional cash transfer programmes for child health, growth, and development: an analysis of Mexico's Oportunidades. *The Lancet*, 371(9615), 828–37.

Ford, N.D., Patel, S.A., and Narayan, K.M.V. (2017). Obesity in low- and middle-income countries: burden, drivers, and emerging challenges. *Annual Review of Public Health*, **38**, 145–64.

Frenk, J., et al. (2006). Comprehensive reform to improve health system performance in Mexico. *Lancet (London, England)*, 368(9546), 1524–34.

Gill, I. and Kharas, H. (2007) *An East Asian Renaissance: Ideas for Economic Growth.* Washington, DC: World Bank. [Online] Available at: https://openknowledge.worldbank.org/handle/10986/6798 (Accessed: 1 November 2019)

Gilson, L., et al. (2017). Everyday resilience in district health systems: emerging insights from the front lines in Kenya and South Africa. *BMJ Global Health*, 2(2), e000224.

Ginsburg, C., et al. (2018). Association between internal migration and epidemic dynamics: an analysis of cause-specific mortality in Kenya and South Africa using health and demographic surveillance data. *BMC Public Health*, 18(1), 918.

Gosselin, R., et al. (2009). Injuries: the neglected burden in developing countries. *Bulletin of the World Health Organization*, 87(4), 246–46.

Groenewald, P., et al. (2017). Burden of disease. In: Massyn, N., et al. (eds.) *District Health Barometer 2016/2017*. Durban: Health Systems Trust, pp. 206–26.

Grover, A. and Singh, R. (2015). Urban pollution, health, and sustainable development in Delhi. In: Condie, J. and Cooper, A.M. (eds.) *Dialogues of Sustainable Urbanisation: Social Science Research and Transitions to Urban Contexts*. University of Western Sydney.

Health Effects Institute (2019) *State of Global Air 2019. Special Report.* Boston, MA. [Online] Available at: https://www.stateofglobalair.org/report

Hong, P.K. (2009). Health care in Malaysia: the dynamics of provision, financing and access—edited by Chee Heng Leng and Simon Barraclough. *Singapore Journal of Tropical Geography*, 30(1), 144–7.

Hosegood, V., et al. (2007). Revealing the full extent of households' experiences of HIV and AIDS in rural South Africa. *Social Science & Medicine*, 65(6), 1249–59.

Hosseinpoor, A. R., et al. (2016). Data resource profile: WHO Health Equity Monitor (HEM). *International Journal of Epidemiology*, 45(5), 1404–5e.

Hu, J. and O'Neill, L. (2019). Antimicrobial resistance: more quick action is needed in BRICS and MINT economic transition countries. *BMJ Opinion*. [Online] Available at: https://blogs.bmj.com/bmj/2019/10/02/antimicrobial-resistance-more-quick-action-is-needed-in-brics-and-mint-economic-transition-countries/

Institute for Health Metrics and Evaluation (IHME) (2018). *India Profile, IHME.* [Online] Available at: http://www.healthdata.org/india

Institute for Health Metrics and Evaluation (IHME) (2018a). *Brazil Profile, IHME*. [Online] Available at: http://www.healthdata.org/brazil
Institute for Health Metrics and Evaluation (IHME) (2018b). *China Profile, IHME*. [Online] Available at: http://www.healthdata.org/china
Institute for Health Metrics and Evaluation (IHME) (2018c). *Global Burden of Disease (GBD), IHME*. [Online] Available at: http://www.healthdata.org/gbd
Institute for Health Metrics and Evaluation (IHME) (2018d). *Mexico Profile, IHME*. [Online] Available at: http://www.healthdata.org/mexico
Institute for Health Metrics and Evaluation (IHME) (2018e). *Russian Federation Profile, IHME*. [Online] Available at: http://www.healthdata.org/russia
Institute for Health Metrics and Evaluation (IHME) (2018f). *South Africa Profile, IHME*. [Online] Available at: http://www.healthdata.org/south-africa
International Monetary Fund (2000). *Transition Economies: An IMF Perspective on Progress and Prospects*. [Online] Available at: https://www.imf.org/external/np/exr/ib/2000/110300.htm
Klein, E.Y., et al. (2019). Tracking global trends in the effectiveness of antibiotic therapy using the Drug Resistance Index. *BMJ Global Health*, 4(2), e001315.
Kruk, M.E., et al. (2015). What is a resilient health system? Lessons from Ebola. *Lancet (London, England)*, 385(9980), 1910–12.
Kruk, M.E., et al. (2018). High-quality health systems in the sustainable development goals era: time for a revolution. *The Lancet*, 6(11), e1196–252.
Kusuma, D., et al. (2017). New evidence on the impact of large-scale conditional cash transfers on child vaccination rates: the case of a clustered-randomized trial in Indonesia. *World Development*, **98**, 497–505.
Labrique, A.B., et al. (2018). Best practices in scaling digital health in low and middle income countries. *Globalization and Health*, 14(1), 103.
Laxminarayan, R., et al. (2016). Access to effective antimicrobials: a worldwide challenge. *Lancet*, 387(10014), 168–75.
Leon, D.A. and Walt, G. (2000). Poverty, inequality, and mental health in developing countries. In: Leon, D. and Walt, G. (eds) *Poverty, Inequality, and Health*. Oxford: Oxford University Press.
Lewis, T., et al. (2012). E-health in low- and middle-income countries: findings from the Center for Health Market Innovations. *Bulletin of the World Health Organization*, 90(5), 332–40.
Liu, J.X. and Modrek, S. (2016). Evaluation of SMS reminder messages for altering treatment adherence and health seeking perceptions among malaria care-seekers in Nigeria. *Health Policy and Planning*, 31(10), 1374–83.
Lurie, M.N., et al. (2003). Who infects whom? HIV-1 concordance and discordance among migrant and non-migrant couples in South Africa. *AIDS*, 17(15), 2245–52.
M'yambo Magodoro, I., et al. (2019). Engagement in HIV care as an avenue to primary care in rural South Africa: results from the Vukuzazi population health platform. *Open Forum Infectious Diseases*, 6(Supplement_2), S13–S14.
Manne-Goehler, J., et al. (2017). The ART advantage: health care utilization for diabetes and hypertension in rural South Africa. *JAIDS Journal of Acquired Immune Deficiency Syndromes*, 75(5), 561–67.
Manyanga, T., et al. (2014). The prevalence of underweight, overweight, obesity and associated risk factors among school-going adolescents in seven African countries. *BMC Public Health*. BioMed Central, 14(1), 887.
Mason, J.B., et al. (2006). Community health and nutrition programs. In: Jamison, D., et al. (eds.) *Disease Control Priorities in Developing Countries* (2nd ed.). New York: Oxford University Press.
Massuda, A., et al. (2018). The Brazilian health system at crossroads: progress, crisis and resilience. *BMJ Global Health*, 3(4), e000829.
McEwen, B.S. and Seeman, T. (1999). Protective and damaging effects of mediators of stress. Elaborating and testing the concepts of allostasis and allostatic load. *Annals of the New York Academy of Sciences*, **896**, 30–47.
Miller-Petrie, M., Pant, S., and Laxminarayan, R. (2018). Drug resistant infections. In: Holmes, K., et al. (eds.) *Disease Control Priorities (3rd ed.): Volume 6, Major Infectious Diseases*. Washington, DC: World Bank.
Miranda, J.J., et al. (2019). Understanding the rise of cardiometabolic diseases in low- and middle-income countries. *Nature Medicine*, 25(11), 1667–79.
Murthy, R.S. (2017). National Mental Health Survey of India 2015–2016. *Indian Journal of Psychiatry*, 59(1), 21–6.
Ogedegbe, G., et al. (2014). Task shifting interventions for cardiovascular risk reduction in low-income and middle-income countries: a systematic review of randomised controlled trials. *BMJ Open*, 4(10), e005983.
Patel, V., et al. (2018). The *Lancet* Commission on global mental health and sustainable development. *Lancet*, 392(10157), 1553–98.
Patil, R.R. (2014). Urbanization as a determinant of health: a socioepidemiological perspective. *Social Work in Public Health*, 29(4), 335–41.
Patton, G.C., et al. (2016). Our future: a *Lancet* commission on adolescent health and wellbeing. *The Lancet*, 387(10036), 2423–78.
Perez, F. and Villegas, M.V. (2015). The role of surveillance systems in confronting the global crisis of antibiotic-resistant bacteria. *Current Opinion in Infectious Diseases*, 28(4), 375–83.
Preston, S.H. (1975). The changing relationship between mortality and level of economic development. *Population Studies*, 29(2), 231–48.
Pritchett, L. and Summers, L.H. (1993). *Wealthier is Healthier*. 1150. Washington, DC: World Bank. [Online] Available at: http://documents.worldbank.org/curated/en/684651468741004317/pdf/multi0page.pdf
Sánchez-Romero, L.M., et al. (2016). Projected impact of Mexico's sugar-sweetened beverage tax policy on diabetes and cardiovascular disease: a modeling study. *PLOS Medicine*, 13(11), e1002158.
Saxena, A., et al. (2019). The distributional impact of taxing sugar-sweetened beverages: findings from an extended cost-effectiveness analysis in South Africa. *BMJ Global Health*, 4(4), e001317.
Schuster, R.C., et al. (2018). Performance-based financing empowers health workers delivering prevention of vertical transmission of HIV services and decreases desire to leave in Mozambique. *International Journal of Health Policy and Management*, 7(7), 630–44.
Seebregtsi, C., et al. (2016). MomConnect: an exemplar implementation of the Health Normative Standards Framework in South Africa. *South African Heath Review*, pp. 125–36. [Online] Available at: https://www.hst.org.za/publications/South African Health Reviews/11 MomConnect An exemplar implementation of the Health Normative Standards Framework in SA.pdf
Shei, A., et al. (2014). The impact of Brazil's Bolsa Família conditional cash transfer program on children's health care utilization and health outcomes. *BMC International Health and Human Rights*, **14**, 10.
Sum, G., et al. (2019). Implications of multimorbidity patterns on health care utilisation and quality of life in middle-income countries: cross-sectional analysis. *Journal of Global Health*, 9(2), 020413.
Swinburn, B.A., et al. (2019). The global syndemic of obesity, undernutrition, and climate change: the *Lancet* Commission report. *The Lancet*, 393(10173), 791–846.
Tangcharoensathien, V., Thwin, A.A., and Patcharanarumol, W. (2017). Implementing health insurance for migrants, Thailand. *Bulletin of the World Health Organization*, 95(2), 146–51.
Task Force on Fiscal Policy for Health (2019). *Health Taxes to Save Lives: Employing Effective Excise Taxes on Tobacco, Alcohol, and Sugary Beverages*. New York: Bloomberg Philanthropies. [Online] Available at: https://www.bloomberg.org/program/public-health/task-force-fiscal-policy-health
The Academy of Medical Sciences (2018). *Multimorbidity: A Priority for Global Health Research*. [Online] Available at: https://acmedsci.ac.uk/file-download/82222577
The World Bank in Middle Income Countries (2018). [Online] Available at: https://www.worldbank.org/en/country/mic/overview

Thornicroft, G., et al. (2017). Undertreatment of people with major depressive disorder in 21 countries. *The British Journal of Psychiatry*, 210(2), 119–24.
UNDP (2009). *Human Development Report 2009: Overcoming Barriers: Human Mobility and Development.* [Online] Available at: http://hdr.undp.org/sites/default/files/reports/269/hdr_2009_en_complete.pdf
United Nations Department of Economic and Social Affairs Population Division (2017). *International Migration Report 2017: Highlights (ST/ESA/SER.A/404)*. New York. [Online] Available at: https://www.un.org/en/development/desa/population/migration/publications/migrationreport/docs/MigrationReport2017_Highlights.pdf
United Nations Population Fund (n.d.). *Demographic Dividend, United Nations Population Fund.* [Online] Available at: https://www.unfpa.org/demographic-dividend
Watkins, D.A., et al. (2017). *Universal Health Coverage and Essential Packages of Care, Disease Control Priorities: Improving Health and Reducing Poverty*. Washington, DC: The International Bank for Reconstruction and Development/The World Bank.
Watts, N., et al. (2015). Health and climate change: policy responses to protect public health. *The Lancet*, 386(10006), 1861–914.
Watts, N., et al. (2018). The 2018 report of the Lancet Countdown on health and climate change: shaping the health of nations for centuries to come. *The Lancet*, 392(10163), 2479–514.
Webster, J., et al. (2017). South Africa's salt reduction strategy: are we on track, and what lies ahead? *South African Medical Journal*, 107(1), 20–1.
Wei, Z., et al. (2013). Service utilization for mental problems in a metropolitan migrant population in China. *Psychiatric Services*, 64(7), 645–52.
Welaga, P., et al. (2009). Coming home to die? the association between migration and mortality in rural South Africa. *BMC Public Health*. BioMed Central, 9(1), 193.
WHA58.33 (2005). *Sustainable Health Financing, Universal Coverage and Social Health Insurance. Fifty-eighth World Health Assembly*. [Online] Available at: https://apps.who.int/iris/bitstream/handle/10665/20383/WHA58_33-en.pdf?sequence=1
Wittenberg, M., Collinson, M., and Harris, T. (2017). Decomposing changes in household measures: household size and services in South Africa, 1994–2012. *Demographic Research*, **37**, 1297–326.
World Bank (2017). *Drug-Resistant Infections: A Threat to Our Economic Future.* Washington, DC. [Online] Available at: http://documents.worldbank.org/curated/en/455311493396671601/pdf/executive-summary.pdf
World Bank (2019). *World Bank Country and Lending Groups, World Bank Data Help Desk.* [Online] Available at: https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups
World Health Organization (2007). *Taking Stock: Task Shifting to Tackle Health Worker Shortages.* [Online] Available at: https://www.who.int/healthsystems/task_shifting/TTR_tackle.pdf?ua=1
World Health Organization (2008). *Task Shifting: Rational Redistribution of Tasks Among Health Workforce Teams: Global Recommendations and Guidelines*. Geneva: WHO. [Online] Available at: https://www.who.int/healthsystems/TTR-TaskShifting.pdf?ua=1
World Health Organization (2010) *WHO Technical Manual on Tobacco Tax Administration*. Geneva: WHO. [Online] Available at: https://apps.who.int/iris/bitstream/handle/10665/44316/9789241563994_eng.pdf?sequence=1
World Health Organization (2014) *Antimicrobial Resistance: Global Report on Surveillance*. Geneva: WHO. [Online] Available at: https://apps.who.int/iris/bitstream/handle/10665/112642/9789241564748_eng.pdf?sequence=1
World Health Organization (2017). *mDiabetes, An Innovative Programme to Improve the Health of People with Diabetes in Senegal | WHO | Regional Office for Africa*. Geneva: WHO. [Online] Available at: https://www.afro.who.int/news/mdiabetes-innovative-programme-improve-health-people-diabetes-senegal
World Health Organization (2018a). *Global Antimicrobial Resistance Surveillance System (GLASS) Report: Early Implementation 2016–2017*. Geneva: WHO. [Online] Available at: https://www.who.int/glass/resources/publications/early-implementation-report/en/
World Health Organization (2018b). *Technical Series on Primary Health Care: Antimicrobial Resistance and Primary Health Care*. Geneva: WHO. [Online] Available at: https://apps.who.int/iris/bitstream/handle/10665/328084/WHO-HIS-SDS-2018.57-eng.pdf
World Health Organization (n.d., a) *Global Health Observatory data repository: Health Equity Monitor Database*. Geneva: WHO. [Online] Available at: http://apps.who.int/gho/data/node.main.nHE-1540?lang=en
World Health Organization (n.d., b) *SDG 3: Ensure Healthy Lives and Promote Wellbeing for All at All Ages*. Geneva: WHO. [Online] Available at: https://www.who.int/sdg/targets/en/
Zhou, M., et al. (2019). Mortality, morbidity, and risk factors in China and its provinces, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. *The Lancet*, 394(10204), 1145–58.

1.5

Globalization

Kelley Lee

Introduction

‘Globalization’ is a term associated with varied changes to our world, and the net balance of the costs and benefits of these changes to specific individuals and populations remains fiercely debated. There is now substantial scholarship on the links between globalization and public health, explaining how increasing flows of people, other life forms, economic activity, and information worldwide are influencing health determinants and outcomes (Lee 2003b; Labonte et al. 2012; Hanefeld 2015). The greater scale and geographical reach of these flows pose three core challenges for the public health community: how can the evidence base concerning the causal relationship between globalization and health continue to be strengthened; what effective policy responses are needed to optimize globalization’s benefits, and minimize its costs, to population health; and how can these policy options be practically and effectively implemented?

This chapter begins by defining the concept of globalization, and the nature of the changes taking place in the contemporary world as a result. The most prominent feature of globalization—the emergence of a global economy—is then described. The chapter then focuses on how selected public health functions might take account of the changes posed by globalization. Many of the challenges raised in this chapter are addressed in more detail elsewhere in this textbook, reflecting how globalization has come to influence the full spectrum of public health functions. The chapter concludes by considering the prospects for global health governance as the means for strengthening collective action to protect and promote population health. While the public health community is often on the frontline of responding to health risks arising from globalization, it has played only a

limited role to date managing and shaping globalization processes and outcomes. The chapter concludes that addressing the public health impacts of globalization through effective collective action remains essential to the future sustainability of globalization.

What is globalization?

The widespread use of the term ‘globalization’ has been accompanied by variation in its definition. In many cases, the term ‘global’ is used interchangeably with ‘international’ (subjects that concern two or more countries). At other times, globalization is used to refer to specific phenomena such as the spread of fast food restaurants beyond western societies, the worldwide popularity of Hollywood films, or the vast quantities of information digitally available through the internet. Alongside definitional vagueness lie marked differences in how globalization is normatively assessed. Some observers believe globalization as highly desirable for generating economic growth and prosperity for millions (Feachem 2001; Bhalla 2018). Others argue that globalization is a contemporary form of colonialism creating new patterns of inequality within and across countries (Cornea 2001; Labonte et al. 2011). There are contrasting values and assumptions underpinning these starkly different perspectives, resulting in highly contested debates within the research, policy, and advocacy communities.

While it is beyond the scope of this chapter to review the nuances of these debates in detail, an important starting point is to approach the term with a critical lens. In the broadest sense, globalization has become widely understood as the closer integration or interconnectedness of human societies across geographical or territorially-defined boundaries. While this has arguably been occurring over millennia, what is distinct since the late twentieth century has been the vast increases in the types, scale and speed of transboundary flows, and the reach of their consequences. Held et al. (1999) write that it is this greater intensity and extensity of interconnectedness of human societies that distinguish contemporary globalization.

More specifically, the flows associated with globalization can be understood as leading to three types of changes to human connectedness (Lee 2003a). *Spatial change* refers to how people organize and interact across physical or territorial space. The concept of a ‘global village’ derives

from the increased capacity for people to conduct social relations across vast distances because of the greater availability and affordability, for example, of information technologies and long-haul flights. Similarly, the reorganization of the world economy by globalization has enabled ready consumer access to many goods and services produced in faraway locations. This 'death of distance' (Cairncross 2001) includes the formation of new human geographies, such as virtual communities in social media, by which individuals come together across territorial space to interact, exchange, and form identities. In short, the spatial changes resulting from globalization are profoundly shifting how people define, occupy, and navigate geography.

A second, and closely related consequence of globalization, is *temporal change* which concerns how we think about and experience time. The modern world is characterized by an acceleration of the timeframe in which many phenomena occur. For instance, global financial transactions such as foreign exchange, buying and selling equities (stocks and shares), and securing credit can take place in a matter of seconds. New modes of transportation have enabled larger numbers of people to travel greater distances in shorter amounts of time. Mechanization and digitalization have sped up processes of resource extraction, manufacturing, and distribution. The capacity for certain tasks to be done faster, in turn, has driven accelerated change in some parts of society such as business, communications and transportation (Friedman 2012).

Third, globalization is leading to *cognitive change*—how we think about ourselves and the world around us. The production, dissemination, and adoption of new knowledge, ideas, values, and beliefs has become global as a consequence of mass and social media (including the advertising industry), research, and educational institutions, thinktanks, consultancy firms, religious organizations, and political parties. The ascendance of English, as the preferred language for diplomacy, business, and science, is also a reflection of cognitive globalization. Huntingdon (2011) cast such changes within a terrain of epic conflict, a historic 'clash of civilizations' between western and other world views. In contrast, many raise concerns that globalization is leading to western-dominated cultural homogenization, threatening diversity by eradicating local cultures, languages, and belief systems. Others argue that globalization is enabling cultural links across diverse societies to be negotiated, focused on shared ideas and values such as human rights, gender equity, and environmental and labour standards.

Still others see new cultures emerging, a global mélange from processes of hybridization and fusion (Pieterse 2012). Together, spatial, temporal, and cognitive are closely intertwined, creating complex and dynamic impacts on societies worldwide, with both positive and negative consequences.

As well as the nature of the changes caused by globalization, the drivers of globalization have also elicited substantial debate. Mass access to certain technologies, led by information and communication technologies, and transportation, are clear enablers of transboundary flows (Rossignoli et al. 2017). For example, declines in transport costs per ton by sea, due to larger ships, more efficient container management, and consolidation of the industry has been central to the large-scale shift in manufacturing (off-shoring) to lower-cost countries (Wright 2017). Similarly, between 1997 and 2015, the Consumer Price Index for personal computers and peripheral equipment declined by 96% (US Department of Labor 2015).

For some writers, however, technology is the enabler, but not driver, of globalization. They argue that the underlying factors driving such technological changes are economic in nature. The global spread of capitalism has been spurred, on the one hand, by producers seeking cheaper inputs (i.e. raw materials, labour, research and development, transport, and communications), more efficient (and greatest) economies of scale, and ultimately profitable markets. Billions of consumers around the world, on the other hand, fuel this process by demanding higher quality, quantity, and range of goods and services at lower prices. The resulting economic transactions, guided by what eighteenth-century economist Adam Smith called the ‘invisible hand’ of the market, are described as the real force behind globalization (Schuh 2007; Dicken 2015). Globalization, in this sense, is seen as a largely economic and market-driven process of change.

A further perspective rejects globalization as essentially technology or economics-driven process which implies a degree of rationality and progress. Instead, current forms of globalization are arguably driven by particular ideologically-based values and beliefs broadly referred to as neoliberalism. The global spread and dominance of this ideology from the 1980s, and its embedding within global institutional arrangements has, for instance, defined the industrial policies facilitating the development of such technologies (e.g. the promotion of an information economy through deregulation and privatization of the telecommunications sector), and their dissemination for particular purposes (e.g. deregulation of financial

markets). Neoliberalism has also defined economic policies which encourage trade liberalization, market-based competition, and foreign direct investment (e.g. tax incentives), while advocating a minimal role for the state at the expense of social welfare and environmental protections (Slobodian 2018).

Differing perspectives about the drivers of globalization, in turn, underpin varying beliefs about whether, on balance, global changes are leading to net costs or benefits to people and societies. 'Globalists' (supporters of contemporary globalization) anticipate positive benefits overall—a world of shared identities, greater efficiency and productivity, economic growth, and rising prosperity. While a period of transition is recognized, during which accrued benefits may be unevenly distributed, including temporary inequalities in wealth within and across countries, globalists believe that the globalization process is progressive in the longer term, bringing the greatest benefits to the greatest numbers of people. In sharp contrast, opponents hold that there are fundamental flaws underpinning the market-driven logic of contemporary globalization. This logic is producing stark inequalities in wealth and power which are neither temporary nor just. The result is a world increasingly structurally skewed in favour of a relative minority (e.g. those with access to technology, education, capital, and gainful employment) at the expense of a majority disproportionately bearing the costs of globalization (e.g. those with insecure or unsafe employment, low wages, low education). Thus, while globalization may be generating large-scale wealth, critics challenge the assumption that this new wealth will eventually 'trickle down' the global pecking order. Without a strong commitment to redistributive policies, and meaningful social and environmental protections, neo-liberal globalization is seen as widening the gap between the 'haves' and 'have nots' (Mittelman 2002). The greater concentration of the world's wealth in ever fewer hands—the richest 1% increased their share of the world's wealth between 2008 and 2016 from 42.5% to 50.1% (Credit Suisse Research Institute 2017)—is seen as evidence of the structural inequities created by contemporary globalization.

From a public health perspective, there is substantial evidence that globalization is creating diverse and complex changes to health determinants, resulting in both positive and negative health outcomes for specific individuals and populations. Policies to minimize the costs to and maximize the benefits for health, as well as measures to ensure that these

costs and benefits to public health are equitably distributed within and across societies, are arguably central to the longer-term sustainability of globalization (Bacchetta and Jansen 2011). Towards the development of effective policy measures, a fuller understanding of the public health impacts of globalization is required.

The public health implications of a global economy

Economic globalization is characterized by the steady restructuring of production, distribution, and consumption into systems organized on a worldwide scale. While economists disagree about the precise timing of this restructuring, most agree that this has led to a fundamental transformation of the world economy by the latter part of the twentieth century. Key features of this 'global shift' have been the liberalization of capital flows; rise of transnational corporations in major industrial sectors; and the emergence and spread of global production networks. Each of these has implications for public health determinants and outcomes.

The first major pillar of the globalized world economy, the liberalization of capital flows, was introduced in the United States from the mid-1970s. Historically, banks, insurance companies, investment companies, and brokerage firms have been subject to heavy regulation. Deregulation of financial markets, such as the removal of restrictions on the types of securities that financial institutions could trade, levels of interest that could be paid on specific types of securities and bank accounts, and types of institution entitled to act as financial intermediaries, was introduced to increase competition and encourage capital flows within and across countries. Information and communication technologies enabled high-speed electronic-based transactions which eventually linked financial markets across countries. The result has been a globally integrated financial market capable of 24-hour trading in foreign exchange, commodities, capital, and other financial assets (Valdez 2006). The world's financial assets have correspondingly boomed, from US $136 trillion in 2005 to US $294 trillion in 2014 (Witkowski 2015). By 2020, this is expected to rise to US $371 trillion, with one-third accounted for by emerging economies (Roxburgh et al. 2011).

A second key pillar is the increased international trade of goods and services. While trade has been the lifeblood of commerce for thousands of

years, it is the global restructuring and integration of production, accompanied by the increased scale and scope of trade, which characterizes the world economy since the late twentieth century. Historically, trade among countries has been dominated by raw materials and natural resources (e.g. oil, timber), commodities (e.g. grains, metals), or manufactured products (e.g. textiles and clothing, food products). In 1948, the General Agreement on Tariffs and Trade (GATT) was established as an agreement under which signatories could negotiate reductions in tariffs on traded goods. Eight trade rounds were carried out between 1948 and 1994 by which thousands of tariff reductions were negotiated. During this period, membership grew from 23 countries in 1947 to 125 countries in 1994, thus establishing a worldwide trading system.

The growth of international trade led to renewed support for a permanent organization, resulting in the creation of the World Trade Organization (WTO) in 1995. As of 2018 there are 164 member states and 23 observer states, many of which are seeking accession. Moreover, the WTO's remit embraces trade in services, agriculture, intellectual property rights, government procurement, and other areas (Wilkinson 2014). The overall effect has been that world trade grew faster than world output. Between 2006 and 2016, for example, world merchandise exports increased from US $8 to US $11 trillion (37.5%), and commercial services exports from US $2.9 to US $4.8 trillion (66%) (WTO 2017a), between 1.1 and 2 times faster than global gross domestic production (GDP). The global financial crisis beginning in 2008 led to a significant downturn in trade volumes, with trade eventually growing slower than global GDP in 2016. Nevertheless, since 2017 the overall trend in world trade has once again been growth (WTO 2017b; WTO 2018).

The global restructuring of production and exchange processes has been an integral part of this boom in international trade. In many sectors, transnational corporations (TNCs) have emerged which have relocated components of their business to different parts of the world. Thus, resource extraction may take place in one country, manufacturing in another, and research and marketing in still others (Dicken 2015). The targeted consumers have also changed, with TNCs seeking markets worldwide for global brands. According to the European Brand Institute (2017), the ten most valuable global brands are Apple, Google, Microsoft, Coca Cola, AT&T, Amazon, Johnson & Johnson, China Mobile, Verizon, and Facebook.

Global economic restructuring can be observed in many industrial sectors posing distal or proximal impacts on health determinants and outcomes. One important example is the pharmaceutical industry which has become dominated globally by large companies formed through mergers and acquisitions. This includes the buying of domestic generic and manufacturing companies to access emerging markets, and the consolidation of the pharmaceutical and biotechnology industries. Table 1.5.1 describes the largest pharmaceutical companies in 2016 by revenue. In 1992, the ten largest companies accounted for one-third of global revenue (Tarabusi and Vickery 1998). By 2016, this had increased to 40%. This consolidation of the market has been driven by intense competition, the high cost of research and development (R&D), economies of large-scale manufacturing, and the need for capacity to market and distribute worldwide (Wright 2017). For example, the reported cost of developing and bringing a new prescription drug to the market is now around US $2.6 billion, a 145% increase since 2003 (Mullin 2014). Success in the pharmaceutical sector amid globalization requires large companies with global reach and access to considerable resources.

Table 1.5.1 World's ten largest pharmaceutical companies (2016)

Company	HQ	Revenue (millions US$)	Growth over previous year (%)
Pfizer	United States	52,824	+8
Novartis	Switzerland	48,518	-2
Merck	United States	39,807	+1
Roche	Switzerland	38,369	+5
GlaxoSmithKline	United Kingdom	37,929	+7
Sanofi	France	35,633	-1
Johnson & Johnson	United States	33,464	+7
Gilead	United States	30,390	-7
Abbvie	United States	25,638	+12
AstroZeneca	United Kingdom	23,002	-7

Source: data from Wright, T., Top 20 Pharma and BioPharma Report, *Contract Pharma*, 25 July 2017, Copyright © Rodman Media. All Rights Reserved. Available from https://www.contractpharma.com/issues/2017-07-01/view_features/top-25-pharma-and-biopharma-report-476409.

The alcoholic beverages industry, worth an estimated US $645 billion in 2017 and projected to reach US $1,594 billion by 2022 (Allied Market Research 2017), is another health-related example of the global restructuring of the world economy. As market saturation and increased regulation may moderate growth in North America, Europe, and Australia/New Zealand, the industry has turned to emerging markets notably in Asia (Esser and Jernigan 2018). The ability to compete globally, in turn, is prompting mergers and acquisitions, and the increasing domination in parts of the industry by transnational alcohol companies (TACs). For example, beer manufacturing is led by Anheuser-Busch InBev, Heineken and Carlsberg which produced 38% of global supply in 2016. The spirits and wine sectors are less concentrated to date, but further globalization of production, distribution, and consumption is expected to be spurred by intensifying competition for

new markets, new product development and marketing, and trade and investment liberalization (Mitchell and Casben 2016).

The direct public health consequences of economic globalization arise from changes in the production, distribution, and consumption of either health promoting or health harming products and services. Where a product is inherently health harming, such as tobacco, an increase in risk is posed to public health. This is not withstanding the use of non-combustible nicotine products such as electronic cigarettes, increasingly produced and marketed by the tobacco industry itself, to potentially mitigate although not remove such risks if used exclusively instead (National Academies of Sciences, Engineering and Medicine 2018). For other potential health harming products, such as alcoholic beverages, food and drink, the level of risk depends on the specific product or amount consumed. Globalization, in this case, may create a mixture of benefits and harms. For example, many products of the pharmaceutical industry are essential to healthcare worldwide. The increased production and affordability of such products, through improved economies of scale to serve larger markets, increased spending on R&D, and innovations in distribution would accrue benefits to health. However, the competitive pressures caused by economic globalization can also lead 'Big Pharma' to unduly focus on the most profitable products and markets to recover investment or maximize returns for shareholders. This can skew production decisions away from diseases and populations deemed to offer insufficient financial returns (Trouiller et al. 2002), and allow companies to price their products beyond the affordability of many who need them. This lack of access to patent-protected medicines has led to ongoing tensions worldwide between ensuring an economic return for Big Pharma and enabling equitable access to improve health outcomes. The Treatment Action Campaign initiated in South Africa in 1998 by HIV/AIDS activists, which has pressed the government to make various drug treatments more readily available in the country (Heywood 2009), and the licensing of generic drug production by the Thai government despite industry pressure to enforce patent protections under the TRIPS agreement (Yamabhai et al. 2011), illustrate this tension between private and public interests.

More indirectly, the global economy can adversely impact human health by enabling conditions that pose new risks. These risks can arise in several ways. First, economic globalization has encouraged unprecedented

population mobility. Population movements are not new—humankind has been on the move since *Homo erectus* migrated out of Africa in about 1 million BC. What characterizes population mobility since the late twentieth century has been its volume, geographical reach, speed, and frequency. Globalization is what Castles and Miller (2009) call the 'age of migration'. In 2017, the number of international migrants reached 258 million people, an increase from 173 million in 2000. By 2017, 244 million people lived outside their country of birth, compared with 156 million in 1990 (UNDESA 2017). The reasons why people migrate are diverse. For the relatively educated, highly skilled, and mobile, economic globalization offers new training and employment opportunities. For the lower educated and skilled, work opportunities in places of 'global employment' may beckon, such as the offshore factories of TNCs, fields of cash crops grown by the agri-food industry, or the extraction sites of the natural resources industries. Changing patterns of population mobility and settlement may contribute benefits to health through, for example, providing new training opportunities to health workers or enabling vulnerable populations to move from locations posing risks. Increased population mobility may also generate new risks by, for example, heightening the risk of communicable disease outbreaks where refugee settlements do not have adequate housing, water, and sanitation; facilitating the spread of communicable disease outbreaks globally via long distance travellers; or weakening healthcare systems in low and middle-income countries (LMICs) through the depletion of domestic healthcare workers.

Second, economic globalization can increase environmental health concerns, ranging from local, small-scale factors (e.g. household exposures) to widespread exposures affecting whole populations. The possible impacts of globalization on environmental health are wide-ranging, potentially affecting populations at the global, regional, community, occupational, and household levels. These impacts can be direct, such as through the dumping of hazardous waste across borders, relocation of risky occupations (such as ship breaking or asbestos disposal), and damage caused by cross-border pollution or contamination. Economic globalization, in which companies compete on a worldwide scale to increase returns, is raising concerns that profit-maximization motives can undermine environmental health. Governments, seeking to attract foreign direct investment or unleash market competition to fuel economic growth and generate employment, are accused

of engaging in a 'race to the bottom' by offering reduced tax rates, lower wages, or weaker environmental, labour, and health and safety protections (Mosley and Uno 2007). A change in investment or lending policy at the global level can also have health consequences for local communities. For example, policy conditions set by the World Bank, or requirements to repay substantial sums of foreign debt, can restrict the public expenditure of borrowing countries on environmental protection or investment in basic infrastructure, such as water and sanitation. A cholera outbreak in South Africa in 2000–2001, which led to around 120,000 cases and 265 deaths, has been blamed on the introduction of user fees following the privatization of water utilities as part of the country's structural adjustment programme. As water supplies were cut to poor people unable to pay for the new charges, many resorted to using polluted river water (Anonymous 2006). The World Bank (2000) argues that '[e]very society has to decide for itself on the relative value it places on economic output and the environment'. Others argue, however, that there are substantial inequities in terms of who bears the environmental costs of globalization (Wiedmann and Lenzen 2018).

Third, the rapid growth, open nature, and regulatory challenges of the global economy have led to a substantial growth in illegal products, many of which pose risks to human health. The illicit tobacco trade, for instance, is believed to account for 9 per cent of retail sales, increasing access to lower-priced and unregulated cigarettes (WHO n.d.). Around 5 per cent of the global population aged 15–64 have used illicit drugs at least once in 2015. The production, distribution, and consumption of such drugs, since the late twentieth century, have been transformed by transnational networks of criminal organizations (UN Office on Drugs and Crime 2017). A related problem is the global business of counterfeiting which was worth US $1.2 trillion in 2017 (Research and Markets 2018). Many counterfeits are health-related or affecting products, notably pharmaceuticals, and food and beverage products. WHO estimates that one in ten medical products circulating in low- and middle-income countries is either substandard or falsified (WHO 2017a). Counterfeit baby milk powder containing melamine reportedly caused 50 child deaths and acute malnutrition in more than one hundred others in rural China (Watts 2004). Deaths from counterfeit alcohol products have led to hundreds of deaths in countries worldwide including Indonesia (Faridz and Griffiths 2018), Russia (Reevell 2016), Ireland (McNamee 2016), and Mexico (ACS Sensors 2018). Economic globalization

has flourished through the creation of special economic zones, global financial systems, worldwide transportation networks, and internet sales. The same systems have been harnessed by criminal organizations to produce and distribute illicit goods with health harming effects. As discussed next, in relation to global health governance, within efforts to strengthen the regulatory frameworks managing globalization, there is much need for increased measures to protect and promote population health.

Globalization and essential public health functions

Public health is concerned with protecting and promoting the health of whole populations within and across countries. Since the late twentieth century, there have been efforts in many countries to reinvigorate public health systems. The Institute of Medicine, for example, initiated debate in the United States through its reports, *The Future of Public Health* (Institute of Medicine 1988), and *The Future of the Public's Health in the 21st Century* (Institute of Medicine 2002), with the latter identifying the need to address 'emerging threats' caused by rapid changes to the nation's social, cultural, and global contexts, scientific and technological advances, and demographic shifts (Institute of Medicine 2002). Given an inadequate capacity to response to the SARS (severe, acute respiratory syndrome) in 2003–2004, the Canadian government formed the Public Health Agency of Canada (PHAC). Correspondingly, the creation of the European Centre for Disease Control (ECDC) in 2004 was to strengthen the capacity of European Union countries to respond to infectious disease outbreaks. Over the past decade, there has also been similar investments to build public health systems in such countries as India (Planning Commission of India 2011), Brazil (Jerberg 2008), and China (Liang and Burns 2017).

While disease outbreaks prompted renewed attention to public health systems in many countries, WHO and national policy makers now recognize the need to adapt a fuller range of essential public health functions (EPHFs) to meet the challenges of a globalizing world (Martin-Moreno et al. 2016). As described earlier, spatial, temporal, and cognitive changes arising from globalization are having important impacts on a broad range of population health determinants and outcomes. The remainder of this chapter describes how globalization has implications for three EPHFs identified by the Pan American Health Organization (PAHO 2002)—monitoring, evaluation, and

analysis of health status; public health surveillance; and health promotion—and how the public health community might effectively adapt to these impacts (see Box 1.5.1).

Box 1.5.1 Essential public health functions

EPHF 1 Monitoring, Evaluation, and Analysis of Health Status
EPHF 2 Public Health Surveillance, Research, and Control of Risks and Threats to Public Health
EPHF 3 Health Promotion
EPHF 4 Social Participation in Health
EPHF 5 Development of Policies and Institutional Capacity for Planning and Management in Public Health
EPHF 6 Strengthening of Institutional Capacity for Regulation and Enforcement in Public Health
EPHF 7 Evaluation and Promotion of Equitable Access to Necessary Health Services
EPHF 8 Human Resources Development and Training in Public Health
EPHF 9 Quality Assurance in Personal and Population-based Health Services
EPHF 10 Research in Public Health
EPHF 11 Reduction of the Impact of Emergencies and Disasters on Health

Source: data from PAHO (2002). Essential Public Health Functions. In *Public Health in the Americas*. Washington DC: Pan American Health Organization, p. 67.

Monitoring, evaluation, and analysis of health status

Assessing the health of a given population is the starting point for a wide range of public health actions including priority setting, resource allocation, the design and implementation of interventions, and programme evaluation. There are well-recognized steps for assessing population health status:

Define the purpose of the assessment.
Define the population concerned and any comparator population.
Define the aspects of health to be considered.
Identify and review existing data sources.
Select the most appropriate existing data (Gentle 2001).

However, assessing population health status within the context of globalization presents two key challenges. First, there remains variable

capacity within and across countries to collect and manage the essential health information needed. In many LMICs, in particular, inadequate health information systems results in data of variable quality or incomplete regarding selected populations and health conditions. According to WHO (2006a):

> A country health information system comprises the multiple subsystems and data sources that together contribute to generating health information, including vital registration, censuses and surveys, disease surveillance and response, service statistics and health management information, financial data, and resource tracking. The absence of consensus on the relative strengths, usefulness, feasibility, and cost-efficiency of different data collection approaches has resulted in a plethora of separate and often overlapping systems.

To address this problem, the Health Metrics Network (HMN) was formed in 2005 as a global partnership hosted by WHO to advocate for the strengthening of health information systems in LMICs. Funded by the Gates Foundation and other donors, the partnership was comprised of governments, multilateral and bilateral development agencies, foundations, global health initiatives, and technical experts that aimed to increase the availability and use of timely, reliable health information. In 2013, the WHO Executive Board decided to discontinue the HMN because the network was deemed to have ‘fulfilled its time-limited and catalytic role’ of ‘raising the profile of health information systems, helping countries to assess and strengthen their systems and accelerating improvements to country civil registration and vital statistics systems’ (WHO 2013). The focus has turned to harnessing the capacity of global information and communication systems for a wide range of health applications including monitoring, evaluation, and analysis of health status (de Fatima Marin et al. 2017). While improvements in health information systems have been achieved across many LMICs, including the innovative application of mobile technologies, resource constraints continue to pose the largest problem.

A second challenge is the limitations of existing health information systems for capturing health status and needs that do not conform with, or cut across, national boundaries. For each country, health data is collected and managed by local and national health departments and associated institutions. The WHO Statistical Information System (SIS), now part of the Global Health Observatory (GHO), in turn, collates this data to produce aggregated health indicators, mortality and health status, disease statistics, health system statistics, risk factors and health service coverage, and

inequities in health from member states. These data are compiled in the *World Health Statistics*. Governments, donors, and other organizations may use these data to identify health priorities.

The impacts of globalization on health, however, are by definition transcendent of national borders such that new health and disease patterns may be emerging that may not be captured by aggregated or disaggregated national data. A good example is the complex pattern of rising overweight and obesity rates. In high-income countries, obesity is rising rapidly across all socioeconomic groups but is particularly associated with social deprivation. In the United States, for example:

> CSD [community socioeconomic deprivation] was associated with higher BMI [body mass index] at average age [of study cohort] (10.7 years) and with more rapid growth of BMI over time. Children born in communities with greater CSD had steeper increases of BMI at younger ages. Those born into the poorest communities displayed sustained accelerated BMI growth. CSD remained associated with BMI trajectories after adjustment for a measure of household socioeconomic deprivation. (Nau et al. 2014)

In France, Romon et al. (2005) found that genetic predisposition influences the prevalence of obesity and changes in BMI among children from the higher social class. For children within the lowest social class, which has seen an increase in BMI across the whole population, environmental factors appear to have played a more important role. In LMICs, the total number of obese or overweight people has grown rapidly in such diverse countries as Kenya, China, India, and Brazil, notably among a rising middle class. This includes a 10-fold increase in obese children and adolescents in LMICs (NCD Risk Factor Collaboration 2017). At the same time, many urban populations within the lower social classes are experiencing rising rates of overweight and obesity, now exceeding rates of undernutrition (Ford et al. 2017). Trends in over/undernutrition, in other words, are complex and dynamic, requiring detailed and comparable data across population groups defined along variables such as age, gender, socioeconomic status, geography, and occupation.

The challenges of monitoring, evaluating, and analysing health status, within the context of globalization, can be illustrated by a wide range of other examples. The emergence of a global economy dominated by TNCs, for example, has restructured employment around global value chains as leading modes of production, distribution, and consumption. This has led to the employment of millions of factory workers throughout the world, creating

new sources of employment, but often under conditions of weak social and environmental protections (Phillips 2017). How might we measure the changing health status of these workers over time and place? What public health needs do these workers share across domestic jurisdictions which arise, not from their nationality, but from their conditions of employment? The unprecedented movement of people across national borders also requires appropriate ways to understand the health status and needs of varied types of migrants such as tourists, refugees, overseas workers, and trafficked people. Some patterns of migration will not be captured by aggregated national data, thus obscuring the public health needs arising for home and host populations. Similarly, the public health impacts on population health from climate change need to use measures beyond national data given the planetary scale of the environmental effects taking place. These examples illustrate the need for methods that identify new population groups arising from globalization's impacts which do not conform to national borders, and enhanced health information systems which generate corresponding data to monitor, evaluate, and analyse the health status of these populations.

Public health surveillance, research, and control of risks and threats to public health

The just-described need for genuinely global health data, beyond aggregated national level data, underpins the core public health function of preventing, controlling, and responding to public health risks. Traditionally, this function has focused on disease outbreaks and epidemics. Historically, risks from disease-causing microbes have long mirrored the migration and settlement patterns of human species. The shift from hunting-gathering to agrarian societies and the domestication of animals, between 8000 and 3500 BC, for example, brought new risks from zoonotic diseases (Swabe 1999). The so-called 'Columbian exchange', following the arrival of Christopher Columbus in the Americas, is noted for the widespread exchange of plants, animals, foods, human populations (including settlers and slaves), and ideas that followed. Diseases, such as cholera, influenza, measles, malaria, smallpox, and tuberculosis, were also exchanged, resulting in the decimation of Indigenous peoples (Crosby 1972). The influenza pandemic of 1918–1919, which killed around 25 million people worldwide, was undoubtedly

enabled by the large-scale population movements and devastation caused by the First World War. The changing structures of human societies over time and place, in short, give rise to particular health risks.

In this context, contemporary globalization can be understood as altering the structure of human societies once again in ways that pose new risks to human health. First, globalization may alter the *population size* at risk due to patterns of settlement. The world's urban population will almost double, from 3.3 billion in 2007 to 6.3 billion in 2050, with most of this growth in cities within the developing world. Rapid urbanization has been accompanied, in many cases, by poverty, inadequate housing, limited access to clean water and sanitation, or basic healthcare. This has created larger populations now vulnerable to infectious disease risks (Alirol et al. 2010). One of the factors attributed to the unprecedented magnitude of the Ebola virus outbreak in West Africa during 2014–2015, for instance, is that the continent has had the world's highest urban growth rate for 20 years. Rapid urbanization, prompted in large part by war and poverty, may explain how Ebola shifted from relatively small outbreaks at village level to a much larger outbreak putting millions at risk (Mackenzie et al. 2014).

Moreover, globalization has increased the size of populations at risk due to far greater mobility of human and animal species. Diseases have a greater capacity to spread more rapidly and further afield in a more interconnected world. Some infections, such as tuberculosis and cholera, are historically associated with poverty but may now pose new patterns of risk due to globalization (Gushulak and MacPherson 2004). United Nations peacekeepers from Nepal arriving in earthquake-affected Haiti in 2010 are alleged to have caused a devastating cholera outbreak within an already highly vulnerable population (Gladstone 2017). In other ways, globalization is 'democratizing' disease risks. The spread of SARS in 2002–2003, from China to around 25 countries within weeks, posed a threat to populations rich or poor in a world of migrant workers, business travellers, and long-haul flights. The human health risks potentially arising from highly pathogenic pandemic influenza would be truly global by virtue of the widespread lack of immunity, and rapid person-to-person transmission facilitated by highly mobile human populations. Further concerns relate to the global spread of antimicrobial drug-resistant organisms and drug-resistant infectious diseases by mobile human populations. As MacPherson et al. (2009) write: 'Current population movement dynamics rapidly and effectively link regions of

marked health disparity, and these linkages can be associated with risk for importation of drug-resistant infectious diseases.'

Second, globalization may influence the *prevalence* (number of people with a disease within a given population) and *incidence* (number of new cases of a disease in a specified period of time per total population) of certain diseases. De Vogli and Birbeck (2005) argue that vulnerability to HIV/AIDS, tuberculosis, and malaria is closely linked to poverty, gender inequality, development policy, and health sector reforms that involve user fees and reduced access to care. Dorling et al. (2006) similarly argue that 'global inequality in wealth will have compounded the effects of AIDS on Africa', notably caused by globalization-related policies such as structural adjustment programmes. Changing employment patterns and conditions, arising from economic globalization, have also had epidemiological consequences. Moody (2007) describes the globalization of the mining industry and its impact on local communities. Basu et al. (2009) observe how tuberculosis incidence among southern African miners is ten times greater than the general population, with migration to and from mines amplifying tuberculosis epidemics in the general population.

Third, globalization is increasing the *risks from emerging and re-emerging diseases*. Environmental degradation, such as rainforest destruction and intense farming methods, can bring human and animal populations into closer contact. Sixty per cent (60%) of emerging infectious diseases that affect humans are zoonotic (originating in animals), and more than two-thirds of those originate in wildlife including HIV and AIDS, West Nile, Ebola virus, Lyme disease and possibly SARS-CoV-2. The detection and rapid spread of a novel coronavirus, the Middle East Respiratory Syndrome coronavirus (MERS-CoV), from Saudi Arabia to South Korea in 2012, illustrated what Gannon (2015) calls 'a general archetype for emerging infectious diseases: zoonotic in origin, carried across continents by an infected air traveller, initially undiagnosed, bringing fear and economic disruption, and spreading most quickly through vulnerable populations'. Moreover, evidence by Smith et al. (2014) suggest that 'the total number and diversity of outbreaks, and richness of causal diseases increased significantly' from 1980 to 2013.

This heightening of public health risks within a globalizing world has prompted efforts to improve capacities in surveillance, research, and control measures. Globalization offers opportunities, for instance, to harness

information and communication technologies to enable faster, cheaper, and more efficient gathering and worldwide sharing of knowledge. ProMED-mail (see Box 1.5.2) and regional disease surveillance networks have facilitated the collection and reporting of epidemiological data. The lessons learned from SARS prompted renewed WHO efforts to revise the International Health Regulations (IHR), which came into effect in 2007, including the harnessing of a broader range of information sources. The expanded scope of the IHR beyond a handful of named diseases, to the broader concept of 'public health emergencies of international concern' also reflected recognition of the need for collective action on a more diverse range of global health risks. Global technologies are also enhancing research capacities in data analysis, such as genomics, and the use of so-called 'big data' source including the innovative harnessing of non-health global scale data. For example, the use of airline travel data to model transmission dynamics and potential impact of the West African Ebola virus outbreak provided lessons for enhancing real-time epidemiological assessment in future outbreaks via improved data collection and analyses (Cori et al. 2017). There remain critical challenges to harnessing big data, to improve the capacity of countries to respond to global public health risks, namely better understanding of what big data is most useful, how to accurately interpret such data for public health purposes, and how to protect individual privacy and the proprietary interests of data holders (Lazer and Kennedy 2015).

Box 1.5.2 Programme for Monitoring Emerging Diseases (ProMED-mail)

ProMED—the Program for Monitoring Emerging Diseases—is an internet-based reporting system dedicated to rapid global dissemination of information on outbreaks of infectious diseases and acute exposures to toxins that affect human health, including those in animals and in plants grown for food or animal feed. Electronic communications enable ProMED to provide up-to-date and reliable news about threats to human, animal, and food plant health around the world, 7 days a week. By providing early warning of outbreaks of emerging and re-emerging diseases, public health precautions at all levels can be taken in a timely manner to prevent epidemic transmission and to save lives.

ProMED is open to all sources and free of political constraints. Sources of information include media reports, official reports, online summaries, local observers, and others. Reports are often contributed by ProMED subscribers. A team of expert human, plant, and animal disease moderators screen, review, and investigate reports before posting to the network. Reports are distributed by email

to direct subscribers and posted immediately on the ProMED web site. ProMED currently reaches over 70,000 subscribers in at least 185 countries.

Reproduced with permission from the International Society for Infectious Diseases, *About ProMED-mail*, Copyright © 2018 International Society for Infectious Diseases. All rights reserved. Available from http://www.promedmail.org/aboutus/.

Global health promotion

There is now substantial evidence of a marked rise in the global burden of non-communicable diseases (NCDs) since the 1980s. This includes an almost quadrupling of the total number of adults worldwide with Type 2 diabetes (NCD Risk Factor Collaboration 2016) since 1980, and a tripling of the rate of obesity since 1975 (WHO 2017b). In 2017, NCDs killed 40 million people worldwide and accounted for 70% of all deaths. Cardiovascular diseases accounted for most NCD deaths (17.7 million people annually), followed by cancers (8.8 million), respiratory diseases (3.9 million), and diabetes (1.6 million) (WHO 2017c). WHO predicts that, while infectious diseases like malaria, tuberculosis, and HIV/AIDS will largely decline, cancers and ischaemic heart disease, as well as cerebrovascular disease, will continue to steadily rise from 2015 to 2030 (WHO 2018a).
The clarity of evidence about this rapid rise in NCDs across societies at all levels of development, and within populations spanning varying levels of socioeconomic status, has been attributed in part to changes associated with globalization. As Gostin (2014) writes:

> Rapid travel, mass migration and the globalization of culture are known to fuel the spread of infectious diseases. The same factors are also increasing the incidence of non-communicable diseases (NCDs) such as cancer, diabetes and heart disease. As developing countries prosper, these conditions are a by-product of rising air pollution, physical inactivity, and consumption of alcohol, tobacco and excess calories.

This association, in turn, has prompted greater efforts to strengthen collective action on NCDs including global health promotion. Health promotion is the process of enabling people to increase control over, and to improve, their health (WHO 1986). Interventions address three areas of activity to prevent disease and promote the health of a community: communication, service delivery, and enabling factors (individual and structural) which influence the chance of an intention being carried out and a behaviour sustained. Individual enabling factors include communication

skills and literacy; structural enabling factors include social supports, income level and cultural attitudes (Walley et al. 2001, p. 147). A global approach to health promotion takes account of the ways in which globalization may be influencing the social determinants of health, and in turn health behaviours, as well as, offers opportunities for providing appropriate interventions. The Bangkok Charter for Health Promotion in a Globalized World adopted in 2005 recognized the major challenges posed by globalization, and the actions and commitments needed to address the determinants of health in a globalized world by reaching out to people, groups, and organizations that are critical to the achievement of health. To galvanize concerted action by governments, three UN General Assembly High-Level Meetings have been held on the prevention and control of NCDs. The first adopted a Political Declaration on NCDs (2011), and the second followed with an Outcome Document on NCDs (2014), with the target of reducing premature NCD mortality by 25% by 2025 (WHO 2018b). However, the virtual absence of heads of state at both meetings was a disappointment for the NCD Alliance (largely composed of civil society organizations). A 2017 progress report revealed 'uneven and insufficient' progress in most countries in meeting globally agreed targets including a substantial number of countries without NCD national plans and targets (Ravelo 2017). To address the lack of political will, and propose 'bold and practical actions steps to scale up prevention and treatment of NCDs and provide health for all' (WHO 2018c), WHO appointed an Independent High-Level Commission on Noncommunicable Diseases, co-chaired by the heads of state of Uruguay, Sri Lanka, Finland, Russia, and Pakistan. The challenges faced by the Commission in its deliberations, and for societies worldwide seeking to tackle NCDs, is well-illustrated by two of the leading risk factors—obesity and tobacco use.

The rapid increase in overnutrition and obesity, among adults and children, has attracted substantial public health concern in recent years (Fig. 1.5.1). According to WHO (2017b), around 1.9 billion people worldwide are overweight, of which around 650 million people worldwide are obese (BMI >30). At least 340 million school-age children are overweight or obese. The number of overweight or obese infants and young children (aged 0 to 5 years) increased from 32 million globally in 1990 to 41 million in 2016, and expected to rise to 70 million by 2025 (WHO 2017d). Importantly, while obesity rates are higher in wealthier countries than LMICs, these changes

span populations in diverse settings and income levels, including such wide-ranging countries as India, Nigeria, Mexico, and China. The association to globalization of obesity and overweight populations began to be identified in the 1990s. For example, Drewnoski and Popkin (1997, p. 32) write:

> Whereas high-fat diets and Western eating habits were once restricted to the rich industrialized nations . . . the nutrition transition now occurs in nations with much lower levels of gross national product (GNP) than previously . . . First, fat consumption is less dependent on GNP than ever before. Second, rapid urbanization has a major influence in accelerating the nutrition transition.

Prentice (2006, p. 93) attribute these trends to the globalization of food production and lifestyles:

> The pandemic is transmitted through the vectors of subsidized agriculture and multinational companies providing cheap, highly refined fats, oils, and carbohydrates, labour-saving mechanized devices, affordable motorized transport, and the seductions of sedentary pastimes such as television. This trend has been linked to the globalisation of sedentary lifestyles alongside changes in food production and consumption.

Beyond anecdotal and descriptive evidence, Goryakin et al. (2015) undertook extensive econometric analyses of several datasets, using a series of new proxies for different dimensions (social, political, and economic) of globalization potentially affecting overweight in up to 887,000 women aged 15–49 living in 56 countries between 1991 and 2009. They found that both political and especially social globalization dimensions appear strongly positively related to the greater overweight risk, while the association with economic globalization countries is dependent on ‘country-specific factors such as economic development and infrastructure, education, attractiveness of economies to investors, as well as the size of the market’ Goryakin et al. (2015, p. 74). Overall, this evidence supports the hypothesis that the multitude of yet interrelated changes in diet, levels of physical activity, and the built environment associated with globalization are substantial and significant explanations for the sharp rise in obesity and overweight worldwide since the 1980s.

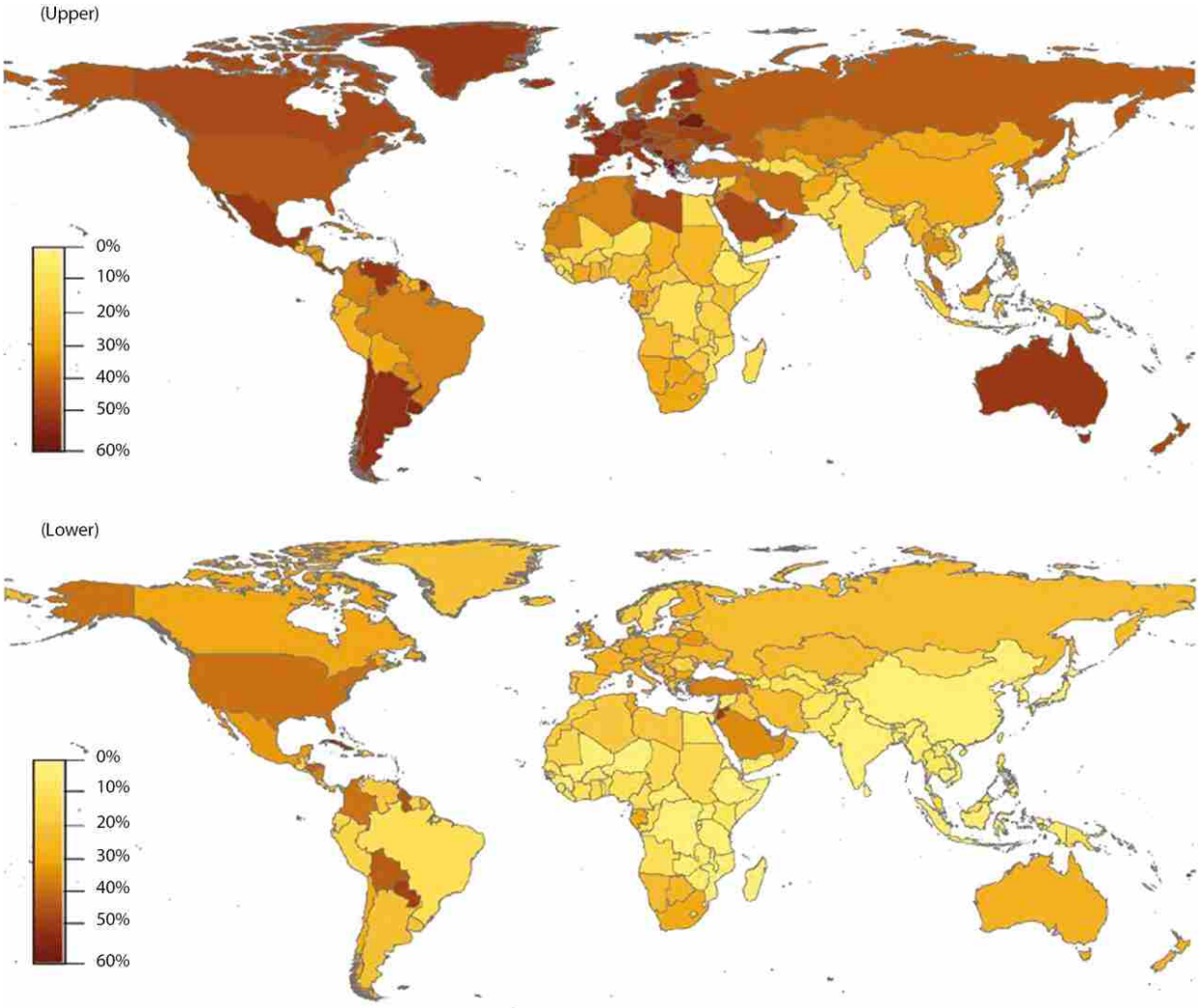


Fig. 1.5.1 Worldwide age-standardized prevalence of overweight (upper) and obesity (lower) in adults 20 years and older by country in 2005.

Reproduced with permission from Kelly, T. Global burden of obesity in 2005 and projections to 2030. *International Journal of Obesity*, 9(32). Copyright © Springer Nature. Available at https://www.nature.com/articles/ijo2008102

The few measures adopted to address so-called 'globesity' (Costa-Font and Mas 2016) to date have focused on modifying individual behaviours or lifestyles. These measures emphasize public health education led by healthy eating initiatives and the promotion of increased physical activity. However, more comprehensive approaches to global health promotion are increasingly recognized as necessary. This includes addressing the structural factors that constrain or enable healthy eating, and adopting both supply and demand-side measures concerning what and how food is produced, marketed, and consumed. For example, the Institute of Medicine (2005) report, *Food*

Marketing to Children and Youth: Threat or Opportunity? recognizes that dietary patterns begin in childhood and are shaped by the interplay of genetics and biology, culture, and values, economics, physical and social environments, and commercial media environments. The report includes a comprehensive review of the scientific evidence on the influence of food marketing on diets and diet-related health of children and youth. It concludes that environments supportive of good health require leadership and action from the food, beverage, and restaurant industries; food retailers and trade associations; the entertainment industry and the media; parents and caregivers; schools; and governments. Prior to the Second UN High-level Meeting on NCDs in 2014, WHO Director-General Margaret Chan stated, 'The globalization of unhealthy lifestyles is by no means just a technical issue for public health. It is a political issue. It is a trade issue. And it is an issue for foreign affairs' (WHO 2014). Similarly, following its initial technical consultation in 2018, the WHO Independent High-level Commission on NCDs has argued for a whole-of-society approach (WHO 2018b).

The regulation of the agri-food industry amid globalization, however, remains a major challenge. The concentration of ownership in the food industry, to be dominated by a small number of large transnational food companies, has been achieved by greater economies of scale, lower-cost production methods, and heavy marketing of high-profit foods. This has led to a substantial increase in the production and consumption of highly processed foods which contain industrially-produced ingredients such as high-fructose corn syrup and palm oil. The marked rise in hypertension, heart disease, and type 2 diabetes worldwide, coinciding with the ready availability and affordability of food products produced from such ingredients, continues to be studied (Bray 2013; Ismail et al. 2018). In the meantime, debates about the regulation of such ingredients have intensified. Many public health advocates call for governments to adopt improved and mandatory labelling, and even restrictions on their use. At the global level, WHO called for the adoption of a Global Strategy on Diet, Physical Activity, and Health. However, these efforts were hindered by the agri-food industry's opposition to explicit recommendations on healthy levels of salt, sugar, and fats intake, and subsequently given no regulatory 'teeth' by member states (Lang et al. 2006; Johnson 2012). Instead, the agri-food industry has promoted voluntary self-regulation such as limits on the targeted marketing

of processed food products at children and corporate social responsibility initiatives. The industry has also emphasized the importance of protecting consumer choice, and the primary importance of non-food factors such as physical activity as a health concern. However, Lang et al. (2006) assessed compliance by 25 of the world's largest food manufacturers and retailers to recommendations concerning ingredients, advertising, portion size, and labelling. Only a small proportion of companies voluntarily reduced salt, sugar, and fats in their products. In Spain, León-Flández et al. (2017) found a decline in compliance with the Spanish code of food and drinks advertising directed at children under 12 since 2008. In the US, as Kelly (2005) advises:

> So far the U.S. government has declined to regulate the aggressive ways in which food producers market high-energy, low-nutrition foods to young people. That public-health responsibility has been left to an industry-created scheme of self-regulation that is deeply flawed; there is a compelling need for government involvement.

Similar lessons for global health promotion can be drawn from tobacco control.

Like the agri-food industry, the tobacco industry has globally restructured and is now dominated by a small number of large TNCs. Industry consolidation has been achieved, for example, by the establishment of global value chains, development of 'global drive brands', and intense marketing in emerging markets (Lee et al. 2016). As WHO Director-General Gro Harlem Brundtland (2000, p. 2) stated, in initiating the Framework Convention on Tobacco Control (FCTC) negotiation process in the 1990s:

> [O]ver the past fifteen years, we have seen that modern technology, has limited the effectiveness of national action. Tobacco advertising is beamed into every country via satellite and cable. Developing countries are the subject of massive marketing campaigns by international tobacco companies. In the slipstream of increasing global trade, new markets are opened to international tobacco companies which see these emerging markets as their main opportunity to compensate for stagnant or dwindling markets in many industrialized countries.

The negotiations faced intense efforts by the tobacco industry to limit stronger regulation by participating in some national delegations, lobbying leaf-producing countries and farmers, orchestrating criticism of WHO, and even challenging the science on tobacco and health (Wipfli 2015). Despite industry opposition, the FCTC was signed in 2003 and came into effect in 2005. As of 2018, 181 countries are parties to the treaty. Importantly, the FCTC remains the most prominent example of collective action across WHO member states to address a significant public health problem worsened by

globalization. This required participation by a broad spectrum of national, regional, and global stakeholders. The mobilization of public health advocates worldwide, led by the Framework Convention Alliance, was greatly facilitated by the use of the internet (focused on the Globalink network) which enabled stakeholders to follow negotiations, organize advocacy activities, and share experiences (Collin et al. 2004). While the final treaty was weaker on some key challenges posed by globalization than initially hoped, such as setting aside tobacco from international trade agreements, the possible negotiation of additional protocols was to address these transboundary issues. To date, only the FCTC Protocol to Eliminate Illicit Trade in Tobacco Products has been signed, and will come into effect when ratified by 40 states parties (WHO 2015).

In summary, globalization has coincided with the rise of NCDs worldwide across diverse populations, posing new challenges for public health. In many cases, powerful vested interests within key sectors of the world economy oppose stronger action to prevent and control NCDs. Initiating collective action, such as the WHO Global Action Plan for the Prevention and Control of NCDs 2013–2020, can serve to make NCDs a higher priority. However, past experience shows that international agreements are only the starting point. Effective implementation of commitments requires sustained mobilization of political will, resources, and technical capacity to translate strategies into results (Ravelo 2018). A key lesson from efforts to date to address the risk factors of obesity and tobacco use in some countries, is that globalization needs to be better harnessed for health promotion. Worldwide consumer boycotts of TNCs, such as Nestlé and McDonalds, have been organized via social media by civil society organizations to pressure companies to change their practices (Swinburn et al. 2015). The public health community has also drawn attention to the dangers of illicit products through global campaigns (Fig. 1.5.2). These examples of collective action across societies, to harness globalization to counter continued high rates of tobacco use, obesity, and other risk factors behind the rising trends in NCDs, offer important examples to build upon.

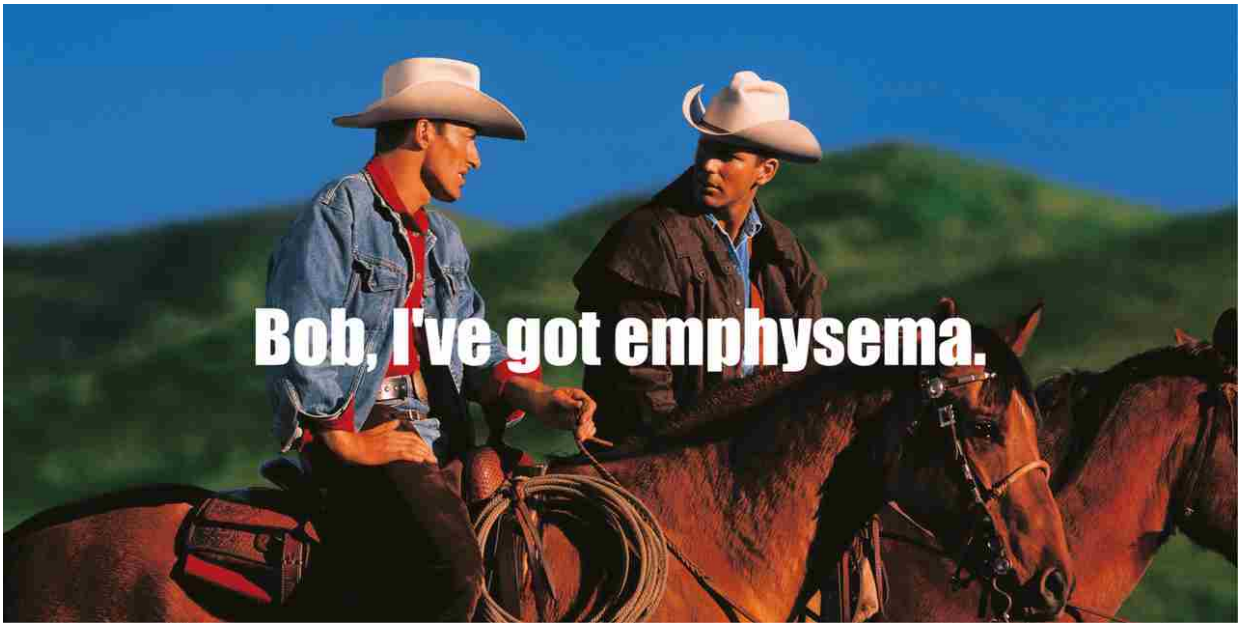


Fig. 1.5.2 World No Tobacco Day 2015.
From https://blogs.worldbank.org/health/world-no-tobacco-day-2015-illicit-trade-and-taxes

Harnessing global information and communication technologies for public health

A defining and enabling feature of contemporary globalization has been rapid advances in information and communication technologies (ICTs), spurring the transformation of societies worldwide. From the personal computer in the 1980s, the progressive development of ICT devices with ever faster processing speeds and larger storage capacity has enabled a wide range of applications at affordable prices. Investments in ICTs not only generate economic impacts, they served as core infrastructure for the global economy as a whole. Global ICT networks underpin global supply chains, transportation systems, 24-hour financial trading and vital information networks (Baller et al. 2016).

The globalization of ICTs may lead to profound changes in the health sector including research, diagnostics, patient care, planning and management, informatics, and health promotion. Often known as eHealth (or electronic health), technological change has been spurred by the increased complexity of health systems, need to manage ever larger and more complex datasets, rising cost of healthcare, need to provide longitudinal and comprehensive health records for individual patients, and increased mobility

and/or geographically dispersal of many patient populations and healthcare providers. ICTs have been identified as particularly effective, for example, at reaching more individuals with effective health-related advice and information at a very low cost (Bloom et al. 2017).

The use of mobile and wireless technologies in health (known as mHealth), in particular, has grown rapidly with advances in technologies and applications (or 'apps'), the declining cost of handsets and data transmission, and growth cellular network coverage. According to the International Telecommunication Union (ITU), subscriptions grew 20% annually from 2012 to 2017, reaching 4.3 billion globally by 2018 (ITU 2017). LMICs have seen the highest rate of growth where the potential for mHealth to improve healthcare, hindered in the past by a lack of infrastructure, is most promising (Silva et al. 2015).

Social media, 'interactive platforms via which individuals and communities create and share user-generated content' (Kietzmann et al. 2011, p. 241), has also seen the blossoming of health-related applications. A systematic review of social media and public health messages finds that, as more consumers turn to the internet for health-related information, health organizations have begun to use social media for connecting with the public For example, health insurance companies are applying social media, health gaming apps, and other mobile apps to increase patient engagement in their own healthcare. Evidence suggests considerable reach associated with social media applications, and the potential for engaging specific target audiences, although efficacy of public health messaging requires further investigation through controlled studies (Ventola 2014).

Overall, ICTs offer much promise to enhancing core public health functions. There continues to be substantial investment in such technologies by the public and private sectors, and their application to a wide range of health applications. However, systematic reviews of the use of eHealth technologies in behavioural interventions (Oosterveen et al. 2017) and mental health (Stratton et al. 2017), for instance, cast doubt on the longer-term efficacy and cost-effectiveness of using such technologies.

Importantly, there has also been an increased use of ICTs for commercial purposes to shape consumer choices and behaviours. The online marketing, promotion and distribution of health-affecting products and services has spread rapidly and, in most cases, remain weakly regulated. The global reach of direct-to-consumer marketing of pharmaceuticals (genuine and

counterfeit), supplements, food and beverages, and tobacco products (including electronic nicotine delivery systems) via the internet has been a particular source of regulatory concern (Freeman et al. 2014). For national authorities, the challenge of regulating the internet arises from the global nature of ICTs. How to strengthen global regulatory mechanisms, notably in countries where such goods originate, has been the focus of growing policy debate (Institute of Medicine 2012). Collective action to regulate the global use of ICTs, to protect and promote population health, remains much needed.

Strengthening the governance of global public health

From international to global health governance

Governance broadly concerns the agreed actions, rules, and means adopted by a society to promote collective action and solutions in pursuit of common goals. Governance takes place whenever people seek to organize themselves to achieve a shared end through agreed rules, procedures, and institutions. This can take place at different levels of decision-making and action. In public health, if a local community decides to initiate a campaign to slow traffic speed and improve road safety, this requires some form of governance to organize the effort. If a national policy is adopted to improve food labelling and increase the taxation of unhealthy foods, governance is needed to set the agreed labelling standards and rates of taxation, as well as enforce compliance. If a global campaign is initiated to strengthen tuberculosis control, an agreed form of governance is needed to take decisions, for example, on agreed treatment, resource mobilization, and implementation of agreed actions across countries. To what extent is there need for more effective governance to protect and promote public health within an increasingly globalized context?

The existing institutional arrangements for collective action among countries have historical roots dating from the International Sanitary Conferences of the nineteenth century. This series of meetings, largely dominated by European countries and the US, focused on protecting their economic and political interests from epidemic diseases such as cholera and plague. The institutions eventually created, such as the *Office International d'Hygiène Publique*, were primarily concerned with collecting and disseminating epidemiological data on these diseases (Fidler 2001). The

creation of the WHO as a specialized agency of the United Nations in 1948 was intended to universalize membership, and broaden the scope of, international health cooperation. Its objective, 'the attainment by all peoples of the highest possible level of health', was reflected in the vast array of programmes initiated under WHO's auspices (Lee 2008).

Recent decades have seen challenges to WHO's designated role as 'the directing and co-ordinating authority on international health work' (WHO Constitution, Article 2(a)). In part, this has arisen from differences in perspective on whether WHO should be biomedically (disease) focused, or whether the organization should address the broad social determinants of 'health for all'. These debates have been accompanied by rapid changes in WHO's operating environment, with the ascendance of new institutional players with often competing resources, technical expertise, and normative perspectives (McInnes and Lee 2012). From the 1980s, the World Bank became a major influence as the biggest source of financing for health development (Buse 1994), while other UN bodies such as the UN Children's Fund and UN Development Programme (UNDP) expanded their health portfolios. The 1990s saw the creation of numerous global public–private partnerships for health which attracted additional resources, but rendered the policy environment far more complex and crowded (Buse and Walt 2000). Charitable foundations, led by the Bill and Melinda Gates Foundation, have since become the biggest players in the funding of health development (Birn 2014). This influx of institutions and resources has, on the one hand, reflected the higher priority to health development given by governments, corporations, and civil society organizations (CSOs). On the other hand, there is substantial evidence of overlapping mandates, duplication of effort, and, above all, a lack of consensus about how to effectively tackle the collective public health challenges posed by globalization (Gostin 2017).

The governance of global public health can thus be understood as undergoing a period of transition, from international (state-based) to global (state and non-state-based) collective action. In principle, leading institutions formally responsible for governing public health remain governmental. Public health authority lies within the ministries of health of each WHO member state, and collaboration across countries and regions is expected to take place on a wide range of functions through governmental bodies. However, as described in this chapter, intensified flows of people, other life forms, goods and services, capital, and knowledge are creating

transboundary health determinants and outcomes that can transcend state capacity to manage them. This is due, in part, to differences in perspective about what priority goals and actions should be pursued, the capacity of governments to act, and even a willingness to act collectively. Where governments cannot or will not assume a global health governance role, non-state actors in the form of CSOs, charitable foundations, and even corporations have been looked to as potentially filling this governance gap. Many scholars of global governance point to the need for new institutional arrangements to enable more effective action in the face of urgent need. Examples described in this chapter—such as tobacco control, obesity, and infectious disease outbreaks—demonstrate the need for innovation but also the political challenges posed.

Strengthening legal and regulatory frameworks for global health

Globalization is now an established subject of discussion and debate at public health meetings around the world, and there has been no shortage of commitments to addressing its impacts. Efforts to do so through stronger regulatory frameworks have yielded mixed results. The FCTC is seen as a successful example of WHO member states coming together to address a major global public health problem through the creation of a legal instrument requiring signatories to adopt a minimum package of measures. The revision of the IHR in 2005 was also an important step towards the creation of legal frameworks for global health groups (GHGs). Collective action was furthered on access to medicines through the 2001 Doha Declaration on Intellectual Property Rights and Public Health, followed in 2003 by the WTO's Decision on Implementation of Paragraph 6, both intended to clarify flexibilities in the Agreement on Trade-Related Intellectual Property Rights permitting member states to take measures, such as compulsory licensing, that improves access to medicines (Kerry and Lee 2007). The Report of the WHO Commission on Intellectual Property Rights, Innovation and Public Health (WHO 2006b) sought to address the tensions between economic and health equity goals by reviewing 'current incentive and funding regimes, including intellectual property rights, designed to

stimulate the creation of new medicines and other products, and facilitate access to them' (p. iv).

Beyond the health sector, however, to what extent is there recognition of the need to govern globalization differently? The foremost challenge, perhaps, remains the governance of the world economy in ways that are ecologically sustainable, and bring the most benefits to the most people. The global financial crisis beginning in 2007 revealed fundamental regulatory weaknesses—the availability of easy credit, incentives to undertake high-risk lending and borrowing, poor fiscal management by governments, and rampant speculative investing. The reforms that followed have sought to 'fix' selected parts of the global financial system. However, critics argue that the parts addressed are those which enable the 'haves' in the global economy to have more. The further concentration of global wealth in ever fewer hands —1% of the world's population owned 50% of global wealth by 2017 (an increase from 42.5% in 2008)—sharply questions the social, political, and environmental sustainability of current forms of economic globalization (Credit Suisse 2017). The rise in populist movements worldwide, many opposing globalization, reflect this discontent.

In relation to public health, more fundamental reforms in the governance of the global economy are required which recognize the profound consequences of insecure employment, stark inequalities in wealth, large-scale corporate tax evasion, and environmental degradation on human health. As heads of state, ministers of trade and investment, and corporate executives ponder the future of economic globalization, the public health community can seek to play a bigger role in creating the necessary checks and balances for the global economy to function sustainably. Public health representation in policy debates on trade and finance to date has been limited. WHO, public health institutions, and civil society organizations have remained at the margins of key decision-making forums such as the WTO, World Economic Forum and Group of Seven (G7) meetings. The substantial concerns by citizens and consumers worldwide about the social and environmental footprints of global products and services, the looming economic time bomb of NCDs, and the potential costs to the global economy of a highly lethal infectious disease pandemic may be leverage points for finding common ground between public health and economic policy. A global public goods for health approach, for example, addresses globalization and health from an economic perspective. The concept

identifies where a good or service (such as knowledge of an infectious disease outbreak), which would be of benefit globally, will not be produced or disseminated if left to the market because no one can be excluded from accessing the good, and thus no charge can be levied for its use and no costs recouped. There may even be economic disincentives to produce certain goods, such as outbreak reports, because of resultant negative economic consequences. At the national level, the production of these goods is usually ensured by government intervention, but at the global level there remains no 'global government' to undertake this role. Certain functions of public health may be classed as global public goods (e.g. immunization programmes, disease surveillance), which require collective action to overcome market failures (Smith et al. 2005). The concept might thus be an appealing rationale to non-health policymakers shaping global change. The potential role of health impact assessment for informing policy proposals is another such approach (Friedman and Burci 2017). Most importantly, the long-term sustainability of globalization requires public health to be a core part of any efforts to build better global governance.

Conclusion

This chapter has described key ways in which globalization is relevant to the theory and practice of public health. There are both risks and opportunities arising from the complex and diverse changes being created by globalization, with current forms of market-dominated globalization characterized by an inequitable distribution of winners and losers. For the public health community, there is a need to understand and contribute to more effective management of the rapid changes taking place. Greater attention to these public health impacts will, in turn, contribute to more sustainable forms of globalization.

REFERENCES

ACS Sensors (2018). 'Sniffing' out counterfeit liquors. *ACS News*, 10 January. [Online] Available at: https://www.acs.org/content/acs/en/pressroom/presspacs/2018/acs-presspac-january-10-2018/sniffing-out-counterfeit-liquors.html

Alirol, E., Laurent, G., Stoll, B., Chappuis, F., and Loutan L. (2010). Urbanisation and infectious diseases in a globalised world. *The Lancet*, **10**, 131–41.

Allied Market Research (2017). *Alcohol Beverages Market Expected to Reach $1,594 Billion, Globally, by 2022—Allied Market Research*. PR Newswire, 5 April. [Online] Available at: https://www.prnewswire.com/news-releases/alcoholic-beverages-market-expected-to-reach-1594-billion-globally-by-2022---allied-market-research-618354513.html

Anonymous (2006). Report: water problems remain in rural areas. *Mail & Guardian*, 12 July.

Bacchetta, M. and Jansen, M. (2011). *Making Globalization Socially Sustainable*. WTO/ILO, Geneva. [Online] Available at: http://www.wto.org/english/res_e/booksp_e/glob_soc_sus_e.pdf

Baller, S., Dutta, S., and Lanvin B. (eds.) (2016). *The Global Information Technology Report 2016*. Geneva: World Economic Forum. [Online] Available at: http://www3.weforum.org/docs/GITR2016/WEF_GITR_Full_Report.pdf

Basu, S., Stuckler, D., Gonsalves, G., and Lurie, M. (2009). The production of consumption: addressing the impact of mineral mining on tuberculosis in southern Africa. *Globalization and Health*, 5(11), 1–8.

Bhalla, S. (2018). *The New Wealth of Nations*. Delhi: S&S India.

Birn, A.E. (2014). Philanthrocapitalism, past and present: the Rockefeller Foundation, the Gates Foundation, and the setting(s) or the international/global health agenda. *Hypothesis*, 12(1), e8–e34.

Bray, G.A. (2013). Potential health risks from beverages containing fructose found in sugar or high-fructose corn syrup. *Diabetes Care*, 36(1), 11–12.

Brundtland, G.H. (2000). *Opening Statement*. First meeting of Intergovernmental Negotiating Body, Framework Convention on Tobacco Control, Geneva, 16 October. [Online] Available at: http://www.who.int/director-general/speeches/2000/english/20001016_tobacco_control.html

Buse, K. (1994). Spotlight on international organizations: the World Bank. *Health Policy and Planning*, **9**, 95–9.

Buse, K., and Walt, G. (2000). Global public–private partnerships: part I—a new development in health? *Bulletin of the World Health Organization*, 78(4), 549–61.

Cairncross, F. (2001). *The Death of Distance: How the Communications Revolution Is Changing Our Lives*. Cambridge, MA: Harvard Business Press.

Castles, S., and Miller, M. (2009). *The Age of Migration, International Population Movements in the Modern World* (4th ed.). London: Macmillan.

China (2009). *Deepening the Reform of Health-care System*. Beijing: Central Committee of the Communist Party of China and the State Council.

Collin, J., Lee, K., and Bissell, K. (2004). Negotiating the framework convention on tobacco control: the politics of global health governance. In: Wilkinson, R. and Murphy C. (eds.) *The Global Governance Reader*, pp. 254–73. London: Routledge.

Cori, A., et al. (2017). Key data for outbreak evaluation: building on the Ebola experience. *Philosophical Transactions of the Royal Society B*, 372(1721), 20160371.

Cornea, G.A. (2001). Globalization and health: results and options. *Bulletin of the World Health Organization*, 79(9), 834–41.

Costa-Font, J. and Mas, N. (2016). 'Globesity'? The effects of globalization on obesity and caloric intake. *Food Policy*, **64**, 121–32.

Credit Suisse Research Institute (2017). *Global Wealth Report 2017*. Geneva. [Online] Available at: https://www.credit-suisse.com/corporate/en/articles/news-and-expertise/global-wealth-report-2017-201711.html

Crosby, A. (1972). *The Columbian Exchange: Biological and Cultural Consequences of 1492*. Westport, CT: Greenwood Press.

De Vogli, R. and Birbeck, G.L. (2005). Potential impact of adjustment policies on vulnerability of women and children to HIV/AIDS in sub-Saharan Africa. *Journal of Health Population and Nutrition*, **23**, 105–20.

De Fatima Marin, H., Massad, E., Gutierrez, M.A., Rodrigues, R.J., and Sigulem D (2016). *Global Health Informatics: How Information Technology Can Change Our Lives in a Globalized World*.

London: Academic Press.
Dicken, P. (2015). *Global Shift, Mapping the Contours of the World Economy* (7th ed.). London: Guilford Press.
Dorling, D., Shaw, M., and Davey Smith, G. (2006). Global inequality of life expectancy due to AIDS. *BMJ*, **332**, 662–64.
Drewnoski, A. and Popkin, B. (1997). The nutrition transition: new trends in the global diet. *Nutrition Reviews*, 55(2), 31–43.
Esser, M.B. and Jernigan, D.H. (2018). Policy approaches for regulating alcohol marketing in a global context: a public health perspective. *Annual Review of Public Health*, **39**, 385–401.
European Brand Institute. *Global Top 100 Brand Corporations 2017*. Vienna: European Brand Institute. [Online] Available at: https://www.rankingthebrands.com/PDF/Global%20Top%20100%20Brand%20Corporations%202017,%20European%20Brand%20Institute.pdf
Faridz, D. and Griffiths, J. (2018). More than 80 dead from drinking bootleg alcohol in Indonesia. *CNN*, 12 April. [Online] Available at: https://www.cnn.com/2018/04/12/health/indonesia-fake-alcohol-intl/index.html
Feachem, R.G.A. (2001). Globalization is good for your health, mainly. *BMJ*, 323(7311), 504–6.
Fidler, D. (2001). The globalization of public health: the first 100 years of international health diplomacy. *Bulletin of the World Health Organization*, 79(9), 842–9.
Ford, N., Patel, S., and Narayan, K. (2017). Obesity in low- and middle-income countries: burden, drivers, and emerging challenges. *Annual Review of Public Health*, **38**, 145–64.
Freeman, B., et al. (2014). Digital junk: food and beverage marketing on Facebook. *American Journal of Public Health*, 104(12), e56–e64.
Friedman, E. and Burci, G.L. (2017). Taking health in all policies seriously: health impact assessments, WHO leadership, and international law. *Lancet Global Health*, 5 September. [Online] Available at: https://marlin-prod.literatumonline.com/pb-assets/Lancet/langlo/TLGH_Blogs_2013-2018.pdf
Friedman, T. (2012). *The Lexus and the Olive Tree, Understanding Globalization* (2nd ed.). New York: Picador.
Gannon, S. (2015). What MERS shows us about emerging diseases. *Centre for Strategic & International Studies*, 12 June. [Online] Available at: https://www.csis.org/blogs/smart-global-health/what-mers-shows-us-about-emerging-diseases
Gentle, P. (2001). Assessing health status. In: Pencheon, D., Guest, C., Melzer, D., and Muir Gray J.A. (eds.) *Oxford Handbook of Public Health Practice*, pp. 28–30. Oxford: Oxford University Press.
Gladstone, R. (2017). U.N. brought cholera to Haiti: now it is fumbling its effort to atone. *The New York Times*, 26 June. [Online] Available at: https://www.nytimes.com/2017/06/26/world/americas/cholera-haiti-united-nations-peacekeepers-yemen.html
Goryakin, Y., Lobstein, T., James W.P.T., and Suhrcke, M. (2015). The impact of economic, political and social globalization on overweight and obesity in the 56 low and middle income countries. *Social Science & Medicine*, **133**, 67–76.
Gostin, L.O. (2014). Non-communicable diseases: healthy living needs global governance. *Nature*, 511(7508), 147–9.
Gostin, L.O. (2017). The World Health Organization's historic moment of peril and promise: reimaging a global health agency fit for purpose in the 21st century. *Global Health Governance*, 30 May. [Online] Available at: http://blogs.shu.edu/ghg/2017/05/30/commentary-the-world-health-organizations-historic-moment-of-peril-and-promise-reimaging-a-global-health-agency-fit-for-purpose-in-the-21st-century/
Gushulak, B.D. and MacPherson, D.W. (2004). Globalization of infectious diseases: the impact of migration. *Clinical Infectious Diseases*, 38(12), 1742–8.

Hanefeld, J. (ed.) (2015). *Globalization and Health* (2nd ed.). Maidenhead: Open University Press.
Held, D., McGrew, A., Goldblatt, D., and Perraton, J. (1999). *Global Transformations*. Stanford, CA: Stanford University Press.
Heywood, M. (2009). South Africa's Treatment Action Campaign: Combining Law and Social Mobilization to Realize the Right to Health. *Journal of Human Rights Practice*, 1(1), 14–26.
Huntingdon, S. (2011). *The Clash of Civilizations and Remaking of World Order*. New York: Simon and Schuster.
IMAP (2010). *Pharmaceuticals and Biotech Industry Global Report—2011*. [Online] Available at: http://www.imap.com/imap/media/resources/IMAP_PharmaReport_8_272B8752E0FB3.pdf
Institute of Medicine (1988). *The Future of Public Health*. Washington, DC: National Academy of Sciences.
Institute of Medicine (2002). *The Future of the Public's Health in the 21st Century*. Washington, DC: National Academy of Sciences.
Institute of Medicine (2005). *Food Marketing to Children and Youth: Threat or Opportunity?* Washington, DC: National Academy of Sciences.
Institute of Medicine (2012). *Ensuring Safe Foods and Medical Products Through Stronger Regulatory Systems Abroad*. Washington, DC: National Academy of Sciences.
Ismail, S.R., Maarof, S.K., Ali, S.S., and Ali, A. (2018). Systematic review of palm oil consumption and the risk of cardiovascular disease. *PLoS One*, 13(2), e0193533.
ITU (2017). ITU releases 2017 global information and communication technology facts and figures. *ITU News*, 31 July. [Online] Available at: http://news.itu.int/itu-releases-2017-global-information-and-communication-technology-facts-and-figures/
Jerberg, C. (2008). Flawed but fair: Brazil's health system reaches out to the poor. *Bulletin of the World Health Organization*, 86(4), 248–9.
Johnson, T. (2012). *UN High-Level Meeting on NCDs: One Year Later, Interview with Thomas Bollyky, Council on Foreign Relations*, 19 September. [Online] Available at: http://www.cfr.org/global-health/un-high-level-meeting-ncds-one-year-later/p29090
Kelly, D. (2005). To quell obesity, who should regulate food marketing to children? *Globalization and Health*, 1(9). [Online] Available at: http://www.globalizationandhealth.com/content/1/1/9
Kerry, V.B. and Lee, K. (2007). TRIPS, the Doha declaration and paragraph 6 decision: what are the remaining steps for protecting access to medicines? *Globalization and Health*, 3(3), 1–12.
Kietzmann, J.H., Hermkens, K., McCarthy, I.P., and Silvestre, B.S. (2011). Social media? Get serious! Understanding the functional building blocks of social media. *Business Horizons*, 54(3), 241–51.
Labonte, R., Mohindra K., and Schrecker, T. (2011). The growing impact of globalization for health and public health practice. *Annual Review of Public Health*, **32**, 263–83.
Labonte, R., Schrecker, T., Packer, C., and Runnels, V. (2012). *Globalization and Health: Pathways, Evidence and Policy*. London: Routledge.
Lang, T., Rayner, G., and Kaelin, E. (2006). *The Food Industry, Diet, Physical Activity and Health: A Review of Reported Commitments and Practice of 25 of the World's Largest Food Companies*. London: Centre for Food Policy, City University.
Lazer, D., and Kennedy, R. (2015). What we can learn from the epic failure of Google Flu Trends. *Wired*, 1 October. [Online] Available at: https://www.wired.com/2015/10/can-learn-epic-failure-google-flu-trends/
Lee, K. (2003a). *Globalization and Health: An Introduction*. London: Palgrave Macmillan.
Lee K. (ed.) (2003b). *Health Impacts of Globalization: Towards Global Governance*. London: Palgrave Macmillan.
Lee, K. (2008). *The World Health Organization*. London: Routledge.
Lee, K., Eckhardt, J., Holden, C. (2016). Tobacco industry globalization and global health governance: towards an interdisciplinary research agenda. *Palgrave Communications*, **2**, 16037. [Online] Available at: https://www.nature.com/articles/palcomms201637.epdf

León-Flández, K., et al. (2017). Evaluation of compliance with the Spanish Code of self-regulation of food and drinks advertising directed at children under the age of 12 years in Spain, 2012. *Public Health*, **150**, 121–9.
Liang, X.F., and Burns, L.R. (2017). China's public health system and infrastructure. In: Burns, L.R., and Liu, G.G. (eds.) *China's Healthcare System and Reform*. Cambridge: Cambridge University Press, pp. 75–116.
Martin-Moreno, J., Harris, M., Jakubowski, E., and Kluge, H. (2016). Defining and assessing public health functions: a global analysis. *Annual Review of Public Health*, **37**, 335–55.
McInnes, C. and Lee, K. (2012). *Global Health and International Relations*. Oxford: Polity.
Mackenzie, D., Skettt, P., and Wilson, C. (2014). Deadliest Ebola outbreak being driven by urbanisation. *New Scientist*, 6 August. [Online] Available at: https://www.newscientist.com/article/mg22329812-500-deadliest-ebola-outbreak-being-driven-by-urbanisation/
McNamee, G. (2016). Counterfeit alcohol has been linked to a number of suspicious deaths in Ireland. *The Journal*, 19 November. [Online] Available at: http://www.thejournal.ie/counterfeit-alcohol-revenue-3086874-Nov2016/
MacPherson, D., et al. (2009). Population mobility, globalization, and antimicrobial drug resistance. *Emerging Infectious Diseases*, 15(11), 1727–37.
Mitchell, A.D. and Casben, J. (2016). Trade law and alcohol regulation: what role for a global Alcohol Marketing Code? *Addiction*, 112(S1), 109–16.
Mittelman, J.H. (2002). Making globalization work for the have nots. *International Journal on World Peace*, 19(2), 3–25.
Moody, R. (2007). *Rocks and Hard Places: The Globalization of Mining*. New York: Zed Books.
Mosley, L. and Uno, S. (2007). Racing to the bottom or climbing to the top? Economic globalization and collective labor rights. *Comparative Political Studies*, 40(8), 923–48.
Mullin, R. (2014). Cost to develop new pharmaceutical drug now exceeds $2.5b. *Scientific American*, 24 November. [Online] Available at: https://www.scientificamerican.com/article/cost-to-develop-new-pharmaceutical-drug-now-exceeds-2-5b/
National Academies of Sciences, Engineering, and Medicine (2018). *Public Health Consequences of E-Cigarettes*. Washington, DC: The National Academies Press.
Nau, C., et al. (2014). Community socioeconomic deprivation and obesity trajectories in children using electronic health records. *Obesity*, 23(1), 207–12.
NCD Risk Factor Collaboration (2016). Worldwide trends in diabetes since 1980: a pooled analysis of 751 population-based studies with 4.4 million participants. *Lancet*, 387(10027), 1513–30.
NCD Risk Factor Collaboration (2017). Worldwide trends in body-mass index, underweight, overweight, and obesity from 1975 to 2016: a pooled analysis of 2416 population-based measurement studies in 128.9 million children, adolescents, and adults. *Lancet*, **390**, 2627–42.
Oosterveen, E., Tzelepis, F., Ashton, L., and Hutchesson, M.J. (2017). A systematic review of eHealth behavioural interventions targeting smoking, nutrition, alcohol, physical activity and/or obesity for young adults. *Preventive Medicine*, **99**, 197–206.
PAHO (2002). Essential Public Health Functions. In: *Public Health in the Americas*. Washington DC: Pan American Health Organization. [Online] Available at: http://www1.paho.org/hq/dmdocuments/2010/EPHF_Public_Health_in_the_Americas-Book.pdf
Phillips, N. (2017). Power and inequality in the global political economy. *International Affairs*, 93(2), 429–44.
Pieterse, J.N. (2012). *Globalization and Culture: Global Mélange* (3rd ed.). London: Rowman and Littlefield.
Planning Commission of India (2011). *High Level Expert Group Report on Universal Health Coverage for India*. New Delhi. [Online] Available at: http://planningcommission.nic.in/reports/genrep/rep_uhc0812.pdf

Prentice, A. (2006). The emerging epidemic of obesity in developing countries. *International Journal of Epidemiology*, 35(1), 93–9.
Ravelo, J.L. (2018). High-level meeting on NCDs holds promise, and pitfalls. *Devex*, 22 February. [Online] Available at: https://www.devex.com/news/high-level-meeting-on-ncds-holds-promise-and-pitfalls-92101
Reevell, P. (2016). 72 dead in Siberia after drinking counterfeit alcohol. *ABC News*, 22 December. [Online] Available at: https://abcnews.go.com/International/72-now-dead-siberia-drinking-counterfeit-alcohol/story?id=44342514
Research and Markets (2018). *Global Brand Counterfeiting Report 2018.* Dublin: Research and Markets.
Romon, M., Duhamel, A., Collinet, N., and Weill, J. (2005). Influence of social class on time trends in BMI distribution in 5-year-old French children from 1989 to 1999. *International Journal of Obesity*, **29**, 54–9.
Rossignoli, C., Virili, F., and Za, S. (2017). *Digital Technology and Organizational Change: Reshaping Technology, People, and Organizations Towards a Global Society.* New York: Springer.
Roxburgh, C., Lund, S., Dobbs, R., Manyika, J., and Wu, H.H. (2011). *The Emerging Equity Gap: Growth and Stability in the New Investor Landscape*. McKinsey Global Institute. [Online] Available at: http://www.mckinsey.com/insights/global_capital_markets/emerging_equity_gap
Schuh, A. (2007). Brand strategies of Western MNCs as drivers of globalization in Central and Eastern Europe. *European Journal of Marketing*, 41(3/4), 274–91.
Silva, B.M., Rodrigues, J.J.P.C., de la Torre Diez, I., Lopez-Coronado, M., and Saleem K. (2015). Mobile-health: a review of current state in 2015. *Journal of Biomedical Informatics*, **56**, 265–72.
Slobodian, Q. (2018). *Globalists: The End of Empire and the Birth of Neoliberalism*. Cambridge: Harvard University Press.
Smith, R., Beaglehole, R., Woodward, D., and Drager, N. (eds.) (2005). *Global Public Goods for Health, Health Economic and Public Health Perspectives*. Oxford: Oxford University Press.
Smith, K.F., Goldberg, M., Rosenthal, S., and Ramachandran, S. (2014). Global rise in human infectious disease outbreaks. *Journal of the Royal Society Interface*, 11(101), 20140950.
Stratton, E., et al. (2017). Effectiveness of eHealth interventions for reducing mental health conditions in employees: a systematic review and meta-analysis. *PLoS One*, 12(12), e0189904.
Swabe, J. (1999). *Animals, Disease, and Human Society: Human-Animal Relations and the Rise of Veterinary Medicine*. London: Routledge.
Swinburn, B., et al. (2015). Strengthening of accountability systems to create healthy food environments and reduce global obesity. *Lancet*, 385(9986), 2534–45.
Tarabusi, C. and Vickery, G. (1998). Globalization in the pharmaceutical industry. *International Journal of Health Services*, 28(1), 67–105.
Trouiller, P., Olliaro, P., Torreele, E., Orbinski, J., Laing, R., and Ford, N. (2002). Drug development for neglected diseases: a deficient market and a public-health policy failure. *The Lancet*, 359(9324), 2188–94.
UNDESA (2017). *International Migration Report.* New York: United Nations Department of Economic and Social Affairs. [Online] Available at: http://www.un.org/en/development/desa/population/migration/publications/migrationreport/docs/MigrationReport2017_Highlights.pdf
UN Office on Drugs and Crime (2017). *World Drug Report 2017*. Vienna: UNDOC. [Online] Available at: http://www.unodc.org/wdr2017/field/Booklet_2_HEALTH.pdf
US Department of Labor (2015). Long-term price trends for computers, TVs, and related items. Washington DC, 13 October. [Online] Available at: https://www.bls.gov/opub/ted/2015/long-term-price-trends-for-computers-tvs-and-related-items.htm
Valdez, S. (2006). *An Introduction to Global Financial Markets*. London: Palgrave Macmillan.

Ventola, C.L. (2014). Social media and health care professionals: benefits, risks, and best practices. *Pharmacy and Therapeutics*, 39(7), 491–99, 520.
Walley, J., Wright, J., and Hubley, J. (2001). *Public Health: An Action Guide to Improving Health in Developing Countries*. Oxford: Oxford University Press.
Watts, J. (2004). Chinese baby milk blamed for 50 deaths. *The Guardian*, 21 April. [Online] Available at: http://www.guardian.co.uk/china/story/0,7369,1196996,00.html
Wiedmann, T. and Lenzen, M. (2018). Environmental and social footprints of international trade. *Nature*, 30 April. [Online] Available at: https://www.nature.com/articles/s41561-018-0113-9
Wilkinson, R. (2014). *What's Wrong with the WTO and How to Fix It*. London: Polity Press.
Wipfli, H. (2015). *The Global War on Tobacco*. Baltimore, MD: Johns Hopkins University Press.
World Health Organization (WHO) (n.d.). *Illicit Trade in Tobacco: A Summary of the Evidence and Country Responses.* Geneva: WHO. [Online] Available at: http://www.who.int/tobacco/economics/illicittrade.pdf
World Health Organization (WHO) (1986). *The Ottawa Charter for Health Promotion*. First International Conference on Health Promotion, Ottawa, 21 November 1986.
World Health Organization (WHO) (2006a). *Health Metric Network (HMN) Workshop—Better Health Information Systems*. Geneva: Health Metrics Network. [Online] Available at: http://www.who.int/healthmetrics/news/20061027/en/index.html
World Health Organization (WHO) (2006b). *Public Health Innovation and Intellectual Property Rights*. Geneva: WHO.
World Health Organization (WHO) (2013). WHO's arrangement for hosting health partnerships and proposals for harmonizing WHO's work with hosted partnerships. *EB132/5 Add.9*. Geneva: WHO. [Online] Available at: http://apps.who.int/gb/ebwha/pdf_files/EB132/B132_5Add9-en.pdf
World Health Organization (WHO) (2014). UNGA NCD Review 2014. [Online] Available at: http://www.who.int/nmh/events/2014/unga-review.pdf?ua=1
World Health Organization (WHO) (2015). *The Protocol to Eliminate Illicit Trade in Tobacco Products: An Overview.* WHO FCTC. [Online] Available at: http://www.who.int/fctc/protocol/Protocol_summary_en.pdf
World Health Organization (WHO) (2017a). 1 in 10 medical products in developing countries is substandard or falsified. *News Release*, 28 November. Geneva: WHO. [Online] Available at: http://www.who.int/en/news-room/detail/28-11-2017-1-in-10-medical-products-in-developing-countries-is-substandard-or-falsified
World Health Organization (WHO) (2017b). Obesity and overweight. *Fact Sheet*, 18 October. Geneva: WHO. [Online] Available at: http://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight
World Health Organization (WHO) (2017c). Noncommunicable diseases. *Fact Sheet*, 21 June. Geneva: WHO. [Online] Available at: http://www.who.int/en/news-room/fact-sheets/detail/noncommunicable-diseases
World Health Organization (WHO) (2017d). Facts and figures on childhood obesity. *Commission on Ending Childhood Obesity*. Geneva: WHO. [Online] Available at: http://www.who.int/end-childhood-obesity/facts/en/
World Health Organization (WHO) (2018a). Projections of mortality and causes of death, 2015 and 2030. *Health Statistics and Information Systems.* Geneva: WHO. [Online] Available at: http://www.who.int/healthinfo/global_burden_disease/projections/en/
World Health Organization (WHO) (2018b). *Report of the Technical Consultation*. Geneva: WHO. [Online] Available at: http://www.who.int/ncds/governance/high-level-commission/HLC_Final_report_of_the_Technical_Consultation_21-22_March_2018-CORR1.pdf?ua=1
World Health Organization (WHO) (2018c). World leaders join new drive to beat noncommunicable diseases. News Release, 1 March. [Online] Available at:

http://www.who.int/mediacentre/news/releases/2018/world-leaders-ncds/en/
Witkowski, W. (2015). Global stock market cap has doubled since QE's start. *MarketWatch*, 12 February. [Online] Available at: https://www.marketwatch.com/story/global-stock-market-cap-has-doubled-since-qes-start-2015-02-12
World Bank (2000). *Is Globalization Causing a* 'Race to the Bottom' *in Environmental Standards?* Briefing Papers, April. Washington DC: PREM Economic Policy Group and Development Economics Group. [Online] Available at: http://www1.worldbank.org/economicpolicy/globalization/documents/AssessingGlobalizationP4.pdf
Wright, R. (2017). Container shipping faces critical moment after years of losses. *Financial Times*, 21 February. [Online] Available at: https://www.ft.com/content/8b633cfa-e7f0-11e6-967b-c88452263daf
World Trade Organization (WTO) (2017a). *Trends in World Trade: Looking Back Over the Past Ten Years*. Geneva: WTO. [Online] Available at: https://www.wto.org/english/res_e/statis_e/wts2017_e/WTO_Chapter_02_e.pdf
World Trade Organization (WTO) (2017b). *Trade Recovery Expected in 2017 and 2018, Amid Policy Uncertainty.* Press Release, 12 April. Geneva: WTO. [Online] Available at: https://www.wto.org/english/news_e/pres17_e/pr791_e.htm
World Trade Organization (WTO) (2018). *Strong Trade Growth Continues but Momentum May Soften in Q2, Trade Indicators Suggest*. Press Release, 17 May. Geneva: WTO. [Online] Available at: https://www.wto.org/english/news_e/news18_e/wtoi_17may18_e.htm
Yamabhai, I., Mohara, A., Tantivess, S., Chaisiri, K., and Teerawattananon, Y. (2011). Government use licenses in Thailand: an assessment of the health and economic impacts. *Globalization and Health*, **7**, 28.

SECTION 2

Determinants of health and disease

2.1 Determinants of health: overview
Matthew Fisher, Belinda Townsend, Patrick Harris, Ashley Schram, and Fran Baum

2.2 Politics of public health
Kaitlyn B. McBride and Linda Rosenstock

2.3 Poverty, justice, and health
Ronald Labonté, Fran Baum, and David Sanders

2.4 Socioeconomic inequalities in health in high-income countries: the facts and the options
Frank J. van Lenthe and Johan P. Mackenbach

2.5 Reducing health inequalities in developing countries
Romulo Paes-Sousa, Paulo M. Buss, and Mauricio L. Barreto

2.6 Genomics and public health
Veron Ramsuran and Tulio de Oliveira

2.7 Water and sanitation
Thomas Clasen

2.8 Food and nutrition
Roger Shrimpton, David Sanders, and Anne Marie Thow

2.9 Climate change and human health
Alistair Woodward and Alex Macmillan

2.10 Behavioural determinants of health and disease
Lawrence W. Green, Kristin S. Hoeft, and Robert A. Hiatt

2.11 How access to healthcare affects population health
Yvonne Inall, Rachel Lamdin Hunter, Stephen Leeder, and Angela Beaton

2.1

Determinants of health

Overview

Matthew Fisher, Belinda Townsend, Patrick Harris, Ashley Schram, and Fran Baum

Introduction

Determinants of health have been defined as ‘the range of personal, social, economic and environmental factors which determine the health status of individuals or populations’ (Nutbeam and World Health Organization 1998). An understanding of determinants of health accepts that human health is affected by multiple factors operating at different levels and interacting in complex ways. Determinants of health are classically grouped into four categories of biological, behavioural, environmental, and social determinants (Remais and Jackson 2016), although other classifications are available which incorporate categories such as ‘knowledge, attitudes and beliefs’ (Australian Institute of Health and Welfare 2008). Determinants of health are understood through a combination of research methods, investigating associations between specified factors and disease or disease risk in populations, and investigating mechanisms by which particular factors affect the health status of individuals (Rose 1985).

Biological determinants

Biological determinants are endogenous features of the human body that affect health. They include the genetic structures that play a central role in the development of body and brain in early life and may render a person more or less susceptible to particular diseases. Age is also identified as a biological determinant of health recognizing, for example, that deterioration of cell functions in old age may increase susceptibility to disease. Arguably,

the category of biological determinants is limited by its attention on physical attributes and overlooks the role that neuropsychological states play in affecting health; a matter we will discuss further later in the chapter.

Behavioural determinants

Behavioural determinants notably include the widely recognized ‘lifestyle’ factors of tobacco smoking, use of alcohol and other drugs, and diet and exercise. Tobacco smoking is known to increase risk of a range of respiratory and cardiovascular diseases and several forms of cancer, and contributes to around six million premature deaths globally each year (World Health Organization 2015). Diet and exercise affect health in a range of ways, but most notably influence the development of overweight and obesity, which increase risk of conditions such as type-2 diabetes, cardiovascular diseases, musculoskeletal disorders, and some cancers. Prevalence of obesity worldwide had tripled since the 1970s, and represents a growing risk to health in many countries (World Health Organization 2017). As we discuss later, this category is regarded by some as too limited because it overlooks the role that social and commercial factors play in influencing health behaviours (Baum and Fisher 2014).

Environmental determinants

Environmental determinants are factors in the physical environment that affect health, including features of human-made environments such as air pollution or availability of sewerage systems, and feature of natural environments such as weather events, insect-borne disease vectors, or availability of fresh water. The characteristics of foods such as their nutritive properties or presence of food-borne pathogens are important environmental determinants of health. For example, a lack of naturally occurring iodine in soil—and thus in the food it grows—impairs mental function in adults and during pregnancy can compromise foetal development. Diarrhoeal disease caused by contaminated water or food is a major cause of illness or death in children under 5 years, especially in low-income countries (LICs). Clearly, human activities are increasingly shaping ‘natural’ environments in a variety of ways with implications for human and ecological health such as land clearing and agriculture, commercial fishing, and use of fossil fuels

contributing to global warming. Air pollution is an environmental determinant of growing concern. Recent analysis from the World Health Organization (WHO) indicates that each year around 4.2 million deaths result from ambient air pollution and 3.8 million deaths from household air pollution from cooking fuels (World Health Organization 2018).

Social determinants: an introduction

Social determinants of health (SDH) have been defined as including both the daily social and economic conditions in which people live, and the higher-level socioeconomic and political factors that affect distribution of power, money, and resources, within or between countries (Commission on the Social Determinants of Health 2008). Factors in daily life that affect health include living and working conditions, availability of healthy foods, and access to healthcare. The higher-level factors, such as the decisions of governments or the structures of a market economy, are understood to *distribute* the social conditions that positively or negatively affect health; working through key mediating mechanisms of power, income/wealth, education, and employment (Commission on the Social Determinants of Health 2008). Where these processes produce inequalities in access to conditions supportive of good health, or in exposure to conditions that adversely affect health, then (socially determined) inequalities in health are the likely result. For example, political and economic conditions within a particular country may be such that relatively wealthy, urban populations enjoy better access to good quality food, medical care, and housing than impoverished rural populations, leading to higher rates of infant mortality in the latter group. This multilevel understanding of SDH is illustrated in Figure 2.1.1.

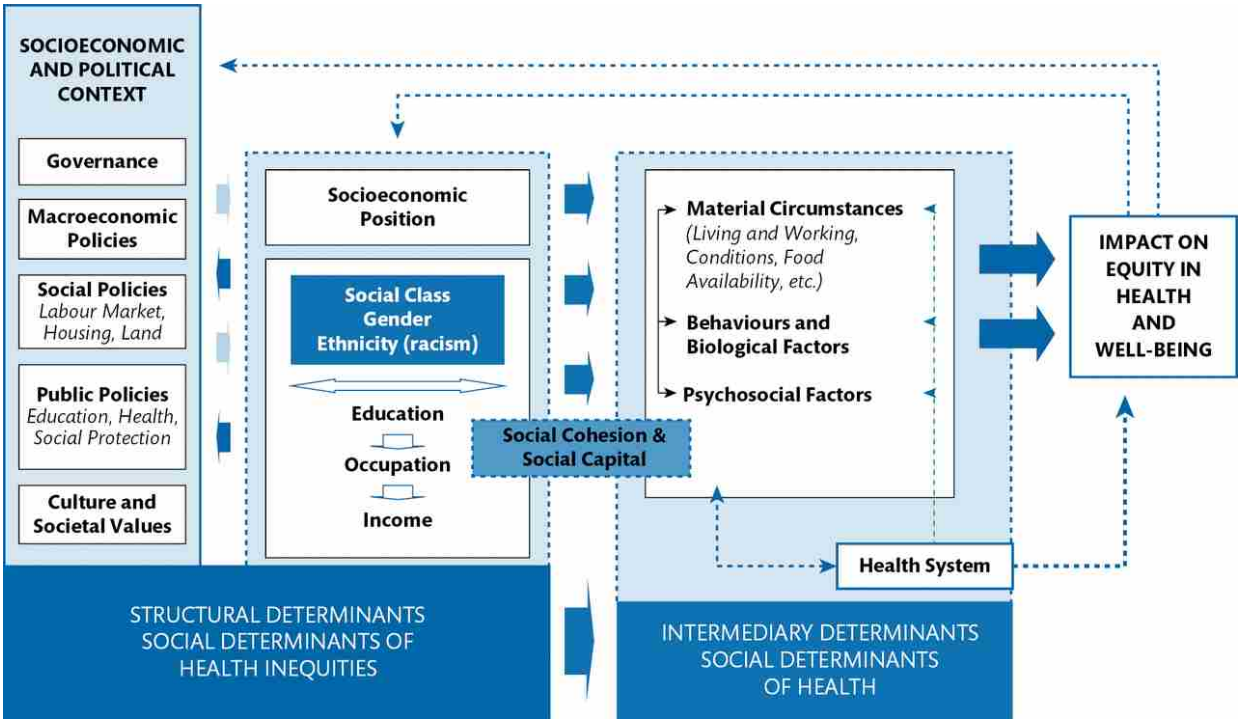


Fig. 2.1.1 Social determinants of health and health inequities.
Adapted with permission from Solar, O. & Irwin, A. (2010). A Conceptual Framework for Action on the Social Determinants of Health: Social Determinants of Health Discussion Paper 2. Geneva: World Health Organization. Copyright © 2010 WHO.

Not all health inequalities between populations are socially determined. Women have much higher rates of breast cancer than men but this is thought to result from biological differences, rather than differences in social conditions. (However, equally, socioeconomic factors may determine that some women have access to screening or treatments for breast cancer while others do not.) Health inequalities that *are* understood as socially determined may also be regarded as unfair or unjust. Those taking this view argue that health inequalities arising from structural social and economic inequalities could be reduced or eliminated, if governments and societies made different choices. To not do so, then, is regarded as unfair to those who (unnecessarily) are subsequently subjected to higher rates of disease and shorter lives than others (Whitehead 1992). Those who accept such arguments may refer to socially determined health inequalities as ‘health inequities’ (Baum 2016) to signal their ethical stance; a position we will adopt in this chapter. Such arguments can appeal to evidence showing that across different countries both overall health outcomes and the extent of health inequalities vary widely, in ways that relate to differences in political,

social, or economic conditions (Commission on the Social Determinants of Health 2008). In the next section we discuss some commonly identified SDH in more detail.

Social determinants of health

Many features of human societies are identified as SDH (Marmot et al. 2008). Here we discuss some standard items in classifications of SDH and some issues that have come to prominence more recently, as shown in Table 2.1.1. It is useful to note that there are different pathways and mechanisms by which social factors get 'under the skin' to affect individual or population health. Some of these are predominantly physical in nature, such as poor working conditions in mining leading to exposure to dust, contributing to lung disease. Other pathways involve psychological states. In particular, recurrent exposure to social environments perceived as threatening or demanding can cause psychological distress and chronic arousal of stress physiology, leading to adverse effects on mental or physical health (Fisher and Baum 2010).

Table 2.1.1 Social determinants of health

• Healthcare services	• Determinants of mental health
• Environments for early child development	• Urbanization and health
• Income, education, and employment status	• Globalization and commercial determinants of heath
• Employment conditions	• Social determinants of health behaviours
• Gender	• Climate change and SDH
• Social and cultural determinants of Indigenous health	• Distribution of power and resources

Healthcare services

Access to healthcare services is widely recognized as an SDH, supporting health and preventing disease by providing trained advice and support, preventative interventions such as vaccination, and medical treatments for disease or injury; when these are needed. Research suggests that universal access to primary healthcare (PHC) is an important step in improving population health and health equity in all countries (World Health Organization 2008). A comprehensive approach to PHC combines first-level medical care with strategies to promote health, prevent disease, engage communities, and address other determinants of health within local spaces (Baum 2016).

Environments for early child development

Environments affecting the early emotional, cognitive, behavioural, and physical development of children, say from conception to 7 years of age, are crucial in determining health prospects over the life-course (Irwin et al. 2007). Social factors involved here include: access to material resources during pregnancy to promote good health; opportunity to undertake and continue breastfeeding; nurturing and stimulation; interaction to develop socialization and language; housing conditions; and access to clean water (Maggi et al. 2010). Higher-level economic and social policies shaping issues such as minimum wages, parental leave provisions and welfare provision, or inequalities in income, differentially affect the lives and circumstances of families and other proximal conditions of child development, contributing to inequities in child health (Irwin et al. 2007).

Income, education, and employment status

These three factors are central to contemporary understanding of SDH and health inequities within and between countries (Baum 2016). Clearly they are related: for example, education is likely to affect employment status, which affects income, which in turn affects educational opportunities across generations. Income, education, and employment are key mechanisms by which hierarchical societies distribute resources, thereby affecting differences in living and working conditions, and ability to afford essential goods and services. Some evidence suggests that, in addition, the psychological experience of being in a lower position within a status

hierarchy can itself have adverse consequences for health (Wilkinson and Pickett 2009). Income, education, and employment status are commonly used as measures of socioeconomic status (SES), and research shows that health inequities do not just manifest as poorer health among those at the 'lowest' end of the SES scale, but often follow a gradient pattern whereby average health status decreases incrementally at each step down the SES scale, from 'highest' to 'lowest' (Fryers et al. 2003; Marmot and Friel 2008). Furthermore, among developed countries those with higher levels of inequality tend to have worse health across all SES levels, compared to more equal societies (Wilkinson and Pickett 2009). Thus, arguably, social and health inequities are a concern for all. Figure 2.1.2 presents an example of social gradients in child mortality within five countries. Note that there are also inequities between these countries.

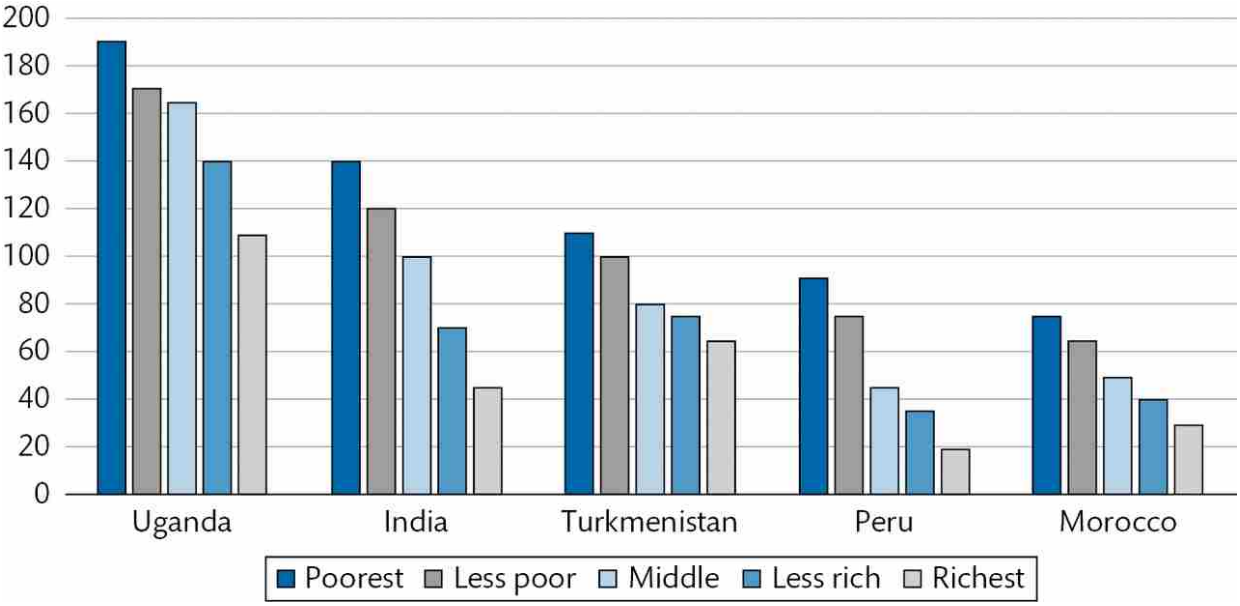


Fig. 2.1.2 Under-5 mortality rate per 1000 live births by level of household wealth in five countries, between 1998 and 2004.

Reprinted from Gwatkin, D.R. et al (2007). Socio-economic Differences in Health, Nutrition, and Population within Developing Countries: An Overview. Washington DC: © World Bank. License: Creative Commons Attribution license (CC BY 3.0 IGO). This is an adaptation of an original work by World Bank—Health, Nutrition & Population group. Responsibility for the views and opinions expressed in the adaptation rests solely with the author or authors of the adaptation and are not endorsed by any member institution of the World Bank Group.

Employment conditions

Aside from the role of employment in determining SES, the conditions of employment and unemployment are also important SDH. Work may provide many benefits for health, including financial security, social relationships, a sense of purpose, and opportunities for learning. However, workplace conditions may also involve exposure to health-damaging substances, physical demands causing injury, or psychosocial demands causing chronic stress (Marmot et al. 2008). In many countries the nature of work is changing under pressures from (inter alia) globalization, neoliberal 'austerity' politics, financial crises, technological change, declines in trade unions, and growth in casual and informal work and self-employment. Economic inequalities between countries create incentives for permanent or temporary migration to seek work, and these workers may be at higher risk of exploitation. According to the International Labour Organization (2015), in 2013 around 150 million people migrated for reasons of work. These changes risk an expanding divide between the 'haves and have-nots' in work status, security, and remuneration—within and between countries—with the potential to exacerbate health inequities.

Gender

In most parts of the world women tend to live longer than men despite reporting more ill health. Explanations for women's longer lifespans have included biological differences, the more dangerous nature of male employment, the tendency for women to seek healthcare more readily, and the greater social attachments women have (Baum 2016). However, gender inequities unfavourable to women have a range of adverse impacts on health by placing women and girls in cultural, social, or economic positions of lesser status compared to their male counterparts. Gender inequities affect health in a variety of ways, contributing to violence against women, discrimination against female children, unfair divisions of work, lower wages, and restricted access to education and healthcare (Marmot et al. 2008).

Social and cultural determinants of Indigenous health

Indigenous peoples are defined as the descendants of 'those who inhabited a country or . . . region at the time when people of different cultures or ethnic origins arrived, [who then] became dominant through conquest, occupation, settlement or other means' (United Nations Permanent Forum on Indigenous Peoples 2018). Although data on the health status of Indigenous peoples is uneven, in countries where it is available the story is often the same: Indigenous peoples have significantly worse health than the non-Indigenous population (Anderson et al. 2016). While Indigenous health is no doubt influenced by 'general' determinants such as income, scholars now identify a range of determinants affecting Indigenous peoples more particularly. These include health-promoting factors such as strong cultural identity and community control over resources and, on the negative side of the ledger, processes of colonization such as dispossession of traditional lands, discriminatory policies such as removal of children, and exposure to racism (Carson et al. 2007).

Determinants of mental health

Common forms of mental illness (including substance abuse disorders) account for around 20% of the total global burden of disease in terms of years lived with disability (Whiteford et al. 2013), equating to many millions of people whose ability to gain an education, keep a job, or maintain social relationships may be compromised. Depressive disorders alone are a major and growing cause of disease burden globally and contribute to the approximately 800,000 suicides that occur each year. Many of the SDH we have discussed may adversely affect mental health by causing chronic stress (Fisher and Baum 2010). Social factors affecting use of alcohol and other drugs also play a significant part. Chronic physical ill health may also contribute to poor mental health. Major events such as wars or natural disasters are likely to have significant impacts on stress and mental health by destroying basic infrastructure and causing dislocation of populations. As with other health conditions, rates of common forms of mental illness are typically inversely related to SES (Fryers et al. 2003).

Urbanization and health

A growing proportion of populations globally and within countries are living in cities. In 2008, for the first time in human history, the majority of the world's population lived in urban environments, and the decade since has seen an upsurge in interest in the links between cities and health. There is now broad acceptance that urbanization itself is a determinant of health and of health inequities within and between urban populations.

To understand urbanization as a determinant of health, there is now evidence of the full causal chain from urban system policies to health outcomes such as injuries, respiratory diseases, stress and mental illness, and infectious disease; and to quality of life outcomes such as well-being and liveability (Giles-Corti et al. 2016). Risk exposures cluster around environmental conditions (air, noise, and traffic), human factors (social isolation and safety), and behaviours (healthy eating, physical activity) (Giles-Corti et al. 2016). Although these factors have tended to be described as health-related features of the 'built environment', health is also affected by the social and economic make-up of cities. Here it is important that basic goods such as sanitation, education, healthcare, and social services are provided in ways, 'that are socially cohesive, that are designed to promote good physical and psychological well-being and that are protective of the natural environment' (Friel et al. 2011, p 863). At the same time it is now understood that urbanization is in large part contributing to global issues like climate change and migration.

In terms of action, the alignment of the determinants of health within urban planning policy systems is crucial. A century ago planning was strongly aligned with public health objectives to prevent the spread of infectious disease. This close relationship was not sustained, however, as planning became a specific policy sector in most developed countries with narrower aims related to setting strategic direction for cities, assessing and approving developments, and other local issues to do with zoning and urban design. While planning and public health are gradually realigning, full realization of the potential of urban planning to prevent disease and reduce health inequity requires substantial shifts in the policy systems that govern the built environment (Friel et al. 2011). From an SDH perspective, the challenge is to align the broad interests of a public health agenda with those of urban planning—tricky in countries where there are well-developed urban planning sectors, and even more difficult in resource-poor settings where regulatory frameworks governing city planning may not exist. Evidence on

health impacts alone is unlikely to suffice given that city planning is closely linked to competition between cities globally and thus is strongly influenced by local politics and perceived economic imperatives. Rather than prioritizing health, concepts like ‘agglomeration’ have come to dominate urban planning, whereby urbanization creates local centres for production activity that require servicing to function effectively. However it is important to recognize that considerations of health and economic vitality are not incompatible, given that well-planned cities that are more liveable and equitable in terms of offering jobs, goods, and services are likely to be more competitive economically.

Questions of scale ought to dominate thinking about urbanization as a determinant of health. Urbanization itself is uneven across countries even as cities become central to the global economy; massive slums dominate some cities and can even exist alongside affluent suburbs, whereas others are characterized by urban sprawl. Within cities the need to plan at varying scales is fundamental to managing urbanization and population growth at regional and city levels down to local places and spaces. These issues of scale are replicated in the various governance arrangements that dominate how cities are developed and planned. Local government has long been recognized as politically powerful, for instance by the Healthy Cities movement, and the ‘local’ scale is seen by some as an important venue for action on SDH and health equity, able to mitigate some effects of determinants operating at a national or global scale. There are also a range of actors—public and private—engaged in planning for and delivering urban environments ranging from large city-wide infrastructure projects to precincts to buildings. Action to ensure these varying governance regimes and networks continue to engage with ideas about urbanization as profoundly influencing health are essential.

Globalization and commercial determinants of health

Globalization facilitates increasing interconnectedness between national economies, societies, and cultures, shaping the upstream determinants of health. Arguably, globalized economic activity has produced some benefits for health through increased living standards, although this proposition is contested (Baum 2016). However, globalization has also facilitated widening income inequality. Income inequality has increased in nearly all regions of

the world over recent decades, although at different speeds (Alvaredo et al. 2018). Globalization therefore presents risks and challenges for population health and health equity within and between countries.

In the last two decades, intensified movement of capital and expanded trade and investment has facilitated the growth and size of transnational corporations (TNCs); those firms with manufacturing, production, and marketing operations across multiple countries. TNCs like Coca-Cola have production sites on every continent and extensive supply chains across the world. TNCs have also become significant economies on their own. In 2016, of the top 100 world economies, only 31 were countries. Most of the world's largest economies are now TNCs such as Walmart, Royal Dutch Shell, and Exxon Mobile, and technology firm Apple.

TNCs have become powerful actors, with significant economic resources, influencing global policymaking. For example, global pharmaceutical firms played a key role in shaping the World Trade Organization's *Trade-Related Aspects of Intellectual Property Rights Agreement* (TRIPS) which requires signatories to provide a minimum of 20-year protections on patents (Sell 2003). TNCs, alongside other industry bodies, have pushed for an expansion of the issues covered under trade and investment agreements between countries. While trade agreements once covered commitments to reduce tariffs, they increasingly include rules for domestic regulatory issues, such as intellectual property, competition, technology, and rights for investors.

There are several pathways through which these trade agreements can impact on health. Trade agreements can have a positive effect on health through increased access to technologies and health-promoting products and services. But they can also create negative consequences. The extension of intellectual property rights in trade deals, for example, can delay the market entry of cheaper generic medicines, reducing affordable access (Townsend et al. 2016). Trade agreements also promote greater market liberalization which facilitates the availability, affordability, and accessibility of unhealthy commodities such as ultra-processed food, tobacco, and alcohol—three key risk factors for non-communicable diseases (NCDs). Simultaneously, these agreements secure greater access for TNCs in domestic regulatory policymaking, which can increase the time and resources required for governments to regulate these products, as well as the influence of industry on the selected regulatory approach (Schram et al. 2017). Trade deals can therefore serve as enabling mechanisms for the commercial drivers of

NCDs. The costs associated with the negative effects of products such as cheap ultra-processed foods (most significantly an increase in NCDs) is externalized to governments.

The ‘commercial determinants of health’ is a recent term in public health used to refer to the practises and strategies of corporations involved in driving risk factors for NCDs (Kickbusch et al. 2016). In addition to influencing trade and investment deals, companies that manufacture and trade in unhealthy commodities such as tobacco, alcohol, and ultra-processed food utilize a number of strategies to promote greater uptake of their products and undermine public health regulation. Documented strategies by firms include marketing strategies targeting young people, lobbying, and donating to political representatives, partnering with governments, funding biased research, co-opting health professionals and policymakers to promote industry objectives, and sponsoring campaigns to favourably influence public opinion (Kickbusch et al. 2016). The regulation of unhealthy commodities industries has therefore become a key challenge for global public health. Public health researchers have criticized the continuing participation of alcohol and food industries in global health policymaking. For example, despite strong financial motivation to promote the consumption of sugary drinks, Pepsi-Co and Coca-Cola contributed to the recommendations of the 2011 UN High-Level Meeting on NCDs (Buse et al. 2017).

Public health campaigns have successfully intervened in the commercial determinants of health, most notably in tobacco control. Following decades of collecting and disseminating evidence and advocacy from health practitioners and policymakers, the first global treaty on tobacco—the WHO *Framework Convention on Tobacco Control* (FCTC)—was adopted in 2003 and has since been ratified by 181 countries. The FCTC explicitly prevents parties from engaging with tobacco companies, and offers a strong example of regulation to prevent industry interference. The case of tobacco illustrates that public health advocacy can be channelled to counter industry power and profits.

Social determinants of health behaviours

Ultimately, improving the health of populations may require changes in behaviour. On the face of it doing this appears relatively straightforward, in

that there are a range of behaviour change techniques to assist. However, behaviours are socially structured and so it follows that changing these behaviours requires changing the structures within which behaviours occur. Despite mounting evidence to support the importance of structure in shaping behaviours, behaviourism is still often promoted as a solution to 'lifestyle' diseases while paying little attention to structures. Ideologies influence these policy choices, as Tesh (1988) noted some time ago. Her analysis of US public policy demonstrates that there is a strong ideology of individualism which undermines appreciation of the role of structural factors, meaning that behaviourism is often privileged despite a weak evidence base for its effectiveness. There is a strong inherent logic to behaviour change strategies (Baum and Fisher 2014). If the problem of smoking is seen as one of people choosing to smoke and obesity as one of people overeating, then telling them not to do this seems to make sense.

There are many theories grounded in social psychology which aim to assist individuals to change their behaviours. These include social learning theory (Bandura 1977) and the health belief model (Becker 1974). Each tackles individuals' underlying beliefs and their motivations to change their behaviours. Such theories underpinned a generation of heart health programmes in the 1970s and 1980s. These started with great promise but in the main have been shown not to be successful. Their major limitation is that these models pay scant attention to the social, economic, and cultural environments in which people's behaviours occur. One of the few heart health programmes to succeed was the Finnish North Karelia programme. Its evaluation reported greater changes in chronic heart disease (CHD) risk factors than for the Finnish population as a whole, and in CHD male mortality (Puska et al. 2009). The relative success in North Karelia was not mainly a result of individual increase in health knowledge, but rather of community-level strategies—including provision of PHC services and involvement of community organizations—and policy changes, such as the reduction of fat in food.

The accumulated experiences with unsuccessful health promotion programmes directed at individual behaviour change reinforces the importance of the social determinants of behaviour. Many behaviours judged to be bad for health are often those that help people cope with stressful circumstances or are simply a response to an unhealthy environment. Thus the high smoking rates among Australian Aboriginal people reflect, in part, a

means of coping with the racism they experience in their daily life. The increase in overweight and obesity is primarily because the marketing and supply of high fat and sugar foods in many countries has made these foods attractive, available, and affordable. As we saw in the previous section, trade agreements have made it much easier for transnational food companies to gain access to food markets and undermine healthy food supply. Thus individual behaviours reflect social, cultural, economic, and physical environments in which people live. Changing behaviours requires changing these environments.

Changing environments is particularly important in achieving health equity. Behaviour makes the most difference to people's health status when other conditions in their life are favourable. So, if someone is not poor in relation to the society they are living in, if they are not living in absolute poverty, have a supportive social network, are reasonably free of disease, then behaviour change strategies might make some difference to their health status; but such strategies alone are likely to increase rather than decrease health inequities.

Climate change and social determinants of health

Anthropogenic global warming leading to climate change is having significant, adverse effects on human health that are only likely to increase in the future. While climate change is an environmental rather than a social determinant of health, human societies as currently configured are the primary cause of global warming, and the effects of climate change also intersect with SDH in ways that significantly shape health and health equity outcomes (McMichael 2017). Climate change can adversely affect health in a variety of ways including: the frequency and intensity of extreme weather events; sea level rise and associated flooding in low-lying islands or coastal areas; changing vectors of contagious disease; loss of arable land; and reduced access to fresh water. In many cases the health impacts of these events are likely to fall most heavily on countries, and population groups within countries, that are also subject to socially determined health inequities. Social impacts of climate change include the possibility of environmental refugees, forced to move in response to extreme weather events or changed environmental conditions, who will be highly vulnerable to poor health. Health impacts of climate change may also increase demand

on under-resourced healthcare systems in LICs. For all these reasons, climate change—intersecting with SDH—presents risks of increased health inequities globally and within countries over time.

The complexity of determinants and their interactions

The complex and multiscaled nature of determinants of health presents a number of complications for understanding and action, whether by researchers, healthcare practitioners, civil society organizations, governments, or international agencies.

Temporality

In epidemiology, the concept of relative risk is used to express a causal link between a particular social or environmental determinant and the health of an individual or population group: that is, the likelihood or risk that people exposed to condition *x* will go on to develop disease *y*, relative to others not exposed. For example, in Australia, a current smoker is nine times more likely to develop lung cancer than a non-smoker. However, these equations on their own do not convey the temporal nature of the relationships between determinants and health or disease (Remais and Jackson 2016). A clinically recognizable effect on health arising from exposure to a particular determinant may take years to emerge. Exposure of young children to abuse or neglect is understood to increase the risk of conditions such as depressive disorders, drug use, and sexually transmitted infection (Norman et al. 2012), but these conditions generally develop in adolescence or adulthood. Along the way, other mediating factors may intercede to mitigate or exacerbate the risk. This kind of life-course perspective offers significant potential for primary prevention of multiple health problems, and possible gains in health equity, by acting on the common, underlying risk factor of child maltreatment (Irwin et al. 2007). However, for researchers, long time-lags between exposures and health effects complicate the challenge of investigating and demonstrating the links (Remais and Jackson 2016).

The dynamic, temporal nature of the relationships between determinants and health are also expressed over longer periods, and at larger scales. A combination of improvements in medical care, urban environments, nutrition, and living conditions is likely to have contributed to the so-called

epidemiological transition in many countries, where rates of infectious diseases (IDs) have declined, while NCDs have increased (Wahden 1996). In part, the increased prevalence of NCDs is a by-product of longer lifespans and ageing populations; because incidence of NCDs (such as heart disease and cancer) increases later in life. However, IDs do not go away, and may continue to have significant, adverse impacts among populations subject to disadvantage, producing what has been referred to as a 'double burden of disease', while having only small effects on overall population health (Wahden 1996).

Multiple scales and multiple determinants

As was alluded to earlier, relationships between determinants of health and health problems can be understood as spanning multiple mechanisms and events operating at different scales; from individual biology, psychology, and behaviour to social or environmental conditions at a local, national, and international scale. This understanding is represented in the nested models of SDH presented in the literature, such as Dahlgren and Whitehead's famous 'rainbow' model (2006). Thus, the health status of a particular individual or population group at any point in time is very likely to have been shaped by multiple factors. This presents challenges for research because more factors affecting health will be drawn into consideration, the farther 'out' from individual biology and behaviour one looks. The link between smoking (as a behavioural determinant) and risk of lung cancer can be established fairly readily, but if we ask why one population group has a smoking rate of 20% and another a rate of 5%, the determinants to consider expand accordingly. The nested nature of SDH also presents challenges for action. Healthcare practitioners seeking to reduce smoking rates might find it easier to focus on approaches targeting individual behaviour change such as nicotine therapy or health literacy, rather than worrying about broader determinants such as the lobbying practices of tobacco TNCs (Remais and Jackson 2016). Policymakers aiming to address increased rates of type-2 diabetes linked to obesity might find it easier to implement a social marketing campaign encouraging healthy food choices, than to tackle powerful interests determining the availability and marketing of ultra-processed, energy dense foods. The general argument from an SDH perspective is to look 'upstream' from biology and behaviour and address the higher-level environmental and

social determinants, the ‘causes of the causes’ (Commission on the Social Determinants of Health 2008); because they are the key to prevention rather than cure, and because even modest changes in social or environmental factors operating across whole populations will produce far larger gains for population health than interventions with high-risk individuals (Rose 1985).

In response to the complexity of social determinants, integrative theories have been put forward, explaining how the many factors, operating at different levels interact to produce health inequities. For example, Singh-Manoux and Marmot (2005) argue that health inequities are socially produced (and reproduced) via processes of ‘socialization’, whereby social class positions people within certain social milieu, which shape their psychology, attitudes, behaviours, and living conditions related to health.

Cross-boundary effects

Understanding environmental and SDH means that questions about the role and responsibilities of national governments cannot be avoided (Commission on the Social Determinants of Health 2008). However, it is also the case that many public health issues transcend national borders and policies and require processes of global governance (Marmot et al. 2008), presenting challenges to international organizations such as the United Nations and its health body the WHO. No public health challenge exemplifies this level of international and global concern more than the profound and entrenched social and economic inequalities within and between nations (Alvaredo et al. 2018), which are the underlying cause of health inequities. Similarly, climate change is surely a matter of global concern requiring a global response, including not just the obvious imperative to reduce greenhouse gas emissions but also consideration of the potentially complex repercussions of national policy responses. For example, policy decisions by wealthy nations to increase use of bio-fuels may create incentives in LMICs to shift agricultural production from food crops to bio-fuel crops such as palm oil. This in turn can lead to increased prices for food staples on global markets, contributing to undernutrition among people with lower incomes (Remais and Jackson 2016).

Another major public health issues transcending national boundaries is the movement of migrants and refugees across borders, whether stimulated by the desire to find work or to escape oppression or the violence of war. It is

estimated that the Syrian civil war alone has led to over 5 million refugees leaving the country and another 6 million people displaced within Syria. The repercussions of these events for health will inevitably be substantial and long-lasting, in Europe and the Middle East, and around the world.

Political challenges

National governments now operate within a global, neoliberal economic environment that prioritizes free trade, private profit, and economic growth and promotes these as vehicles for meeting human needs and reducing poverty. In this context, the role and capacity of national governments to regulate markets in ways conducive to public health, or tax and spend to deliver universal public services that contribute to health may be compromised in a variety of ways. Depending on the political and economic circumstances of the country in question, these challenges may take a variety of forms: governments adopting neoliberal or antidemocratic political values, weak or stagnating public revenues, influence of the private sector on policy, out-migration of skilled workers, or domestic political corruption just to name a few. Here we discuss two challenges for national governments in low, middle, and high income countries in taking action on SDH to improve public health and reduce health inequities.

Economic growth

The Preston Curve (Bloom and Canning 2007) indicates that, for LICs in particular growth in national income from a low base is often associated with marked increases in average life expectancy (although the benefits decline as national income continues to increase). Such health gains may well be influenced by improvements in determinants such as nutrition, access to clean water and safe sanitation, and PHC (Bloom and Canning 2007). The pathway to economic growth for LICs—as perceived by a national government, or promoted by international agencies—has commonly been seen as one of foreign investment for industrialization of mining, agriculture, or manufacturing for export, facilitated by low labour costs, urbanization, fossil fuel-based electricity generation, limited regulation and environmental protections, or tax concessions (Remais and Jackson 2016). Development loans to LICs from agencies such as the World Bank and the International

Monetary Fund have cultivated this approach by imposing conditions of economic liberalization and reduced government spending. However, this neoliberal model of development may involve adverse changes in determinants of health and health equity such as: poor nutrition or reduced access to healthcare; increased income inequalities; environmental degradation and pollution; poor urban living conditions; low wages and poor working conditions; displacement of population groups; and loss of agricultural production to meet domestic need.

Healthy public policy

The accumulated evidence on SDH along with a concern for health equity leads compellingly to the conclusion that health as a policy concern cannot be confined to the health sector alone but should involve all sectors and levels of government (Marmot et al. 2008). Such thinking has led to innovative strategies such as the 'Health in All Policies' approach in South Australia and elsewhere (Baum 2016). However, calls for a whole-of-government approaches to health face both institutional and political obstacles. On the former point, one issue identified is the 'siloed' structure of large government agencies, meaning the structures of decision-making, funding and accountability are organized vertically within departments and thus discourage 'horizontal' collaboration. On the political front, a raft of incentives may come into play encouraging decision makers (despite the evidence) to adopt a narrow biomedical and behaviourist view of health and illness, and thus focus their policy attention for 'health' on systems of acute care for illness, or on individualized, behaviour change strategies such as social marketing. As noted earlier the incentives at work here may variously include: (again) neoliberal, individualist political values; the power and lobbying efforts of TNCs in sectors such as food, soft drink or alcohol; and the influence of the medical profession or the media on health policy (Baum and Fisher 2014; Tesh 1988).

Conclusion

In this chapter we have provided an overview of contemporary evidence and understanding of biological, psychological, behavioural, social, economic, and environmental determinants of health. Identifying particular categories

of determinants can be very useful for analysis and research. Research focused on isolating and understanding the effects of particular factors on the health of individuals or populations has contributed to rich and detailed body of evidence on determinants of health and health inequities now available (Commission on the Social Determinants of Health 2008). Given this evidence, it is clear that governments and societies concerned with improving health cannot limit their attention to healthcare services, important as these may be. Neither can they reasonably believe that the adverse health impacts of (inter alia) tobacco, ultra-processed food, or alcohol are adequately understood as purely matters of individual 'lifestyle' choice. We have discussed the complexity of determinants and their interactions, and this might seem to suggest that determinants of health and health inequities are just too complex for governments to address successfully. However, this is not so: although the political challenges are no doubt significant, in principle what is required is to understand that addressing determinants of health to improve population health and reduce health inequities require multiple forms of action, at different scales, in different domains, and giving attention to different phases of the lifespan.

REFERENCES

Alvaredo, F., Chancel, L., Piketty, T., Saez, E., and Zucman, G. (2018). *World Inequality Report*. Paris; Berkeley, CA: World Inequality Lab.

Anderson, I., et al. (2016). Indigenous and tribal peoples' health (The Lancet–Lowitja Institute Global Collaboration): a population study. *The Lancet*, 388(10040), 131–57.

Australian Institute of Health and Welfare (2008). *Australia's Health 2008*. Cat no AUS 99, Canberra: Australian Institute of Health and Welfare.

Bandura, A. (1977). *Social Learning Theory*. Englewood Cliffs, NJ: Prentice Hall.

Baum, F. (2016). *The New Public Health* (4th ed.). Melbourne: Oxford University Press.

Baum, F. and Fisher, M. (2014). Why behavioural health promotion endures despite its failure to reduce health inequities. *Sociology of Health and Illness*, 36(2), 213–25.

Becker, M.H. (1974). The health belief model and personal health behaviour. *Education Monographs*, 2(4), 324–508.

Bloom, D.E. and Canning, D. (2007). Commentary: the Preston Curve 30 years on: still sparking fires. *International Journal of Epidemiology*, 36(3), 498–9.

Buse, K., Tanaka, S., and Hawks, S. (2017). Health people and healthy profits? Elaborating a conceptual framework for governing the commercial determinants of non-communicable diseases and identifying options for reduced risk exposure. *Globalization and Health*, **13**, 34. DOI 10.1186/s12992-017-0255-3

Carson, B., Dunbar, T., Chenhall, R.D., and Bailie, R. (eds.) (2007). *Social Determinants of Indigenous Health*. Crows Nest, NSW: Allen and Unwin.

Commission on the Social Determinants of Health (2008). *Closing the Gap in a Generation: Health Equity Through Action on the Social Determinants of Health*. Geneva: World Health

Organization.
Dahlgren, G. and Whitehead, M. (2006). *European Strategies for Tackling Social Inequities in Health: Levelling up* (Part 2). Copenhagen: World Health Organization Europe.
Fisher, M. and Baum, F. (2010). The social determinants of mental health: implications for research and health promotion. *Australia and New Zealand Journal of Psychiatry*, 44(12), 1057–63.
Friel, S., et al. (2011). Addressing the social and environmental determinants of urban health equity: evidence for action and a research agenda. *Journal of Urban Health*, 88(5), 860–74.
Fryers, T., Melzer, D., and Jenkins, R. (2003). Social inequalities and the common mental disorders: a systematic review of the evidence. *Social Psychiatry and Psychiatric Epidemiology*, **38**, 229–37.
Giles-Corti, B., et al. (2016). City planning and population health: a global challenge. *The Lancet*, 388(10062), 2912–24.
Gwatkin, D.R., Rutstein, S., Johnson, K., Suliman, E., Wagstaff, A., and Amouzou, A. (2007). *Socio-economic Differences in Health, Nutrition, and Population within Developing Countries: An Overview*. Washington DC: © World Bank. License: Creative Commons Attribution license (CC BY 3.0 IGO) [This is an adaptation of an original work by World Bank—Health, Nutrition & Population group. Responsibility for the views and opinions expressed in the adaptation rests solely with the author or authors of the adaptation and are not endorsed by any member institution of the World Bank Group.]
International Labour Organization (2015). *ILO Global Estimates on Migrant Workers*. Geneva: International Labour Organization.
Irwin, L.G., Siddiqi, A., and Hertzman, C. (2007). *Early Childhood Development: A Powerful Equaliser*. Geneva: World Health Organization.
Kickbusch, I., Allen, L., and Franz, C. (2016). The commercial determinants of health. *The Lancet*, 4(12), e895–6.
Maggi, S., Irwin, L.J., Siddiqi, A., and Hertzman, C. (2010). The social determinants of early child development: an overview. *Journal of Paediatrics and Child Health*, 46(11), 627–35.
Marmot, M. and Friel, S. (2008). Global health equity: evidence for action on the social determinants of health. *Journal of Epidemiology & Community Health*, **62**, 1095–7.
Marmot, M., Friel, S., Bell, R., Houweling, T.A.J., and Taylor, S. (2008). Closing the gap in a generation: health equity through action on the social determinants of health. *The Lancet*, 372(9650), 1661–9.
McMichael, A.J. (2017). *Climate Change and the Health of Nations: Famines, Fevers, and the Fate of Populations*. New York: Oxford University Press.
Norman, R.E., Byambaa, M., De, R., Butchart, A., Scott, J., and Vos, T. (2012). The long-term health consequences of child physical abuse, emotional abuse, and neglect: a systematic review and meta-analysis. *PLoS Medicine*, 9(11), e1001349.
Nutbeam, D. and World Health Organization (1998). *Health Promotion Glossary*. WHO/HPR/HEP/98.1. Geneva: World Health Organization.
Puska, P., Vartiainen, E., Laatikainen, T., Jousilahti, P., and Paavola, M. (2009). *The North Karelia Project: From North Karelia to National Action*. Helsinki: National Institute for Health and Welfare.
Remais, J.V. and Jackson, R.J. (2016). Determinants of health: overview. In: Detels, R., Gulliford, M., Karim, Q.A., and Tan, C.C. (eds.) *Oxford Textbook of Global Public Health* (6th ed.). Oxford: Oxford University Press.
Rose, G. (1985). Sick individuals and sick populations. *International Journal of Epidemiology*, 14(1), 32–8.
Schram, A., et al. (2017). A conceptual framework for investigating the impacts of international trade and investment agreements on noncommunicable disease risk factors. *Health Policy and Planning*, 33(1), 123–36.

Sell, S. (2003). *Private Power, Public Law: The Globalization of Intellectual Property Rights.* Cambridge: Cambridge University Press.

Singh-Manoux, A. and Marmot, M. (2005). Role of socialization in explaining social inequalities in health. *Social Science & Medicine*, **60**, 2129–33.

Solar, O. and Irwin, A. (2010). *A Conceptual Framework for Action on the Social Determinants of Health: Social Determinants of Health Discussion Paper 2.* Geneva: World Health Organization. [Figure 2.1.1 was drawn from this publication, Executive Summary, p. 6, Copyright 2010.]

Tesh, S. (1988). *Hidden Arguments: Political Ideology and Disease Prevention Policy.* New Brunswick, NJ: Rutgers University Press.

Townsend, B., Gleeson, D., and Lopert, R. (2016). The regional comprehensive economic partnership agreement, intellectual property and access to medicines. *Asia Pacific Journal of Public Health*, 28(8), 682–93.

United Nations Permanent Forum on Indigenous Peoples (2018). *Who are Indigenous Peoples? Indigenous Peoples, Indigenous Voices: Fact Sheet.* New York: United Nations.

Wahden, M.H. (1996). The epidemiological transition. *Eastern Mediterranean Health Journal*, 2(1), 8–20.

Whiteford, H.A., et al. (2013). Global burden of disease attributable to mental and substance use disorders: findings from the Global Burden of Disease Study 2010. *The Lancet*, 382(9904), 1575–86.

Whitehead, M. (1992). The concepts and principles of equity and health. *International Journal of Health Services*, 22(3), 429–45.

Wilkinson, R. and Pickett, K. (2009). *The Spirit Level: Why More Equal Societies Almost Always Do Better.* London: Penguin Books.

World Health Organization (2008). *The World Health Report 2008: Primary Health Care Now More Than Ever.* Geneva: World Health Organization.

World Health Organization (2015). *WHO Global Report on Trends in Prevalence of Tobacco Smoking*. Geneva: World Health Organization.

World Health Organization (2017). *Obesity and Overweight.* Fact sheet. Geneva: World Health Organization.

World Health Organization (2018). *Air Pollution.* Health topics. Geneva: World Health Organization.

2.2

Politics of public health

Kaitlyn B. McBride and Linda Rosenstock

Introduction to the Politics of Public Health

The aim of public health is to improve and protect individuals and their communities, as well as promote evidence-based policy to drive decision-making across the larger health system. Researchers in academic institutions, government agencies, and the private sector provide the science to help policymakers make evidence-based health policy decisions, such as interventions to increase smoking cessation, or policies to improve access to preventive care services. Despite the proliferation of evidence-based policy research, and significant advances in both the prevention and treatment of disease, there has been consistent efforts to suppress and distort scientific research, as well as prevent the creation, implementation, and enforcement of laws to protect the public health of individuals and their communities.

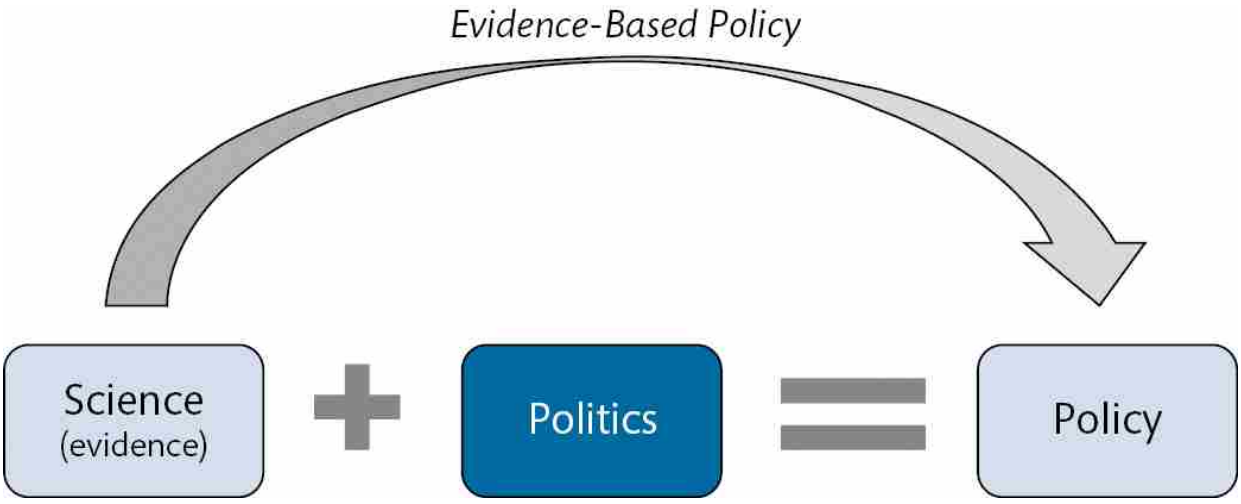


Fig. 2.2.1 A formula to guide the analysis of politics in the policy process.

Evaluating politics in the policy process can be reduced to the following components of the formula: Science + Politics = Policy (see Fig. 2.2.1). In an ideal world, science (or evidence) drives policy, with politics playing a minimal or enabling role. However, this formula recognizes that public health policy is not only decided based on science (or evidence), but science itself is increasingly vulnerable to the political process, and to attacks by a wide array of players (e.g. politicians, industry leaders, and interest groups) that have vested interests to undermine scientific evidence in order to thwart the actions that would credibly follow that knowledge. Put differently, politics are key—and can be systematically analysed—in advancing or thwarting evidence-based policy.

Unfortunately, the adoption of evidence-based research and practices is increasingly subjected to complex strategies—including economic manipulation, delays, harassment, and rhetoric—which weaken evidence-based policy (Rosenstock and Lee 2002; Rosenstock 2006). The politics in this arena involves significant trade-offs, and political leaders may elect to not support objective or scientific evidence, due to many factors, including biases (e.g. ideology and religion), fear of losing political capital (e.g. votes from constituents), loss of financial or political support from powerful organizations, or strong partisanship.

Framework for evaluating the politics of the health policy process

To examine the politics of the health policy process, our framework is a guide to evaluating how the role of politics influences the adoption of public health policy. To analyse this process, we will assess specific examples in a case-based approach, by expanding on the following domains:

Define the problem and evidence for policy interventions (The What)
Understand the roles and the players (The Who)
Understand motivations (The Why)
Understand strategies and tactics (The How)

The problem and the evidence (The What)

Defining the public health problem is the first step in 'the what'. For example, the problem might be that preventive care services are not sufficiently utilized. Once the problem has been identified, it is imperative that there is strong empirical evidence in support of a public health strategy to advance and improve the public health. For example, reducing an individual's out-of-pocket expenses (e.g. through added fees such as deductibles or co-pays) is an effective strategy for increasing preventive care utilization.

In this context, the collection of accurate data and objective, unbiased research is essential for establishing evidence-based policy (Choi, 2005). Additional components of high-quality scientific evidence include a sound analysis, comprehensive synthesis of findings, appropriate translation of results, and accessibility of information, either for policymakers, government officials, or for the general public (Choi 2005). An evidence-based approach in health policy can lead to both direct and indirect benefits, such as increased utilization of preventive health services, improved productivity among the workforce, and better quality of care for patients (Brownson et al., 2009).

The politics of the health policy process can best be analysed in focusing on steps 2–4 of the framework: The Who, The Why, and The How:

The players (The Who)

Policymaking is influenced by a myriad of players, including elected political leaders**,** public health officials, corporations, key interest groups, the media, and research institutions. In some settings, judicial review is key.

In general, policymaking is often driven by legislators and elected officials who shape the law. Understanding who the key players are, and how they work together, or in opposition to one another, is a key step in recognizing the role of politics in the policy process.

Motivations (The Why)

The motivation for pursuing a public health policy should ideally be driven by evidence-based public health research that results in policy that advances the public good. But health policy is also driven by economic, ideological, religious, and emotionally-held beliefs that are not always based on sound science. As such, motivations beyond those to increase the public good are key to understanding the role of vested interests—where one's interest, such as economic, may override the drive towards evidence-based policy.

For example, in many countries, elected leaders have prioritized ideology over sound empirical evidence in support of needle-exchange programmes (NEPs), a harm reduction strategy aimed to reduce the harms accompanied by risky health behaviours and illicit drug use. Arguments against needle-exchange programmes have been tied to moral and religious philosophies, which purport that harm reduction approaches, such as NEPs, condone amoral behaviour, and increase violence in communities. Despite the abundant evidence in support of NEPs in improving outcomes and decreasing disease burden, public officials continue to misrepresent and ignore scientific evidence for ideological purposes (Hwang 2007).

Laws restricting access to abortion are another example of public health policy driven by political, religious, and moral beliefs, rather than scientific evidence. Research has shown that laws restricting access to abortion and reproductive healthcare are harmful to women's health and well-being, and actually increase risk of injury or death; however, many countries have laws that criminalize abortion, as well as restrict access to reproductive health services, such as family planning (Berer 2017).

Strategies and tactics to undermine science and evidence-based policy (The How)

Strategies and tactics are key to advancing evidence-based policy, but all too often they are employed to thwarting policy that would protect the public

good. One key strategy is simply to attack the underlying science (or evidence), taking advantage of the distraction that claiming scientific uncertainty creates (Rosenstock and Lee 2002). A key example is climate change, where some question the mainstream of scientific opinion in order to undermine steps to mitigate climate-related adverse health effects. Economic manipulation is another tactic commonly used by politicians, interest groups, and corporations. For example, insufficient funding of regulatory and health agencies can limit the effectiveness of the policy apparatus. In addition, industry-sponsored research—especially in the US—is playing an increasingly large role as a source of funding, with a range of potential effects of altering the nature of the research itself to altering the dissemination of research results. In particular, the tobacco industry has used its financial ties to pressure health insurance companies to reduce consumer incentives to quit smoking (Shamasunder and Bero 2002). Due to these financial conflicts of interests, industry-sponsored research may lead to a suppression of study findings, prioritization of projects to produce economic gains, or biased interpretation of results by investigators (Okike et al. 2008).

Corporations, politicians, and vested interest groups also use delay and harassment tactics to forestall the release of sound science. Delay tactics includes funding of parallel studies (with the aim to confound the understanding of the original data), threats to sue, initiation of litigation, delay of hearing of the facts, and legislative barriers to prevent policies from being implemented. These vested interest groups have also been known to harass researchers and federal agencies to suppress scientific findings and prevent publications that support research in opposition to their economic interests (Rosenstock and Lee, 2002).

Conclusion

To be able to respond to attacks on science that are the underpinning of blocking evidence-based policy, the scientific community must increase its awareness of potential manipulation, distortion of empirical evidence, and conflicts of interest. This practice will hopefully encourage researchers and investigators to not only be vigilant about the potential consequences of undertaking industry-sponsored research but to understand their role in advocating for evidence-based policy.

Harm reduction strategies: needle-exchange programmes

An estimated 13 million people worldwide inject drugs—and 10% of HIV infections are attributed to injection drug use (IDU) (World Health Organization 2016). Use of contaminated needles is the primary mode of transmission of blood-borne viruses, such as HIV, hepatitis C (HCV), and hepatitis B (HBV), and is a significant contributor to worldwide morbidity and mortality (Degenhardt et al. 2017).

NEPs have shown to be an effective strategy to reduce transmission of infection among injection drug users. NEPs were first introduced in Amsterdam in the early 1980s by a group known as the Junkie Union. The primary purpose was to prevent the transmission of hepatitis B and HIV among injection drug users (Jarlais 2006). Studies later demonstrated a reduction in drug injection and HIV transmission among the programme's participants (O'Gurek and Kirchner 2010). NEPs were quickly adopted throughout Europe as a principal HIV prevention initiative and began to spread in US cities amidst the growing HIV/AIDS epidemic in the 1980s. During this period, NEPs gained support from several global institutions and non-governmental organizations, including the World Health Organization (WHO) and the Joint United Nations Programme on HIV/AIDS.

NEPs vary considerably in their design, services provided, programme policies, and daily operations (Fernandes et al. 2017). In addition to giving clean needles, the majority provide free testing for infectious disease (e.g. HIV and HCV), advice on safe injection and disposal practices, referrals to drug treatment programmes, such as opioid substitution therapy, and other supplies, including alcohol pads, and condoms.

NEPs are considered a harm reduction initiative aimed at mitigating risk among IDUs and reducing transmission of infectious disease (Clark and Fadus 2009). The term 'harm reduction' was first introduced in 1974 by the 20th WHO Expert Committee on Drug Dependence, referring to strategies aimed at reducing adverse consequences associated with drug use (Ball 2007). The concept and theoretical framework for the application of harm reduction strategies began initially with a defined focus on drug-related abuse; however, harm reduction policies have more recently been applied in response to broader public health issues ad not only drug-related harm, including alcohol abuse and sex trafficking (Myers et al. 2004).

From a public health perspective, harm reduction is a strategy that aims to reduce the harms that are accompanied by risky health behaviours. In the case of NEPs, these initiatives are a proven, practical public health approach to reduce injection-related infections, while not leading to an increase in injecting drug use (Witteveen and Schippers 2006). The science behind the effectiveness of NEPs has been well-established. Several studies have shown that these programmes lead to positive health outcomes among NEP participants, including reductions in needle sharing and injection frequency. Reviews of international evidence by the WHO and Institute of Medicine (IOM) have consistently found evidence in support of the effectiveness of NEPs in reducing the spread of HIV, and other risk behaviours, including syringe sharing among IDUs.

NEPs have been implemented in several countries, including Australia, Canada, Norway, Portugal, Spain, France, Italy, United Kingdom, China, and Malaysia. Despite the proliferation of empirical evidence and global support for NEPs– significant challenges remain in the global health response to the rapid spread of disease among IDUs. In 2016, of the 158 countries with reported injecting drug use, only 90 countries have implemented NEPs (Stone 2016).

The harm reduction strategy remains extremely controversial and underfunded in parts of the world, including Eastern Europe, Asia, and particularly in the United States. Opponents argue that NEPs lead to an increase of illicit drug use and perpetuate the cycle of addiction (Wodak and Cooney 2006). Others critique the philosophy of harm reduction, stating that harm reduction strategies condone immoral and illicit behaviour. These opponents support incarceration and punishment for IDUs, although there is no empirical evidence to suggest that this strategy leads to decreased use of illicit drugs, reduces crime, or prevents the spread of disease.

Case study Federal funding for needle-exchange programmes in the United States

In the United States, opposition to NEPs has historically been rooted in conservative and Republican ideology and moral philosophy, which purports that harm reduction initiatives such as NEPs, promote drug use, increase injection of illicit drugs, and lead to increased violence in communities. Although hundreds of needle-exchange programmes operate across the country, the United States has one of the lowest rates of needle-exchange availability and is one of the only countries to ban federal

funding for NEPs (Coleman, 2015). Consequently, ongoing programmes have limited resources to provide basic services, and must rely on private donations to operate. Lack of federal funding has also stalled widespread adoption of NEPs.

The debate over federal funding for NEPs has remained a steady battle between Republicans and Democrats since the introduction of NEPs in the 1980s. In 1988, during the height of the HIV epidemic in the United States, a federal funding ban for NEPs was put in place by Congress, which could only be lifted if it was determined that NEPs were effective 'in reducing drug abuse and transmission of HIV infection' (US Code Title 42: The public health and welfare). This 1988 ban set the stage for decades of policy driven by politics, not science, and an ideological war against providing clean needles for IDUs (Weinmeyer, 2016). Since the introduction of the federal funding ban, efforts by Democratic leaders to finance NEPs with federal dollars were defeated by five presidential administrations, regardless of the growing body of scientific evidence in support of NEPs, and increases in in intravenous drug use accompanied by a rise in transmission of HIV and HCV through shared needles (Mathers et al. 2013).

Fear tactics

In the United States the sharing of needles is due to fear of arrest for needle possession or violation of state laws and ordinances (Vlahov and Junge 1998).

Some law enforcement authorities resorted to scare tactics to demonstrate their opposition to needle exchanges. Officers threatened patrons at local NEPs and IDUs were afraid of being arrested if found exchanging old syringes for new ones. Other law enforcement associations lobbied against any legislation that would protect IDUs against the possession of illicit syringes (Cohen, 2003; Beletsky et al. 2011).

Attacks on science

Conservative interest groups and policy organizations have continued to oppose harm reduction policies with arguments tied to symbolism and moral philosophy. For example, the Drug Free America Foundation is a well-known group particularly adamant against harm reduction initiatives, including NEPs. They framed their arguments on the notion that needle exchanges promote drug use and enable drug users (Drug Free America Foundation, 2020). Other conservative groups also propagated the perception that injection drug users were not as important or worth saving as other members of society (Brownstein, 2016).

Opponents of NEPs also reference and misinterpret studies to show the failures by NEPs to curb drug use, even though subsequent research always upheld the views of the scientific community (Drug Policy Alliance, 2012). One study conducted in Montreal was repeatedly mentioned since it indicated that NEPs increased HIV transmission rates. However, the authors later claimed that the participants were not a representative population, and a large majority had a higher risk of infection (O'Gurek and Kirchner 2010). Despite the negation, Republicans, and opposition groups used these studies to add greater legitimacy to their support of the federal ban.

Policy window for federal funding

Republican leaders avoided addressing the growing public health crisis or legislation associated with NEPs, until the epidemic hit in their own states among their own constituents. At this time, the United States was experiencing rapidly increasing addiction rates, accompanied by a surge in HCV and HIV infections, predominantly affecting the Midwestern and Southern states, especially Kentucky, Indiana, and Tennessee (Zibbell et al. 2015). These states also had the strictest laws against NEPs and prevention initiatives like HIV testing.

In early 2015, Scott County, a region in southeastern Indiana, had a massive outbreak of HIV infections linked to IDUs injecting prescription opioids (Conrad et al. 2015). Republican Governor

Mike Pence authorized a needle-exchange programme to stall the HIV outbreak, effectively overriding state laws that prohibit syringe programmes (Rich and Adashi 2015). During the same period, Kentucky saw a surge of hepatitis C infections in nearly 54 counties, also linked to IDU. With this latest outbreak, a window of opportunity opened for bipartisan legislation to end the decades long ban for NEPs. Two Republican leaders from Kentucky ensured language was included in the 2016 legislative spending bill to end the federal funding ban. Some members of Congress who had historically supported the ban, finally acknowledged that legislation was necessary to counter the growing public health crisis.

At first glance, it finally looked like a win for advocates and Democrats, however, the federal funding ban was not *fully* repealed. The bill currently includes language that will allow state and public health departments to use federal funds for NEPs, 'as long as the federal funds do not purchase the syringes' (Consolidated Appropriations Act 2016). Federal funds for NEPs can also only be used in areas with high rates of IDU, deemed eligible by the Centers for Disease Control and Prevention (CDC). Furthermore, the bill does not guarantee any more federal funds for NEPs, just a shift in funding.

Although though it took nearly three decades, the lifting of the ban in 2016 signals a shift in conservative politics towards implementing evidence-based preventive harm reduction policies.

The politics of women's reproductive health

Tremendous improvements globally have been documented in recent decades in women's reproductive health and related quality of care. From 1990 to 2015, maternal mortality dropped by about 44% worldwide (Alkema et al. 2016). Despite this progress, significant challenges remain in meeting the reproductive healthcare needs for women and girls. Access to family planning services, including contraception, is still significantly limited in many low- and middle-income countries, and those countries bear a disproportionate burden of maternal mortality and morbidity. Recent estimates indicate that nearly 810 women die every day from preventable causes associated with pregnancy and childbirth, such as high blood pressure, delivery complications, and unsafe abortions; and approximately 94% of maternal deaths worldwide occur in low and-middle income countries (World Health Organization 2019).

The accessibility and availability of quality family planning and reproductive health services, including safe abortion services and effective contraceptive methods, are key components to preventing maternal mortality and morbidity (Filippi et al. 2016). A recent study of 172 countries found contraceptive use in one year avoided approximately 272,040 maternal deaths (a 44% reduction), and would have prevented an additional 102,000 deaths (29% reduction) if the unmet need for contraceptives been addressed (Ahmed et al. 2012). Research has further shown that the use of effective

family planning and contraception methods decreases the number of unintended pregnancies, including among adolescent girls, results in significantly fewer unsafe abortions, and is also an effective approach to reducing costs associated with HIV/AIDS (World Health Organization 2018a).

Although access to effective sexual and reproductive healthcare has been shown to reduce national and global rates of maternal mortality, as well as pregnancy-related health risks, many countries fail to provide these services, and women face significant barriers to accessing them. It is estimated that about 225 million women of reproductive age in developing countries lack access to birth control and effective family planning services primarily due to cultural or religious opposition, poor access to quality health services, or restrictive laws and regulations (Sedgh et al. 2016).

The US, because of its own domestic policy debates, has played a disproportionate role in curtailing access to and information about safe abortions globally. One policy, created by executive mandate of the US President without congressional approval, and that affects countless women worldwide, is the 'Mexico City Policy', also known as the Global Gag Rule. This rule initially implemented under the Reagan administration in 1984, curtailed US Agency for International Development family planning funds going directly or indirectly to non-US-based non-governmental organizations (NGOs) (international or regional NGOs located outside the United States and local NGOs) that provided information efforts, or counselling about or access to safe and legal abortions (Crane and Dusenberry 2004). The Gag Rule policy established a dangerous precedent for curtailing women's rights for safe and effective sexual and reproductive healthcare worldwide—effects well beyond the intended restriction on provision of abortion information and services—and which has caused significant damage towards women's access to services.

Research on the impact of the Global Gag Rule found evidence that the Mexico City Policy, because of its overall effect on the NGO community including limiting access to family planning services, is associated with increases in abortion rates in several sub-Saharan African countries. One study found that abortion rates rose dramatically in countries with high exposure to the Gag Rule (namely, those most vulnerable to the funding restrictions) compared to those with low exposure (see Fig. 2.2.2) (Bendavid et al. 2011).

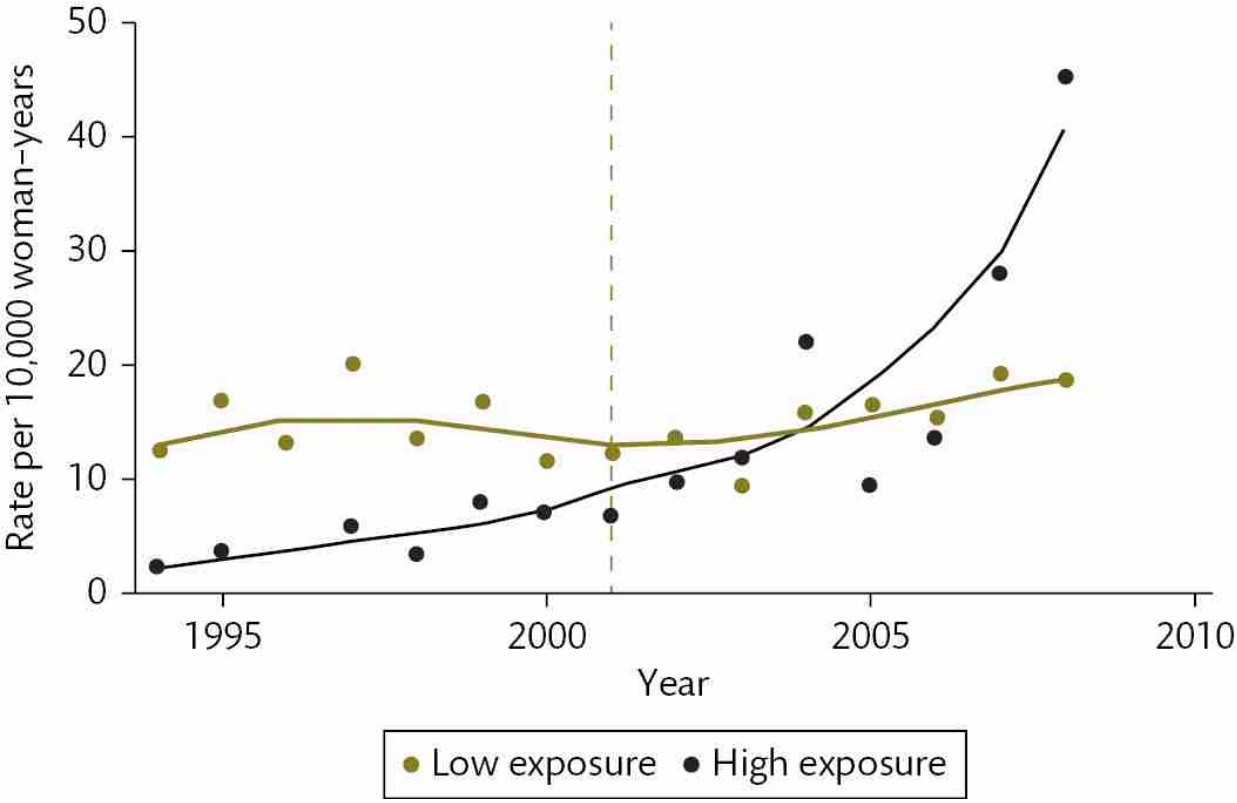


Fig. 2.2.2 Abortion rates in 20 sub-Saharan African countries, by exposure to Gag Rule, 1994–2008. Reprinted from Bendavid, E., Avila P., and Miller, G. United States aid policy and induced abortion in sub-Saharan Africa. *Bull World Health Organ*. 2011. Vol. 89. Number 12. doi: 10.2471/BLT.11.091660. Figure used under Creative Commons Attribution 3.0 IGO licence (CC BY 3.0 IGO).

The Gag Rule has historically been supported by Republican Presidents (and removed by Democratic ones), reflective of highly partisan views (in part driven by moral and religious ideologies) of women's reproductive rights in the United States. The policy was recently reinstated by President Trump in January 2017 (Congressional Budget Office 2017), and as with many such policies in this administration, broadened even further than previous norms. Currently, the scope of the Gag Rule goes beyond US AID family planning funding (about $700 Million USD per annum) to include most US Government global health funding (about $9 Billion USD per annum). The consequences of this broadly expanded Global Gag Rule are as yet unmeasured, but are likely to include far more negative consequences than during previous Republican administrations.

Case study Reproductive health rights for women in the US

Despite formidable odds, President Barack Obama signed the Patient Protection and Affordable Care Act (ACA) on 23 March 2010. This historic legislation also marked a new era of partisanship in Washington as well as the beginning of endless Republican congressional efforts to repeal the law. Prior to the implementation of the ACA, access to healthcare and preventive services was a significant challenge for millions of Americans. Following the law's passage, millions more gained health insurance, along with coverage for preventive care services without out-of-pocket costs to the consumer. The original legislation did not include key preventive service coverage for women. This was addressed by an amendment to the ACA that resulted in the formation of a review process led by the Institute of Medicine (now, National Academy of Medicine) to identify gaps in coverage for women that should be included as benefits (Council 2011). After a contentious roll out of these added benefits, the small but vocal opposition was largely focused on the decision to include the full range of US Food and Drug Administration (FDA) approved contraceptives. Soon coined by opponents as the 'contraception mandate' (no more a mandate than any of the other listed evidence-based preventive measures similarly offered without cost-sharing, such as screening colonoscopy or mammography), the inclusion of family planning services as an insurance benefit has been strongly opposed by most Republicans, some employers, and many social conservative and religious organizations.

The science

There is significant research to indicate that expanded family planning coverage has largely benefited women and teens of reproductive age in the United States. Following the implementation of the policy of expanded insurance coverage, the number of US abortions decreased by more than 132,000 between 2011 and 2014 (the abortion rate fell to 14.6 abortions for every 1,000 women between the ages of 15–44, a 14% drop since 2011) (Dreweke 2017). In addition to the health benefits of reductions in unintended pregnancies and concomitant abortions, there are other health benefits including decreased maternal morbidity and likely benefits to both mother and child of more optimal birth spacing (Cleland et al. 2012). Contraception use also confers several economic benefits; for example, prior to the ACA's mandate, out-of-pocket expenses for contraceptives was a large portion of women's medical expenditures, amounting to about 21 percent of their total out-of-pocket healthcare spending (Taylor and Mhatre, 2017). Recent evidence indicates that women saved $1.4 billion in out-of-pocket payments in 2013 resulting from the ACA's inclusion of family planning as a covered preventive service (Polsky 2015).

The players (The Who)

The ACA family planning benefits expansion drew a broad array of opponents despite broad support among the population. Opponents include religious organizations (notably, Catholic-affiliated organizations and employers), antiabortion groups, and the majority of Republican lawmakers. Reproductive health access opposition was driven by several components, based on religious, ideological, and political grounds.

Those in support of greater access to contraception and reproductive health rights for women in the United States tend to include most Democrats along with women's health advocacy groups, and pro-choice organizations. Universal access to contraception in the United States is also largely supported by key professional and academic organizations, including the National Academy of Medicine, the American Academy of Pediatrics, the American Academy of Family Physicians, the American College of Obstetricians and Gynecologists, and the American Public Health Association.

United States Supreme Court and the ‘contraception mandate’

The Supreme Court has played a significant role in the contraception debate in the United States In 2012, Hobby Lobby, an arts and crafts company founded and owned by the Evangelical Christian Green family, filed a lawsuit against the United States over the new regulation requiring health insurance be provided to cover the full range of contraception methods, citing objections to the inclusion of emergency contraception (often known as Plan B or the ‘morning after’ pill) and intrauterine devices (IUDs). In the Supreme Court Case (Burwell vs. Hobby Lobby, decided in July of 2014), a divided court 5–4 struck down the ‘contraception mandate’ for family-owned businesses, based on alleged violation of the federal Religious Freedom Restoration Act (Luchenister 2015). This was the first time in US history that the court ruled that a private corporation can claim to have religious beliefs and deny employees a federal benefit.

Attacks on science

Several tactics were used by opponents to undermine the science behind the family planning rule; challengers attempted to weaken support for the evidence-based policy with conservative rhetoric and delay tactics through litigation, motivated by ideology and partisan politics. For example, religious organizations and social conservatives, as well as the owners of the Hobby Lobby, erroneously equated emergency contraception drugs with abortion (Carroll 2014). Although this was unsupported by any empirical evidence, religious leaders and conservative lawmakers continue to spread the notion that certain family planning methods and contraception options, such as IUDs, are a form of abortion and thus constitute a violation of faith; this rhetoric was used to garner support from their constituent-base and antiabortion activist groups.

The future of women’s reproductive health rights

Although the majority of the American public support mandated coverage of birth control in health plans, the law’s family planning inclusion as a preventive benefit remains a politically polarizing and controversial topic, primarily along partisan lines (Moniz et al. 2014). Republican party leaders and lawmakers are increasingly gaining support to restrict reproductive rights for women; Republican-leaning states have also slowly eroded women’s access to safe abortions and reproductive health rights in recent years; between 2011 and 2016 states enacted nearly 334 abortion restrictions (Nash 2016). There are now new efforts to reduce reproductive health funding and access for women under President Trump, and conservative lawmakers continue to attack the ACA’s contraception coverage and funding for women’s health. Although not a problem unique to the United States, as other countries are in general broadening access to such services, the United States is experiencing a serious erosion of women’s sexual health and reproductive rights.

Evidence-based policy challenged by politics: income inequality and health

Despite recent increases in overall economic growth, global income and wealth inequality have been rising steadily since 1980; for example, the wealthiest 1% have increased their total wealth twice as much as the poorest 50% of the world’s population (Alvaredo et al. 2018). In some countries,

including India, China, and the United States, growth in income inequality between 1980 and 2016 has increased exponentially, as a result of various factors- including educational inequalities, tax policies, and deregulation. Conversely, the rise in inequality has grown only moderately across continental Europe, primarily attributed to social, educational, and tax-related policies that favoured the middle- and lower-classes (Alvaredo et al. 2018).

There are several methods for assessing the level of inequality in society. One widely accepted global measure is the GINI index, which is used for evaluating the level of income inequality within and across different countries (De Maio 2007); the index is a statistical measure to represent income (or wealth) distribution in a country. A coefficient of 1 (or 100%) represents maximum inequality (e.g. all wealth in one individual), while perfect equality is represented by a coefficient of 0.

Using the GINI index from World Bank estimates, South Africa and Zambia, as shown in Table 2.2.1, are among the most unequal countries in terms of income distribution—while Ukraine, Sweden, and Norway are among the most equal nations in the world. China and the United States have similar levels of income inequality and fall in the middle of country rankings in terms of inequality (World Bank 2018).

Table 2.2.1 Global income inequality using the GINI Index

Country	GINI %
South Africa	63.0
Zambia	57.1
Brazil	51.3
China	42.2
United States	41.5
United Kingdom	33.2
Sweden	29.2
Norway	27.5
Ukraine	25.0

Lower % reflects less inequality.

Source: data from World Bank Development Research groups GINI Index estimates (2012–2016).

Substantial evidence shows that poverty in and of itself is a significant risk factor for poor health. In recent decades, evidence is accumulating that shows income and wealth inequality, in addition to poverty *per se*, are independent drivers of mortality and morbidity (Pickett and Wilkinson 2015; Williams and Rosenstock 2015). The additional role of another socioeconomic status indicator, namely, wealth and income inequality, is not surprising in the context of a Social Determinants of Health Framework, which posits a wide array of factors beyond biology and access to healthcare that have outsized influence on health throughout the life cycle (Friedman and Starfield 2003).

A WHO Commission employed this framework to elucidate the main pathways through which social, economic, and political mechanisms influence socioeconomic status and social positions, inequities in education and household wealth, and how this shapes health status and vulnerabilities to specific health conditions. Social position and socioeconomic status are also influenced by income, since poor health status may affect employment, or other opportunities in various social and economic structures (Solar and Irwin 2010).

A substantial body of literature demonstrates that from a population level-perspective, economic and social deprivation have a direct link on incidence of disease. Several studies have evaluated the relationship between low socioeconomic status and increased risk of morbidity and mortality (Brown et al. 2004; Bosworth, 2018). This evidence is reflected in life expectancy across countries- income inequality (as measured by the GINI coefficient) is one factor in explaining differences in life expectancy. Individuals living in nations with lower levels of inequality have been shown to have longer life expectancies (Deaton 2003). In 2016, the average life expectancy in Sweden is 82.4 years (one of the most equal countries), the US average life expectancy is 78.5, and South Africa is 63.6 years (one of the most unequal countries) (World Health Organization 2015).

Evidence also shows that child and adolescent mortality is higher in countries with increased income inequality. One study found, using the GINI coefficient, that income inequality was positively associated with increased all-cause and communicable disease mortality in both males and females (Ward and Viner 2017). Additionally, researchers have also found evidence to suggest direct links between wealth inequities and mental health and overall psychological well-being; a recent review evaluating the association between income inequality and depression found that countries with larger gaps between the rich and poor have a higher incidence of mental illness (Patel et al. 2018).

In sum, the evidence supports that higher stress and depression, and higher rates of morbidity and mortality, occur among countries with significant income inequity. Evidence further indicates that greater income equality leads to slower economic growth, and leaves countries more vulnerable to political corruption.

Case study The politics of US income inequality in a global context

By several measures, income inequality in the United States has risen since 1980. The wealth gap in the United States has also increased steadily in the past few decades, and is currently the highest in the country's history. As of 2016, the wealthiest 1% of Americans own nearly 40% of the nation's wealth (Wolff 2017).

The relation between income inequality in select developed countries and an index of social, economic and health outcomes (including mental illness, obesity, imprisonment, education, and social mobility) is shown in Figure 2.2.3. There is a clear correlation between rising inequality and poor outcome, with the United States an outlier in both measures.

The US is the only advanced economy without universal healthcare; the country also performs poorly on several indicators, compared to other nations, including maternal mortality, and life

expectancy. Among the 37 countries in the Organization for Economic Cooperation and Development (OECD), the US ranks 28th in life expectancy at birth (OECD 2020).

In addition to the growing gap between rich and poor, the US is also experiencing a dramatic reduction in the size and standing of the middle class. Middle-class Americans have experienced a much greater decline compared to other advanced economies across continental Europe (Kochhar 2017). Research also shows growing income disparities by race and ethnic groups—the racial and ethnic disparity in wealth holdings widened considerably between 2007 and 2016 (Wolff 2017).

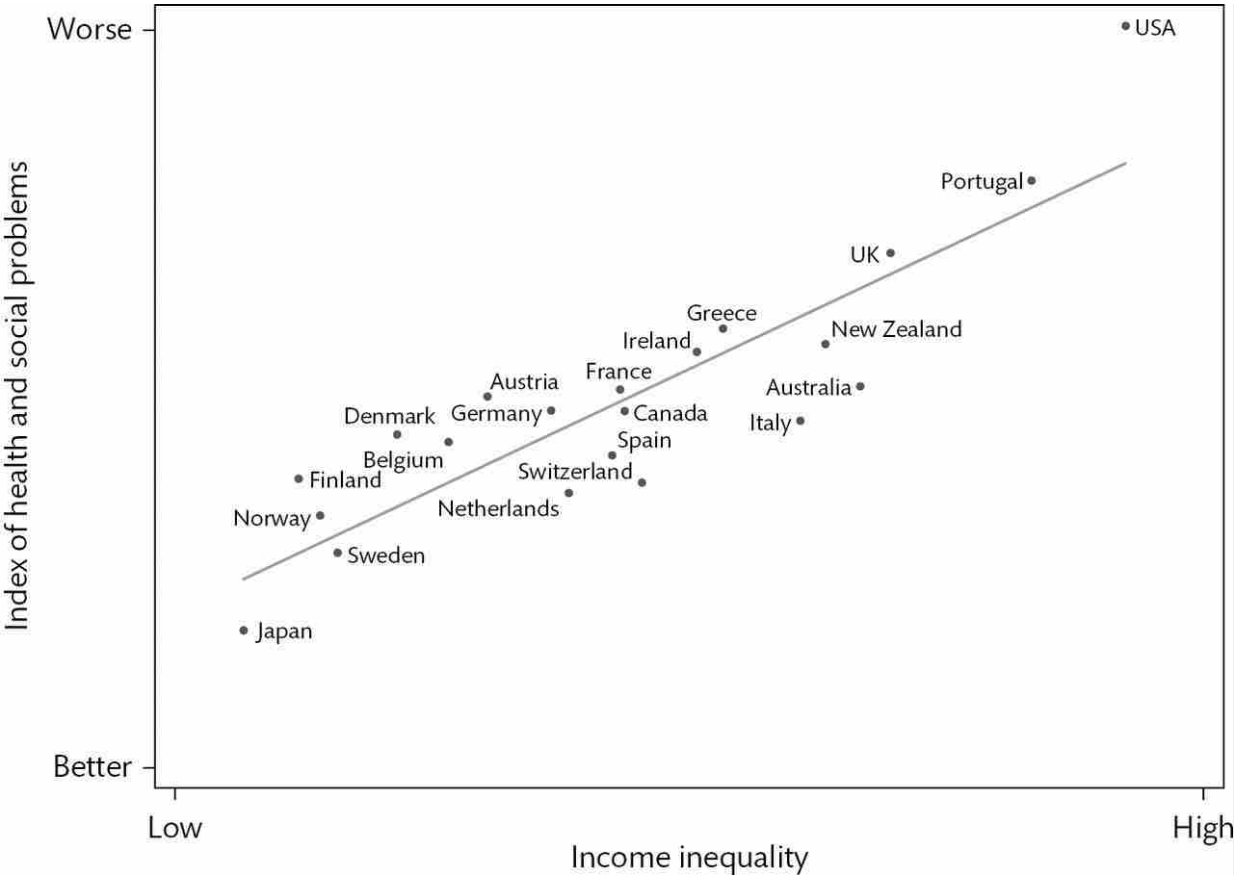


Fig. 2.2.3 Index of health and social problems.

Reprinted from Wilkinson RG, Pickett KE. 2009. The Spirit Level: Why More Equal Societies Almost Always Do Better. London: Penguin with permission author.

The political context of rising inequality in the US

Although inequality has been rising for decades in the US among both Republican and Democratic administrations, there is a clear partisan divide about perception of the issue and responses to it. Republicans are more

likely to associate a person's financial standing to his/her own hard work, and are also less likely to favour taxes for the wealthy that mitigate inequality, or advocate for policies to raise the minimum wage; however, Democratic counterparts tend to argue that social and economic determinants, and institutional factors embedded in society, such as structural racism, are the primary drivers of poverty and inequality. Democrats also favour policies aimed to support lower and middle-class families (Smith 2017).

Partisan divide over solutions for inequality

This partisan divide towards solutions to address the widening income gap are ever more apparent in two recent legislative bills: The Patient Protection and ACA, passed in 2010, and the recent Republican Tax Bill of 2017.

The ACA was passed by President Obama and Democrats in 2010. This remarkable legislative achievement—the first major health and social legislation in decades—had a significant impact on inequality for the nation. The goal of the ACA was to make health insurance more affordable for low- and moderate-income families. Following the law's implementation, nearly 20 million more Americans gained health insurance, reducing the uninsured rate to the lowest level on record (Cohen and Martinez 2016). It also reduced inequality (increased high earning taxes) and redistributed benefits, made many more families financially secure (blunting for lower-income shocks of healthcare cost), and lowered the federal deficit. A recent report found healthcare access for people in lower socioeconomic strata improved following ACA implementation and reduced disparities, with the gap in insurance coverage between higher- and lower-income households falling (Griffith et al. 2017).

In contrast, President Trump and Republican lawmakers passed the 'Tax Cuts and Job Act of 2017', the largest tax reform in decades, which effectively redistributed wealth to leading corporations and the wealthiest Americans. President Trump and Republicans claimed several victories in passing this bill: lowered the individual top tax to 37% (from 39.5%), lowered corporate tax rate to 21% from 35%, and repealed the ACA individual mandate (the mandate not only assured through penalties broad participation and hence stabilization in the insurance markets, but also banned health insurance companies from discriminating against individuals

with pre-existing conditions). Although the 2017 tax legislation does reduce taxes, if only slightly, across all income groups, the bill will continue to increase income inequality; higher income households will receive larger average tax cuts as a percentage of after-tax income; the biggest cuts as a share of income will go to taxpayers in the 95th to 99th percentiles of the income distribution, and the wealthiest will receive the largest cuts of all (Tax Policy Center 2018).

Solutions to address global inequality

Substantial evidence indicates income inequality across regions and countries is driven by a multitude of social and political factors—as well as disparities in educational and economic opportunities. The downstream effects of increasing inequity not only harm individuals and their surrounding communities, as evidenced by worse health outcomes, increased crime rates and political instability, but impede investments in education and training programmes for future generations.

To address the rapidly increasing structural inequities, the WHO recently identified several evidence-based policy options, with a particular focus to improve social determinants of health within the four following areas: childhood education and development, fair employment, social protection (e.g. cash transfers), and living and environmental conditions (e.g. housing affordability, legal protections) (Saunders et al. 2017). The various mechanisms to address systematic inequities in education, health, food security, etc., integrate several policy frameworks; however the underlying construct through which equity can be obtained is through short- and long-term investments in social determinants of health. With a more explicit policy focus to address upstream factors, such as economic stability and physical environment, regions and countries will see improvements in long-term economic growth, as well as improvements in health outcomes for the population (Nichols and Taylor 2018).

Conclusion

Decision makers can mitigate threats to science (the public health evidence-base) by recognizing that broad-based support and public education are key to enacting and enforcing effective health policies, on both a domestic and

global scale. The mainstream media as a purveyor of trusted information to the public needs to be protected from politically motivated attacks (e.g. in both the US and globally, the rhetoric in some quarters about false vaccine-related threats to health are aided when the public no longer knows who to trust).

The role of science in public health today is critical for evidence-based policy, and crucial for improving public health worldwide. Advancement of policy based on sound science will continue to be necessary to combat growing global public health crises and respond effectively to new public health threats. Understanding the players, their motivations, and tactics used, is fundamental to analysing the role of politics in the process.

REFERENCES

Ahmed, S., Li, Q., Liu, L., and Tsui, A.O. (2012). Maternal deaths averted by contraceptive use: an analysis of 172 countries. *The Lancet*, **380**, 111–25.

Alkema, L., et al. (2016). Global, regional, and national levels and trends in maternal mortality between 1990 and 2015, with scenario-based projections to 2030: a systematic analysis by the UN Maternal Mortality Estimation Inter-Agency Group. *The Lancet*, **387**, 462–74.

Alvaredo, F., et al. (2018). *World Inequality Report 2018*. Cambridge, MA: Harvard University Press.

Ball, A.L. (2007). HIV, injecting drug use and harm reduction: a public health response. *Addiction*, **102**, 684–90.

Beletsky, L., Grau, L. E., White, E., Bowman, S., and Heimer, R. (2011). The roles of law, client race and program visibility in shaping police interference with the operation of US syringe exchange programs. *Addiction*, **106**, 357–65.

Bendavid, E., Avila, P., and Miller, G. (2011). United States aid policy and induced abortion in sub-Saharan Africa. *Bulletin of the World Health Organization*, **89**, 873–80c.

Berer, M. (2017). Abortion law and policy around the world: in search of decriminalization. *Health and Human Rights*, **19**, 13–27.

Brown, A.F., et al. (2004). Socioeconomic position and health among persons with diabetes mellitus: a conceptual framework and review of the literature. *Epidemiologic Reviews*, **26**, 63–77.

Brownstein, Joseph. (2014). “Opposition to Clean Needles for Addicts: Symbolism Over Science?” Aljazeera America. Online. http://america.aljazeera.com/articles/2014/4/18/needle-exchange-programs.html

Bosworth, Barry. (2018). Increasing disparities in mortality by socioeconomic status. *Annual Review of Public Health*, **39**:237–251.

Brownson, R. C., Fielding, J. E., & Maylahn, C. M. (2009). Evidence-based public health: a fundamental concept for public health practice. *Annual Review of Public Health*, **30**, 175–201.

Carroll, A.E. (2014). How the hobby lobby ruling could limit access to birth control. *New York Times*. [Online] Available at: https://www.nytimes.com/2014/07/02/upshot/how-hobby-lobby-ruling-could-limit-access-to-birth-control.html

Carroll, Aaron E. (2014). How the Hobby Lobby Ruling Could Limit Access to Birth Control. New York Times.

Choi, B.C. (2005). Peer reviewed: twelve essentials of science-based policy. *Preventing Chronic Disease*, **2**, A16.
Clark, P.A. and Fadus, M. (2009). Federal funding for needle exchange programs. *Medical Science Monitor*, **16**, PH1–PH13.
Cleland, J., Conde-Agudelo, A., Peterson, H., Ross, J., and Tsui, A. (2012). Contraception and health. *The Lancet*, **380**, 149–56.
Cohen, Jonathan. United States, Injecting Reason: Human Rights and HIV Prevention for Injection Drug Users: California, a Case Study. Vol. 15. No. 2. Human Rights Watch, 2003.
Cohen, R.A. and Martinez, M.E. (2016). *Health Insurance Coverage: Early Release.* Washington, DC: National Center for Health Statistics.
Coleman. (2015). Austin. "Needle Exchange Programs." The Council of State Governments. Accessed https://knowledgecenter.csg.org/kc/content/needle-exchange-programs
Congressional Budget Office (2017). *Repealing the Individual Health Insurance Mandate: An Updated Estimate.* [Online] Available at: https://www.cbo.gov/publication/53300
Conrad, C., et al. (2015). Community outbreak of HIV infection linked to injection drug use of oxymorphone—Indiana, 2015. *Morbidity and Mortality Weekly Report*, **64**, 443–4.
Consolidated Appropriations Act (2016). HR 2029, PL 114–13, 114th Cong. at p. 257 (Dec. 18, 2015).
Council, N.R. (2011). *Clinical Preventive Services for Women: Closing the Gaps.* Washington, DC: National Academies Press.
Crane, B.B. and Dusenberry, J. (2004). Power and politics in international funding for reproductive health: the US Global Gag Rule. *Reproductive Health Matters*, **12**, 128–37.
De Maio, F.G. (2007). Income inequality measures. *Journal of Epidemiology & Community Health*, **61**, 849–52.
Deaton, A. (2003). Health, inequality, and economic development. *Journal of Economic Literature*, **41**, 113–58.
Degenhardt, L., et al. (2017). Global prevalence of injecting drug use and sociodemographic characteristics and prevalence of HIV, HBV, and HCV in people who inject drugs: a multistage systematic review. *The Lancet Global Health*, **5**, e1192–207.
Dreweke, J. (2017). US abortion rate reaches record low amidst looming onslaught against reproductive health and rights. *Guttmacher Policy Review*, **20**.
Drug Free America Foundation. (2020). Position Statements. Statement on So-Called "Harm-Reduction" Policies. https://www.dfaf.org/position-statements/
Drug Policy Alliance. (2012). Top Five Studies Misused by Syringe Opponents. https://harmreduction.org/hrc/wp-content/uploads/2012/01/TalkingPointsFiveStudies-DPA.pdf
Fernandes, R.M., et al. (2017). Effectiveness of needle and syringe programmes in people who inject drugs–an overview of systematic reviews. *BMC Public Health*, **17**, 309.
Filippi, V., Chou, D., Ronsmans, C., Graham, W., and Say, L. (2016). Chapter 3. Levels and causes of maternal mortality and morbidity. In: Black, R.E., et al. (eds.) *Disease Control Priorities* (3rd ed.). Washington (DC): The International Bank for Reconstruction and Development/The World Bank.
Friedman, D.J. and Starfield, B. (2003). Models of population health: their value for US public health practice, policy, and research. *American Journal of Public Health*, **93**, 366–9.
Griffith, K., Evans, L., and Bor, J. (2017). The Affordable Care Act reduced socioeconomic disparities in health care access. *Health Affairs*, **36**, 1503–10.
Hwang, S.W. (2007). Science and ideology. *Open Medicine*, **1**, e99.
Interview Survey (2016). [Online]. Available at: https://www.cdc.gov/nchs/nhis/releases.htm
Taylor J. and Mhatre N. (2017). Contraceptive Coverage Under the Affordable Care Act. Online. https://www.americanprogress.org/issues/women/news/2017/10/06/440492/contraceptive-coverage-affordable-care-act/

Jarlais, D.C.D. (2006). Mathilde KRIM, amfAR, and the prevention of HIV infection among injecting drug users: a brief history. *AIDS Patient Care & STDS*, **20**, 467–71.
Kochhar, R. (2017). Middle class fortunes in Western Europe. LIS Working Paper Series.
Luchenister, A.J. (2015). A new era of inequality: hobby lobby and religious exemptions from anti-discrimination laws. *Harvard Law & Policy Review*, **9**, 63.
Mathers, B.M., Degenhardt, L., Bucello, C., Lemon, J., Wiessing, L., and Hickman, M. (2013). Mortality among people who inject drugs: a systematic review and meta-analysis. *Bulletin of the World Health Organization*, **91**, 102–23.
Moniz, M.H., Davis, M.M., and Chang, T. (2014). Attitudes about mandated coverage of birth control medication and other health benefits in a US national sample. *JAMA*, **311**, 2539–41.
Myers, T., Aggleton, P., and Kippax, S. (2004). Perspectives on harm reduction: editorial introduction. *Critical Public Health*, **14**, 325–8.
Nash, E. (2016). *Laws Affecting Reproductive Health and Rights: State Trends at Midyear, 2016.* New York: Guttmacher Institute
Nichols, L.M. and Taylor, L.A. (2018). Social determinants as public goods: a new approach to financing key investments in healthy communities. *Health Affairs*, **37**, 1223–30.
OECD. (2020). Life expectancy at birth (indicator). doi: 10.1787/27e0fc9d-en
O'Gurek, D.T. and Kirchner, J.T. (2010). Needle exchange for HIV/AIDS: an effective public health strategy. *Journal of Lancaster General Health*, **5**, 9.
Okike, K., Kocher, M.S., Mehlman, C.T., and Bhandari, M. (2008). Industry-sponsored research. *Injury*, **39**, 666–80.
Patel, V., Burns, J. K., Dhingra, M., Tarver, L., Kohrt, B.A., and Lund, C. (2018). Income inequality and depression: a systematic review and meta-analysis of the association and a scoping review of mechanisms. *World Psychiatry*, **17**, 76–89.
Pickett, K.E. and Wilkinson, R.G. (2015). Income inequality and health: a causal review. *Social Science & Medicine*, **128**, 316–26.
Polsky, N.V.B.A.D. (2015). Women saw large decrease in out-of-pocket spending for contraceptives after ACA mandate removed cost sharing. *Health Affairs*, **34**, 1204–11.
Position Statements. Statement on So-Called "Harm-Reduction" Policies. https://www.dfaf.org/position-statements/
Rich, J.D. and Adashi, E.Y. (2015). Ideological anachronism involving needle and syringe exchange programs: lessons from the Indiana HIV outbreak. *JAMA*, **314**, 23–24.
Rosenstock, L. (2006). Protecting special interests in the name of 'good science'. *JAMA*, **295**, 2407–10.
Rosenstock, L. and Lee, L.J. (2002). Attacks on science: the risks to evidence-based policy. *American Journal of Public Health*, **92**, 14–18.
Saunders, M., Barr, B., McHale, P., and Hamelmann, C. (2017). *Key Policies for Addressing the Social Determinants of Health and Health Inequities*. Copenhagen: WHO Regional Office for Europe.
Sedgh, G., L. Ashford, and R. Hussain. "Unmet Need for Contraception in Developing Countries: Examining Women's Reasons for Not Using a Method. Guttmacher Institute, New York." New York NY: The Alan Guttmacher Institute (2016).
Shamasunder, B. and Bero, L. (2002). Financial ties and conflicts of interest between pharmaceutical and tobacco companies. *JAMA*, **288**, 738–44.
Smith, S. (2017). Why people are rich and poor: republicans and democrats have very different views. Pew Research Center. [Online] Available at: http://www.pewresearch.org/fact-tank/2017/05/02/why-people-are-rich-and-poor-republicans-and-democrats-have-very-different-views
Solar, O. and Irwin, A. (2010). *A Conceptual Framework for Action on the Social Determinants of Health*. Geneva: WHO. [Online] Available at:

https://www.who.int/sdhconference/resources/ConceptualframeworkforactiononSDH_eng.pdf?ua=1
Stone, K. (2016). *The Global State of Harm Reduction 2016*. London: Harm Reduction International.
Tax Policy Center (2018). *Analysis of the Tax Cuts and Jobs Act*. [Online] Available at: https://www.taxpolicycenter.org/feature/analysis-tax-cuts-and-jobs-act
US Code Title 42: The Public Health and Welfare Title 42: The public health and welfare. *Section 139a, State Plans for Medical Assistance*, 783.
Vlahov, D. and Junge, B. (1998). The role of needle exchange programs in HIV prevention. *Public Health Reports*, **113**, 75.
Ward, J.L. and Viner, R.M. (2017). The impact of income inequality and national wealth on child and adolescent mortality in low- and middle-income countries. *BMC Public Health*, **17**, 429.
Weinmeyer, Richard. (2016). Needle exchange programs' status in US politics. *AMA Journal of Ethics* 18.3: 252–257.
Williams, J.A.R. and Rosenstock, L. (2015). Squeezing blood from a stone: how income inequality affects the health of the American workforce. *American Journal of Public Health*, **105**, 616–21.
Witteveen, E. and Schippers, G. (2006). Needle and syringe exchange programs in Amsterdam. *Substance Use & Misuse*, **41**, 835–6.
Wodak, A. and Cooney, A. (2006). Do needle syringe programs reduce HIV infection among injecting drug users: a comprehensive review of the international evidence. *Substance Use & Misuse*, **41**, 777–813.
Wolff, E.N. (2017). *Household Wealth Trends in the United States, 1962 to 2016: Has Middle Class Wealth Recovered?* National Bureau of Economic Research. [Online] Available at: https://www.nber.org/papers/w24085
World Bank (2018). *GINI Index (World Bank Estimate) | Data*. [Online] Available at: https://data.worldbank.org/indicator/SI.POV.GINI?
World Health Organization (WHO) (2015). Global health observatory (GHO) data. [Online] Available at: http://www.who.int/gho/tb/en
World Health Organization (WHO) (2016). *People Who Inject Drugs*. [Online] Available at: http://www.who.int/hiv/topics/idu/about/en/
World Health Organization (WHO) (2018a). *Family Planning/Contraception*. [Online] Available at: http://www.who.int/news-room/fact-sheets/detail/family-planning-contraception
World Health Organization (WHO) (2019). *Maternal Mortality*. [Online] Available at: http://www.who.int/news-room/fact-sheets/detail/maternal-mortality
Zibbell, J.E., et al. (2015). Increases in hepatitis C virus infection related to injection drug use among persons aged ≤ 30 years-Kentucky, Tennessee, Virginia, and West Virginia, 2006–2012. *Morbidity and Mortality Weekly Report*, **64**, 453–8.

2.3

Poverty, justice, and health

Ronald Labonté, Fran Baum, and David Sanders

Introduction to poverty, justice, and health

> Poverty has to be understood not just as a disadvantaged and insecure economic *condition*, but also as a shameful and corrosive social *relation*. (p. 7, emphasis in original)
>
> Lister, R., *Poverty*, Polity Press, Copyright © 2004.

Poverty has long been a concern in public health. Not only do people living in poor circumstances generally suffer higher burdens of disease, but during periods of infectious epidemics or pandemics they are perceived as posing a risk to others as well. Studies even suggest that high poverty and disease rates together can contribute to social unrest and 'failed' (or failing) states (Cheek 2001; Hotez 2002; Singer 2002; Peterson and Shellman 2006). Apart from these more 'public' (or population) health risks, the persistence of poverty has also led many public health theorists and practitioners to ask why poverty continues to exist in times of great global wealth, and to argue that a socially just society would seek to eliminate poverty. The reasons for the persistence of poverty lie in how societies' economic and political systems allocate the opportunities and resources people need to be 'non-poor' and healthy, or intervene to reduce inequities (avoidable inequalities) in their access to services and resources. Thus, any discussion of poverty, and of its impacts on health, unavoidably intersects with understanding how inequalities arise in the distribution of income and wealth, and of the material and psychosocial resources these socioeconomic privileges accord. As the British economist-cum-philosopher, Adam Smith, known best for his writings extolling the virtues of free markets, noted:

> Wherever there is great property, there is great inequality. For one very rich man, there must be at least five hundred poor, and the affluence of the few supposes the indigence of the many. (Smith 1776, p. 419)

To further our understanding we need to answer the question of what poverty is exactly? How is it affecting health, and ill health affecting poverty? And what theories or practices of justice should guide public health interventions based on poverty's persisting health risks? These three questions form the basis of the chapter that follows.

What is poverty?

At first glance, it may seem simple to define poverty: the inability of people to provide for the basic necessities of life. But there are actually many forms of poverty, all of them important to public health and to public policy; and the concept of poverty itself is 'highly contested' (Alcock 1993). Its contestation is not simply theoretical (what does it mean?) or empirical (how do we measure it?) but also political, since it is the same sociopolitical structures and economic arrangements that give rise to both poverty and affluence.

Absolute poverty

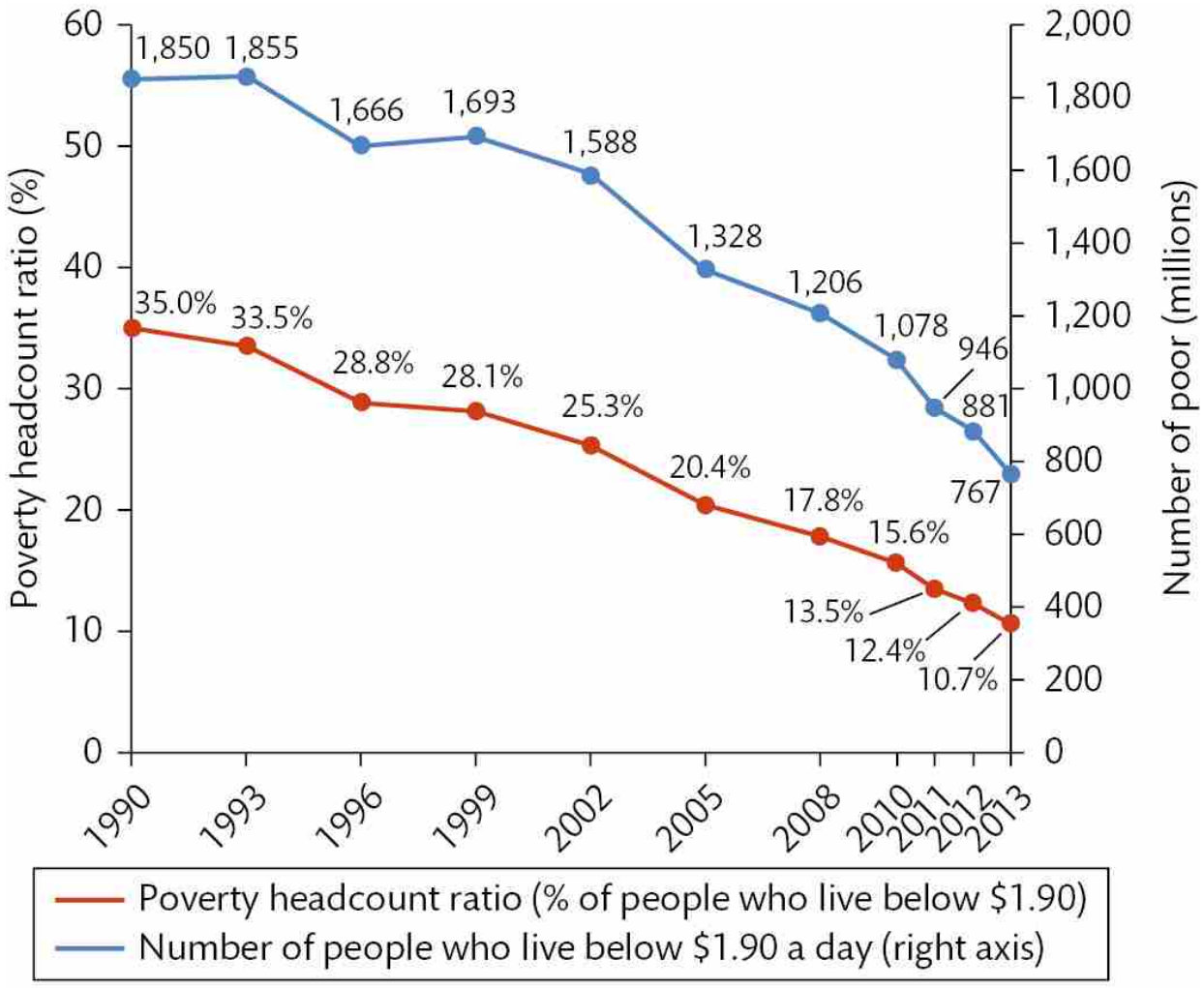
There are people in the world so hungry, that God cannot appear to them except in the form of bread.

Mahatma Gandhi

Absolute poverty is foundational to an understanding of the concept, and is commonly considered to be the lack of a minimum income necessary for simple survival based upon the price of a basket of 'essential' goods and services (Masters and Wickstrom 2006; Saunders et al. 2008). The World Bank standardized poverty measures globally through its 'dollar-a-day' poverty rate. The dollar-a-day level, close to the national poverty line in India, was subsequently raised to $1.25/day and then revised upwards once again to $1.90/day, which remains the World Bank's headline measure of 'extreme' or absolute poverty (Ferreira and Sanchez 2017). Extreme poverty, or what is now referred to as the 'lower bound' rate (Ravallion and Chen 2017), describes subsistence consumption at a level that is minimally life-supporting and 'exceptionally frugal . . . even by the standards of the world's poorest countries' (Chen and Ravallion 2012, p. 1). It remains associated with comparatively low life expectancies and high infant and maternal

mortality rates. (We discuss the relationship between poverty and health later in this chapter.)

Global poverty at the $1.90/day extreme level has fallen consistently since the baseline year of 1990, both proportionately (the ‘headcount index’) and in absolute numbers (Fig. 2.3.1) (World Bank 2016), often attributed to globalization-associated economic growth (Santarelli and Figini 2002; World Bank 2002). There are important regional variations. Poverty rates and absolute numbers fell most dramatically in East Asia and China. Poverty rates and absolute numbers rose in Latin America and the Caribbean during the 1980s and 1990s—the ‘lost decades’ that many attribute to structural adjustment programmes imposed by the World Bank and International Monetary Fund (Grindle 1996)—before beginning to fall in the 2000s. The recent decline in poverty in this region is a result of economic growth combined with redistributive social policies and programmes (Birdsall et al. 2011). Poverty rates in sub-Saharan Africa fell only slightly over this 30-year period, and only since the early 2000s, and failed completely to keep pace with population growth with the number of people living in extreme poverty rising from 276 million in 1990, to 389 million in 2013 (latest year available) (World Bank 2016).



Source: Latest estimates based on 2013 data using PovcalNet (online analysis tool), World Bank, Washington, DC, http://iresearch.worldbank.org/PovcalNet/.
Note: Poverty is measured using the US$1.90-a-day 2011 PPP poverty line.

Fig. 2.3.1 Trends in the Global Poverty Headcount Ratio and the Number of the Global Poor, 1990–2013.

Reprinted from World Bank. 2016. Poverty and Shared Prosperity 2016: Taking on Inequality. Washington, DC: World Bank. doi:10.1596/978-1-4648-0958-3. License: Creative Commons Attribution CC BY 3.0 IGO.

These data mean that the not very ambitious Millennium Development Goal of halving the number of people living in extreme poverty between 1990 and 2015 has been achieved. Of comparative significance is that the value of the global economy more than quadrupled during this same time period—from US $18 trillion in 1980 (De Long 1998) to US $80 trillion in

2011 (Central Intelligence Agency 2013), indicative that very little 'trickle down' of the benefits of that growth reached the 'bottom billion' (Collier 2007). Neither does halving these rates say anything of the adequacy of the poverty lines themselves, to which other analysts have responded with estimates of an 'ethical poverty line'.

The ethical poverty line

The ethical poverty line was first established by working backwards from countries with an average life expectancy at birth of 74 years (considered an ethical minimum) to the average level of consumption associated with such a life expectancy (around $3/day when the study was first made) (Edward 2006). More recent estimates of the income (measured as consumption) required for this still modest outcome (since many in high-income countries can expect to live well past 80 or even 90 years by 2030 (Kontis et al. 2017)) range between US$5/day and US$7.40/day (Hickel 2016).

The 2015 Sustainable Development Goals (SDGs) established a slightly more ambitious target than the MDGs: eliminate extreme poverty worldwide by 2030. But what of poverty at levels that approach the 'ethical' minimum? Using the US$5/day level (the minimum the United Nations Conference on Trade and Development in 2013 considered to qualify as a poverty line, and below which over 4.3 billion people presently live), and based on current poverty reduction trends, when 2030 dawns: 15% in Latin America and the Caribbean would still be poor, rising to 30% in East Asia and the Pacific, 50% in the Middle East and North Africa, and an astounding 90% in South Asia and sub-Saharan Africa (United Nations Conference on Trade and Development 2018). At the current rate by which the global economy has been reducing poverty, it would take 200 years to eliminate poverty at the US$5/day level, and over 300 years at the US$7.40/day level (Woodward and Simms 2015). Even if these more appropriate absolute poverty targets were eventually achieved, doing so under the current norms of economic growth would almost certainly destroy the ecological resources required for life long before achieving this effect.

Finally, the World Bank's estimates of people living in absolute poverty are generally based on what people report consuming. Since many of the world's poor consume healthcare for which they have to pay, this can result in medical poverty, when catastrophic illness forces people to sell their

assets and exhaust their savings to pay for treatment. These people may lift themselves out of 'consumption' poverty (since they are consuming healthcare at a high level) but leave themselves in income poverty (since they no longer have money for other goods or services). In 2005, the World Health Organization (WHO) estimated that:

> Each year 100 million people slide into poverty as a result of medical care payments. Another 150 million people are forced to spend nearly half their incomes on medical expenses. That is because in many countries people have no access to social health protection — affordable health insurance or government-funded health services. (WHO 2005).

This estimate is still considered to be largely accurate (Aizenman 2017), although there is some evidence that the number may be declining slightly (Wagstaff et al. 2018). Recognizing the seriousness of medical poverty, in December 2012, the United Nations General Assembly passed a resolution calling on Member States to develop universal health coverage systems to prevent significant payments at the point of delivery and to pool risks to avoid catastrophic healthcare spending and impoverishment (United Nations General Assembly 2012). This 2012 resolution has since become embodied in the 2015 SDG 3 (Good Health and Wellbeing), and its Target 3.8:

> Achieve universal health coverage, including financial risk protection, access to quality essential healthcare services and access to safe, effective, quality, and affordable essential medicines and vaccines for all. Achieving universal health coverage (UHC) has also become a platform initiative of the WHO.

Absolute (extreme) poverty informs public health by way of a presumption that people living at or below such poverty lines lack access to sufficient material resources to develop or maintain their health. This level of poverty represents a form more commonly associated with poor groups in low-income and middle-income countries than with poor groups in high-income countries. This poverty concept nonetheless applies to high-income countries where some households live in 'deep poverty', where income levels are 75% or less of nationally established poverty lines and are considered inadequate to meet even basic needs of food, shelter, and clothing (Beiser et al. 2002). Conservative think-tanks in high-income countries have sometimes argued that absolute poverty, which is indifferent to measures of income or wealth distribution, should be the only policy concern. Not all agree. Apart from growth-related poverty reduction being much less under conditions of high-income inequality (Bourguignon 2004), most of the

world's absolute poor now live in middle-income countries such as India and China, making governance, taxation, and redistribution of paramount importance in reducing levels of absolute poverty (Sumner 2010).

Recognizing that there is no longer a bifurcated world of 'developed' and 'developing' countries, the World Bank has organized nations into four groups based on their average per capita incomes (adjusted for purchasing power parity). Building upon a study that averaged national poverty line rates for countries at different development levels for which data were available (Jolliffe and Beer Prydz 2016), the World Bank defined poverty as $1.90/day for 33 low-income countries, $3.20/day for 32 lower-middle-income countries, $5.50/day for 32 upper-middle-income countries, and $21.70/day for 29 high-income countries (Ferreira and Sanchez 2017). In 2017, it also created what it called a poverty level 'upper bound', 'a schedule of weakly relative lines that rise with the … national poverty lines' (Ravallion and Chen 2017, p. 8) and adjusted to account for the effects of the 'very high incomes' of the top 1% (and even the top 0.1% and 0.01%) which has captured much of the world's income gains since 1980 (Alvaredo et al. 2018). Using the upper bound measure, world poverty since 1990 declined very little, and actually continued to increase (in actual numbers of poor people) in Latin America, the Middle East and North Africa, South Asia, and sub-Saharan Africa (Ravallion and Chen 2017).

Relative poverty

> Poverty means going short materially, socially and emotionally. It means spending less on food, on heating and on clothing than someone on an average income—above all, poverty takes away the tools to build the blocks for the future. It steals away the opportunity to have a life unmarked by sickness, a decent education, a secure home. (p. 3)
>
> Reproduced from Oppenheim C. and Harker L., Poverty: the facts, Child Poverty Action Group, London, UK, Copyright © 1990.

The World Bank's use of 'weakly relative lines' of poverty introduces the concept of relative poverty, and the idea that it is not enough simply to be able to meet the material necessities of life. One is poor if one lacks the resources required to engage in some meaningful ways in society and its sundry activities (social, economic, political, cultural, and recreational). That is, relative poverty is concerned with living standards—which vary by place and over time, but which convey the notion of comparatively full social functioning. This concept of relative poverty as the inability to function

socially to some normative standard can be traced in Western thought as far back as Aristotle (Sen 2000). Peter Townsend was one of the first of the recent theorists to criticize the policy use of absolute poverty.

He, and others since, argued for a concept of ‘relative poverty’ based less on minimal needs for survival and more on needs conditioned by societal expectations—that is, on what people require to participate fully in society (Nunes 2008; Eurostat 2010). This concept of relative poverty is actually reasonably old even in liberal economic thought, having been famously proposed by Adam Smith in his statement that ‘a creditable day-labourer would be ashamed to appear in public without a linen shirt, the want of which would be supposed to denote that disgraceful degree of poverty’ (Labonté et al. 2012, p. 15).

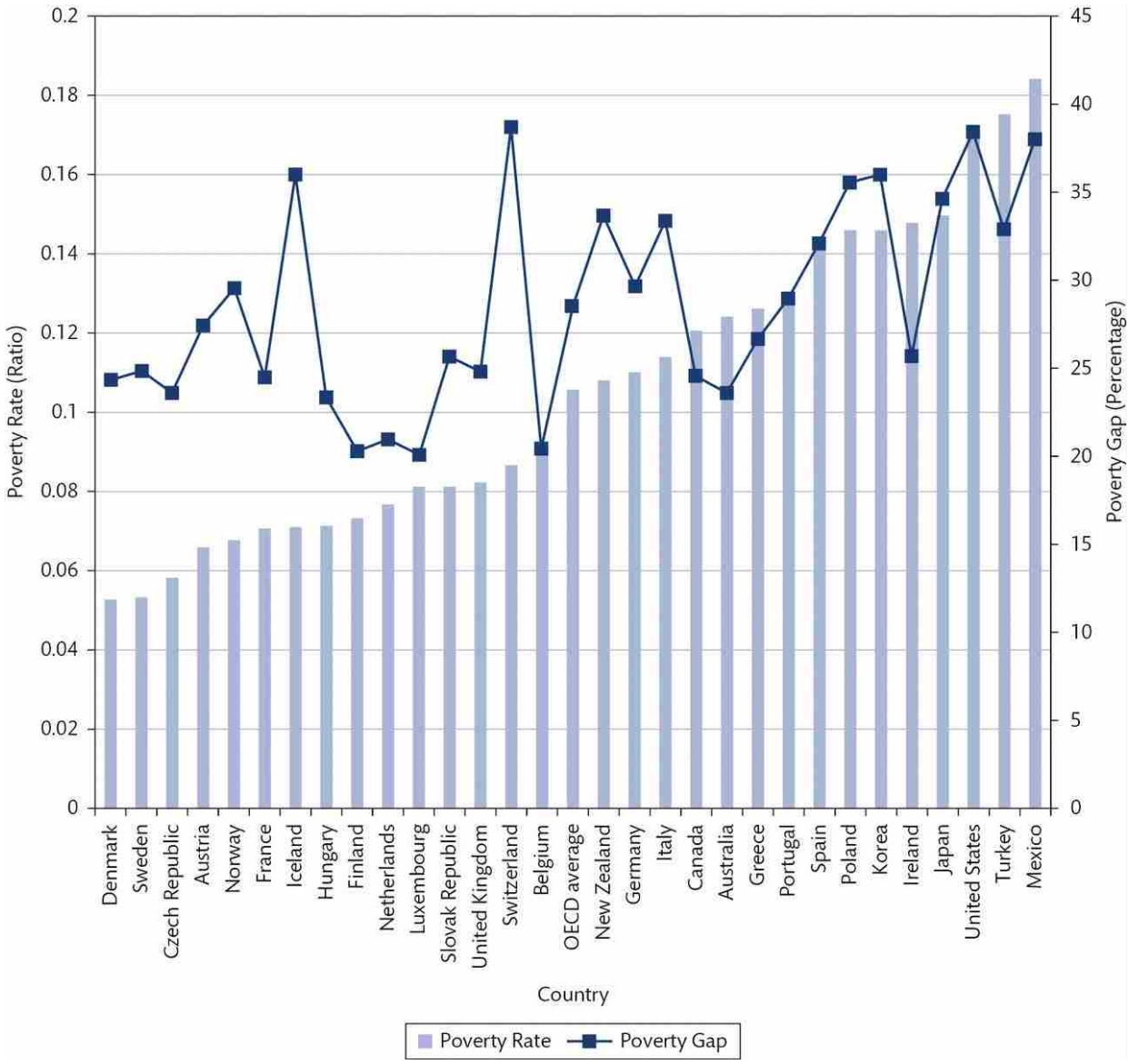


Fig. 2.3.2 Poverty rate and poverty gap, selected OECD countries.

Adapted with permission from Organisation for Economic Co-operation and Development (PECD), *OECD Factbook 2010: Economic, Environmental and Social Statistics*, OECD Publishing, p. 237, Copyright © 2010, DOI: 10.1787/factbook-2010-en.

As Martin Ravallion, the World Bank's expert on global poverty and lead analyst for the absolute measures discussed earlier, points out, this passage from Smith implies that 'certain socially-specific expenditures are essential for social inclusion, on top of basic needs for nutrition and physical survival' and that 'the way this idea is implemented in practice is to set a "relative poverty line" that is a constant proportion of average income for the country' (Ravallion 2011).

The more common relative poverty measure is having a net household income (one that takes into account government cash transfers and taxes, and that is adjusted for family size and often also for urban or rural location) that is below 50% of the country median. This measure is commonly used for comparative studies within or across high-income countries, i.e. those belonging to the Organisation for Economic Co-operation and Development (OECD). Figure 2.3.2 provides an example using OECD countries, which includes the 'poverty gap'—the median of how far below the relative poverty line a household falls and a measure of 'deep poverty' (OECD 2010). Figure 2.3.3 focuses on relative child poverty in economically advanced countries (UNICEF 2017) defined as living in a household with income level lower than 60% of the median. A relative concept of poverty informs public health by expanding the range of goods and services that members of a household should be able to access to experience, in Ravallion's terms, social inclusion, and the mental and physical health benefits such inclusion might garner. At the same time, the monetization of the concept (pegging it to a percentage of average or median income) has been criticized since it represents a form of poverty that will never disappear. As average or median incomes rise in a population, there will always be some who will fall below these cut-offs. It is possible for absolute poverty to fall within a country or region, while relative poverty rises; since the relative poverty cut-offs can be affected by disproportionate increases in income among a minority of the population, even if the material conditions of those at the bottom of the income gradient improve. As an earlier UNICEF Innocenti Research Centre (2012) warns, relative poverty is not a good metric for comparison across countries except for those wealthier nations where falling below 50% of the median income does risk social exclusion (a concept we define later in this chapter).

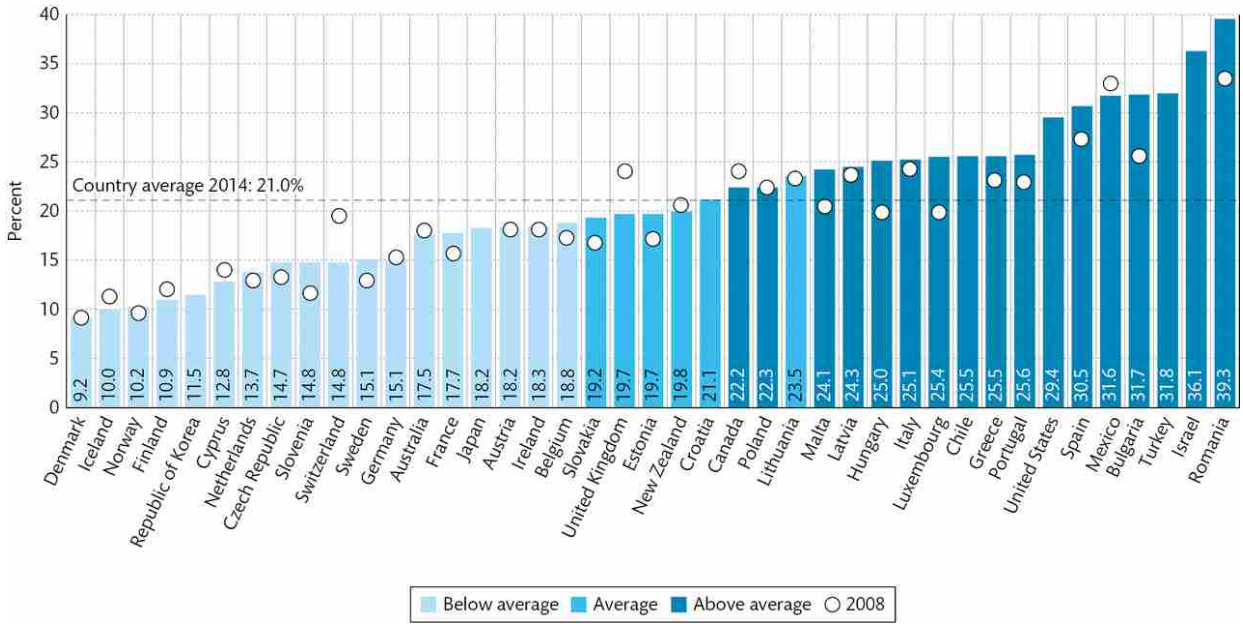


Fig. 2.3.3 Percentage of children aged 0–17 living in a household with income lower than 60% of the median, 2014 and 2008.

Reprinted from UNICEF Office of Research (2017). 'Building the Future: Children and the Sustainable Development Goals in Rich Countries', *Innocenti Report Card 14*, UNICEF Office of Research—Innocenti, Florence.

Poverty as lack of capabilities

> The capability approach . . . builds a bridge between absolute and relative concepts of poverty. It is absolute in identifying the capabilities or functioning that people should experience, but relative in the income or resources that may be required.

Reproduced from Labonté, R. et al., Indicators of Social Exclusion and Inclusion: A Critical and Comparative Analysis of the Literature, E/Exchange Working Paper Series, PHIRN, University of Ottawa, Canada © 2012, with permission from the author.

While relative poverty presents a more nuanced understanding of 'being poor' than does absolute poverty, it still relies upon indicators of income. Amartya Sen (1985), in his pioneering capability approach to poverty, focused attention on the infrastructural enabling conditions ('capabilities') that allow people to attain desirable states ('functioning'), without which a minimally dignified human existence is not possible. This has brought in different kinds of indicators that are much more transparent and direct reflections of people's actually achieved levels of well-being and ability 'to accomplish what [they] value' (Alkire 2002, p. 6, cited in Nunes 2008). Income is simply a means to a valued and socially purposeful end. Sen

considered indicators such as life expectancy, literacy, and infant mortality to be important proxies for capabilities (at least at a population level) and more useful than income alone to identify poor from non-poor groups or countries. He proposed that indicators of these other characteristics should be aggregated into a more complex poverty measure (Nunes 2008). This approach became the basis of the Human Development Index (HDI), first released in 1990 and issued annually by the United Nations Development Programme (UNDP), and which combines data on health, education, and living standards (this last dimension using the income indicator of log gross domestic product (GDP)/capita) (UNDP 2011). The capability approach from which the HDI arose has since been considerably elaborated (Ranis et al. 2007), including the UNDP's creation of a multidimensional poverty index (MPI) in which additional weighted measures have been added to its HDI (Box 2.3.1).

The key contribution the capability approach makes to public health is that it de-couples meaningful social participation (the concern of relative poverty) and, in broader conceptual terms, the notion of 'development' itself, from a purely or primarily economic growth/income growth model. It identifies a range of goods and services for which social (state) provision is important due to their cost or to market failures in ensuring equitable access. The capability approach arose, in part, to challenge the dominance in public policy of the GDP measure, which critics (Daly and Cobb 1990; Fioramonti 2017) argued was a deeply flawed indicator of how well countries were doing in providing for the general health or welfare of their citizens.

Box 2.3.1 Multidimensional poverty index

1. Education (each indicator is weighted equally at 1/6):
    - Years of schooling: no household member has completed 5 years of schooling.
    - School attendance: no school attendance in years 1–8.
2. Health (each indicator is weighted equally at 1/6):
    - Child mortality: if any child has died in the family.
    - Nutrition: if any adult or child for whom there is nutritional information is malnourished.
3. Standard of living (each indicator is weighted equally at 1/18):
    - Electricity: the household has no electricity.
    - Drinking water: the household lacks access to clean drinking water within a 30 minute walk from home.
    - Sanitation: the household has no adequate sanitation or only a shared toilet.
    - Flooring: the household has a dirt, sand, or dung floor.
    - Cooking fuel: the household cooks with wood, charcoal, or dung.
    - Assets: the household does not own more than one of: radio, TV, telephone, bike, motorbike, or refrigerator, and does not own a car or tractor.

Reproduced with permission from Alkire, S. and Santos, M.E., *Measuring acute poverty in the developing world: Robustness and scope of the Multidimensional Poverty Index*, OPHI Working Paper 59, Oxford, UK, Oxford Poverty and Human Development Initiative, Copyright © 2013, also available from http://www.ophi.org.uk/measuring-acute-poverty-in-the-developing-world-robustness-and-scope-of-the-multidimensional-poverty-index-2/

While Sen (1999) argued effectively for the obligations on states to provide a minimum basket of resources allowing people to develop their capabilities (hence their choice of ‘functionings’, including choices concerning their health), his collaborator in developing this theory, Martha Nussbaum, attempted to identify the contents of that basket:

Life: being able to live to the end of a human life of normal length; not dying prematurely, or before one's life is so reduced as to be not worth living.
Bodily health: being able to have good health, including reproductive health, to be adequately nourished; to have adequate shelter.
Bodily integrity: being able to move freely from place to place; being able to be secure against assault, including sexual assault, having opportunities for sexual satisfaction and for choice in matters of reproduction.
Senses, imagination, and thought: being able to use the senses, to imagine, think, and reason—and to do these things in a way informed and cultivated by an adequate education, including but not limited to literacy and basic mathematical and scientific training. Being able to use one's mind in ways protected by guarantees of freedom of expression with respect to both political and artistic speech. Being able to search for ultimate meaning of life in one's own way.
Emotion: being able to have attachments to things and people outside ourselves.
Practical reason: being able to form a conception of the good and to engage in critical reflection about planning one's life; Nussbaum recognizes that this entails protection for the liberty of conscience and religious observance.
Affiliation: being able to live with and towards others; to engage in various forms of social interaction; having the social bases of self-respect and non-humiliation; being able to be treated as a dignified being including protection against discrimination on the basis of race, sex, sexual orientation, religion, caste, ethnicity, or national origin.
Other species: being able to live with concern for and in relationship to the world of nature.
Play: being able to laugh, or play, to enjoy recreational activities.
Control over one's environment both politically and materially, including having the right to political participation and having property rights on an equal basis with others, having the right to seek employment on an equal basis with others (Nussbaum 2000, pp. 77–80).

Defining poverty as capability-failure, especially given the long list of positively stated attributes of 'flourishing' that Nussbaum provides, nonetheless risks taking policy attention away from income poverty. In commodified, wage-based market systems, income (and its quantity) represents both choice (capabilities) and power (Lister 2004). Money may not be everything, but adequate income (however adequacy is determined)

remains basic to any understanding of poverty. We note that in some low-income countries market systems are less entrenched and access to the capabilities through subsistence farming and strong cohesive communities remain important protections against absolute poverty.

Poverty as deprivation

> Deprivation . . . exists when people lack or are denied resources to participate in social and economic activities in the society in which they live.
>
> Reproduced from Labonté, R. et al., *Indicators of Social Exclusion and Inclusion: A Critical and Comparative Analysis of the Literature*, E/Exchange Working Paper Series, PHIRN, University of Ottawa, Canada © 2012, with permission from the author.

Deprivation entered the poverty literature in the late 1960s (Yitzhaki 1979). Like poverty, deprivation has both absolute and relative conceptualizations. As an absolute condition, deprivation is an undesired state that arises when people are unable to access necessary goods and services due to a lack of resources (Hallerod 1996); much depends here on where one draws a line under 'necessary' which immediately takes us into the terrain of relative deprivation. As a relative condition, deprivation exists when people lack or are denied resources to participate in social and economic activities in the society in which they live (Bailey et al. 2004). One example is the UNICEF Innocenti Research Centre's measure of 'child deprivation' based upon the percentage of children aged 1–16 lacking two or more items from a list of 14 considered essential to acceptable child functioning in economically advanced countries (Fig. 2.3.4) (UNICEF Innocenti Research Centre 2012). Like relative poverty, the emphasis of relative deprivation is on social participation. The notion of being 'denied' resources for such participation is captured evocatively in deprivation as an 'enforced lack of socially perceived essentials' (Saunders et al. 2008, p. 175). Deprivation, like multidimensional poverty, can be multiple, when people lack access to sufficient levels of several attributes important to social participation, such as income, housing, healthcare, and education. Moreover, deprivation is an enforced state; that is, its origins can be traced directly to certain political, economic, or policy choices made by governments and which are entrenched in economic systems such as capitalism or, in more contemporary parlance, neoliberal globalization.

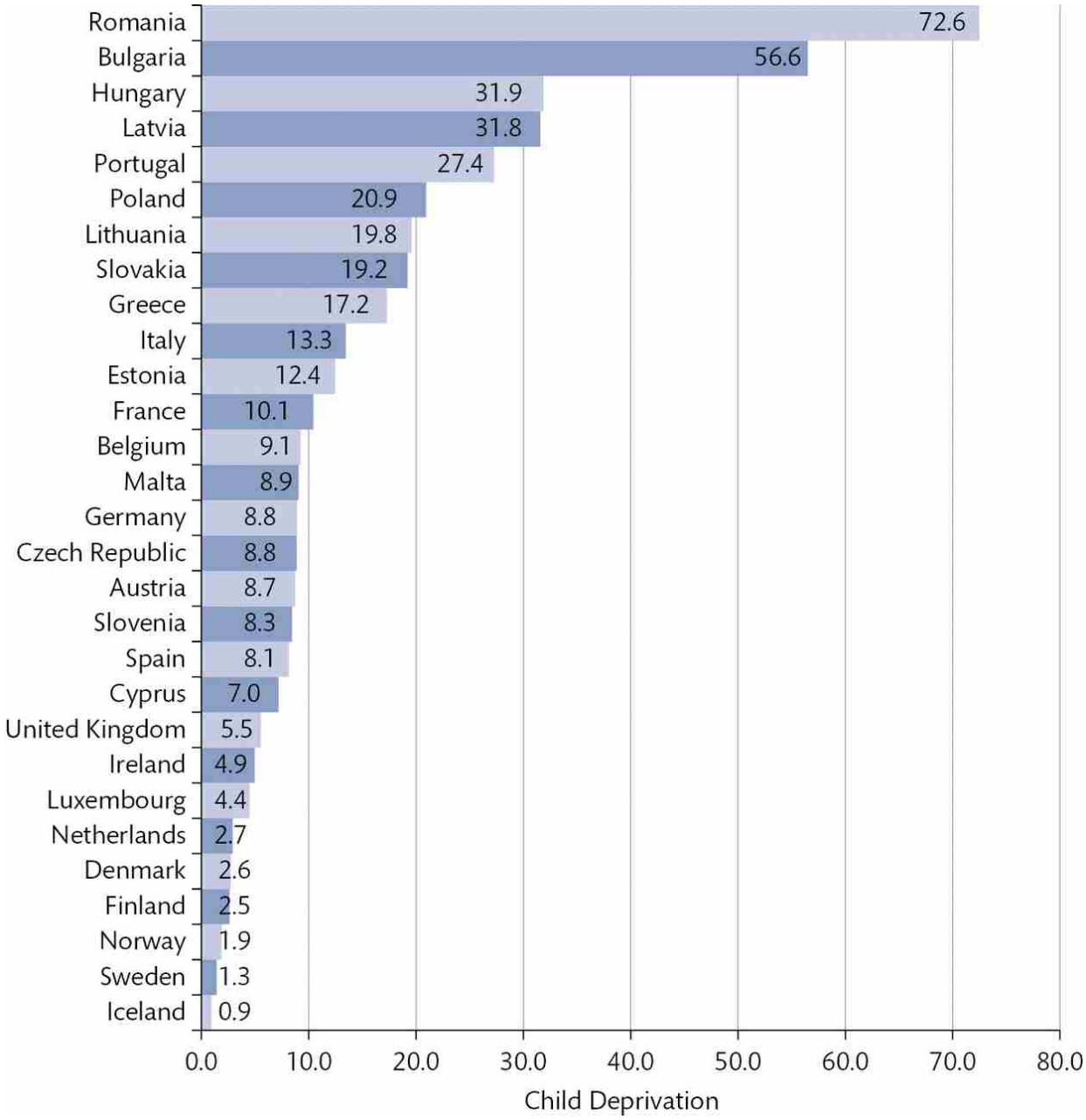


Fig. 2.3.4 Child deprivation in selected countries. Data for 2009.
Adapted with permission from UNICEF Innocenti Research Centre, *Measuring Child Poverty: New League Tables of Child Poverty in the World's Rich Countries*, p. 2, UNICEF Innocenti Research Centre, Innocenti Report Card 10, Florence, Italy, Copyright © 2012.

While deprivation first emerged from the poverty literature, there are important conceptual differences between the two terms. A person may be poor in terms of income but not deprived in terms of, for example, housing,

healthcare, education, or transportation, to the extent these are publicly subsidized or provided. Conversely, a person could experience deprivation in multiple dimensions while not being income poor, such as experiencing discrimination by virtue of one's gender, ethnicity, or sexual orientation (Saunders et al. 2008). This intersectionality of social identity and health risk has become increasingly invoked as a way of incorporating a broader understanding of poverty and inequality, by exploring 'simultaneous intersections between aspects of social difference and identity (e.g. as related to meanings of race/ethnicity, indigeneity, gender, class, sexuality, geography, age, disability/ability, migration status, religion) and forms of systemic oppression (e.g. racism, classism, sexism, ableism, homophobia)' (Springer et al. 2012, p. 1661). It shifts an understanding of poverty well away from matters of simple material goods and physical functioning, which some critics argue deflects attention away from the financial needs of the many who live in absolute or relative poverty (Nolan and Whelan 1996), to a much broader concern with multiple forms of what has been termed social exclusion.

Poverty as social exclusion

> Poverty and social exclusion refer to when people are prevented from participating fully in economic, social and civil life and/or when their access to income and other resources (personal, family, social and cultural) is so inadequate as to exclude them from enjoying a standard of living and quality of life that is regarded as acceptable by the society in which they live. In such situations people often are unable to fully access their fundamental rights. (p. 11)

Reproduced from European Commission, *Draft Joint Report on Social Exclusion*, European Commission, Brussels, Copyright © 2001.

Social exclusion extends beyond poverty and deprivation concepts by incorporating a broader understanding of social processes and their consequences (Bailey et al. 2004). Both poverty and deprivation are components of social exclusion but, as Berger-Schmitt and Noll (2000) point out, income poverty can be both a cause and a consequence of social exclusion. But income poverty is only one possible cause or consequence. As with deprivation, one can be wealthy (not poor) and still experience social exclusion, as has been the case with LGBTQ+ (lesbian, gay, bisexual, transgender, queer/questioning) people in many of the world's countries (Estivill 2003). Social exclusion, although sharing much in common with Townsend's concept of relative deprivation, differs in the emphasis it places

on non-material aspects, from ‘prospects and networks’ (which work to reduce exclusion) to prejudice and discrimination (which work to increase it) (Saunders and Wong 2009).

Social exclusion first arose in research and policy discourse in France in the early 1970s, and was used both to explain the effects of extreme economic restructuring on social solidarity beyond what poverty and deprivation alone could describe, and to assist in developing policies to promote social cohesion within the European Union (Gore and Figueiredo 1997; Aasland and Flotten 2001; Rawal 2008). Rene Lenoir in 1974 was the first to coin the term as a way to define the condition of a large portion of the population that included the poor, handicapped, vulnerable, aged, abused children, substance abusers, and other marginalized groups or ‘social misfits’ (de Haan 1999; Sen 2000). A broader conceptualization was later proposed to incorporate the ‘social, economic, political and cultural systems which determine the integration of a person in society’ (Walker 1997, cited in Lessof and Jowell 2000, p. 5)—that is, the social processes of exclusion.

Economic exclusion is a term frequently used synonymously with social exclusion. Dertwinkel (2008, p. 4) argues that the ‘meaning of social exclusion has always been an economic and less a political or cultural one’ and defines economic exclusion as a state of ‘non-participation in or blocked access to the labour market, public services, finance, housing, educational and health sector, among other possibilities’ (Labonté et al. 2012, pp. 19–20). This emphasis on labour market attachment is most often encountered in European writing on social exclusion, even when other dimensions of exclusion are considered. Part of the reasoning behind this emphasis is that it creates greater social cohesion and reduces social conflict, given evidence that high levels of youth unemployment and poverty are associated in many countries with gang participation or rebel group involvement, motivated primarily by a desire to earn income (World Bank 2011). Other theorists, however, give greater attention to the relational aspects of social exclusion, linking exclusion to inadequate social participation, lack of social integration, and lack of power (Room 1995, p. 5). As Sen (2000, p. 6) puts it, social exclusion (unlike poverty or deprivation) gives ‘a central role to relational connections’. Work, like the income it generates, is simply one of the means of making such connections. This emphasis was echoed in the final report of the Social Exclusion Knowledge Network of the WHO Commission on Social Determinants of Health which distinguished between

‘social exclusion as a state experienced by particular groups (common in policy discourse) as opposed to the relational approach . . . [in which] exclusion is viewed as a dynamic, multidimensional process driven by unequal power relations’ (Popay et al. 2008, p. 7). This process results in unequal resources, reduced capabilities, and fewer claims on human rights leading to inequalities in access to entitled services. This perspective on social exclusion also points the policy spotlight at unfair economic and social structures rather than on deficiencies of individuals.

Gough et al. (2006) discuss social exclusion in terms of the ways in which it is an inevitability of neoliberal capitalism. They acknowledge that it is a highly ambiguous concept (Gough et al. 2006, p. 49), and one that can be used to support a neoliberal status quo by regarding as unproblematic the ‘inclusion’ of elite and wealthy groups whose privileges and practices contribute to the exclusion of others. They maintain that a broad notion of social exclusion should be centrally concerned with social oppression particularly those of ‘race’, gender, sexuality, disability, and age. Crucially for the discussion in this chapter they note that the discourse of inclusion avoids confronting the question of justice and that ‘one can be included in any social arrangement in an unjust fashion’ (Gough et al. 2006, p. 4). As Levitas noted:

> The term social exclusion is intrinsically problematic. It represents the primary division in society as one between an included majority and an excluded minorityThe solution is a transition across the boundary to become an insider rather than an outsider in a society whose structural inequalities remain largely uninterrogated. (Levitas 2005, p. 7)

Poverty summarized

Taking account of these various attempts to define and understand poverty, the concept might usefully be summarized as an absolute and comparative lack of resources and capabilities, comprised principally of the following:

Adequate material resources (food, shelter, clothing, and other normatively defined social necessities).
Income to acquire such necessities (generated through employment or government subsidies and transfers).
Access to formal labour markets or post-retirement or child-rearing income support (to provide adequacy and security of income, as well as access to social relations).
Access to adequate and affordable housing (extending beyond basic shelter to incorporate aspects of overcrowding, disrepair, insecure tenancy), and to services basic to health (e.g. sanitation, water).
Access to adequate and affordable educational and healthcare opportunities (to improve access to formal labour markets and reduce barriers to broader forms of social participation).
Socially conditioned and structured forms of discrimination (which reduce access to formal labour markets, educational and healthcare opportunities, and broader forms of social participation).
Power or voice to influence the policy choices of governments affecting all of the aforementioned conditions (which extends to lack of political freedoms or human rights) (Labonté et al. 2012, p. 15).

There are three cautions about poverty that need issuing. The first is that poverty mobility can be dynamic, with people moving in and out of income- or consumption-based poverty (whether absolute or relative) even within short time periods. The chronic poor are those who ‘remain trapped in situations and relationships which produce deprivation and capability losses over long periods’ (Policy Analysis Discussion Group 2012, p. 6). The second is that people living in poverty, however conceptualized or measured, may not view themselves as poor, deprived, or without capabilities. ‘Subjective’ poverty differs from its objective measure, and may reflect people’s perception of their agency, their rejection of the stigmatizing label of being one of ‘the poor’, or an internalized resignation to the inevitability of their circumstances. The third is that, just as poverty is now regarded as multidimensional, people’s identities and capabilities are multiple. A focus only on people’s relative deprivation without appreciating their relative capabilities risks creating or reinforcing resignation to their circumstances, rather than mobilization to change them.

Finally, persistently high rates of poverty (at the not-so-extreme level) need to be seen within a context of rising income and wealth inequalities.

Since the 1980s the income share of the top 10% has risen sharply in North America, China, India, and Russia; remained flat (but at excessively high existing levels) in sub-Saharan Africa, Brazil, and the Middle East; and increased slowly in Europe (UNCTAD 2018). When global inequality is mapped, it becomes an 'elephant curve', with the top 1% accumulating most of the income gains (Fig. 2.3.5). Given the environmental impossibility of eliminating poverty at any level through conventional 'trickle-down' economics, it is reassuring that one of the SDGs (SDG 10) specifically commits the world's governments 'to reduce inequality between and within countries' (Sustainable Development Knowledge Platform 2017). One of the goal's targets calls for sustained income growth for the bottom 40% (within countries and globally) at a rate greater than the average. Without addressing the capture of global income by an elite transnational capitalist class (the 1%, but more accurately the 0.01%), however, the bottom 40% might experience a relatively higher-income growth *rate* than the average, but in absolute dollar terms the gap between richest and the rest could continue to worsen. The imperative for national and global systems redistribution of income assumes paramount importance.

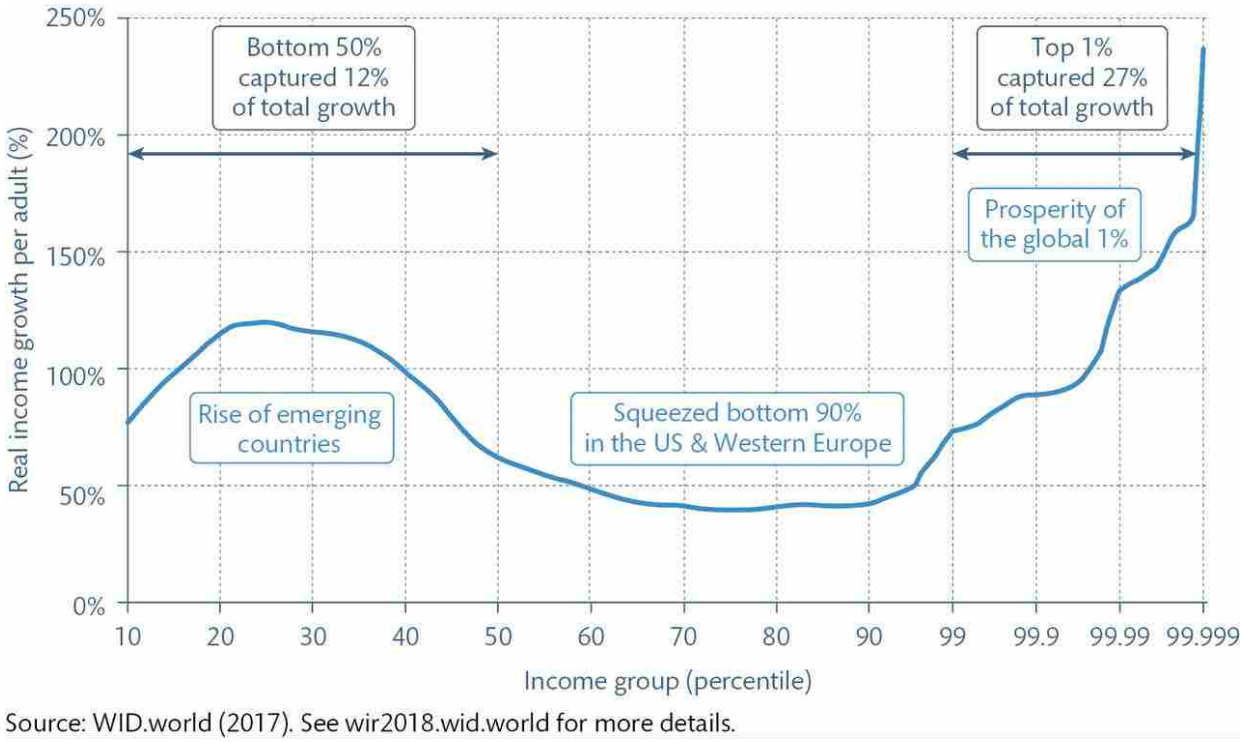


Fig. 2.3.5 The elephant curve of global inequality and growth, 1980–2016.
Reproduced from Alveredo F *et al.* (2017) 'World Inequality Report 2018: Executive Report' World Inequality Lab, 2017, under the Creative Commons License CC BY 4.0.

How does poverty affect health and ill health affect poverty?

> Whatever is going round, people in lower social positions tend to get more of it, and to die earlier even after adjustment for the effects of specific individual or environmental hazards. (p. 1355)

Reproduced with permission from *Social Science & Medicine*, Volume 31, Issue 12, Evans R. and Stoddart G., Producing Health, Consuming Health Care, pp. 1347–63, Copyright © 1990.

On average, the lower the socioeconomic position, the lower one's health status. Socioeconomic status (SES) is usually measured by income, occupation, or education, but also includes stratification by social class, gender, and race/ethnicity—examples of the intersectionality of different deprivations discussed earlier. One of the key findings in recent years has been that differences between social position and health status follow a gradient, a step-wise pattern between where one sits in a SES hierarchy and most health outcomes. Although not all health conditions follow a socioeconomic gradient (a few diseases are more common in higher than in

lower SES groups) the pattern is robust, observed in many countries, persists over time, and is evident for all age groups and for both men and women. Research has also shown that health inequities accumulate over a lifetime and that disadvantage at each stage of the life course affects health (Lynch et al. 1997; Smith et al. 1997, 1998). Adverse social and economic conditions in the early years of life are increasingly being shown to have a lifelong impact on health (Lynch and Smith 2005; Hertzman and Boyce 2010; Power et al. 2012). Much of this textbook is concerned with the determinants of the gradient in health; here, we recount briefly the three major explanatory models for this gradient: natural or social selection, cultural/behavioural, and materialist or structural (Baum 2008).

Natural or social selection

The social selection argument suggests that healthier people become wealthier people, enjoying greater social mobility than their poorer fellows. There is some evidence that better health is associated with social mobility, especially at younger ages; but this accounts for only a small portion of health differences between social groups (Whitehead 1992; Power et al. 1996). The social selection and mobility argument comes with its inverse: people are poor because they are unhealthy. Again, there is some evidence for this, particularly in low-income countries. We have already noted the problem of medical poverty, where the absence of affordable healthcare pushes sick people into poverty. There is also evidence that people living in countries with high burdens of particular diseases (such as HIV/AIDS, tuberculosis, or malaria) and with poor social protection safety nets (such as income transfers or sickness and unemployment benefits) can suffer greater poverty due to their illness reducing their ability to earn income. Poor health also contributes to poverty through its debilitating effects of inadequate nutrition on individual development and later economic productivity (Horton and Ross 2003; Sanders and Chopra 2007; Baum 2008). Early insults to the growth and brain development of children, even with intensive interventions later in life, are partially irreversible. In many cases, the damage is done even before the child is born. Undernutrition among mothers is a major reason why babies are born with a low birth weight, a result of small maternal size at conception and low pregnancy weight gain or inadequate food and energy intake during pregnancy (Kramer 1987). Malnutrition

among infant girls is therefore one of the main routes for the intergenerational transmission of poverty (Horton and Ross 2003).

That sick people become poor, in other words, reflects the same dynamics of social stratification and material circumstances that explain why poor people are more likely to become ill in the first place. This accounts for why some wealthy but highly unequal countries with limited social protection can have poorer population health outcomes than poorer but more equal countries with high levels of social programme spending (notably for primary healthcare, education, women's literacy, land reforms, and physical infrastructure) (Balabanova et al. 2011). Sen (1999, pp. 335–6) showed this very clearly when he compared the life expectancies in the early 1990s of US whites with US African Americans, and with Indians in Kerala and the Chinese. While African Americans are very many times richer in income terms than the people in China and Kerala (even after correcting for cost of living differences) they are less likely to live as long. Sen explains this by reference to 'the social arrangements and community relations such as medical coverage, public healthcare, elementary education, law and order, and prevalence of violence'. In other words: there is nothing 'natural' about 'social selection'; such selection arises as an effect of how the political and economic systems of countries are structured.

Cultural/behavioural explanations

A more commonly encountered reason for the SES gradient in health is that poorer people make unhealthier lifestyle choices. This is not a new argument. During the early decades of the nineteenth-century Industrial Revolution in Europe, with its dislocating impacts on rural agricultural livelihoods and explosion of ghettoes in rapidly expanding cities (a phenomenon now characteristic of many developing countries in late twentieth- and early twenty-first-century globalization) the poor were seen as a palpable disease threat to the rich. Their poverty, in turn, was not regarded as an effect of their economic dislocations, but as 'the interaction of the twin problems of indolence and vice' (Alcock 1993, p. 12). The British 'father' of public health, Sir Edwin Chadwick, in his reforms of the poor laws requiring urban paupers to earn their minimal allowance in workhouses, and so to become 'deserving poor', argued that:

> The population so exposed is less susceptible to moral influences, and the effects of education are more transient than with a healthy population; these adverse circumstances tend to produce an adult population short-lived, improvident, reckless and intemperate, and with habitual avidity for sensual gratifications. (Chadwick 1930)

Even while acknowledging the 'adverse circumstances' of the poor, Chadwick posits that the poor health of the poor results primarily from their lack of the proper behaviours and morality of the healthier and wealthier. A more socially critical observer of the same era, Friedrich Engels, agreed with Chadwick that alcohol was fairly prevalent among the poor labouring classes, but that 'it was almost their only source of pleasure':

> The working man comes home from his work tired, exhausted, finds his home comfortless, damp, dirty, repulsive; he has urgent need of recreation, he must have something to make … the prospect of the next day endurable. (Engels 1845/1999, p. 113)

The same dichotomy between a behavioural and a socially critical account of the poverty/health relationship persists today. Many (although not all) unhealthy behaviours do tend to be more prevalent as one moves down the social gradient. Even in developing countries undergoing nutritional transitions (where traditional diets are being replaced by increased consumption of high-sugar/refined carbohydrate, energy-dense, low-nutrient food commodities), what might begin as a status choice among a rising middle class quickly becomes a necessity choice (due to availability and cost) for poorer social classes (Drieskens et al. 2009; Monteiro et al. 2010; Stuckler and Nestle 2012). The same is true for smoking, originally more common among elite social classes, and, to lesser degrees, alcohol and other drug use (Drieskens et al. 2009). But explaining these behaviours as 'cultural' effects of poorer groups (their improvident or reckless 'lifestyle'), or as their lack of knowledge or motivation to adopt healthier patterns, ignores the social contexts that condition and constrain individual choice (Layte and Whelan 2009). It also fails to account for the excess disease burden faced by many poorer groups, even after accounting for all of the individual/behavioural risk factors (Lynch and Smith 2005).

Materialist or structural explanations

The limitation of a narrowly behavioural understanding of the poverty/health relationship draws attention to how the material conditions of poverty affect health. Here the enumeration of, and evidence for, the health determining

effects of these conditions is plentiful, reflecting many of the material circumstances identified in this chapter's discussion of poverty and deprivation: access to income, housing, employment, nutritious food, safe water and sanitation, unemployment protection, health services, education services, and other socially provided resources that allow for the development of capabilities. These structural factors (so named because they reflect embedded economic and power relations within societies) are also related to what have been called psychosocial risk factors: the isolation, low social support, poor social networks, high self-blame, and low perceived power that are more frequent among poorer and less educated groups (Najman 1994; Lynch and Smith 2005; Kawachi et al. 2008). Racism has also been shown to have a powerful effect on health status (Krieger 2000; Paradies 2006; Ziersch et al. 2011), psychosocial and material risks cluster together, and poorer groups (those lower along the SES gradient) are more likely to experience all three. In countries where material inequalities have been significantly reduced through welfare measures, for example, health inequalities experienced by poorer groups are more likely to be a result of greater behavioural and psychosocial risks; although, as Johan Mackenbach suggests, 'the persistence of health inequalities in modern European welfare states can partly be seen as a failure of these states to implement more radical redistribution measures' (Mackenbach 2012, p. 761; see also Chapter 2.4). Behavioural, psychosocial, and material risks matter and interact with each other (Krieger 2000; Paradies 2006; Ziersch et al. 2011).

What theories of justice should guide our interventions?

The public health concern with justice, poverty, and health enjoys a rich history. Over 150 years ago the young Prussian physician, Rudolf Virchow, hired by the government to recommend how to control a typhoid epidemic among poor coal miners, famously called for massive social reforms that ranged from progressive taxation and organized food cooperatives, to improved wages and working conditions, strong local government, improved education, and an end to church interference urging the poor to suffer now and to reap their rewards in heaven. Friedrich Engels, writing around the same time as Virchow, published his lengthy analysis of The Condition of the Working Class in England, leading to his co-authorship with Karl Marx of the influential Communist Manifesto with its call for a revolutionary

social justice. Some of this early concern with poverty and its social and physical ills continued into the late nineteenth and into the twentieth century, although it became muted by the rise of 'scientific medicine' and its promise of a cure for most diseases. Supported by industrialist philanthropies (notably in the Americas, by The Rockefeller Foundation), scientific medicine's emphasis on cure rather than prevention deflected much public health attention away from the social roots of disease (Brown 1979; Tesh 1988). Poverty as a public health issue occasionally surfaced, especially during times of economic recession or depression when high poverty rates were associated with the spread of diseases such as tuberculosis and cholera. But it was in Latin America that the radicalism of nineteenth-century public health activism became institutionalized in the practice of what has been called 'social medicine' (Waitzkin 2005). One of its major proponents was the Chilean physician (and later socialist President) Salvador Allende who, as health minister in the late 1930s and 1940s, approached health inequities in Chile as a direct result of historic underdevelopment, international economic dependency, and foreign debt. He argued that the 'medico-social problems of the country … require precisely the solution of the economic problems' (Waitzkin 2005, p. 740). Allende's reforms, echoing those of Virchow, included wage improvements, wealth redistribution, land reforms, improved food security measures, and publicly supported housing development.

More recently, the WHO Commission on Social Determinants of Health (2008) took a clear stand that inequities in health (avoidable inequalities) were unacceptable and the goal of policy should be to eliminate these and so argued for closing the life expectancy gap in a generation.

These are bold public health proposals, generally framed in broadly stated concerns with justice (fairness). But how should we determine when poverty's health risks become unacceptable? What theories of social justice might guide our public health interventions or advocacy?

Equality of opportunity or equality of outcome?

Social justice theory is generally associated with European societies and particularly with struggles during the industrial revolution and the emergence of socialist, social democratic, or other models of redistributive welfare states: 'Social justice is not possible without strong and coherent

redistributive policies conceived and implemented by public agencies' (United Nations Division for Social Policy 2006, p. 6). Social justice theory is essentially concerned with equity, or fairness. On this basis, it is argued that social justice (equity) is a universal concern, since all social arrangements, to be legitimate and to function at all, must attend to issues of equality (Sen 1992). But there are subtleties to how equity is conceived with two main dimensions: equality of opportunity, achieved through procedural justice or 'horizontal equity' in which equals are treated the same; and equality of outcome, achieved through substantive justice or 'vertical equity' in which people are treated differently according to their initial endowments, resources, privileges, or rights. Both equalities (opportunity, outcome) are ideal types; neither exists in 'true' form. They represent aspirational ideals of what societies strive to create for their members (fairness in outcomes) and how they believe this should be accomplished (fairness in opportunity) (Labonté 2010, p. 84).

There is a political, as well as philosophical, difference in these two broad social justice streams (Labonté 2010, p. 84). Recent decades of global market integration and the collapse of Soviet-style 'socialism' have been accompanied by the increasing dominance of neoliberal ideology which has supplanted concerns with social justice by an exclusive focus on economic growth which, it is suggested, will eventually result in greater and better-distributed wealth—the trickle-down theory.

In turn, proponents of more open markets and conventional approaches to growth and development emphasize equality of opportunity, with only residual attention to equality of outcome. This was the position taken by the World Bank's 2006 World Development Report on the theme of equity and development, which was less supportive of post-market income redistribution to achieve greater equality, instead favouring greater individual equality of opportunity through inter alia 'equality before the law, equal enforcement of personal and property rights, non-discriminatory institutions, and equal access to public services and infrastructure' (World Bank 2006, pp. 18–19). The slight nod to a concern with outcomes was reference to avoidance of absolute deprivation (Labonté 2010, p. 84).

Theories of justice

Moral defence for some mitigation of health and social inequalities arising from poverty is a recurrent theme in much contemporary Western philosophy. Peter Singer (1972) posits 'a Greater Moral Evil Principle', that it is both just and of collective benefit to act to relieve poverty and deprivation if, in doing so, we do not sacrifice something of comparable moral significance. Around the same time John Rawls published his highly influential Theory of Justice. Standing behind a 'veil of ignorance' as to their social standing at birth, Rawls argues that people would choose a justice that guaranteed a minimum of primary goods that any rational person would choose as basic to their needs. This justice theory builds upon two principles. The first principle is the 'priority of the equal' (basic liberties), which roughly equates with individuals having civil and political rights that protect them against excess authoritarian rule by the state. The second principle is based on legal equality of opportunity, which roughly equates with individuals having economic, cultural, and social rights that obligate states to provide certain goods or services (the minimum resources or capabilities required to exercise one's functioning). His second principle also invokes the 'difference principle': that inequality in the distribution of social and economic goods ('primary goods') are allowable only to the extent that they also improve the lot of the least advantaged. The difference principle obliges a degree of state interventions of redistribution and regulation, although Rawls did not believe that the extreme differences in wealth and power that markets create was of moral concern provided the conditions of the least advantaged improved. He also emphasized the centrality of better procedural justice (Schaefer 2007; Labonté 2010, p. 85).

Rawls' justice theory is located within the social contract school, which views nation states as the primary actors in international relations. He held that poverty in low-income countries was primarily an effect of domestic policies and practices with little international or global causality. Pogge (2002), drawing on cosmopolitan arguments, challenges Rawls on this conclusion, as well as on evidence that poverty cannot be de-linked from global economic institutions and actors. In doing so, he extends Rawls' basic justice theory to a global level, contending that there are not simply 'positive duties' to assist those in need, but moral obligations (negative duties) to prevent harm (Pogge 2004). Pogge's theory of relational justice (Pogge 2002) is based on three lines of argument:

The radical inequalities observed between peoples and nations today are partly an effect of a violent history in which some gained at the expense of others. While we individually cannot be held responsible for the actions of our forebears in this 'conquest', as moral persons we can be held accountable for rectifying the vast disparities in initial conditions that this history has created.

Not only does procedural justice by itself fail to account for these vast disparities in initial conditions; it is impossible to conceive of these disparities existing on the scale that they do without 'an organized state of civilization' (Pogge 2005) to uphold them. Both procedural and substantive injustices thus endure.

There is evidence that economic institutions operating on an international scale (the 'organized state of civilization') have been complicit in upholding these injustices. There are also feasible alternatives to these economic institutions that would reduce the 'radical inequality' of persisting poverty. Persons involved in upholding these institutions are thus implicated in creating subsequent ill health, even though they may be half-way around the world (Pogge 2004).

Pogge concludes that: 'we are *harming* the global poor if and insofar as we collaborate in imposing an *unjust* global institutional order on them'; and proceeds to offer an evidence-informed argument to establish that present global institutional rules and procedures are unjust (Pogge 2005, emphasis in original). The justice implication is one of immediately engaging in 'rectification' through strengthened human rights and more progressive systems of global resource redistribution; but also an obligation to change the very rules of economic governance in order to overcome the historic and radical inequalities in initial conditions (Labonté 2010, pp. 85–6).

A similar theory of social justice has been argued by Iris Marion Young. Accepting that there are demonstrable 'structural injustices' that 'put large categories of persons under a systematic threat of domination or deprivation of the means to develop and exercise their capacities, at the same time as these social processes enable others to dominate' (Young 2006, p. 114), Young argues that 'all agents who contribute by their actions to the structural processes that produce injustice have responsibilities to work to remedy these injustices' (Young 2006, pp. 102–3). Her 'social connection model' of responsibility extends these moral obligations beyond those more directly (causally) liable to a larger body of people who may be only indirectly

connected. An example she offers is responsibility for retaining developing country ‘sweatshops’, which includes those who shop at retailers sourcing their goods from such shops as well as the retailers and factory owners themselves. The limitation of Young’s argument is that its extensive diffusion of responsibility may limit the ability to enforce changes in practice.

The role of human rights

Several justice theories make reference to human rights. Not only do human rights conventions, ratified by most of the world’s states, represent a consensus on citizen rights and state obligations (Mack 2009), but there is widespread agreement that ‘poverty is the principal cause of human rights violations in the world’ (Office of the High Commissioner for Human Rights 1999, para. 9) and that ‘poverty constitutes a denial of human rights’ (Office of the High Commissioner for Human Rights 1999, para. 1). First established after the Second World War with the UN Declaration on Human Rights, several subsequent conventions have been agreed upon that detail specific citizen rights and state duties. These include the Covenant on Civil and Political Rights (intended to protect individuals against abuses of state authority) and the Covenant on Economic, Social and Cultural Rights (intended to guarantee individuals certain entitlements from the state); as well as specific treaties on the rights of children and women, racial minorities, indigenous peoples, and numerous other facets of social justice. Together these treaties comprise an ‘International Human Rights Framework’ which, although lacking enforcement measures, empowers different UN committees and ‘special rapporteurs’ to review countries’ compliance with their legally binding obligations. Central to concerns with the poverty/health relationship is Article 12 of Covenant on Economic, Social and Cultural Rights (technically known as the Right to the Highest Attainable Standard of Physical and Mental Health, see Chapter 3.3) and its General Comment 14, which identifies a broad range of actions required for the progressive realization of this right. Health is considered a basic right, since it is foundational to the enjoyment of most other human rights.

The right to health, when adopted in countries’ own constitutions or laws, however, can be and has been interpreted in legal decisions or national policy as an individual right only. Individual rights to treatment have been

used to force public payment of costly medicines with opportunity costs to other facets of public health access of greater importance to the poor (Hogerzeil et al. 2006; Gianella-Malcam et al. 2009; Yamin and Parra-Vera 2009). International legal scholars argue that human rights emphasis, instead, should be placed on populations who are poor and living in vulnerable circumstances. This requires greater attention to collective rights. Collective rights are implied in General Comment 14 on the right to health (Committee on Economic Social and Cultural Rights 2000), and are explicit in the Declaration on the Right to Development (Office of the High Commissioner for Human Rights 1986). Though the right to development is not a binding treaty, it is considered to have some standing in international human rights law (Aguirre 2008) and has strong normative support through UN agencies especially in the context of the Millennium Development Goals.

Some activist scholars and civil society organizations argue against the present international emphasis on human rights for its lack of class and political economy analysis. Others contend that human rights, with their individualistic and legalistic focus, 'are not sufficient to serve as ethical criteria in solving the fundamental problems of the absolute poor' partly because within these static statements, the 'poor are not invited . . . to articulate their moral needs and rights for themselves' (Mack 2009, p. 11). At the same time, human rights are considered the most globalized social justice statements of the post-Second World War era; and human rights treaties are being used to advance policies to improve health and to assist people living in poverty (Dasgupta 1995; Schrecker et al. 2010).

Inequalities in initial conditions

Confronting inequalities in initial conditions (the lives people are born into) is vital especially given the evidence cited earlier on the lifetime impact of adverse living circumstances in early life. A focus on initial conditions begins to blur the earlier distinctions between equality of opportunity (procedural justice) and equality of outcome (substantive justice). Equality of opportunity, to be just, requires vertical equity: a disproportionate provision of public goods and capability resources for those whom history's conquests, and today's political institutions, place in highly unequal initial conditions. Departing from the redistributive minimalism of the World

Bank's 2006 World Development Report, a later report of the Bank's Latin American Development Forum argues that:

> A better understanding of the importance of inequality of opportunity in the determination of inequality of outcomes may change attitudes towards redistribution. People dislike and consider unfair inequalities associated with differences in circumstances, which many argue should be compensated for by society. By highlighting that component of inequality attributable to circumstances, this type of analysis can help build a social and political consensus on … the best means for addressing inequality of opportunity. (p. 50)

Reproduced with permission from Ricardo Paes de Barros et al., Measuring inequality of opportunities in Latin America and the Caribbean, World Bank Publications, Washington, USA, Copyright © 2009 The International Bank for Reconstruction and Development/The World Bank.

What remains at issue is the extent of moral (or legal) obligation for amelioration of gross inequalities in initial conditions that create 'shortfall inequalities in central health capabilities' (Ruger 2008, p. 440). Is there an ethically defensible scale of rectification?

There is no answer to this question, apart from the imperative to seek one. In this quest, norms of procedural justice re-assume importance. Boggio (2009), in an argument for why international organizations and those within them have an ethical obligation to act to redress systematic health inequalities arising from poverty, addresses how such policy decisions can be made in a just manner. He identifies three basic principles for an 'ethically-informed deliberative process': publicity (transparency in process, a comprehensible rationale, and public argument and evidence); relevance (trust in actors/institutions by recipients, opportunity for wide participation, and interventions based on recipients' needs, values and aspirations); and revisability (policies and programmes can be challenged over time and improved, and individuals and institutions can be held accountable) (Boggio 2009). Several of these conditions are similar to principles of good governance widely held by governments and multilateral organizations (Governance Institute of Australia 2017; Lee et al. 2009); that is, they can be considered as having a broad normative base.

What public policies will reduce poverty?

> The last decade of research on the causes of chronic poverty has shown that a key factor in keeping certain categories or social groups in a state of chronic poverty (transmitted through the generations) is the operation of power relations which stigmatise the people involved, undermine

> their confidence, and systematically close off options for individual or collective advancement . . . A stronger focus on structural inequalities, on systematic obstacles to redistributive processes and on social groups and interactions between them is part of this. (p. 6, emphasis in original)

Reproduced with permission from Policy Analysis Discussion Group, *Understanding Poverty and Well-Being: A Note with Implications for Research and Policy*, Overseas Development Institute, London, UK, Copyright © 2012.

The first issue public health confronts in grappling with the poverty, health, and justice relationship is simply this: if it is possible to improve the health of those living in poverty without necessarily reducing income inequalities or relative poverty, is this sufficient? Some of the justice theories cited earlier might be content with saying yes to this, as their emphasis was more on absolute than on relative poverty. More recent poverty theories (capabilities, social exclusion/inclusion), however, give greater emphasis to the core value of ensuring human dignity or 'flourishing'. There is no justice if one simply survives; one must be able to live with purpose and meaning. Public health concern with poverty, then, extends beyond simply the specific disease risks poverty creates, to the existence of poverty itself—both absolute and relative.

The existence of poverty is inherently political, and there are different political theories for how poverty might be reduced. Alcock (1993) distils these to four dominant models: neoliberalism, conservatism, social democratic, and socialist, to which Gough et al. (2006) add a fifth: the associational.

The neoliberal model is associated with policy responses to the declining profit rates, economic recessions, and developing world debt problems that arose in the 1970s and 1980s. Promoted by conservative governments in the United Kingdom, United States, and Germany (in the 1970s, the major economic powers in the world), neoliberalism is based on a belief that free markets, sovereign individuals, free trade, strong property rights, and minimal government interference is the best means to enhance human well-being. Its policy platform rested on what became known in later years as the 'Washington Consensus': privatization, deregulation, tax reform (lower corporate and marginal rates), deficit reduction, and trade and financial liberalization. In general, the neoliberal view is that governments should avoid poverty reduction as a policy matter, and instead rely upon economic growth through stronger free markets and economic liberalization to 'trickle down' and lift the poor out of their condition. Extreme neo-liberals have argued that the state should not interfere at all with conditions of poverty

(Murray 1984), although this was rarely pursued as a policy option. Instead, neoliberal-dominated governments tend towards a 'welfare minimalism', with benefits or entitlements at a level that prevents the worst forms of deprivation, but that is rarely sufficient for 'flourishing' or 'health capabilities'. Increasingly punitive conditionalities are demand of people receiving welfare payments which increase the stigma and hardship of relying on the welfare system.

The conservative model overlaps with the neoliberal, although it holds to the importance of some state intervention to ameliorate the worst inequities arising from markets, and specifically market failures (Alcock 1993). Persisting chronic poverty is symptomatic of market failures, implying that free markets and economic growth are unable, by themselves, to create equal opportunities for all. Conservative policies are not aimed at preventing poverty, which would require strong regulation of markets, but with relieving the worst of its effects. The approach is selective (means-tested benefits), often punitive (distinguishing between the 'deserving' and the 'undeserving' poor, the latter being people who could work even if their employment kept them in relative poverty), and minimalist (offering supports well below the minimum wage that could be earned in work, thus creating an incentive to 'labour market attachment').

The social democratic model, historically associated most with northern European countries, but also found in other parts of the world, emphasizes universalism in its poverty reduction policies. Social democracies are often highly interventionist in markets to prevent poverty in the first instance, as well as using high taxation and social spending to reduce residual levels of poverty. As with the previous two models, there is still an emphasis on labour market participation and an expectation that those who can engage in work will do so. But there is also greater emphasis on extending positive incentives to people entering lower-waged positions to overcome what has been called the 'welfare wall', a term that describes the loss of non-income benefits when moving off social assistance to minimum waged positions. Although universalism in benefits and entitlements helps to build cross-class solidarity, the non-income benefits of social spending (e.g. in healthcare, education, transportation services, and the like) can be disproportionately 'captured' by wealthier social groups. This has led to calls for 'proportional universalism' in which policies are constructed such that those in greater need enjoy disproportionate levels of benefits (Marmot 2010); and for some

degree of targeting of benefits to maximize their poverty-reducing efficiencies (OECD 2008).

The associational model also reflects opposition to neoliberalism and is based on the need to involve people in economic models and reduce the extent to which they are alienated by them (Gough et al. 2006). Building strong social capital, voluntary organizations, and civil society are considered to be the key poverty-reducing components. Micro credit banks such as the Grameen Bank are an example in low-income countries where (mainly) women can obtain loans to set up their own business. A further example is the Indian Self-Employed Women's Association, which organizes millions of women to provide their own banks, workers' cooperatives, childcare, and health and unemployment insurance. In rich countries, examples would be locally based bartering systems, food, and worker cooperatives. These initiatives are based on self-help, are democratically run, have a local focus, and depend on forms of collectivity that can be plural, multiple, and diverse. Gough et al. (2006) indicate that the expectation is that states should be subordinate to these myriad local organizations and that political democracy should grow out of the practical economic and social democracy. They go on to note that associationalist projects are usually hemmed in by disciplinarian capital and the localist strategies leave the power structures that create poverty in the first place untouched.

The socialist model, though less prominent in recent decades since the collapse of the Eastern Bloc countries, holds that capitalist market systems inevitably create poverty and that the only enduring means to reduce poverty is to replace capitalism with a socialist system of governance. This implied a radical transformation of capitalist labour markets, which inevitably create unemployment and poverty, to socialist (collective) forms of production and ownership. In the absence of a political revolution, the entrenchment of neoliberal politics and ideology means that a process of transition to a socialist society needs to be envisaged. Moves towards greater egalitarianism include: protection of workers' conditions and rights; progressive taxation and tax enforcement measures for the wealthy and corporations; combating the multiple forms of discrimination in jobs, housing, and public services; withdrawal of public support for private education and health; public provision of good quality education and health services; provision of universal lifelong social security for all citizens; and maintaining public spaces for community use. Many of the critiques of

market economics made by socialism are finding more policy and media traction in the wake of the 2007 global financial and employment crises; accompanied by civil society calls for progressive taxation, stronger market regulations, and enforceable citizen rights to hold government to account (Labonté and Schrecker 2009). The election of Donald Trump as the US President in 2016 has also seen increased protest around the world against the extreme neoliberalism behind the policies his Presidency is implementing. There are also strong protests against the austerity agenda in European countries especially in Spain and Greece.

The role of policy initiatives and civil society mobilization for poverty reduction and fairer societies

Most countries have organized civil society groups, including labour organizations, campaigning for policy actions to reduce poverty. These groups generally blend the broad approaches of the social democratic and revolutionary socialist models. They often provide detailed policy analyses and briefs, and collaborate with some academic researchers to ensure the empirical and theoretical integrity of their work (Global Health Watch 4). Some groups focus on reducing poverty while others argue that poverty reduction will only result from a broader reform of the economic system and the inequities to which it gives rise. There are several examples in history of social and labour movement mobilization forcing changes in government and social policies, including those directly related to poverty reduction and health improvement (Szreter 2003). We offer two here, one a governmental example and the other capturing civil society activism.

Brazil's governmental reforms

Significant gains in poverty reduction, health system development, and health outcomes have been achieved in several countries through progressive social policies, resulting ultimately from political mobilization. Brazil's progress since the mid-1980s over the first 25 years of democratic government provides a contemporary example of progressive change within a market economy (see Chapter 2.5). Infant mortality, for example, fell from 114 per 1,000 live births in 1970 to just 19.3 per 1,000 live births in 2007,

while life expectancy at birth increased by almost 40%, reaching 72.8 years in 2008. It is not possible (or even likely) to attribute these impressive advances to health sector activities alone, since these activities were part of several large-scale social reforms initiated in the country. These reforms led to steep increases in school attendance and decreases in illiteracy rates, both of which are pathways to better health. As well, Brazil's GDP doubled between 1981 and 2008, potentially lifting more families out of income poverty; and its high levels of income inequality fell substantially as a result of such social policies as the Bolsa Família conditional cash transfer programme, which covers 10.5 million families, and increases in the legal minimum wage. Other social determinants of health were also improved substantially, including increases in provision to households of indoor water, sewage disposal, and electricity (Paim et al. 2011). Sustained and significant investment in the creation of a universal health system, the centrepiece of which is the Family Health Programme, is mostly directed at providing more accessible and equitable health services, particularly at the primary level, but also engages community health workers who identify in their localities social factors that negatively affect socioeconomic and health status and enrol other sectors in addressing them (Macinko et al. 2007; People's Health Movement et al. 2011).

The political context of Brazil's reforms dates back to the late 1980s when popular mobilization challenged a conservative government with strong pro-market policies. This mobilization against a military dictatorship brought together oppositional political parties (at the time, illegal), grassroots groups, progressive academics and researchers, and trade unions. This movement catalysed the institution of many progressive health policies. With the election of the Workers' Party in 1989, popular mobilization waned as progressive social policies were introduced, while 'social participation' in local government continues through such bodies as the National Health Council, which plays an ongoing role in democratizing policy development (Paim et al. 2011). In 2013, 'middle-class' protests began once again in Brazil, which some attribute to the country's targeting of many of its social reforms to the poorest sectors of the population. The challenge becomes how to universalize the benefits of these reforms, building the cross-class solidarity that remains important in sustaining support for redistributive social protection programmes (People's Health Movement et al. 2011).

Brazil entered an era of uncertain politics in 2017, following what some have described as a 'judicial coup' that removed the elected President and replaced her with a conservative politician and the subsequent imprisonment of the previous President Lula, in part following disenchantment by higher-income groups in the country with Brazil's policy directions, and widespread concerns over perceived government corruption (Saad-Filho 2016). The country's remarkable health system improvements, and the role of civil society activism in initiating and supporting these, still stands; as does the health and social justice activism of many civil society organizations within Brazil.

Civil society challenges to unhealthy globalization

Another important example of civil society activism that catalysed far-reaching policy change that indirectly impacts on health equity is the campaign against the Multilateral Agreement on Investment (MAI). The Association for the Taxation of Financial Transaction for the Aid of Citizens (ATTAC) was founded in France in 1998 to mobilize support for the Tobin tax on currency transactions as part of a broader movement against inequitable globalization arising from cross-border speculative finance. ATTAC, which has now established itself in many countries, played a leading role in France's decision to withdraw from OECD talks on the MAI, resulting in the failure of the talks (Waters 2004).

The People's Health Movement (PHM), another global civil society group, is a grassroots organization linking health activists around the world. It was formed following the First People's Health Assembly in 2000 and its political position is outlined in the People's Charter for Health (People's Health Movement 2009) and summarized in the Preamble as:

> Health is a social, economic and political issue and above all a fundamental human right. Inequality, poverty, exploitation, violence and injustice are at the root of ill-health and the deaths of poor and marginalized people. Health for all means that powerful interests have to be challenged, that globalisation has to be opposed, and that political and economic priorities have to be drastically changed.

Reproduced with permission from People's Health Movement, The People's Charter for Health, Copyright © 2009.

The Charter goes on to call for much broader political, economic, and social participation from poor people. Since 2000 the PHM has grown to

encompass around 80 national circles of varying size and weight which are organized into regional networks. A Right to Health campaign, whose parameters have been defined at a global level but whose country-level activities vary according to context and conjuncture, has placed the responsibility for poverty reduction and the reduction of health inequities firmly on the shoulders of national states and international organizations. PHM has also collaborated with other organizations to produce five editions of the publication, 'Global Health Watch' (http://www.ghwatch.org), which present alternative and more progressive versions of the World Health Reports produced by the WHO. The GHWs provide a critical 'watch' on international organizations relevant to global health policy and governance including the World Bank, the WHO, and the Bill and Melinda Gates Foundation; offer evidence-informed critiques of neoliberal globalization; and argue strongly for its replacement by a more just and equitable system of global governance and economic system to achieve poverty eradication and health equity.

Conclusion

This chapter has analysed the various definitions of poverty and considered their philosophical and political implications. We have examined a variety of explanations for why poverty affects health adversely. On the basis of this examination we concluded that poverty reflects structural features of global and national economic and social systems and, in particular, that unequal societies give rise to greater poverty and suboptimal health outcomes. We examined the norms of justice that can guide policy interventions to reduce poverty and its causes. We argued that both equality of opportunity and outcome are essential considerations in public policy and that confronting inequalities in initial conditions is vital. Political responses to poverty reflect philosophical and ideological positions. Social democratic and socialist responses place emphasis on tackling the underlying structural causes of poverty and seek to provide non-stigmatizing supports to those living in poverty. Socialism also seeks to replace capitalist modes of production with those that socialize both production and profits and argues that a transition to such an economy would result in a significant reduction in poverty. Despite the efforts of many civil society groups around the world, poverty or near poverty, and its health consequences, although showing recent decline, are

still the reality for about half of the world's population. This fact, together with growing economic inequities, does suggest that a new world economic and social order is required to dramatically reduce or eliminate poverty and the adverse health outcomes to which it gives rise.

ACKNOWLEDGEMENTS

Text extracts from Labonté, R. et al., *Indicators of Social Exclusion and Inclusion: A Critical and Comparative Analysis of the Literature*, E/Exchange Working Paper Series, PHIRN, University of Ottawa, Canada © 2012, reproduced with permission from the author.

Text extracts from Labonté, R., *Global Health Policy: Exploring the Rationale for Health in Foreign Policy*, Globalization and Health Equity Working Papers, University of Ottawa, Canada, Copyright ©2010, reproduced with permission from the author.

REFERENCES

Aasland, A. and Flotten, T. (2001). Ethnicity and social exclusion in Estonia and Latvia. *Europe-Asia Studies*, 53(7), 1023–49.

Aguirre, D. (2008). *The Human Right to Development in a Globalized World*. Aldershot: Ashgate Publishing Company.

Aizenman, N. (2017). *Health Care Costs Push A Staggering Number of People into Extreme Poverty.* National Public Radio: 'Goats and Soda: Stories of Life in a Changing World'. [Online] Available at: https://www.npr.org/sections/goatsandsoda/2017/12/14/569893722/health-care-costs-push-a-staggering-number-of-people-into-extreme-poverty

Alcock, P. (1993). *Understanding Poverty.* London: Macmillan.

Alvaredo, F., Chancel, L., Piketty, T., Saez, E., and Zucman, G. (2018). *World Inequality Report 2018.* World Inequality Lab. [Online] Available at: http://wir2018.wid.world/

Bailey, N., Spratt, J., Pickering, J., Goodlad, R., and Shucksmith, M. (2004). *Deprivation and Social Exclusion in Argyll and Bute: Report to the Scottish Centre for Research on Social Justice*. Glasgow: Scottish Centre for Research on Social Justice, Universities of Glasgow and Aberdeen.

Balabanova, D., McKee, M., and Mills, A. (2011). *'Good Health at Low Cost' 25 Years On: What Makes a Successful Health System?* London: London School of Hygiene & Tropical Medicine.

Baum, F. (2008). The commission on the social determinants of health: reinventing health promotion for the twenty-first century? *Critical Public Health*, 18(4), 457–66.

Beiser, M., Hou, F., Hyman, I., and Tousignant, M. (2002). Poverty, family process, and the mental health of immigrant children in Canada. *American Journal of Public Health*, 92(2), 220–7.

Berger-Schmitt, R. and Noll, H.H. (2000). *Conceptual Framework and Structure of a European System of Social Indicators*. Report No.: 9. Mannheim: Centre for Survey Research and Methodology ZUMA.

Birdsall, N., Lusting, N., and McLeod, D. (2011). *Declining Inequality in Latin America: Some Economics, Some Politics*. Washington, DC: Center for Global Development.
Boggio, A. (2009). Health and development: an ethics perspective. In: Gatti, A. and Boggio, A. (eds.) *Health and Development: Towards a Matrix Approach*. Houndmills: Palgrave Macmillan, pp. 140–52.
Bourguignon, F. (2004). *The Poverty-Growth-Inequality Triangle*. Paper presented at the Indian Council for Research on International Economic Relations, New Delhi, 4 February.
Brown, R. (1979). Exporting medical education: professionalism, modernization and imperialism. *Social Science & Medicine*, 13(A), 585–95.
Central Intelligence Agency (2013). *The World Fact Book*. [Online] Available at: https://www.cia.gov/library/publications/the-world-factbook/geos/xx.html
Chadwick, E. (1930). Edwin Chadwick, Report of Poor Law Commissioner to the British Parliament on Sanitary Conditions, 1842. In: Scott, J.F. and Baltzly, A. (eds.) *Readings in European History Since 1814*. New York: Appelton Century-Crofts Inc.
Cheek, R.B. (2001). Playing God with HIV. *African Security Review*, 10(4), 19–28.
Chen, S. and Ravallion, M. (2012). *An Update to the World Bank's Estimates of Consumption Poverty in the Developing World*. Briefing Note 03-0112. Washington, DC: World Bank.
Collier P. (2007). *The Bottom Billion: Why the Poorest Countries are Failing and What Can Be Done About It*. New York: Oxford University Press.
Commission on Social Determinants of Health (2008). *Closing the Gap in a Generation: Health Equity Through Action on the Social Determinants of Health (Final Report)*. Geneva: World Health Organization.
Committee on Economic Social and Cultural Rights (2000). *Substantive Issues Arising in the Implementation of the International Covenant of Economic, Social and Cultural Rights: General Comment No. 14*. Report No.: E/C.12/2000/4. Geneva: United Nations Economic and Social Council.
Committee on Economic Social and Cultural Rights (2001). *Substantive Issues Arising in the Implementation of the International Covenant on Economic, Social and Cultural Rights: Poverty and the International Covenant on Economic, Social and Cultural Rights*. Report No.: E/C 12/2001/10. New York: UN Economic and Social Council.
Daly, H. and Cobb, J. (1990). *For the Common Good*. Baltimore, MD: Beacon Press.
Dasgupta, P. (1995). *An Inquiry into Well-Being and Destitution*. Oxford: Oxford University Press.
De Barros, R.P. and Ferreira, F.H.G. (2009). *Measuring Inequality of Opportunities in Latin America and the Caribbean*. Washington, DC: World Bank Publications.
De Haan, A. (1999). *Social Exclusion: Towards an Holistic Understanding of Deprivation*. London: Department for International Development.
De Long, B. (1998). *Estimates of World GDP, One Million B.C. Present*. [Online] Available at: http://delong.typepad.com/print/20061012_LRWGDP.pdf
Dertwinkel, T. (2008). *Economic Exclusion of Ethnic Minorities: On the Importance of Concept Specification*. Flensburg: European Centre for Minority Issues.
Drieskens, S., Van Oyen, H., Demarest, S., Van der Heyden, J., Gisle, L., and Tafforeau, J. (2009). Multiple risk behaviour: increasing socio-economic gap over time? *The European Journal of Public Health*, 20(6), 634–9.
Edward, P. (2006). The ethical poverty line: a moral quantification of absolute poverty. *Third World Quarterly*, 27(2), 377–93.
Engels, F. (1999). *The Condition of the Working Classes in England in 1844*. Oxford: Oxford University Press. (Originally published in 1845.)
Estivill, J. (2003). *Concepts and Strategies for Combating Social Exclusion: An Overview*. Portugal: International Labour Office.
European Commission (2001). *Draft Joint Report on Social Exclusion*. Brussels: European Commission.

Eurostat (2010). *Combating Poverty and Social Exclusion. A Statistical Portrait of the European Union*. Luxembourg: Eurostat.
Evans, R. and Stoddart, G. (1990). Producing health, consuming health care. *Social Science & Medicine*, 31(12), 1347–63.
Ferreira, F. and Sanchez, C. (2017). *Let's Talk Development: A richer array of international poverty lines.* The World Bank. [Online] Available at: http://blogs.worldbank.org/developmenttalk/richer-array-international-poverty-lines
Fioramonti, L. (2017). *Wellbeing Economy: Success in a World Without Growth*. South Africa: Pan MacMillan.
Gianella-Malcam, C., Parra-Vera, O., Eli Yamin, A., and Torres-Tovar, M. (2009). Democratic deliberation or social marketing? The dilemmas of a public definition of health in the context of the implementation of Judgment T-760/08. *Health and Human Rights*, 11(1).
Gore, C. and Figueiredo, J.B. (1997). *Social Exclusion and Anti-Poverty Policy: A Debate*. Geneva: International Labour Organization.
Gough, J., Eisenschitz, A., and McCulloch, A. (2006). *Spaces of Social Exclusion*. Abingdon: Routledge.
Governance Institute of Australia. (2017). *The Good Governance Guide: What is Good Governance?* Produced by the Municipal Association of Victoria (MAV), Victorian Local Governance Association (VLGA), Local Government Victoria (LGV) and Local Government Professionals (LGPro). [Online] Available at: http://www.goodgovernance.org.au/
Grindle, M.S. (1996). *Challenging the State: Crisis and Innovation in Latin America and Africa*. Cambridge: Cambridge University Press.
Hallerod, B. (1996). Deprivation and poverty: a comparative analysis of Sweden and Great Britain. *Acta Sociologica*, 39(2), 141–68.
Hertzman, C. and Boyce, T. (2010). How experience gets under the skin to create gradients in developmental health. *Annual Review of Public Health*, **31**, 329–47.
Hickel, J. (2016). The true extent of global poverty and hunger: questioning the good news narrative of the Millennium Development Goals. *Third World Q*, **37**, 749–67.
Hogerzeil, H.V., Samson, M., Casanovas, J.V., and Rahmani-Ocora, L. (2006). Is access to essential medicines as part of the fulfilment of the right to health enforceable through the courts? *The Lancet*, **368**, 305–11.
Horton, S. and Ross, J. (2003). The economics of iron deficiency. *Food Policy*, 28(1), 51–75.
Hotez, P.J. (2002). *Appeasing Wilson's Ghost: The Expanded Role of the New Vaccines in International Diplomacy.* Washington, DC: Chemical and Biological Arms Control Institute.
Jolliffe, D. and Beer Prydz, E. (2016). Estimating international poverty lines from comparable national thresholds. *Journal of Economic Inequality*, 14(2): 185–98.
Kawachi, I., Subramanian, S.V., and Kim, D. (2008). *Social Capital and Health: A Decade of Progress and Beyond*. New York: Springer.
Kontis, V., et al. (2017). Future life expectancy in 35 industrialised countries: Projections with a Bayesian model ensemble. *The Lancet*, **389**, 1323–35.
Kramer, M. (1987). Determinants of low birth weight: methodological assessment and meta-analysis. *Bulletin of the World Health Organization*, 65(5), 663–737.
Krieger, N. (2000). Discrimination and health. In: Berkman, L.F. and Kawachi, I. (eds.) *Social Epidemiology*. Oxford: Oxford University Press, pp. 36–75.
Labonté, R. (2010). *Global Health Policy: Exploring the Rationale for Health in Foreign Policy*. University of Ottawa: Globalization and Health Equity Working Papers.
Labonté, R., Hadi, A., and Kaufmann, X. (2012). *Indicators of Social Exclusion and Inclusion: A Critical and Comparative Analysis of the Literature*. University of Ottawa: E/Exchange Working Paper Series, PHIRN.
Labonté, R. and Schrecker, T. (2009). Rights, redistribution and regulation. In: Labonté, R., Schrecker, T., Packer, C., and Runnels, V. (eds.) *Globalization and Health: Pathways, Evidence*

and Policy. New York: Routledge, pp. 317–33.
Layte, R. and Whelan, C.T. (2009). Explaining social class inequalities in smoking: the role of education, self-efficacy, and deprivation. *European Sociological Review*, 25(4), 399–410.
Lee, K., Koivusalo, M., Ollila, E., Labonté, R., Schuftan, C., and Woodward, D. (2009). Global governance for health. In: Labonté, R., Schrecker, T., Packer, C., and Runnels, V. (eds.) *Globalization and Health: Pathways, Evidence and Policy*. New York: Routledge, pp. 289–316.
Lessof, C. and Jowell, R. (2000). *Measuring Social Exclusion*. Oxford: University of Oxford.
Levitas, R. (2005). *The Inclusive Society?: Social Exclusion and New Labour* (2nd ed.). Basingstoke: Palgrave Macmillan.
Lister, R. (2004). *Poverty*. London: Polity Press.
Lynch, J., Kaplan, G., and Shema, S. (1997). Cumulative impact of sustained economic hardship on physical, cognitive, psychological, and social functioning. *The New England Journal of Medicine*, 337(26), 1889–95.
Lynch, J. and Smith, G.D. (2005). A life course approach to chronic disease epidemiology. *Annual Review of Public Health*, **26**, 1–35.
Macinko, J., de Fátima Marinho de Souza, M., Guanais, F., and da Silva Simoes, C. (2007). Going to scale with community-based primary care: an analysis of the family health program and infant mortality in Brazil, 1999–2004. *Social Science & Medicine*, 65(10), 2070–80.
Mack, E. (2009). Introduction. In: Mack, E., Schramm, M., Klasen, S., and Pogge, T. (eds.) *Absolute Poverty and Global Justice*. Burlington, VT: Ashgate, pp. 1–18.
Mackenbach, J.P. (2012). The persistence of health inequalities in modern welfare states: the explanation of a paradox. *Social Science & Medicine*, 75(4), 761–9.
Marmot, M. (2010). *Fair Society, Healthy Lives: The Marmot Review*. London: The Marmot Review.
Masters, J. and Wickstrom, T. (2006). *Defining and Measuring Poverty: Challenges and Opportunities*. Berkeley, CA: Center for Community Futures.
Monteiro, C.A., Levy, R.B., Claro, R.M., de Castro, I.R.R., and Cannon, G. (2010). Increasing consumption of ultra-processed foods and likely impact on human health: evidence from Brazil. *Public Health Nutrition*, 14(1), 5.
Murray, C. (1984). *Losing Ground: American Social Policy, 1950–1980*. New York: Basic Books.
Najman, J.M. (1994). Class inequalities in health and lifestyle. In: Waddell, C. and Petersen, A.R. (eds.) *Just Health: Inequalities in Illness, Care and Prevention*. Melbourne: Churchill Livingston, pp. 27–46.
Nolan, B. and Whelan, C.T. (1996). *Resources, Deprivation, and Poverty*. Oxford: Clarendon Press.
Nunes, C. (2008). *Poverty Measurement: The Development of Different Approaches and Its Techniques*. Report No.: ECINEQ WP 2008–93. Portugal: Society for the Study of Economic Inequality.
Nussbaum, M.C. (2000). *Women and Human Development: The Capabilities Approach*. Cambridge: Cambridge University Press.
Office of the High Commissioner for Human Rights (1986). *Declaration on the Right to Development*, 41/128. New York: Office of the High Commissioner for Human Rights.
Office of the High Commissioner for Human Rights (1999). *Human Rights and Extreme Poverty. Report of the Independent Expert on Human Rights Submitted to Commission on Human Rights*. 55th Session. Report No.: E/CN 4/1999/48. New York: UN Economic and Social Council.
OPHI and UNDP (2010). *Multidimensional Poverty Index*. Oxford Poverty and Human Development Initiative. University of California, Berkeley, Department of Economics. [Online] Available at: http://www.ophi.org.uk/wp-content/uploads/MPI_One_Page_final_updated.pdf?cda6c1
Oppenheim, C. and Harker, L. (1990). *Poverty: The Facts*. London: Child Poverty Action Group.

Organisation for Economic Co-operation and Development (2008). *Growing Unequal? Income Distribution and Poverty in OECD Countries*. Paris: OECD.

Organisation for Economic Co-operation and Development (2010). Poverty rates and gaps. In: OECD (ed.) *OECD Factbook 2010: Economic, Environmental and Social Statistics*, pp. 236–37. Paris: OECD Publishing.

Paim, J., Travassos, C., Almeida, C., Bahia, L., and Macinko, J. (2011). The Brazilian health system: history, advances, and challenges. *The Lancet*, **377**, 1778–97.

Paradies, Y. (2006). A systematic review of empirical research on self-reported racism and health. *International Journal of Epidemiology*, 35(4), 888–901.

People's Health Movement (2009). *The People's Charter for Health*. [Online] Available at: http://www.phmovement.org/en/resources/charters/peopleshealth

People's Health Movement, Medact, Health Action International, Medico International and Third World Network (2011). Primary health care: a review and critical appraisal of its 'revitalisation'. In: *Global Health Watch 3: An Alternative World Health Report*, pp. 45–61. London: Zed Books.

Peterson, S. and Shellman, S. (2006). *AIDS and Violent Conflict: The Indirect Effects of Disease on National Security*. Working Paper. Williamsburg, VA: College of William and Mary.

Pogge, T. (2002). *World Poverty and Human Rights*. Cambridge: Polity.

Pogge, T. (2004). Relational conceptions of justice: responsibilities for health outcomes. In: Anand, S., Fabienne, P., and Sen, A. (eds.) *Public Health, Ethics and Equity*. Oxford: Clarendon Press, pp. 135–61.

Pogge, T. (2005). World poverty and human rights. *Ethics & International Affairs*, 9(1), 1–7.

Policy Analysis Discussion Group (2012). *Understanding Poverty and Well-Being: A Note with Implications for Research and Policy*. London: Overseas Development Institute.

Popay, J., Escorel, S., Hernández, M., Johnston, H., Mathieson, J., and Rispel, L. (2008). *Understanding and Tackling Social Exclusion*. Final Report to the WHO Commission on Social Determinants of Health from the Social Exclusion Knowledge Network, SEKN. Geneva: SEKN.

Power, C., Matthews, S., and Manor, O. (1996). Inequalities in self rated health in the 1958 birth cohort: lifetime social circumstances or social mobility? *British Medical Journal*, **313**, 449–53.

Power, C., Thomas, C., Li, L., and Hertzman, C. (2012). Childhood psychosocial adversity and adult cortisol patterns. *British Journal of Psychiatry*, 201(3), 199–206.

Ranis, G., Stewart, F., and Samman, E. (2007). Human development: beyond the human development index. *Journal of Human Development*, 7(3), 323–58.

Ravallion, M. (2011). *What Does Adam Smith's Linen Shirt Have to do with Global Poverty? Let's Talk Development*. A blog hosted by the World Bank's Chief Economist. [Online] Available at: http://blogs.worldbank.org/developmenttalk/node/616

Ravallion, M. and Chen, S. (2017). *Welfare-Consistent Global Poverty Measures.* Policy Research Working Paper 8170. Washington, DC: World Bank Group, Development Research Group, Poverty and Inequality Team. [Online] Available at: https://openknowledge.worldbank.org/handle/10986/27977

Rawal, N. (2008). Social inclusion and exclusion: a review. *Dhaulagiri Journal of Sociology and Anthropology*, **2**, 161–80.

Room, G. (1995). *Beyond the Threshold: The Measurement and Analysis of Social Exclusion.* Bristol: The Policy Press.

Ruger, J. (2008). Normative foundations of global health law. *Georgetown Law Journal*, **96**, 423–43.

Saad-Filho, A. (2016). *Watch Out for Judicial Coup in Brazil. Monthly Review* (Magazine). [Online] Available at: https://mronline.org/2016/03/23/sf230316-html/

Sanders, D. and Chopra, M. (2007). Poverty, social inequity, and child health. In: Kibel, M., Saloojee, H., and Westwood, T. (eds.) *Child Health for All* (4th ed.). Oxford University Press Southern Africa Ltd, pp. 22–30.

Santarelli, E. and Figini, P. (2002). *Does Globalization Reduce Poverty? Some Empirical Evidence for the Developing Countries*. Working Paper. Bologna: Dipartimeto Scienze Economiche, Università di Bologna.

Saunders, P., Naidoo, Y., and Griffiths, M. (2008). Towards new indicators of disadvantage: deprivation and social exclusion in Australia. *Australian Journal of Social Issues*, 43(2), 175.

Saunders, P. and Wong, M. (2009). *Still Doing it Tough: An Update on Deprivation and Social Exclusion Among Welfare Service Clients*. Sydney: Social Policy Research Centre, University of New South Wales.

Schaefer, D.L. (2007). Procedural versus substantive justice: Rawls and Nozick. *Social Philosophy and Policy*, 24(1), 164–86.

Schrecker, T., Chapman, A.R., Labonté, R., and De Vogli, R. (2010). Advancing health equity in the global marketplace: how human rights can help. *Social Science & Medicine*, 71(8), 1520–26.

Sen, A. (1985). *Commodities and Capabilities*. Amsterdam: North-Holland.

Sen, A. (1992). *Inequality Re-examined*. Boston, MA: Harvard University Press.

Sen, A. (1999). *Development as Freedom*. Oxford: Oxford University Press.

Sen, A. (2000). *Social Exclusion: Concept, Application and Scrutiny*. Manila: Office of Environment and Social Development, Asian Development Bank.

Singer, P. (1972). Famine, affluence, and morality. *Philosophy & Public Affairs*, 1(1), 229–43.

Singer, P.W. (2002). AIDS and international security. *Survival*, 44(1), 145–58.

Smith, A. (1776). *An Inquiry into the Nature and Causes of the Wealth of Nations*. London: A. and C. Black.

Smith, G.D., Hart, C., Blane, D., Gillis, C., and Hawthorne, V. (1997). Lifetime socioeconomic position and mortality: prospective observational study. *British Medical Journal*, **314**, 547.

Smith, G.D., Hart, C., Blane, D., and Hole, D. (1998). Adverse socio-economic conditions in childhood and cause specific adult mortality: prospective observational study. *British Medical Journal*, **316**, 1635.

Springer, K.W., Hankivsky, O., and Bates, L.M. (2012). Introduction to special issue on gender and health: relational, intersectional, and biosocial approaches to gender and health. *Social Science & Medicine*, 74(11), 1661–6.

Stuckler, D. and Nestle, M. (2012). Big food, food systems, and global health. *PLoS Medicine*, 9(6), e1001242.

Sumner, A. (2010). Global poverty and the new bottom billion: what if three-quarters of the world's poor live in middle-income countries? *IDS Working Papers*, 2010(349), 1–43.

Sustainable Development Knowledge Platform (2017). *Progress of Goal 3 in 2017*. Sustainable Development Goal 3: Ensure healthy lives and promote well-being for all at all ages. Department of Economic and Social Affairs, United Nations.

Sustainable Development Knowledge Platform (2017). *Progress of Goal 10 in 2017*. Sustainable Development Goal 10: Reduce inequality within and among countries. Department of Economic and Social Affairs, United Nations.

Szreter, S. (2003). The population health approach in historical perspective. *American Journal of Public Health*, 93(3), 424.

Tesh, S. (1988). *Hidden Arguments: Political Ideology and Disease Prevention Policy*. New Brunswick, NJ: Rutgers University Press.

UNDP (2011). *Human Development Report 2011: Sustainability and Equity: A Better Future for All*. New York: Palgrave Macmillan.

UNICEF Innocenti Research Centre (2012). *Measuring Child Poverty: New League Tables of Child Poverty in the World's Rich Countries*. Innocenti Report Card 10. Innocenti, Florence: UNICEF Office of Research.

UNICEF Innocenti Research Centre (2017). *Building the Future: Children and the Sustainable Development Goals in Rich Countries.* Innocenti Report Card 14. Innocenti, Florence: UNICEF Office of Research.

United Nations Conference on Trade and Development (UNCTAD) (2018). *World Investment Report 2018: Investment and New Industrial Policies.* New York and Geneva: United Nations.
United Nations Division for Social Policy (2006). *Social Justice in an Open World: The Role of the United Nations*. New York: United Nations Publications.
United Nations General Assembly (2012). *Global Health and Foreign Policy*. Report No.: A/67/L.36. New York: United Nations.
Wagstaff, A., Flores, G., Smitz, M., Hsu, J., Chepynoga, K., and Eozenou, P. (2018). Progress on impoverishing health spending in 122 countries: a retrospective observational study. *The Lancet: Global Health*, 6(2), 180–92.
Waitzkin, H. (2005). Commentary: Salvador Allende and the birth of Latin American social medicine. *International Journal of Epidemiology*, 34(4), 739–41.
Waters, S. (2004). Mobilising against globalisation: Attac and the French intellectuals. *European Politics*, 27(5), 854–74.
Whitehead, M. (1992). The concepts and principles of equity in health. *International Journal of Health Services*, **22**, 429–45.
Woodward, D. and Simms, A. (2006). *Growth Isn't Working: The Unbalanced Distribution of Benefits and Costs from Economic Growth*. London: New Economics Foundation.
World Bank (2002). *Globalization, Growth, and Poverty: Building an Inclusive World Economy*. New York: Oxford University Press.
World Bank (2016). *Taking on Inequality: Poverty and Shared Prosperity.* Washington, DC: World Bank Group, International Bank for Reconstruction and Development. [Online] Available at: http://www.worldbank.org/en/publication/poverty-and-shared-prosperity
World Bank (2006). *World Development Report 2006: Equity and Development*. New York: Oxford University Press for the World Bank.
World Bank (2011). *World Development Report 2011: Conflict, Security, and Development*. Washington, DC: World Bank.
World Health Organization (2005). *Medical Costs Push Millions of People into Poverty Across the Globe*. Geneva: WHO. [Online] Available at: http://www.who.int/mediacentre/news/releases/2005/pr65/en/index.html
Yamin, A.E. and Parra-Vera, O. (2009). How do courts set health policy? The case of the Colombian Constitutional Court. *PLoS Medicine*, 6(2), 147–50.
Yitzhaki, S. (1979). Relative deprivation and the Gini coefficient. *The Quarterly Journal of Economics*, **93**, 321–4.
Young, I.M. (2006). Responsibility and global justice: a social connection model. *Social Philosophy and Policy*, 23(1), 102–30.
Ziersch, A.M., Gallaher, G., Baum, F., and Bentley, M. (2011). Responding to racism: insights on how racism can damage health from an urban study of Australian Aboriginal people. *Social Science & Medicine*, 73(7), 1045–53.

2.4

Socioeconomic inequalities in health in high-income countries

The facts and the options

Frank J. van Lenthe and Johan P. Mackenbach

Introduction to socioeconomic inequalities in health in high-income countries

In the first decades of the twenty-first century, socioeconomic inequalities in health still represent one of the greatest challenges for public health worldwide. This chapter aims to review the available evidence on the magnitude and explanation of these health inequalities (the 'facts'), and to present the available evidence on what can be done to reduce health inequalities (the 'options'). Socioeconomic inequalities in health will be defined as systematic differences in morbidity or mortality rates between people of higher and lower socioeconomic status, as indicated by, for example, level of education, occupational class, or income level.

This chapter only deals with the situation in high-income countries. Wherever possible, we will draw upon international overviews, such as comparative studies, in order to avoid biases related to the selective experiences of single countries. Most of the illustrations will be drawn upon the European experience, which has become very well documented in the past three decades.

An active interest in health inequalities in Europe can be linked to the publication of the Black Report in England in 1980 (Department of Health and Social Services 1980), which has been followed by an enormous rise of descriptive studies in many other countries, testifying to the existence of substantial inequalities in health in all countries with available data. Gradually the emphasis of academic research in this area has shifted from description to explanation, not only to satisfy scientific curiosities, but also to find entry-points for policies and interventions to reduce health inequalities. As a result, our understanding of the causes of socioeconomic inequalities in health has expanded tremendously, and has allowed interested policymakers to start searching for strategies to reduce these inequalities. The World Health Organization has actively promoted the issue of health inequalities by creating a Commission on Social Determinants of Health that has issued a major report urging policy action to 'close the gap in a generation' (Commission on Social Determinants of Health 2008). In some countries political windows of opportunity have arisen which have led to national programmes to tackle health inequalities. Evaluations of the most ambitious of these programmes (the 'English strategy') point to a need to double our efforts to find effective interventions and policies that can be implemented on a sufficiently large scale to achieve population-wide impacts (Mackenbach 2011).

The facts: description

Socioeconomic status (SES) should preferably be measured by individual-level indicators like education (e.g. highest completed level of education), occupation (e.g. occupational class), or income (e.g. household income adjusted for household size). Usually, all these indicators point in the same direction, with health being worse for those in lower socioeconomic groups, but there is often

substantial variation between indicators in the strength of the association, suggesting that the underlying etiological mechanisms are different as well. There is no ‘best single measure’.

The magnitude of socioeconomic inequalities in health can be measured in different ways, ranging from simple measures like rate ratios and rate differences (indicating relative and absolute differences in morbidity or mortality rates between two contrasting groups), to more complex measures like the relative index of inequality and the slope index of inequality (indicating relative and absolute differences in rates between those with the lowest and highest SES, taking all groups into account and calibrating the social position of each group to its relative size) (Mackenbach and Kunst 1997). Although the choice of measure will partly depend on technical considerations, each emphasizes different aspects of health inequalities, often at the expense of others, and it may therefore be important to combine different measures.

Socioeconomic inequalities in health usually present themselves as a gradient, in the sense that there is a stepwise increase of rates of morbidity and mortality with every step down the social ladder, implying that social inequality affects the health of nearly everybody, not only the worst-off.

Mortality

Total mortality

Evidence shows that mortality rates are higher among those in less advantaged socioeconomic groups. According to a cross-national comparative study, the relative index of inequality is greater than 2 for men in a number of European countries in the period around 2005–2009, indicating that mortality rates are twice as high for those in the lowest as compared to the highest educational groups (Table 2.4.1). For women, relative inequalities are reasonably similar or slightly lower (Mackenbach et al. 2016).

Table 2.4.1 Relative indices of inequality and slope indices of inequality of all-cause mortality by education, by country and sex, ca. 1990–4 and ca. 2005–9

	Absolute inequalities							Relative inequalities						
	1990–1994			2005–2009			Change	1990–1994			2005–2009			Change
	SII	95%-CI		SII	95%-CI		p-value	RII	95%-CI	95%-CI	RII	95%-CI	95%-CI	p-value
MEN														
Finland	**1,098**	1,059	1,136	**880**	856	905	**<0.001**	**2.2**	2.1	2.2	**2.5**	2.5	2.6	**0.000**
Sweden	**654**	627	680	**526**	507	544	**<0.001**	**1.8**	1.8	1.9	**2.0**	2.0	2.1	**0.000**
Norway	866	827	905	831	795	864	0.167	**1.9**	1.8	1.9	**2.6**	2.5	2.7	**0.000**
Scotland	**1,634**	1,353	1,890	**591**	427	749	**<0.001**	**3.3**	2.6	4.2	**1.8**	1.5	2.1	**0.001**
England & Wales	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	2.3	2.0	2.7	NA
France	**932**	815	1,056	**718**	621	802	**0.005**	2.3	2.0	2.5	2.2	2.0	2.5	0.579
Switzerland	**788**	760	817	**642**	615.0	667	**0.000**	**2.1**	2.0	2.1	**2.5**	2.4	2.6	**0.000**
Spain (Barcelona)	**956**	893	1,023	**618**	565	669	**<0.001**	2.2	2.1	2.4	2.2	2.0	2.3	1.000
Italy (Turin)	**613**	526	699	**501**	436	567	**0.043**	**1.8**	1.6	1.9	**2.1**	1.9	2.3	**0.019**
Slovenia	**1,412**	1,690	1,876	**824**	1,597.5	1,731.3	**<0.001**	**2.6**	2.4	2.7	**2.3**	2.2	2.4	**0.001**
Lithuania	**1,112**	1,045	1,182	**2021**	1966.9	2,072.0	**<0.001**	**1.9**	1.8	2.0	**3.3**	3.2	3.4	**0.000**
WOMEN														
Finland	**391**	364	418	**357**	340	374	**0.036**	**1.8**	1.7	1.9	**2.2**	2.1	2.3	**<0.001**
Sweden	**368**	349	389	**343**	329	358	**0.046**	**1.9**	1.8	1.9	**2.1**	2.0	2.2	**0.001**
Norway	**420**	391	449	**506**	479	535	**<0.001**	**1.8**	1.7	1.9	**2.6**	2.4	2.7	**0.000**
Scotland	**601**	377	830	**317**	177	461	**0.037**	2.0	1.5	2.6	1.6	1.3	2.0	0.229
England & Wales	NA	NA	NA	401	333	482	NA	NA	NA	NA	2.1	1.8	2.4	NA
France	336	246	423	301	238	367	0.537	2.0	1.6	2.4	2.1	1.7	2.5	0.729
Switzerland	**259**	241	279	**227**	207	247	**0.031**	**1.6**	1.6	1.7	**1.7**	1.7	1.8	**0.006**
Spain (Barcelona)	**350**	301	399	**190**	157	225	**<0.001**	**2.0**	1.8	2.2	**1.7**	1.5	1.9	**0.038**
Italy (Turin)	257	188	322	178	129	223	0.069	1.6	1.4	1.8	1.6	1.4	1.9	1.000
Slovenia	**449**	529	679	**296**	645	751	**0.001**	1.9	1.7	2.0	1.9	1.8	1.9	1.000
Lithuania	**307**	264	350	**708**	679	735	**<0.000**	**1.5**	1.4	1.5	**3.6**	3.4	3.9	**0.000**

Reprinted from Mackenbach J.P et al. (2016) 'Changes in mortality inequalities over two decades: register based study of European countries' *BMJ* 2016;353:i1732, with permission from the BMJ Publishing Group.

The magnitude of these inequalities varies substantially among European countries. For example, in the period around 2005–2009, the relative index of inequality for men is less than 2 in Scotland, whereas it is more than 3 for men in Lithuania. Education-related relative inequalities in mortality are smaller than the European average in all southern European populations and larger than average in most countries in the eastern and Baltic regions. Similar patterns are seen for occupation-related inequalities in mortality among middle-aged men (Mackenbach et al. 2008).

Within Europe, the international pattern observed for relative inequalities in mortality also applies to absolute inequalities in mortality, as indicated by the slope index of inequality. This is because countries with larger relative inequalities in mortality, i.e. those in the eastern and Baltic regions, also tend to have higher average rates of mortality. As a result, the range of variation of absolute inequalities in mortality is even larger than that seen for relative inequalities, suggesting an enormous scope for reducing inequalities in mortality (Mackenbach et al. 2016). While the mortality rates of the higher educated are rather similar in different countries, the mortality rates of the lower educated are much more variable, and higher in countries with larger inequalities in mortality. In other words, the mortality rates among the lower educated drive the magnitude of inequalities in mortality in a country, suggesting that the lower educated are more vulnerable to unfavourable national conditions (van Raalte et al. 2011).

Cause-specific mortality

Across Europe, rates of mortality from ischaemic heart diseases (IHD) are higher among lower as compared to higher socioeconomic groups (Kulhanova et al. 2016). The highest mortality rates in the lowest educational group are observed in the eastern European countries included. In fact, eastern European countries with high inequalities in IHD mortality rates show higher mortality rates in all educational groups. Unfavourable material circumstances, lack of control over life circumstances, and (a culture that approves) hazardous drinking seem to have affected health of the entire population, and those in the lowest education groups specifically (Mackenbach et al. 2008). Southern European countries never witnessed an ‘epidemic’ of IHD mortality, and inequalities in IHD mortality have not arisen.

Consistently higher mortality rates among those in lower as compared to higher socioeconomic groups are also found for stroke. Again, eastern European countries show above-average absolute inequalities in stroke mortality (Mackenbach et al. 2016). Although absolute mortality rates are substantially lower among women, higher absolute inequalities in stroke mortality are found in women in eastern European countries as well (Mackenbach et al. 2008).

Inequalities in mortality from cancer are often large among men, partly because of the excess lung cancer mortality in lower socioeconomic groups (Van der Heyden et al. 2009). Inequalities in cancer mortality are often smaller in women; sometimes they are really small or even have a ‘reverse’ direction, with higher mortality among the higher educated, particularly in Southern Europe (Mackenbach et al. 2008). This is because lung cancer mortality does not (yet) have the same negative association with SES status among women as it has among men, and because the most important cause of cancer mortality among women, breast cancer, often had a ‘positive’ association with SES. However, some evidence suggests a change in the direction of the association, with positive associations no longer being significant (Gadeyne et al. 2017). For lung cancer, there are similar indications for a future change in gradient among women (Van der Heyden et al. 2009).

External causes of injury are another important contributor to socioeconomic inequalities in mortality, particularly (but not exclusively) in Eastern Europe. Mortality from road traffic accidents, from suicides and from homicides is higher in lower socioeconomic groups among men, but not or less clearly so among women (Lorant et al. 2005; Stickley et al. 2012). The larger inequalities in injury mortality in Eastern Europe are likely to be related to higher rates of excessive alcohol consumption (Mackenbach et al. 2008).

Trends

Socioeconomic inequalities in mortality are not a static phenomenon. In addition to the already mentioned change in the direction of the association between SES and breast cancer mortality, this can be illustrated for IHD. In many countries, particularly in the North of Europe, mortality from IHD increased substantially after the Second World War. In the North of Europe, during the 1950s and 1960s IHD mortality was higher in the higher socioeconomic groups, leading to the notion of IHD being a 'manager's disease'. It was only during the 1970s, coinciding with the start of the decline of IHD mortality in the population as a whole, that the current inverse association emerged (Gonzalez et al. 1998).

In the past two decades, absolute declines in all-cause mortality became larger among lower socioeconomic groups, and contributed to the smaller absolute inequalities in mortality in men and women in some North-European countries (Mackenbach et al. 2016). As a result of economic and social developments after political changes around 1990, Eastern European countries witnessed a widening of the gap in death rates (Leinsalu et al. 2009). Recent evidence, however, shows that the initial increase in mortality rates among lower educated men in many countries in Eastern Europe were followed by a decrease around 2008 (Mackenbach et al. 2018).

The latter finding of a decline in mortality among the lower educated across Europe deviates from the situation in the United States. Specifically, mortality rates for white non-Hispanic US 45–54-year-old adults in the lowest educational group increased (Case and Deaton 2015), which has been mainly attributed to increases in mortality from external causes (drug and alcohol poisoning and suicides).

Life expectancy is shorter in lower socioeconomic groups

As a result of these differences in the risk of dying, people from lower socioeconomic groups live considerably shorter lives than those with more advantaged social positions (Stringhini et al. 2017). Again, this difference varies by country. Whereas the difference in life expectancy between primary or lower secondary educated on the one hand and tertiary educated persons on the other hand is relatively small in Italy (2.3 years among men and 0.6 among women), it is large in Lithuania (9.2 years among men and 4.7 among women) (Mäki et al. 2014).

Again, the United States and Europe seem to deviate when it comes to recent trends. A decrease in life expectancy among the lowest educated contributed to an increase in absolute inequalities in life expectancy between 1990 and 2008 in the United States (Olshansky et al. 2012).

Morbidity

Inequalities in generic health indicators

For one indicator, self-assessed health (measured with a single question on an individual's perception of his or her own health), the availability of data is almost as great as that for inequalities in mortality (Fig. 2.4.1).

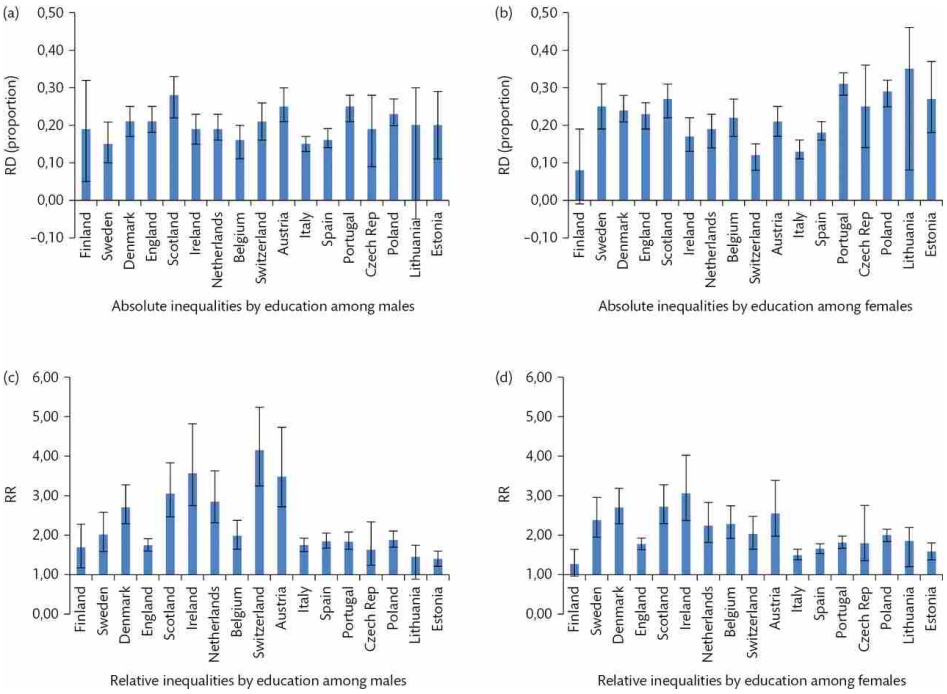


Fig. 2.4.1 Absolute and relative inequalities in self-assessed health across Europe.
Reprinted from Hu, Y. et al. (2016) 'Trends in socioeconomic inequalities in self-assessed health in 17 European countries between 1990 and 2010'. *Journal Epidemiology Community Health* 70: 644–52, with permission from the BMJ Publishing Group Ltd.

The overall pattern is clear again: prevalence rates of less than 'good' self-assessed health are higher in lower socioeconomic groups, resulting in both absolute and relative inequalities (Hu et al. 2016a, b). Although self-assessed health is a good predictor of mortality, trends in inequalities in mortality and in 'less than good' self-assessed health do not correspond exactly. Specifically, whereas absolute inequalities in mortality declined in many countries, they remained rather stable in less than 'good' self-assessed health (Hu et al. 2016a, b).

Inequalities in diseases and disabilities

Socioeconomic inequalities can also be found for many specific indicators, including objective measurements of the incidence or prevalence of diseases and disabilities. Many studies have shown socioeconomic inequalities in the incidence and prevalence of cardiovascular diseases (Manrique-Garcia et al. 2011). Although the size of absolute and relative inequalities seem to differ by diagnosis, substantial inequalities are found for coronary heart diseases, acute myocardial infarction, stroke, and heart failure (Christensen et al. 2016). In addition to a higher incidence, lower socioeconomic groups also tend to have a higher case fatality (Yang et al. 2012).

Similarly, inequalities in cancer mortality reflect inequalities in both incidence and survival. Many cancers, particularly those associated with smoking, excessive alcohol consumption, and other lifestyle factors, such as cancers of the respiratory and upper digestive tract, have a higher incidence in the lower socioeconomic groups, whereas risks of cancer of the prostate, breast, ovary, and colon, and malignant melanoma are often less clear, absent or positive (Aarts et al. 2010). That the prevalence of cancer tends not to be different between socioeconomic groups (Dalstra et al. 2005) is perhaps because longer survival of higher SES patients 'compensates' their lower incidence. Whereas the incidence of breast cancer has been higher for women in higher socioeconomic groups, case fatality rates are higher in lower socioeconomic groups (Lundqvist et al. 2016). Similarly, survival of prostate cancer is lower among patients in lower socioeconomic groups (Klein and Von dem

Knesebeck 2015). As suggested by data on mortality from suicide (see earlier), mental ill health tends to be more prevalent in lower socioeconomic groups (Lorant et al. 2007). The higher prevalence of mental illness in lower socioeconomic groups is likely to have a complex explanation. In psychiatric epidemiology, there is a long tradition of looking at the possible effects of mental health problems on downward social mobility. This 'drift hypothesis' has indeed found some support, for example, in the case of schizophrenia, whose onset usually occurs in adolescence and young adulthood, and which may consequently interfere with school and early work careers. On the other hand, incidence studies have also found higher rates of many mental health problems among those who are currently in a lower socioeconomic position. It seems likely that this at least partly reflects a causal effect, perhaps through a higher exposure to psychosocial stressors and/or a lack of coping resources (Lorant et al. 2007).

As a result of the higher frequency of physical and mental health problems in lower socioeconomic groups, the prevalence of limitations in functioning and various forms of disability also tends to be higher. This applies to many aspects of functioning (mobility, sensory functioning, grip strength, walking speed, etc.) and is particularly evident among the elderly (Stringhini et al. 2018).

'Healthy life expectancy' is shorter in lower socioeconomic groups

The fact that mortality and morbidity rates (among those who are still alive) are higher in lower socioeconomic groups contributes to even larger inequalities in 'healthy life expectancy' (the number of years which people can expect to live in good health). Inequalities in the number of years lived in good health are often seen of more than 10 years (Majer et al. 2011).

The facts: explanation

During the past decades, great progress has been made in unravelling the determinants of health inequalities in high-income countries, and here we will summarize the evidence under three headings: 'selection' versus 'causation', 'specific causal pathways', and 'overarching ideas'.

'Selection' versus 'causation'

Early debates

Selection explanations imply that health determines SES. Because of the occurrence of 'social mobility' (changes in SES during a person's lifetime) there is ample scope for selection processes: health problems may lead to downward social mobility, and upward mobility may be more likely for those with above-average health.

The occurrence of health-related selection is undisputed: during social mobility, some degree of selection on (ill) health does indeed occur, with people who are in poor physical or mental health being more likely to be downwardly mobile (e.g. get a lower status job, or lose income) and less likely to be upwardly mobile (e.g. finish high-level education, or obtain a highly paid job), than people who are in good health. The contribution of health-related selection to the explanation of socioeconomic inequalities in health is small, at least when occupational class or education is used as an indicator of SES (Bartley and Plewis 1997). This is because most health problems occur in late middle or old age, after people have reached their final level of education and main occupational status.

Income, however, is much more sensitive to health status than education or occupational class. For example, poor self-perceived health is an important determinant of transitions into disability pensions, unemployment, and early retirement (Van Rijn et al. 2014), and as such presumably to a

decrease in income. In a study using a more objective indicator of health, employees with an unscheduled hospitalization that could not be postponed experienced a drop in income 2 years later, as compared to matched employees who were not hospitalized. The same absolute reduction in income between the poorest and richest meant that the relative impact of the hospitalization was larger among the poorest (García-Gómez et al. 2013). There is also scope for an impact of what has been called 'indirect selection'—that is, selection on determinants of health. For example, some evidence suggests that obesity affects social mobility, perhaps because of discrimination during recruitment for jobs or promotion (Karnehed et al. 2008).

Selection on the basis of personal attributes, such as cognitive ability, coping styles, control beliefs, personality, and bodily and mental fitness, is also important. These attributes may influence educational and occupational achievement, and at the same time determine later health, either directly or through health-related behaviour. Empirical studies have found associations between many of these personal attributes and social mobility (Lounsbury et al. 2003), and some studies suggest that factors like personality and cognitive ability do indeed contribute to the explanation of health inequalities (Bijwaard et at al. 2015).

Disentangling causal effects

Confounding by both measured and unmeasured variables, and the exclusion of reverse causation are major problems in disentangling the causal impact of SES on health (see Chapter 5.14). Because randomization by SES is not possible, situations have been sought that mimic experimental manipulation (see Chapter 5.10). In lotteries for example, large amounts of money are allocated randomly. Evidence suggests that the impact of winning substantial lottery prizes on mortality, healthcare utilization, and child development outcomes is weak (Cesarini et al. 2016).

Historical changes in the age of compulsory education have been used to study the causal effect of education on mortality. In the United States, between 1915 and 1939, at least 30 states changed their compulsory schooling laws. A comparison of later-life mortality of children attending school just before and just after the changes took effect showed that 1 year of extra schooling reduced mortality substantially (Lleras-Muney 2002). Such results have also been reported for other countries (Kippersluis et al. 2011; Lager and Torssander 2012). A study using state-level school laws and school characteristics to predict years of education, reported that the variation in predicted years of education was inversely related to the risk of dementia (Nguyen et al. 2016). Together, these studies seem to suggest that a causal effect of education is more likely than a causal effect of income and wealth on health outcomes.

Specific causal pathways

The findings summarized in the previous paragraph suggest that socioeconomic inequalities in health at least partly reflect a causal effect of SES on health. The debate on the relative contribution of such causal mechanisms, however, is far from closed. That also applies to the exact nature of these causal mechanisms, which are likely to be largely indirect, by raising the prevalence of specific determinants of morbidity or mortality in lower socioeconomic groups.

Many risk factors for morbidity and mortality are more prevalent in lower socioeconomic groups. Three groups of explanatory factors are likely to explain the bulk of health inequalities in high-income countries: material, psychosocial, and behavioural risk factors (Van Oort et al. 2005). Healthcare may also play a role. Figure 2.4.2 provides a schematic representation of these mechanisms.

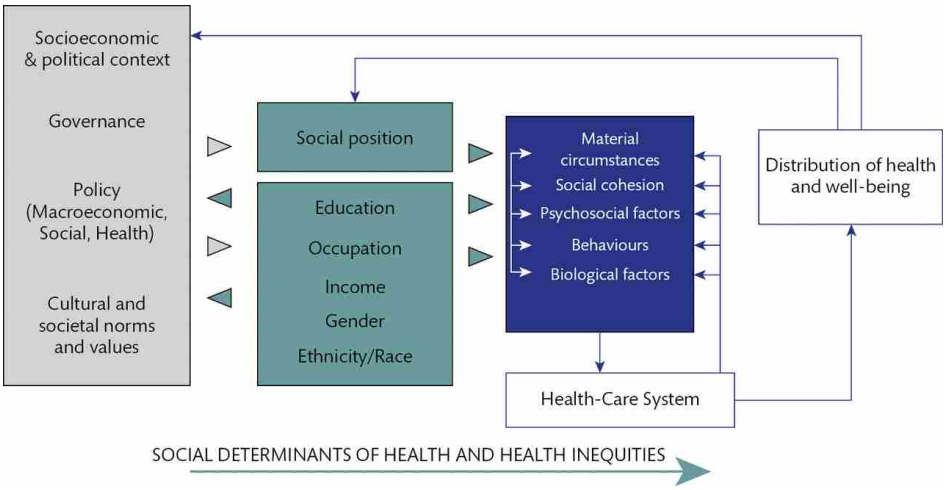


Fig. 2.4.2 A possible model for the explanation of health inequalities.

Reproduced with permission from Commission on Social Determinants of Health, *Closing the gap in a generation: Health equity through the social determinants of health*, World Health Organization; Geneva, Switzerland, Copyright © 2008, available from http://whqlibdoc.who.int/publications/2008/9789241563703_eng.pdf. Adapted with permission from Solar, O. and Irwin, A., *A conceptual framework for action on the social determinants of health: Discussion paper for the Commission on Social Determinants of Health*, World Health Organization; Geneva, Switzerland, Copyright © 2007.

Material factors

There is no doubt that financial resources and related unfavourable material living conditions are very unevenly distributed between people with a lower and higher SES, even in high-income countries with strong egalitarian traditions. For example, in 2014, the average disposable income of the richest ten percent was about 9.5 times higher than that of the poorest ten percent in Europe (OECD 2017).

Financial disadvantage may affect health through various mechanisms. The lowest possible costs for healthy living (as defined by health-based needs for nutrition, physical activity, housing, medical care, social needs, hygiene, and transport) tend to exceed minimum benefit levels (Morris et al. 2007). Yet, for some health behaviours, the cheapest option is also the healthiest, which is obvious in the case of smoking. That smoking is more prevalent among those in the lowest income groups has been attributed, among other factors, to the stress resulting from living in poor material conditions (Beenackers et al. 2018).

Other ‘material’ risk factors that tend to be more prevalent in lower socioeconomic groups include occupational health risks (exposure to chemicals, accident risks, physically strenuous work, etc.) (Costa and D’Errico 2006), health risks related to housing (crowding, dampness, accident risks, etc.) (Dunn and Hayes 1999), and environmental health risks (air pollution, traffic noise, etc.) (Kruize et al. 2007). Some of these factors have been shown to make important contributions to the explanation of health inequalities (Marmot and Wilkinson 2006).

Psychosocial factors

The second group of specific determinants which contribute to the explanation of health inequalities are psychosocial factors. Those with a low SES on average are exposed to more psychosocial stressors, in the form of negative life events, daily hassles, an ‘effort–reward imbalance’ (high levels of effort without appropriate material and immaterial rewards), and a combination of high demands and low control both in the workplace and beyond (Siegrist and Marmot 2004).

At the same time, they also tend to have lower levels of social support (Stansfeld 2006), as well as less effective coping styles (Kristenson 2006) and a weaker sense of control over their life and living conditions (Bosma et al. 1999). This combination of a higher exposure to psychosocial stressors and

less capacity to remove or buffer these exposures may explain part of the higher frequency of health problems in the lower socioeconomic groups. This has been best documented for psychosocial factors related to work organization, such as job strain, which have been shown to play a role in the explanation of socioeconomic inequalities in cardiovascular health (Marmot et al. 2006).

Two possible pathways may be involved. According to the increasingly popular 'scarcity-theory' (Mullainathan et al. 2013) stress from exposure to poor material circumstances (such as financial stress) limits individuals' 'cognitive bandwidth', probably leading to less self-control and ultimately unhealthy behaviour (Beenackers et al. 2018). The second is a biological pathway. The experience of stress affects the neural, endocrine, and immune systems of the body, and chronic stress may lead to maladaptive responses in the form of, for example, high blood pressure, a prolonged high level of cortisol, higher blood viscosity, or a suppression of the immune response, which may in their turn increase the susceptibility to a range of diseases (Brunner 2017). In this context, studies also refer to the concept of 'allostatic load', e.g. the cumulative physiological burden of exposure to stressful circumstances (the bodies' 'wear and tear'). Socioeconomic inequalities in cardiovascular, metabolic, and inflammatory markers used to measure allostatic load, have been explained by stressful living conditions (Robertson et al. 2015), and as such provide an explanation how unfavourable material and psychosocial factors 'get under skin'.

Health-related behaviours

The third group of contributory factors are health-related behaviours, such as smoking, inadequate diet, excessive alcohol consumption, and lack of physical exercise (Oude Groeniger et al. 2016, 2017). In many developed countries one or more of these 'lifestyle' factors are more prevalent in the lower socioeconomic groups. As we have already seen, many of the disease-specific patterns of health inequalities also suggest a substantial contribution of health-related behaviours to inequalities in mortality.

By far the most widely available data on a specific determinant of health inequalities relate to smoking. The prevalence of smoking differs strongly between socioeconomic groups in many European countries (although variations in the magnitude of these inequalities between countries have been reported), so one can safely assume that it plays an important role in generating health inequalities (Huisman et al. 2005). The prevalence of smoking has declined over the past decades in many countries, at least among men. In the first decade of this century, the proportion of quitters among ever-smokers was higher among those in higher than in lower socioeconomic groups, leading to a widening of relative inequalities in smoking (Bosdriesz et al. 2015).

Socioeconomic inequalities in alcohol-attributed mortality suggest an important contribution of excessive alcohol consumption to inequalities in mortality (Probst et al. 2014), particularly in Eastern Europe (Mackenbach et al. 2015a). Indeed, excessive alcohol consumption is more frequent among those with a lower SES, and increases mortality risks (Woods et al. 2018). Alcohol, however, is a complex factor in this context: abstinence, which is more frequent in lower socioeconomic groups, is often associated with a higher mortality risk, as compared to light alcohol consumption in observational epidemiological research (Woods et al. 2018). Moderate alcohol consumption is more frequent in higher socioeconomic groups, and often associated with a lower risk of cardiovascular disease mortality (Ronksley et al. 2011). Methodologically stronger evidence, however, challenges the protective effect of light-to-moderate as compared to non-drinking, and suggests that a lower mortality risk results from residual confounding by for example factors directly related to SES (Holmes et al. 2014).

Socioeconomic inequalities in diet, particularly in fruit and vegetable consumption and fat intake, are well-established (Giskes et al. 2010). Only a few comparative studies have been conducted, and these show that men and women in lower socioeconomic groups tend to less frequently eat fresh vegetables, particularly in the North of Europe. Literature reviews have shown that it is likely that

many other aspects of diet, such as consumption of meat, dairy products, and various fats and oils, also are socially patterned in many European countries, and that these social patterns differ between countries (Lopez-Azpiazu et al. 2003).

Socioeconomic patterns in physical activity depend on the type of activity. A lack of leisure-time physical activity tends to be more common in the lower socioeconomic groups, whereas occupational physical activity is less common among higher socioeconomic groups (Beenackers et al. 2012). High levels of occupational physical activity, however, are related to early mortality (Coenen et al. 2018) and as such may contribute to socioeconomic inequalities in mortality.

A lack of physical activity and an unhealthy diet are the main determinants of obesity. The increase in the prevalence of obesity over the past decades is, beyond doubt, among the major challenges in public health. Interestingly, this is one of the very few health aspects where patterns of social variation are clearer for women than for men. Among women, overweight and obesity are more prevalent in lower socioeconomic groups in European countries with available data, with clearly large inequalities in overweight and obesity in Southern Europe (Roskam et al. 2010). Recent evidence suggests a widening of absolute inequalities in obesity in many countries in Europe, but not of relative inequalities, both among men and women (Hoffmann et al. 2017).

The systematic nature of these differences in health-related behaviour clearly demonstrates that these are not a matter of free choice. While there is evidence of socioeconomic inequalities in intrapersonal determinants of behavioural changes, such as attitudes towards health behaviour (Kamphuis et al. 2009), such findings still do not provide an answer to the question why these determinants are socioeconomically patterned. As a result, the attention shifted towards the role of physical and sociocultural environmental conditions which are at least partly beyond the control of the individual (Van Lenthe et al. 2005).

With regard to the physical ('built') environment, empirical evidence testing the 'deprivation amplification' hypothesis, according to which lower socioeconomic groups are exposed to contextual disadvantages in terms of the availability and accessibility of health enhancing facilities and healthy products produced inconsistent findings, at least in many European countries (Macintyre 2007). There are indications that built environmental characteristics play a more important role in the United States, suggesting again that macro-level processes leading to socioeconomic patterning of these characteristics are involved in generating the inequalities (Cummins and Macintyre 2006). Overall however, there remains an important knowledge gap if and how built environments contribute to inequalities in health behaviour.

With regard to the sociocultural environment, evidence suggests that social capital and social resources, such as membership of support-providing networks and perceived social support, are associated with health behaviour (Conklin et al. 2014). Their contribution to socioeconomic inequalities in health behaviour is less clear. Scholars also explored the role of 'cultural capital', e.g. culture-based resources that shape and influence people's habits, values, norms, knowledge, and preferences, and that are acquired mostly through social learning, with learning conditions varying between socioeconomic groups and milieus (Bourdieu 1984). Bourdieu argued that food and sport preferences may become part of an identity and can be used as a distinction mechanism between social groups. Emerging evidence suggests that cultural capital is a promising explanation for socioeconomic inequalities in healthy behaviour (Oude Groeniger et al. 2017).

Healthcare

Unequal access to effective healthcare could theoretically also play a role in the explanation of health inequalities. Most high-income countries, however, have healthcare financing systems that have reduced inequalities in access to rather low levels. This is clear from studies of healthcare utilization, which typically show that lower socioeconomic groups tend to use more care than higher socioeconomic groups in accordance with their higher levels of need, and that relatively low use of

specialist care in lower socioeconomic groups goes together with relatively high use of primary care (van Doorslaer et al. 2006).

Studies of variations in healthcare outcomes do suggest that there may be inequalities in quality of care that are not captured by studies of global indicators of healthcare utilization. As was mentioned earlier, survival after a cardiovascular event or after a diagnosis of cancer is often shorter for patients in lower socioeconomic groups. Some studies have found higher rates of utilization of surgical and non-surgical interventions among hospitalized patients with a higher socioeconomic position (Stirbu et al. 2012). Like inequalities in survival after myocardial infarction, inequalities in survival of cancer patients are likely to be partly determined by inequalities in comorbidity or other risk factors for mortality independent of care received, but there is some emerging evidence that inequalities in treatment may also play a role, for example, in the case of breast cancer (Aarts et al. 2012).

At the population level, mortality from causes of death that have become amenable to medical intervention also is higher in lower socioeconomic groups. Inequalities in mortality from these conditions are often substantial, and contribute between 11 and 24% to inequalities in partial life expectancy between the ages of 30 and 64 (Stirbu et al. 2010). To what extent these inequalities really reflect inequalities in healthcare access or quality is, however, uncertain. Countries with larger inequalities in healthcare utilization have not been found to also have larger inequalities in amenable mortality (Plug et al. 2012).

A role for genes?

In intriguing question is whether genes contribute to health inequalities. Genetic factors do not easily fit into ‘causation’-type explanations, because temporally a person’s genotype precedes his/her SES, and a causal effect of the latter on the first is therefore logically impossible (Mackenbach 2005). In fact, genes which are both more frequent in lower than in higher socioeconomic groups and predispose to health, would confound the association between SES and health outcomes. This situation is not entirely unlikely: genes may be more frequent in lower socioeconomic groups when they predispose to health outcomes, which subsequently lead to social mobility. For example, genes may predispose to personal attributes such as cognitive ability or personality traits, which may lead to downward social mobility and may affect health. A ‘contribution’ of genes to health inequalities may however, come from ‘gene-environment’ interactions, whereby the health (dis)advantages of higher and lower socioeconomic groups differ between those with different genes. SES may also change gene functioning (e.g. the process of DNA methylation) in absence of changes in DNA sequences (‘epigenetics’), and as such contribute to health inequalities (Fiorito et al. 2017).

Overarching ideas

What has emerged from recent research efforts is a rather complex picture of how individuals in the lower socioeconomic strata are exposed over their lifetime to a wide variety of unfavourable and interacting material, cultural, and psychological conditions, and how these exposures lead to ill health —either directly, or indirectly through unhealthy behaviours or psychosocial stress. Various attempts have been made to combine these factors and pathways into overarching theories. We discuss some of these attempts in the following sections.

Life course perspective

In a life course perspective, health, or disease risk in lower socioeconomic groups later in life result from socially patterned exposures during gestation, childhood, adolescence, young adulthood, and later adult life (see Chapter 5.20). Different life course models have been developed, of which the ‘critical period’ model and the ‘accumulation’ model are most prominent.

In a critical period model, SES has an exclusive effect on health in later life in a short and specific period in life, but not in other periods of life (Mishra et al. 2013). One example of a 'critical period' model relates to the 'developmental origins of health and disease' hypothesis. According to this hypothesis a low birth weight is not only associated with infant health, but also with adult health, e.g. coronary heart disease, stroke, hypertension, and diabetes, probably due to 'fetal programming' of growth patterns and related metabolic and endocrine processes (Barker and Robinson 1992). Robust evidence exists of socioeconomic inequalities in unfavourable birth outcomes, including a low birth weight, preterm birth, babies born small for gestational age (Blumenshine et al. 2010). To the extent that fetal programming results in persistent and structural changes, a low socioeconomic status in pregnancy can be seen as a critical period.

According to the accumulation model, different forms of material and immaterial disadvantage tend to cluster in the same persons, either uncorrelated or with one disadvantage increasing the likelihood of another one at a later point in time. Empirically, accumulation is often tested by investigating dose–response relationships between socioeconomic status at different periods in life and health outcomes in later life (Pollitt et al. 2005). Indeed, risk of disease and mortality increases with an increase in exposure to a low socioeconomic status during life (Mishra et al. 2013).

Accumulation models can be refined to take into account chains of events reinforcing each other. Interestingly, this also resolves the 'selection' versus 'causation' debate, because a low socioeconomic status in one stage of the life course may translate into a health disadvantage in the next, which may then lead to a still lower socioeconomic status some years later, and so on. Health-induced downward social mobility then becomes just another form of accumulation of disadvantage.

Clearly, the implicit assumption in accumulation models of an equally strong impact of socioeconomic status at different periods of life may be invalid. In a 'sensitive period' model exposure to a low socioeconomic status at different periods in life exerts different effects on health in later life. Although disentangling life course models empirically is complex, evidence suggests that a sensitive period model, in which a low socioeconomic status later in life is more strongly related to mortality than a low socioeconomic status earlier in life, best describes the effect of socioeconomic status in mortality across the life course (Mishra et al. 2013).

Fundamental causes

The theory of 'fundamental causes' stipulates that it is the social forces underlying social stratification which ultimately cause health inequalities, and not exposure to the proximal risk factors which are usually studied by social epidemiologists (like smoking, psychosocial stress, working conditions, etc.). According to this theory, the persistence of health inequalities in different time-periods and different national conditions is due to the fact that a person's socioeconomic status provides him or her with 'flexible resources'. These include 'knowledge, money, power, prestige, and beneficial social connections' which can be used 'to avoid disease risks or to minimize the consequences of disease once it occurs' regardless of the prevailing circumstances. The association between SES and health 'is reproduced over time via the replacement of intervening mechanisms', and as opportunities for avoiding disease continue to expand so will health inequalities continue to exist (Phelan et al. 2010).

This theory provides an elegant summary of all the more specific explanations, which fail to explain why health inequalities are seen in all countries with available data, and persist over time, regardless of variations in the proximal determinants of health. Existing empirical tests tend to be in line with the theory (Mackenbach et al. 2015b), Yet, how additional resources of higher socioeconomic groups do protect against prevailing health risk needs further exploration (Mackenbach et al. 2017). The theory is, however, a useful reminder that in the end the existence of health inequalities depends on the existence of socioeconomic inequalities.

Health inequalities and the welfare state

The persistence of socioeconomic inequalities in health is one of the great disappointments of public health. All countries, including those ranking high on indices of economic prosperity and human development, have systematic inequalities in mortality and morbidity between citizens with a higher and a lower socioeconomic position, as indicated by education, occupation, income, or wealth. This also applies to the highly developed 'welfare states' of Western Europe. All Western European countries have created extensive arrangements aiming to reduce socioeconomic inequality and its various consequences. With notable variations all these 'welfare regimes' include measures to redistribute income (e.g. by progressive taxation and social security) and a range of collectively financed provisions (e.g. public housing, education, healthcare, access to culture and leisure facilities) (Esping-Andersen 1990).

There is good evidence that welfare policies have contributed to a reduction of inequalities in income, housing quality, healthcare access, and other social and economic outcomes (Kautto et al. 2001), but they have apparently been insufficient to eliminate health inequalities. The explanation of this paradox has puzzled observers (Bambra 2011), and several hypotheses have emerged. Three circumstances may help to explain the persistence of health inequalities despite attenuation of inequalities in material conditions by the welfare state: (1) inequalities in access to material and immaterial resources have not been eliminated by the welfare state, and are still substantial; (2) due to greater intergenerational mobility, the composition of lower socioeconomic groups has become more homogeneous with regard to personal characteristics associated with ill health; and (3) due to a change in epidemiological regime, in which consumption behaviour became the most important determinant of ill health, the marginal benefits of the immaterial resources to which a higher social position gives access have increased (Mackenbach 2012).

The options: how to build a strategy to reduce inequalities in health?

With all these advances in our understanding of health inequalities, we are now in a much better position than, say, in 1980 when the Black Report appeared (Department of Health and Social Services 1980), to develop rational strategies to tackle them. Bringing together evidence from around the world, the World Health Organization has called upon all countries to 'close the gap in a generation' (Commission on Social Determinants of Health 2008).

Normative assessment of health inequalities

Reducing inequalities in health, by levelling up the health status of those in lower socioeconomic positions, has thus become an important policy goal in many high-income countries. One reason is technical: potential health gains are larger in lower socioeconomic groups, and raising their health status may therefore be a very effective way to improve average population health. Another reason is normative: health inequalities are widely perceived to be 'unjust'. One particularly striking illustration of this perception can be found in the report of the WHO Commission on Social Determinants of Health, which stated that '[p]utting right these inequities is a matter of social justice. Reducing health inequities is an ethical imperative. Social injustice is killing people on a grand scale' (Commission on Social Determinants of Health 2008).

It is important, however, to outline the reasoning behind this move from 'inequalities' (a term with ethically neutral connotations, like 'disparities' or 'variations') to 'inequities' (a term which implies that these variations are 'unjust'). This was first done by Whitehead in a simple but effective scheme consisting of two criteria: 'avoidability' (are health inequalities caused by changeable social

conditions, and therefore unnecessary?) and ‘unfairness’ (do health inequalities result from conditions out of people’s direct control, not from free choice?) (Whitehead 1992).

According to these criteria, health inequalities caused by exposure to health hazards in the environment, or by restricted access to healthcare, would be considered unjust, as would health inequalities caused by health-damaging behaviour restricted by socioeconomic factors. On the other hand, health inequalities caused by natural biological variation (e.g. genetic factors) or freely chosen health-damaging behaviour would not (Whitehead 1992). Application of this scheme thus requires empirical evidence on the determinants of health inequalities, and on the role of free choice versus ‘conditions out of people’s direct control’—evidence that is actually hard to obtain, although it is hard to imagine that systematic differences in behaviour between socioeconomic groups are based on truly free choice.

Attempts to further formalize the normative assessment of health inequalities have tried to use Rawls’s influential ‘theory of justice’ (Daniels et al. 1999) or Sen’s capability theory (Venkatapuram 2011). The latter theory has become especially popular among social epidemiologists because, in contrast to other theories of justice, it gives a central place to health. It argues that all human beings are entitled to a certain set of ‘capabilities’, such as the ability to live a life of normal length, to have good health, to have emotional attachments to others, and to participate effectively in political decisions. From this perspective, tackling health inequalities should be a central objective of all governments pursuing justice (Venkatapuram 2011).

Policy development vis-à-vis health inequalities

In reality, different countries are in widely different phases of awareness of, and willingness to take action on, socioeconomic inequalities in health. Common milestones in policy development are: high-profile independent reports recommending research or policy on health inequalities; national research programmes on health inequalities; government advisory committees recommending policies to reduce health inequalities; and coordinated government action to reduce health inequalities (Mackenbach and Bakker 2003).

Whitehead has proposed a schematic ‘action spectrum’ to characterize the stage of diffusion of ideas on socioeconomic inequalities in health (Fig. 2.4.3). Starting with a primordial stage in which socioeconomic inequalities in health are not even measured, the spectrum covers the stages of ‘measurement’, ‘recognition’, ‘awareness’, ‘denial/indifference’, ‘concern’, ‘will to take action’, ‘isolated initiatives’, ‘more structured developments’, and ‘comprehensive coordinated policy’ (Whitehead 1998).

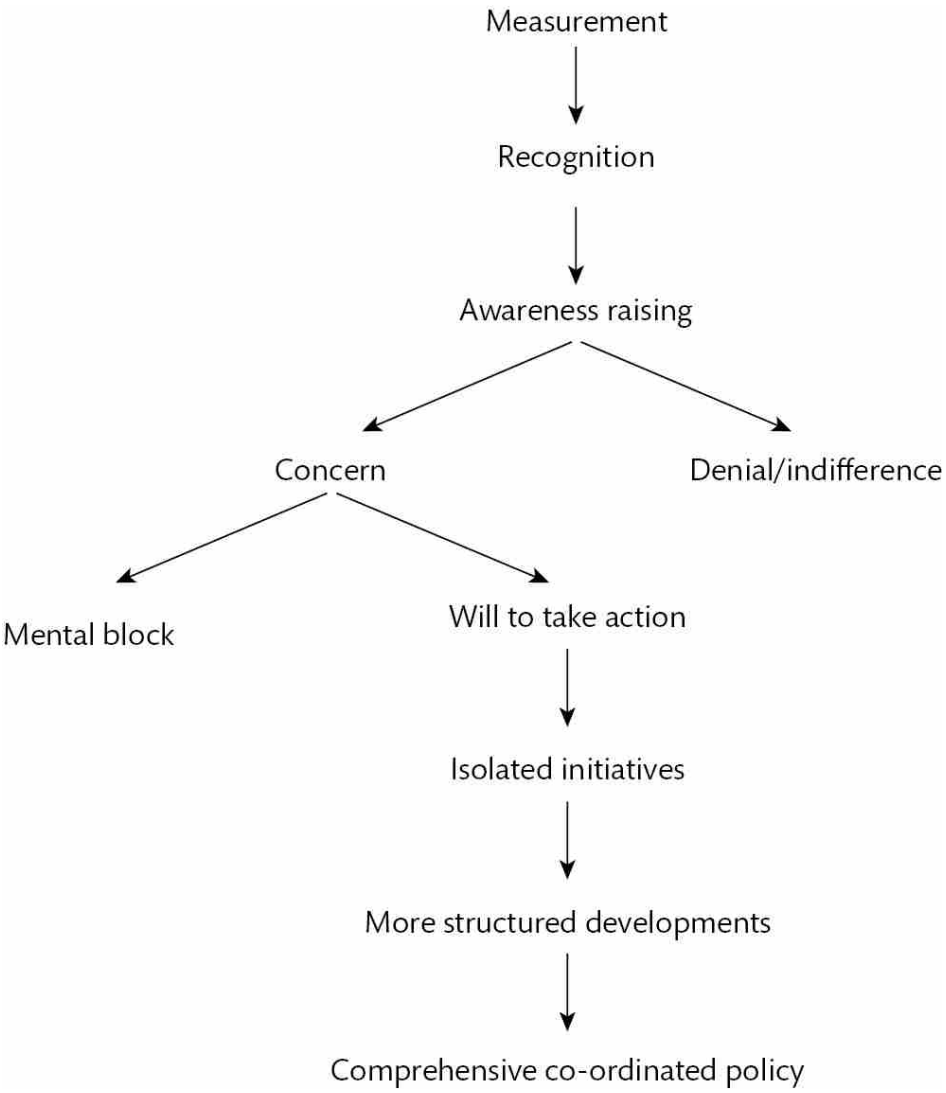


Fig. 2.4.3 Whitehead's action spectrum.
Reproduced from Whitehead, M., Diffusion of ideas on social inequalities in health: a European perspective, *Milbank Quarterly*, Volume 76, Issue 3, pp. 469–92, Copyright © 1998 Milbank Memorial Fund, with permission from John Wiley & Sons, Inc.

Some countries entered a 'more structured developments' stage, with national research programmes as well as high-level advisory committees that have issued comprehensive policy advice on how to reduce socioeconomic inequalities in health, and fewer have started a 'comprehensive coordinated policy' (Whitehead 1998; Mackenbach and Bakker 2003). A rational approach to tackling health inequalities should be based on logical reasoning: dependent on the causes of the problem, and on what is known about possible ways to address these causes, interventions and policies should be developed and tested for effectiveness, and then implemented systematically on the scale required to have population-wide impacts (Whitehead 2007). A possible typology of actions to tackle health inequalities is: (1) strengthening individuals (e.g. improving individuals' abilities to make healthy lifestyle choices); (2) strengthening communities (e.g. community development initiatives to create healthier conditions in neighbourhoods); (3) improving living and working conditions (e.g. improve access to adequate housing and eliminate work-related health risks); and (4) promoting healthy macropolicies (e.g. reduce income inequalities) (Whitehead 2007).

Within high-income countries, there is considerable diversity in the way scientific evidence is being used to underpin policies to reduce health inequalities. Often, most of the evidence used relates to the contribution of specific factors to the explanation of health inequalities, not to the effectiveness of policies and interventions tackling them (Macintyre et al. 2001). It is only rarely that strategies are developed primarily on the basis of evaluation studies of various intervention options (Mackenbach and Stronks 2004). As we will see in the next section the evidence base is still very limited, and opinions differ on what type of evidence is needed to underpin policies and interventions in this field.

There are those who argue that in view of the urgency of starting to tackle health inequalities ('doing nothing is not an option') (Petticrew et al. 2004), one should be prepared to start intervening on the basis of plausibility. Political 'windows of opportunity' are usually short, for example, 4 years at most, and they may be closed before careful evaluation studies have been conducted (Whitehead et al. 2004). A parallel has been drawn with nineteenth-century public health interventions for which controlled intervention studies have never been done, but which were implemented on the basis of plausibility and have proven to be highly successful (Davey Smith et al. 2001). Under the pressure of politicians wanting to see rapid results, the best that can be achieved in terms of scientific evaluation may then be large-scale implementation accompanied by a 'real-time' evaluation study of the intervention, concurrent with its implementation, using some quasi-experimental design (before–after study, interrupted time-series study, and so on) (Macintyre 2003).

On the other hand, there are those who argue that this is a strategy with serious risks. Like in other areas of social and health policy, the actual results of policies and interventions to reduce health inequalities could easily be counterintuitive. There are many historical examples of 'plausible' interventions and policies that did not work, or actually had adverse effects (Macintyre et al. 2001). In addition to that, one could argue that any investment in reducing health inequalities should be justified on the basis of a comparison of its cost-effectiveness with that of other possible investments in health and well-being, and that producing credible evidence is therefore essential (Oliver 2010).

Another issue for debate is what types of evidence are best suited to underpin policy decisions in this area. Clearly, randomized controlled trials will not always be feasible, particularly for the evaluation of policies and interventions that are applied on a population-wide scale. Sometimes, community intervention trials, in which groups of people (school classes, neighbourhoods, etc.) instead of individuals are allocated to the intervention and control condition, will then be a good alternative. But in many circumstances one will have to rely on quasi-experimental or even observational designs to inform policymakers on the effectiveness of new approaches. Controlled before–after studies or interrupted time-series designs could then be used, or observational studies of 'natural experiments', for example, by making comparisons between countries (Thomson et al. 2004). Useful guidance has been developed to support further research in this area (Craig et al. 2011), including the application of techniques applied in economics and econometrics (Hu et al. 2017). Also, a Cochrane and Campbell Equity Methods group has been set up to ensure that research about differential effectiveness of health interventions is available worldwide (http://equity.cochrane.org/).

A complicating factor in evaluating the effectiveness of policies and interventions to reduce health inequalities is that this effectiveness should be measured in terms of favourably changing the distribution of health problems in the population, not of reducing the rate of health problems in a particular group. A 'full' study design therefore requires the measurement, in one or more experimental populations and one or more control populations, of changes over time in the magnitude of health inequalities. Any other design, such as an experimental study of changes over time in the rate of health problems in lower socioeconomic groups only, requires rather strong assumptions to be made, in this case on the absence of health effects in higher socioeconomic groups (Mackenbach and Gunning-Schepers 1997).

Approaches to reduce health inequalities

A number of approaches to reduce health inequalities have been developed, for which there is at least some empirical evidence suggesting that they can help to reduce health inequalities.

In general, there seems to be an 'inverse evidence law', in the sense that more is known about small-scale interventions with potentially little population-level impact, and less is known about interventions and policies with a potentially large impact (Nutbeam 2003). A few promising approaches using different entry-points, are discussed here.

Governmental (social) policies at the national and local level affect education, income, and related material living conditions. If such policies effectively improve the conditions for those in lower socioeconomic groups specifically, and (as a result) lead to improvements in health, health inequalities will be reduced. For example, the United Kingdom introduced a national minimum wage in 1999, resulting in an increase of the lowest incomes. A comparison of those exposed to this increase with a group who earned just more than the minimum wage before and after the introduction or those who experienced no increase, revealed a significant improvement in mental health, but not in health behaviour, in the 'intervention' group (Reeves et al. 2017). Despite the potential to tackle health inequalities via social interventions covering housing, work, transport, unemployment, food, and education, a differential impact on socioeconomic groups is only occasionally studied. To the extent it has been done, work interventions showed some evidence for an effect on reducing inequalities (Bambra et al. 2010).

The important contribution of health behaviour to health inequalities makes health promotion among lower socioeconomic groups a potentially effective approach. One way to classify interventions for this purpose is to assess the degree of choice for individuals, resulting in 'structural' interventions (in which individual choice is removed), 'agento-structural' interventions and 'agentic' interventions (in which individuals are required to make independent choices) (Backholer et al. 2014). Of these, structural interventions seem to be best able to reduce inequalities in health behaviour. Fiscal policies are a good example of a structural intervention with a potentially larger impact on lower than higher socioeconomic groups. Indeed, the effects of raising taxes on tobacco is effective in reducing smoking among lower socioeconomic groups particularly (Main et al. 2008). Recent evidence suggests that a tax on sugar-sweetened beverages is either equally effective across socioeconomic groups or more effective among lower socioeconomic groups (Backholer et al. 2016). Such policies, however, are sometimes criticized for their regressive nature, meaning that the same tax takes a proportionally larger part of lower than higher incomes. One potential way to overcome this criticism is to use these tax revenues for public health investments for the lowest socioeconomic groups. Instead of taxing unhealthy products, there are also indications that lowering costs for healthy products contributes to a reduction of inequalities in health behaviour. A Norwegian experiment showed a large change in the uptake of fruits and a larger decline in the consumption of unhealthy snacks after the introduction of free fruit at schools (Øverby et al. 2012).

Agento-structural interventions target both environmental and individual conditions. They can be implemented in neighbourhoods, but also in schools or workplaces. While the evidence of a reduction of inequalities in health behaviour is not overwhelming, this type of project does have more potential than single interventions. Indeed, an equity specific re-analysis of obesity-related lifestyle interventions showed that intensive and long-lasting community-based projects are potentially able to reduce inequalities in health behaviour (Magnée et al. 2013).

Agentic interventions, entirely relying on individual choice, run the risk of widening instead of closing the gap between socioeconomic groups. This is perhaps, because the cognitive resources needed to fully benefit from these interventions are more prevalent among the higher educated (Magnée et al. 2013).

Comprehensive strategies: evidence from the English strategy

The Black Report, commissioned by a Labour government to investigate the causes and possible remedies for Britain's undiminished social class inequalities in mortality, had proposed a radical change in social and healthcare policies but disappeared into a desk drawer after the Conservatives won the 1979 elections (Department of Health and Social Services 1980). However, research continued and when Labour came back into power almost 20 years later, it again commissioned an expert report to bring together all the evidence—the Independent Inquiry into Inequalities in Health (Department of Health 1998)—and developed a comprehensive programme to tackle health inequalities.

This programme consisted of a range of new government policies including the introduction of a national minimum wage, higher benefits, and pensions, and substantially increased spending on education, housing, urban regeneration, and healthcare. It also announced a number of specific initiatives including the 'Sure Start' programme (free child care, early education, and parent support for low-income families), 'Health Action Zones' (local strategies to improve health in deprived areas), and a series of antitobacco policies (including free nicotine replacement therapy for low-income smokers) (Department of Health 1999). In a further evolution of the programme two quantitative targets (to narrow the socioeconomic gap in life expectancy and that in infant mortality by 10% in 2010) were added, and priority was given to key interventions expected to contribute to closing the life expectancy gap, such as reducing smoking in manual social groups, managing risks for coronary heart disease and cancer (poor diet and obesity, physical inactivity, hypertension), and reducing accidents at home and on the road (Department of Health 2003). The total budget exceeded £20 billion (Table 2.4.2).

The strategy came to an end with the 2010 parliamentary elections and the formation of a Conservative/Liberal Democrat coalition government. An extensive review of the achievements of the programme showed that it had been a mixed success: despite positive effects in some areas (e.g. reduction of child poverty, improvements in housing, better uptake of influenza vaccination) most indicators showed a widening instead of a narrowing of health inequalities, and the programme missed its targets (Department of Health 2010). Further analyses showed that the failure to achieve the targets was due to the fact that the strategy had not sufficiently focused on the most relevant entry-points, that effective policies were unavailable at that point in time, and that the scale of implementation had not been sufficient for achieving population-wide impacts (Mackenbach 2011).

Table 2.4.2 The English programme to tackle health inequalities: commitments and their achievement

'Departmental commitments'	Examples	Budget (2004–2007)[a]	Direct relevance for life expectancy target	Direct relevance for infant-mortality target	Commitments mostly met (2007)[b]
Maternal and child health	Sure Start, chikcare, Welfare Food Scheme	>£2 billion	No	Yes	Yes
Improving life chances for children	Mental health, sports facilities, education	>£2 billion	No	No	Yes
Reducing teenage pregnancy	Sex education, care access	Not specified	No	Yes	Yes
Engaging communities	Neighbourhood Renewal, homelessness reduction	>£1.5 billion	No	No	Yes
Prevention	Smoking cessation, 5 A Day, injury prevention	Not specified	Yes	No	Yes
Primary care	Facilities, breast cancer screening, flu immunization	>£1 billion	Yes	No	Yes
Effective treatment	Access to cancer services, access to coronary heart disease services	>£0.5 billion	No	No	Yes
Underlying determinants	Child poverty, fuel poverty, Pathways to Work	>£15.5 billion	No	No	Yes

[a] '>' sign indicates that amounts exclude commitments for which no budget was specified.

[b] Source: *Tackling Health Inequalities: 2007 States Report of the Program for Action*.

Reproduced from *Journal of Epidemiology Community Health*, Mackenbach, J.P., Volume 65, Issue 7, pp. 568–75, Copyright © 2011, with permission from BMJ Publishing Group Ltd.

These findings were confirmed by a recent analysis, showing that trends in health inequalities did not differ before and after the implementation of the strategy. For most health indicators, changes in trends of health inequalities after 2000 in England were also not significantly different from those seen in Finland, the Netherlands and Italy (Hu et al., 2016a).

Source: *Tackling Health Inequalities: A Program for Action*.

Conclusion

Whether it will actually be possible to substantially reduce socioeconomic inequalities in health remains an open question. The good news, however, is that there has been enormous progress in explanatory research, and that this has identified a large number of targets for policies and interventions to tackle health inequalities. There has also been a beginning of research and development for effective interventions and policies to tackle health inequalities. While this is still a modest beginning, it does put us in a better position to reduce socioeconomic inequalities in health in the coming decades. A number of innovative approaches have been developed for which there is at least some evidence of effectiveness. Comprehensive packages have been developed in several countries that have a sound theoretical basis and clear inspirational value.

Developing effective strategies to reduce health inequalities is a daunting task. No single country has the capacity to contribute more than a fraction of the necessary knowledge. This is a matter not only of restricted manpower or financial resources for research, but also of restricted opportunities for implementing and evaluating policies and interventions. Some policies can be implemented and

evaluated in some countries and not in others, either because they have already been implemented or because they are politically infeasible. International exchange is therefore necessary to increase learning speed.

REFERENCES

Aarts, M.J., et al. (2012). Small but significant socioeconomic inequalities in axillary staging and treatment of breast cancer in the Netherlands. *British Journal of Cancer*, **107**, 12–17.

Aarts, M.J., Van Der Aa, M.A., Coebergh, J.W., and Louwman, W.J. (2010). Reduction of socioeconomic inequality in cancer incidence in the South of the Netherlands during 1996–2008. *European Journal of Cancer*, **46**, 2633–46.

Backholer, K., et al. (2014). A framework for evaluating the impact of obesity prevention strategies on socioeconomic inequalities in weight. *American Journal of Public Health*, **104**, e43–e50.

Backholer, K., et al. (2016). The impact of a tax on sugar-sweetened beverages according to socio-economic position: a systematic review of the evidence. *Public Health Nutrition*, **19**, 3070–84.

Bambra, C. (2011). Health inequalities and welfare state regimes: theoretical insights on a public health 'puzzle'. *Journal of Epidemiology and Community Health*, **65**, 740–5.

Bambra, C., Gibson, M., Sowden, A., Wright, K., Whitehead, M., and Petticrew, M. (2010). Tackling the wider social determinants of health and health inequalities: evidence from systematic reviews. *Journal of Epidemiology and Community Health*, **64**, 284–91.

Barker, D.J.P. and Robinson, R.J. (eds.) (1992). *Fetal and Infant Origins of Adult Disease.* London: British Medical Journal.

Bartley, M. and Plewis, I. (1997). Does health-selective mobility account for socioeconomic differences in health? Evidence from England and Wales, 1971 to 1991. *Journal of Health and Social Behaviour*, **38**, 376–86.

Beenackers, M.A., et al. (2012). Socioeconomic inequalities in occupational, leisure-time, and transport related physical activity among European adults: a systematic review. *International Journal of Behavioural Nutrition and Physical Activity*, 19(9), 116.

Beenackers, M., Oude Groeniger, J., Van Lenthe, F.J., and Kamphuis, C.B.M. (2018). The role of financial strain and self-control in explaining socioeconomic inequalities in health behaviours: the GLOBE study. *European Journal of Public Health*, 28(4), 597–603.

Bijwaard, G., Van Kippersluis, H., and Veenman, J. (2015). Education and health: the role of cognitive ability. *Journal Health Economics*, **42**, 29–43.

Blumenshine, P., Egerter, S., Barclay, C.J., Cubbin, C., and Braveman, P.A. (2010). Socioeconomic disparities in adverse birth outcomes: a systematic review. *American Journal of Preventive Medicine*, **39**, 263–72.

Bosdriesz, J.R., Willemsen, M.C., Stronks, K., and Kunst, A.E. (2015). Socioeconomic inequalities in smoking cessation in 11 European countries from 1987 to 2012. *Journal of Epidemiology and Community Health*, **69**, 886–92.

Bosma, H., Schrijvers, C., and Mackenbach, J.P. (1999). Socioeconomic inequalities in mortality and importance of perceived control: cohort study. *BMJ*, **319**, 1469–70.

Bourdieu, P. (1984). *Distinction: A Social Critique of the Judgement of Taste*. Cambridge, MA: Harvard University Press.

Brunner EJ. (2017). Social factors and cardiovascular morbidity. *Neurosciences & Biobehavioural Review*, **74**, 260–8.

Case, A. and Deaton, A. (2015). Rising morbidity and mortality in midlife among white non-Hispanic Americans in the 21st century. *Proceedings of the National Academy of Science USA*, **112**, 15078–83.

Cesarini, D., Lindqvist, E., Ostling, R., and Wallace, B. (2016). Wealth, health, and child development: evidence from administrative data on Swedish lottery players. *The Quarterly Journal of Economics*, **131**, 687–738.

Christensen, A.V., et al. (2016). Educational inequality in cardiovascular disease depends on diagnosis: a nation-wide register based study from Denmark. *European Journal of Preventive Cardiology*, **23**, 820–33.

Coenen, P., et al. (2018). Do highly physically active workers die early? A systematic review with meta-analysis of data from 193 696 participants. *British Journal of Sports Medicine*, **52**, 1320–6.

Commission on Social Determinants of Health (2008). *Closing the Gap in a Generation. Health Equity Through the Social Determinants of Health*. Geneva: World Health Organization.

Conklin, A.I., et al. (2014). Social relationships and healthful dietary behaviour: evidence from over-50s in the EPIC cohort, UK. *Social Science & Medicine*, **100**, 167–75.

Costa, G. and D'Errico, A. (2006). Inequalities in health: do occupational risks matter? *European Journal of Public Health*, **16**, 340–1.

Craig, P., et al. (2011). *Using Natural Experiments to Evaluate Population Health Interventions: Guidance for Producers and Users of Evidence*. Glasgow: Medical Research Council.

Cummins, S. and Macintyre, S. (2006). Food environments and obesity—neighbourhood or nation? *International Journal of Epidemiology*, **35**, 100–4.

Dalstra, J.A., et al. (2005). Socioeconomic differences in the prevalence of common chronic diseases: an overview of eight European countries. *International Journal of Epidemiology*, **34**, 316–26.

Daniels, N., Kennedy, B.P., and Kawachi, I. (1999). Why justice is good for our health: the social determinants of health inequalities. *Daedalus*, **128**, 215–51.

Davey Smith, G., Ebrahim, S., and Frankel, S. (2001). How policy informs the evidence. *BMJ*, **322**, 184–5.

Department of Health (1998). *Independent Inquiry into Inequalities in Health (Acheson Report)*. London: Department of Health.

Department of Health (1999). *Reducing Health Inequalities: An Action Report*. London: Department of Health.

Department of Health (2003). *Tackling Health Inequalities: A Program for Action*. London: Department of Health.

Department of Health (2010). *Fair Society, Healthy Lives (Marmot Review)*. London: Department of Health.

Department of Health and Social Services (1980). *Inequalities in Health: Report of a Research Working Group (Black Report)*. London: Department of Health and Social Services.
Dunn, J.R. and Hayes, M.V. (1999). Identifying social pathways for health inequalities. The role of housing. *Annals of the New York Academy of Sciences*, **896**, 399–402.
Esping-Andersen, G. (1990). *The Three Worlds of Welfare Capitalism*. Cambridge: Polity.
Fiorito, G., et al. (2017). Social adversity and epigenetic aging: a multi-cohort study on socioeconomic differences in peripheral blood DNA methylation. *Science Reports*, 7(1), 16266.
Gadeyne, S., Menvielle, G., and Kulhanova, I. (2017). The turn of the gradient? Educational differences in breast cancer mortality in 18 European populations during the 2000s. *International Journal of Cancer*, **141**, 33–44.
García-Gómez, P., et al. (2013). Long term spill over effects of health shocks on employment and income. *Journal of Human Resources*, **48**, 873–909.
Giskes, K., Avendano, M., Brug, J., and Kunst, A.E. (2010). A systematic review of studies on socioeconomic inequalities in dietary intakes associated with weight gain and overweight/obesity conducted among European adults. *Obesity Reviews*, **11**, 413–29.
Gonzalez, M.A., Rodriguez Artalejo, F., and Calero, J.R. (1998). Relationship between socioeconomic status and ischaemic heart disease in cohort and case-control studies: 1960–1993. *International Journal of Epidemiology*, **27**, 350–8.
Hoffmann, K., et al., (2017). Trends in educational inequalities in obesity in 15 European countries between 1990 and 2010. *International Journal of Behavioural Nutrition and Physical Activity*, **14**, 63.
Holmes, M.V., et al. (2014). Association between alcohol and cardiovascular disease: Mendelian randomisation analysis based on individual participant data. *British Medical Journal*, 3(49), g4164.
Hu, Y., et al. (2016a). Trends in socioeconomic inequalities in self-assessed health in 17 European countries between 1990 and 2010. *Journal Epidemiology Community Health*, 70: 644–52.
Hu, Y., et al (2016b). Did the English strategy reduce inequalities in self-reported health, obesity and smoking? A 'quadruple-difference' analysis comparing England with three other European countries. *BMC Public Health*, **16**, 865.
Hu, Y., Van Lenthe, F.J., Hoffmann, R., Van Hedel, K., and Mackenbach, J.P. (2017). Assessing the impact of natural policy experiments on socioeconomic inequalities in health: how to apply commonly used quantitative analytical methods. *BMC Medical Research Methodology*, **17**, 68.
Huisman, M., Kunst, A.E., and Mackenbach, J.P. (2005). Educational inequalities in smoking among men and women aged 16 years and older in 11 European countries. *Tobacco Control*, **14**, 106–13.
Kamphuis, C.B., Van Lenthe, F.J., Giskes, K., Huisman, M., Brug, J., and Mackenbach, J.P. (2009). Socioeconomic differences in lack of recreational walking among older adults: the role of neighbourhood and individual factors. *International Journal of Behavioral Nutrition and Physical Activity*, **6**, 1.
Karnehed, N.E., Rasmussen, F., Hemmingsson, T., and Tynelius, P. (2008). Obesity in young adulthood is related to social mobility among Swedish men. *Obesity (Silver Spring)*, **16**, 654–8.
Kautto, M., Fritzell, J., Hvinden, B., Kvist, J., and Uusitalo, H. (eds.) (2001). *Nordic Welfare States in the European Context*. New York: Routledge.
Kippersluis, H.V., O'Donnell, O., and Doorslaer, E.V. (2011). Long run returns to education: does schooling lead to an extended old age? *Journal of Human Resources*, **46**, 695–721.
Klein, J. and Von dem Knesebeck, O. (2015). Socioeconomic inequalities in prostate cancer survival: a review of evidence and explanatory variables. *Social Science and Medicine*, **142**, 9–18.
Kristenson, M. (2006). Socio-economic position and health: the role of coping. In: Siegrist, J. and Marmot, M. (eds.) *Social Inequalities in Health. New Evidence and Policy Implications*, pp. 127–52. Oxford: Oxford University Press.
Kruize, H., Driessen, P.P., Glasbergen, P., and Van Egmond, K.N. (2007). Environmental equity and the role of public policy: experiences in the Rijnmond region. *Environmental Management*, **40**, 578–95.
Kulhanova, I., et al. (2016). The role of three lifestyle risk factors in reducing educational differences in ischaemic heart disease mortality in Europe. *European Journal of Public Health*, **27**, 203–10.
Lager, A.C. and Torssander, J. (2012). Causal effect of education on mortality in a quasi-experiment on 1.2 million Swedes. *Proceedings of the National Academy of Sciences of the United States of America*, **109**, 8461–6.
Leinsalu, M., et al. (2009). Educational inequalities in mortality in four Eastern European countries: divergence in trends during the post-communist transition from 1990 to 2000. *International Journal of Epidemiology*, **38**, 512–25.
Lleras-Muney, A. (2002). *The Relationship Between Education and Adult Mortality in the United States*. NBER Working Paper Series. Cambridge: National Bureau of Economic Research.
Lopez-Azpiazu, I., et al. (2003). Disparities in food habits in Europe: systematic review of educational and occupational differences in the intake of fat. *Journal of Human Nutrition and Dietetics*, **16**, 349–64.
Lorant, V., Croux, C., Weich, S., Deliege, D., Mackenbach, J., and Ansseau, M. (2007). Depression and socio-economic risk factors: 7-year longitudinal population study. *British Journal of Psychiatry*, **190**, 293–98.
Lorant, V., Kunst, A.E., Huisman, M., Costa, G., and Mackenbach, J. (2005). Socio-economic inequalities in suicide: a European comparative study. *British Journal of Psychiatry*, **187**, 49–54.
Lounsbury, J.W., Sundstrom, E., Loveland, J.M., and Gibson, L.W. (2003). Intelligence, 'big five' personality traits, and work drive as predictors of course grade. *Personality and Individual Differences*, **35**, 1231–9.
Lundqvist, A., Andersson, E., Ahlberg, I., Nilbert, M., and Gerdtham, U. (2016). Socioeconomic inequalities in breast cancer incidence and mortality in Europe-a systematic review and meta-analysis. *European Journal of Public Health*, **26**, 804–13.
Macintyre, S. (2007). Deprivation amplification revisited; or, is it always true that poorer places have poorer access to resources for healthy diets and physical activity? *International Journal of Behavioural Nutrition and Physical Activity*, *4*, 32.
Macintyre, S. (2003). Evidence based policy making. *British Medical Journal*, **326**, 5–6.
Macintyre, S., Chalmers, I., Horton, R., and Smith, R. (2001). Using evidence to inform health policy: case study. *British Medical Journal*, **322**, 222–5.

Mackenbach, J.P. (2005). Genetics and health inequalities: hypotheses and controversies. *Journal of Epidemiology and Community Health*, **59**, 268–73.
Mackenbach, J.P. (2011). Can we reduce health inequalities? An analysis of the English strategy (1997–2010). *Journal of Epidemiology and Community Health*, **65**, 568–75.
Mackenbach, J.P. (2012). The persistence of health inequalities in modern welfare states: the explanation of a paradox. *Social Science & Medicine*, **75**, 761–9.
Mackenbach, J.P., et al. (2018). Trends in health inequalities in 27 European countries. *Proceedings National Academy Sciences U S A*, 115(25), 6440–5.
Mackenbach, J.P. and Bakker, M.J. (2003). Tackling socioeconomic inequalities in health: analysis of European experiences. *Lancet*, **362**, 1409–14.
Mackenbach, J.P. and Gunning-Schepers, L.J. (1997). How should interventions to reduce inequalities in health be evaluated? *Journal of Epidemiology and Community Health*, **51**, 359–64.
Mackenbach, J.P. and Kunst, A.E. (1997). Measuring the magnitude of socio-economic inequalities in health: an overview of available measures illustrated with two examples from Europe. *Social Science & Medicine*, **44**, 757–71.
Mackenbach, J.P., et al. (2008). Socioeconomic inequalities in health in 22 European countries. *New England Journal of Medicine*, **358**, 2468–81.
Mackenbach, J.P. and Stronks, K. (2004). The development of a strategy for tackling health inequalities in the Netherlands. *International Journal for Equity in Health*, **3**, 11.
Mackenbach, J.P., et al. (2016). Changes in mortality inequalities over two decades: register based study of European countries. *British Medical Journal*, 11(353), i1732.
Mackenbach, J.P., et al. (2015a). Inequalities in alcohol-related mortality in 17 European countries: a retrospective analysis of mortality registers. *PLoS Medicine*, **1**, 12.
Mackenbach, J.P., et al. (2015b). 'Fundamental causes' of inequalities in mortality: an empirical test of the theory in 20 European populations. *Social Science and Medicine*, **127**, 51–62.
Mackenbach, J.P., et al. (2017). Variations in the relation between education and cause-specific mortality in 19 European populations: a test of the 'fundamental causes' theory of social inequalities in health. *Sociology of Health and Illness*, **39**, 1117–33.
Main, C., et al. (2008). Population tobacco control interventions and their effects on social inequalities in smoking: placing an equity lens on existing systematic reviews. *BMC Public Health*, **8**, 178.
Majer, I.M., Nusselder, W.J., Mackenbach, J.P., and Kunst, A.E. (2011). Socioeconomic inequalities in life and health expectancies around official retirement age in 10 Western-European countries. *Journal of Epidemiology and Community Health*, **65**, 972–9.
Mäki, N.E., et al. (2014). The potential for reducing differences in life expectancy between educational groups in five European countries: the effects of obesity, physical inactivity and smoking. *Journal of Epidemiology*, **68**, 635–40.
Magnée, T., Burdorf, A., Brug. J., et al. (2013). Synthesis of equity-specific subgroup analyses of 26 Dutch intervention studies for obesity prevention. *American Journal Preventive Medicine*, **2013**, 44, e57–e66.
Manrique-Garcia, E., Sidorchuk, A,. Hallqvist, J., and Moradi T. (2011). Socioeconomic position and incidence of acute myocardial infarction: a meta-analysis. *Journal of Epidemiology and Community Health*, **65**, 301–9.
Marmot, M., Siegrist, J., and Theorell, T. (2006). Health and the psychosocial environment at work. In: Marmot, M. and Wilkinson, R.G. (eds.) *Social Determinants of Health* (2nd ed.). Oxford: Oxford University Press, pp. 97–130.
Marmot, M. and Wilkinson, R.G. (eds.) (2006). *Social Determinants of Health*. Oxford: Oxford University Press.
Mishra, G.D., Chiesa, F., Goodman, A., De Stavola, B., and Koupil, I. (2013). Socio-economic position over the life course and all-cause, and circulatory diseases mortality at age 50–87 years: results from a Swedish birth cohort. *European Journal of Epidemiology*, **28**, 139–47.
Morris, J.N., Wilkinson, P., Dangour, A.D., Deeming, C., and Fletcher, A. (2007). Defining a minimum income for healthy living (MIHL): older age, England. *International Journal of Epidemiology*, **36**, 1300–7.
Mullainathan, M.A., et al. (2013). Poverty impedes cognitive function. *Science*, **341**, 976–80.
Nguyen, T.T., et al. (2016). Instrumental variable approaches to identifying the causal effect of educational attainment on dementia risk. *Annals of Epidemiology*, **26**, 71–6.
Nutbeam, D. (2003). How does evidence influence public health policy? Tackling health inequalities in England. *Health Promotion Journal Australia*, **14**, 154–8.
OECD (2017). *Understanding the Socioeconomic Divide in Europe*. [Online] Available at: https://www.oecd.org/els/soc/understanding-the-socio-economic-divide-in-europe-26-january.htm
Olshansky, S.J., et al. (2012). Differences in life expectancy due to race and educational differences are widening, and many may not catch up. *Health Affairs (Millwood)*, **31**, 1803–13.
Oliver, A. (2010). Reflections on the development of health inequalities policy in England. *Health Care Analysis*, **18**, 402–20.
Oude Groeniger, J., et al. (2017). Changing contributions of material and behavioural factors to socioeconomic inequalities in mortality: population based cohort study. *Journal of Clinical Epidemiology*, **91**, 137–45.
Oude Groeniger, J. and Van Lenthe, F.J. (2016). The contribution of time-varying measures of health behaviors to socioeconomic inequalities in mortality: how to understand the underlying mechanisms? *Journal Epidemiology Community Health*, **70**, 1045–8.
Øverby, N.C., Klepp, K.I., and Bere, E. (2012). Introduction of a school fruit program is associated with reduced frequency of consumption of unhealthy snacks. *American Journal of Clinical Nutrition*, **96**, 1100–3.
Petticrew, M., Whitehead, M., Macintyre, S.J., Graham, H., and Egan, M. (2004). Evidence for public health policy on inequalities: 1: the reality according to policymakers. *Journal of Epidemiology and Community Health*, **58**, 811–16.
Phelan, J.C., Link, B.G., and Tehranifar, P. (2010). Social conditions as fundamental causes of health inequalities: theory, evidence, and policy implications. *Journal of Health and Social Behaviour*, 51(Suppl.), S28–40.
Plug, I., et al. (2012). Socioeconomic inequalities in mortality from conditions amenable to medical interventions: do they reflect inequalities in access or quality of health care? *BMC Public Health*, **12**, 346.

Pollitt, R.A., Rose, K.M., and Kaufman, J.S. (2005). Evaluating the evidence for models of life course socioeconomic factors and cardiovascular outcomes: a systematic review. *BMC Public Health*, 20;5:7.
Probst, C., Roerecke, M., Behrendt, S., and Rehm J. (2014). Socioeconomic differences in alcohol-attributable mortality compared with all-cause mortality: a systematic review and meta-analysis. *International Journal of Epidemiology*, **43**, 1314–27.
Reeves, A., McKee, M., Mackenbach, J.P, Whitehead, M., and Stuckler, D. (2017). Introduction of a national minimum wage reduced depressive symptoms in low-wage workers: a quasi-natural experiment in the UK. *Health Economics*, **26**, 639–55.
Robertson, T., Benzeval, M., Whitley, E., and Popham F. (2015). The role of material, psychosocial and behavioral factors in mediating the association between socioeconomic position and allostatic load (measured by cardiovascular, metabolic and inflammatory markers). *Brain, Behaviour and Immunity*, **45**, 41–9.
Roskam, A.J., et al. (2010). Comparative appraisal of educational inequalities in overweight and obesity among adults in 19 European countries. *International Journal of Epidemiology*, **39**, 392–404.
Ronksley, P.E., Brien, S.E., Turner, B.J., Mukamal, K.J., and Ghali, W.A. (2011). Association of alcohol consumption with selected cardiovascular disease outcomes: a systematic review and meta-analysis. *British Medical Journal*, **342**, d671.
Siegrist, J. and Marmot, M. (2004). Health inequalities and the psychosocial environment—two scientific challenges. *Social Science & Medicine*, **58**, 1463–73.
Stansfeld, S.A. (2006). Social support and social cohesion. In: Marmot, M. and Wilkinson, R.G. (eds.) *Social Determinants of Health* (2nd ed.), pp. 148–71. Oxford: Oxford University Press.
Stickley, A., et al. (2012). Socioeconomic inequalities in homicide mortality: a population-based comparative study of 12 European countries. *European Journal of Epidemiology*, **27**, 877–84.
Stirbu, I., et al. (2010). Educational inequalities in avoidable mortality in Europe. *Journal of Epidemiology and Community Health*, **64**, 913–20.
Stirbu, I., Looman, C., Nijhof, G.J., Reulings, P.G., and Mackenbach, J.P. (2012). Income inequalities in case death of ischaemic heart disease in the Netherlands: a national record-linked study. *Journal of Epidemiology and Community Health*, **66**, 1159–66.
Stringhini, S., et al. (2017). Socioeconomic status and the 25 × 25 risk factors as determinants of premature mortality: a multicohort study and meta-analysis of 1.7 million men and women. *Lancet*, **389**, 1229–37.
Stringhini, S., et al. (2018). Socioeconomic status, non-communicable disease risk factors, and walking speed in older adults: multi-cohort population based study. *British Medical Journal*, **360**, k1046.
Thomson, H., et al. (2004). Evaluating the health effects of social interventions. *British Medical Journal*, **328**, 282–5.
Van der Heyden, J.H., et al. (2009). Socioeconomic inequalities in lung cancer mortality in 16 European populations. *Lung Cancer*, **63**, 322–30.
Van Doorslaer, E., Masseria, C., and Koolman, X. (2006). Inequalities in access to medical care by income in developed countries. *Canadian Medical Association Journal*, **174**, 177–83.
Van Oort, F.V.A., Van Lenthe, F.J. Mackenbach, J.P. (2005). Material, psychosocial and behavioural factors in the explanation of socioeconomic inequalities in mortality in the Netherlands: which are more important. *Journal of Epidemiology and Community Health*, **59**, 214–20.
Van Rijn, R.M., Robroek, S.J., Brouwer, S., and Burdorf A. (2014). Influence of poor health on exit from paid employment: a systematic review. *Occupational Environmental Medicine*, **71**, 295–301.
Van Raalte, A.A., et al. (2011). More variation in lifespan in lower educated groups: evidence from 10 European countries. *International Journal of Epidemiology*, **40**, 1703–14.
Van Lenthe, F.J., Brug, J., and Mackenbach, J.P. (2005). Neighbourhood inequalities in physical activity: the role of neighbourhood attractiveness, proximity to local facilities and safety. *Social Science and Medicine*, **60**, 763–75.
Venkatapuram, S. (2011). *Health Justice. An Argument from the Capabilities Approach*. Cambridge: Polity Press.
Whitehead, M. (1992). The concepts and principles of equity and health. *International Journal of Health Services*, **22**, 429–45.
Whitehead, M. (1998). Diffusion of ideas on social inequalities in health: a European perspective. *Milbank Quarterly*, **76**, 469–92, 306.
Whitehead, M. (2007). A typology of actions to tackle social inequalities in health. *Journal of Epidemiology and Community Health*, **61**, 473–8.
Whitehead, M., Petticrew, M., Graham, H., MacIntyre, S.J., Bambra, C., and Egan, M. (2004). Evidence for public health policy on inequalities: 2: assembling the evidence jigsaw. *Journal of Epidemiology and Community Health*, **58**, 817–21.
Woods, A.M., et al. (2018). Risk thresholds for alcohol consumption: combined analysis of individual-participant data for 599 912 current drinkers in 83 prospective studies. *Lancet*, **391**, 1513–23.
Yang, D., Dzayee, D.A., Beiki, O., de Faire, U., Alfredsson, L., and Moradi T. (2012). Incidence and case fatality after day 28 of first time myocardial infarction in Sweden 1987–2008. *European Journal of Preventive Cardiology*, **19**, 1304–15.

2.5

Reducing health inequalities in developing countries

Romulo Paes-Sousa, Paulo M. Buss, and Mauricio L. Barreto

Introduction

Human society is divided into a number of characteristics. Many of these divisions are the result of adaptive, biological, or geographical processes, and others are created in the course of historical, social, economic, or cultural processes. Some of these could be merely differences (e.g. men and women), others are purely social or historical as with the social classes. The importance is that inequalities very often are upgraded to inequities as they define relationships that are essentially based on power and the access to, and the possession of, goods, services, and wealth. Consequently, the fruits of the societal efforts accumulated over generations are unequally distributed (General Assembly 2000; UNDESA 2018). The more visible face of inequality is the existence of people with different level of access to the wealth, and this was observed in all four country income groups—high, upper-middle, lower-middle, and low.

However, an important difference is that social inequalities in both low- and lower-middle-income countries are very often associated with a high level of poverty. While different, there is some overlapping in these two concepts and its consequences and both inequality and poverty has been given very high importance in the study of the social determinants of health. While social inequalities and poverty are both in general higher in lower-middle-income countries when compared with high-income countries, the health literature has in westernized, rich countries emphasized the study of inequalities and health, and the study of poverty and health has great emphasis in lower-middle-income countries. Therefore, while there is an accumulated body of evidence that inequality and poverty have isolated effects on health, in lower-middle-income countries less is known on the

effect of inequalities on health and yet less on the joining effect of inequalities and poverty (UN General Assembly 2000a, b, 2015).

Poverty is an absolute term that defines the lack of resources necessary to have a decent way of life. Certainly, the necessary resource to achieve a decent life vary in time and across different societies. Then, despite the effort to have universal definitions and measures of poverty making easier comparisons among nations, it is in fact very much contextual. In the effort for universal measure, the World Bank uses cut-off values corrected by purchasing power parity-PPP and defines extreme poverty as living on less than US$1.90 per day per capita, and moderate poverty as less than $3.10 a day per capita. In another aspect, inequality is a relational concept related with the distribution of the societal wealth among its members. A commonly used measure of inequality in the GINI Index (GI) that vary from 0 (zero) to 1 (one), where 0 is the situation where the wealth is equally distributed among all members (situation that maybe existed in very primitive societies) and 1 is the maximum of concentration where a member accumulated all the wealth. In present times, HICs with more fair societies have GI below or around 0.3, while the low- and middle-income countries (LMICs) are very often above 0.4 or 0.5, but a few arrive at 0.6 or above. A country can be poorer but less unequal, for instance, India has great poverty than Brazil, but it is less unequal. In 2011, the World Bank estimated 0.549 as the GINI Index of Brazil and 0.341 to India 34.1, while the prevalence of extreme poverty (% of the population living below US$ 1.9/day) was 11.1% and 21.1% in Brazil and India, respectively (UN General Assembly 2012).

As you can see, inequalities within a country relate to the distribution of the accumulated wealth (World Bank, 2018). But, the consolidation of that situation and to transform inequalities in inequities, is dependent on how that society is politically organized and the power and social relations established among its various strata. It is defined by the political organization historically moulded in each society, and how the state redistributed the nation wealth through fiscal and transfer systems, generating greater or lesser distributive distances between existing social groups. Cultural elements are also important regarding amplifying and consolidating some existing inequalities. Inequalities consolidated in the social realm are often visible and sometimes enhanced when the object of study is health, meaning the health conditions or the access to the health services and resources reflect the way the wealth is distributed in that

specific society or nation. Health inequalities generate different chances of exposure to the factors that determine health and disease, different chances about illness and death, as well as unequal possibilities about the ability to take advantage of scientific and technological advances for health. In the same way that social inequalities have persisted in all countries, regardless of their level of development, health inequalities persist in the same fashion (Deaton 2013; Marmot 2015). Given their development stage inequalities are associated with high levels of poverty and extreme poverty and other historical and cultural prejudices that create other dimensions of the inequalities such as the ones related to ethnicity or gender.

Among countries poverty and inequality levels, while correlated, are not similar. Among the lower-middle-income countries, higher levels of inequalities are observed in Latin American countries while higher levels of poverty are observed in other continents in particular, in Africa. It gives a complex picture of the lower-middle-income countries poverty and inequality distribution. For instance, in the effort to understand the reasons why LA countries have great levels of inequalities three axes were explored: its position within the global economic system, the internal colonialism that maintains racial categories and the underdevelopment of state structures (Hoffman and Centeno 2003).

In the current international context of economic globalization, discussions have highlighted inequalities within the same country but also inequalities between countries. Inequalities between countries are related to the differences in economic and social development achieved by different countries, which are generated by the position that these countries have occupied during different phases of history within the global productive system. The latter reflect historical aspects, and the international economic and political environment in relation to each country's share of global resources and development opportunities.

More recently, the concept of global inequality has arisen, which involves the combined effects of these two types of inequalities (Milanovic 2016). Global inequality is the result of inequalities both between and within countries, and it is therefore defined by the interaction of the determinants of each. The availability of international data has made it possible to conduct empirical studies regarding the issue of global inequality. For example, when the GINI Index, which is one of the most frequently used methods to measure social inequality within a country, is

calculated globally it results in even higher levels than those found in nations with the highest levels of social inequality. In recent years, the GINI indices of countries with the highest levels of inequality have been around 0.60 (1 = maximum inequality and 0 = total equality), while the global GINI Index is close to 0.70. The global GINI Index captures the extremes of the poorest strata of the poorest countries (low-income countries) and the richest strata of the richest countries (high-income countries), which translates into a higher level of inequality than when measured in each country separately.

Health inequalities

Poverty associated with an unfair wealth distribution in LMICs effects on health are staggering. It generates an estimated 800 million people chronically hungry, makes one in six children underweight, and more than one-third of deaths among children under 5 years to be attributable to malnutrition (Stiglitz 2012; FAO 2015; Johnston 2015). Insufficient access to safe and nutritious food exists despite the fact that global food production is sufficient to cover 120% of global dietary needs. Within each of these lower-middle-income countries social inequalities are observed among regions, inside regions, inside cities etc, or when societal or individual characteristics are considered (social class, ethnical groups, gender, etc.).

Looking overall the level of inequality (as measured by the GINI Index) is strongly correlated with different health indicators being life expectancy at birth, infant mortality, different infectious diseases incidence, non-communicable chronic diseases (NCD) report, or violence. For instance, the global average of the life expectancy at birth in 1990 was 64 years, while in 2013 that number had increased by seven years to 71 years. However, as averages these values conceal a series of inequalities. For example, at country level, in 2013, the average life expectancy at birth ranged from a minimum of 46 years (38 in 1990) in Sierra Leone (a low-income country), to a maximum of 84 (79 in 1990) in Japan (a high-income country). By 2013, life expectancy had increased in both countries and although the gap has narrowed slightly (from 41 to 38 years) they are still at unacceptable levels (WHO 2015).

Children are especially sensitive to social and environmental adversities. Despite advances in recent decades, it is estimated that 6.3 million children under the age of five died in 2013, mostly due to preventable causes in poor or developing countries. Children in low-income countries from sub-Saharan Africa are 15 times more likely to die before their fifth birthday than children in high-income countries. In 2015, the highest mortality rate was observed in Afghanistan (115 deaths per 1,000 live births) and the lowest was in Monaco (1.8 deaths per 1,000 live births) (Requejo et al. 2015). From 1990, when the Millennium Development Goals were established, to 2015 the global infant mortality rate fell from 62 to 32 deaths per 1,000 live births. Despite this substantial reduction (around 50%), the established target was not reached, which was to reduce the 1990 rate by two-thirds.

Today, infectious diseases continue to be the leading cause of death for children, and they are also a major cause of death in adults. Globally, three of the top ten causes of death are infectious diseases. These also account for 16% of deaths each year. Most of these deaths occur in poor and developing countries, and they are attributable to preventable or treatable diseases such as diarrhoea, respiratory infections, HIV/AIDS, tuberculosis, and malaria. Although there have been significant advances in interventions to prevent and treat most of these diseases, such interventions are not always available for the populations that need them. Taking the example of tuberculosis, which is a disease highly linked to the conditions in which afflicted populations live and whose occurrence changes rapidly when these conditions change, in 2013 it was estimated that nine million people became sick with tuberculosis worldwide, with most of these cases (56%) occurring in Southeast Asia and the Western Pacific. However, Africa had the highest incidence of tuberculosis, with 280 cases per 100,000 inhabitants. Of these cases, around 500,000 were caused by multidrug resistant tuberculosis (MDR-TB) bacilli, which, in addition to causing more severe diseases, often increase treatment costs, making them prohibitive for the majority of patients living in poverty. In the same year, the estimated total number of deaths from tuberculosis was 1.5 million. Of these deaths, more than 95% occurred in developing countries, although the death rate fell by 45% between 1990 and 2013 (WHO 2014). NCDs as a whole are responsible for a significant proportion of the world’s burden of diseases, accounting for almost two-thirds of global deaths (36 of the 57 million

deaths in 2008). The main health problems in this group are cardiovascular diseases, cancers, chronic respiratory diseases (such as chronic obstructive pulmonary disease (COPD) and asthma) and diabetes. Although other diseases, such as mental and neurological diseases (including various forms of dementia), contribute to the high burden of morbidity (lower mortality), they have not been prioritized in global plans. NCDs are increasing rapidly in developing countries, where they impose large-scale human, social and economic costs, many of which could be avoided with well-known, cost-effective, and feasible interventions. Although they were initially associated with wealth, evidence shows that about 80% of NCD deaths occur in lower-middle-income countries. Even in sub-Saharan nations, where communicable diseases, maternal and perinatal issues, and nutritional deficiencies are still more important when taken together, they are in decline, whereas cases of NCDs are growing rapidly. This picture has resulted in projections that by 2030 NCDs will be the most frequent cause of death on the African continent (Rasella et al. 2013; WHO 2013; Worku and Woldesenbet 2015).

It is estimated that more than five million (9%) of deaths occurring globally are related to various forms of violence. Approximately a quarter of these deaths result from suicide and homicide, and traffic accidents are responsible for another quarter. The various types of violence vary in different regions of the world, but, in general, their rates are always higher in poor and developing countries.

Within countries this difference is observed being gradients of health status, measured by different outcomes, or the access and utilization of health services are observed for among regions, among social or ethnical groups.

Development and inequalities in health: the role of the Agenda 2030

The Agenda 2030 recovers a fundamental debate on health justice. The Goal 3 ‘Ensure healthy lives and promote well-being for all at all ages’ implies a health standard which can be globally recognized. The target 3.8 ‘Achieve universal health coverage, including financial risk protection, access to quality essential health care services and access to safe, effective,

quality and affordable essential medicines and vaccines for all' is not explicit in relation to the role of the public sector as the main provider of health goods and services. However, it states that healthcare is local and global social concern, indicating that healthcare is related to social justice. The concept of social justice is explored in Chapter 2.3 ('Poverty, justice and health') in this Textbook (Labonté et al 2015). There is an important convergence for significant part of the moral philosophers in relation to healthcare as an essential component of the well-being and the societies have the moral obligation to make them accessible and affordable for all individuals (Ruger 2004). That leads the debate to another dimension: the need of a global health justice framework, which could direct global institutions and norms towards a globally shared principle of global heath obligations and rights (Ruger 2016).

Health and development are deeply interrelated issues: without a healthy population, a society will hardly achieve development, as will an incomprehensible paradox if a society does not include the health of its population among the main objectives of its development.

Both issues have been on the United Nations' global agenda since its inception in 1945. In the 1960s, anticolonialist liberation struggles, which resulted in the independence of several African countries and other regions, gave renewed impetus to both issues. However, successive economic crises and strategies formulated in the 1970s and 1980s did not live up to the 'right to development' expectations of most underdeveloped countries. In health, the strategy of primary healthcare was drawn up since 1975, with its culmination in the Alma Ata Conference and Declaration in 1978. The relative failure of development policies led the United Nations to make efforts in practically all its agencies to review their policies and practices in the 1990s by holding international conferences to 'prepare the world for the 21st Century' (UNDESA 2018).

In 2000, at its headquarters in New York, the United Nations held the Millennium Summit (General Assembly 2000), which issued the Millennium Declaration (General Assembly 2000) with a set of eight Millennium Development Goals (MDGs) and 21 development goals to be achieved by the year 2015, measured by 60 indicators. Despite some recognized progress to many authors and the United Nations itself the major development goals were not achieved. Despite this, taking on the 'unfinished agenda of the MDGs', the United Nations and its Member

States insisted on the process, replacing in 2015 the Millennium Development Agenda with the 2030 Sustainable Development Agenda (General Assembly 2015), with its 17 goals and 169 targets to be met by 2030, compliance with which will be measured by 232 indicators.

The Rio + 20 Conference, held in Rio de Janeiro in 2012 (UN General Assembly 2012), builds the political foundations of Agenda 2030 and its Sustainable Development Goals (SDG), whose central mantra 'leave no one behind' clearly indicates the radical political rhetoric of the UN and its Member States by the reduction of inequalities between and within countries. The concept of managed sustainable development indicates that it will be realized through the articulation of its three dimensions—economic, social, and environmental—to be realized in the global, national, and local plans. There are reasonable degrees of freedom for each country to seek on the 'menu' of Agenda 2030 and its SDG to set up its own Development Agenda.

It can be said that Agenda 2030 and its SDG are potentially the most relevant global public policy underway to overcome inequalities in society, including in the field of health.

Funding for global development is suggested in the Addis Ababa Agenda for Action (AAAA) (UN General Assembly 2015a, b). The final text of the outcome document, adopted at the Third International Conference on Financing for Development (Addis Ababa, Ethiopia, 13–16 July 2015) and endorsed by the General Assembly in its resolution 69/313 of 27 July 2015, was one of the changes in relation to the Millennium Agenda. While it was basing its funding strategy on building a global development alliance in which official development assistance (ODA) gained importance, the AAAA stipulates that development must be pursued by each country for the openness, albeit asymmetrical, of its markets to international capital investments, which would inevitably lead to development. Numerous experts (including ourselves) doubt that simple trade liberalization can induce sustainable and equitable development, as advocated by Agenda 2030; in fact, the induction and participation of the democratic state is essential in the formulation and implementation of equitable and inclusive development policies.

Overcoming health inequalities is clearly stated on the SDG 3: 'Ensuring health and promoting well-being for all ages', but this concept does not

translate into a set of proposed goals and means of implementation capable of achieving the goal.

This gap between the intention of the goal in health and the means to achieve it is perhaps the main weakness of the Agenda in this area. It is a considerable task for United Nations sectorial agencies in the area—made up of the World Health Organization (WHO), its regionals, and the countries' Ministries of Health—to be able to find, with the support of civil society and users, the ways to adequately translate the politically powerful and far-reaching Objective into achievable goals and indicators that can measure the effective achievement of overcoming health inequalities, particularly in developing countries.

Among the proposed strategies are Health in All Policies (WHO, 2013) with the understanding that given the multidimensional nature of human health, governments need to assess the positive and negative effects of economic, social, and environmental policies on health, including in reducing inequalities. This strategy is supported by the highest policy body for conducting SDG implementation at the global level, the High-Level Political Forum (HLPF), with its 'From silos to integrated policy making' orientation, which explicitly recommends the adoption of integrated multisectoral and harmonious policies for achieving sustainable development, in accordance with the precepts of equity and inclusion.

The political rhetoric of Agenda 2030 and its SDG, including SDG 3, for health, establishes high-level global agreements for reducing health inequalities in developing countries. However, its concrete implementation will depend on the effective political will of the whole society and government—executives, legislators, judiciaries, and civil society—to translate into national development plans designed as a strategy for sustainable development in almost all countries.

Urban transition and health inequality

Demographic and epidemiological transitions are shaping and being shaped by urbanization. The demographic aspects combined with the urban built environment have influenced the conditions of living and, therefore, becoming ill or dying.

Fertility and mortality trends influence the age composition of populations, which, consequently, influences morbimortality patterns. The links between morbimortality patterns and population dynamics are commonly studied within the epidemiological transition theory. In emerging countries for example, three components of the population dynamics (fertility, mobility, and mortality) seem to be strongly associated with urbanization. Urbanization presents controversial impacts on health. If on the one hand the urban citizens are more likely to be reached by health, hygiene policies, and food availability, on the other they are also exposed to risks related to industrial sources, overfeeding, violence, and psychosocial factors. This means that a contradictory scenario tends to take place in the cities, especially, in developing countries (Eckert and Kohler 2014).

Since the mid of 20th Century, the emerging urban population has faced dramatic changes in morbimortality patterns. When compared to rural areas, emerging cities present lower levels of morbimortality due to infectious diseases and undernutrition, while presenting higher levels of morbimortality due to chronic diseases (Eckert and Kohler 2014). In the cities the levels of fertility have declined largely in the last 50 years as well. This suggests that the urban population tend to present better conditions in terms of sanitation and medical care and more exposed to widespread ideas on birth control (Bongaarts 2017).

However, mortality rates due to infectious diseases (including neglected diseases) and undernutrition vary within the city, suggesting that the gains of urbanization are not equally distributed among all settlers. When examining disaggregated data, intraurban variations expose the health inequality patterns across the urban territories (Hotez 2017). Socioeconomic differences and uneven morbimortality patterns have been a usual scenario in emerging cities.

Migrations and the limits of human movements

Since time immemorial, sectors of populations or even whole populations have moved to new destinations for various reasons. In 2013, it was estimated that 232 million people—3.2% of the world's population—lived outside their country of birth, and another 700 million were internal migrants in their own country of birth (UNDESA 2013). The patterns and

motives for these migratory movements have changed greatly over time; however, it is clear that the majority of migrants who cross national borders do so in search of better economic and social opportunities. In recent decades, exacerbated disparities between nations, global economic expansion, geopolitical transformations, wars, ecological disasters, and many other occurrences have had, and will continue to have, a profound impact on people's decisions to move to other nations. The recent phenomenon of the massive migration from some Arab countries to Europe is an example of the explosive and uncontrolled possibilities that the migratory issue can generate (Social Europe 2018).

The migratory issue introduces an important point into the debate on inequalities. Estimates show that social inequalities between countries account for a larger share of global inequalities than inequalities within countries. While inequalities within nations are much more related to class issues and other processes of social stratification, inequality between nations raises the issue of place of birth, or what has been termed the 'citizenship prize', which is related to history and the overall process of development of nations. Taking the question of life expectancy, a child born in Sierra Leone in 1990 only had an average life expectancy of 38 years, which was less than a child born in Japan in the same year (46 years compared with 84 years). Therefore, the issue of where one is born takes on great importance, and the migratory movement is defined by this attempt to change what was to some extent established by one's place and moment of birth. However, in a globalized world, in which capital, commodities, and human beings circulate, there are serious limitations on the movement of people, especially when they wish to move between nations.

Defining contexts on health inequalities

The essence of the health inequality is the uneven variation of health indicators among the individuals due to their social characteristics. That variability can be accessed through the stratification of the individuals according to available social characteristics data. Evans and Brown (2003) described the most used social characteristics in health inequality studies. It is summarized by the acronym PROGRESS, relating to first letter of the following set of variables: Place of residence (rural, urban, etc.); Race or

ethnicity; Occupation; Gender; Religion; Education, Socioeconomic status; and Social capital or resources. That comprehensive list of equity stratifiers for studying health risks and morbimortality does not consider some characteristics that are present in the literature. Age for example can be an equity stratifier if that condition is associated with unfair access to health services, as it occurs in many countries where adolescents have no access to contraceptive methods (WHO 2013). The relevance of each stratifier varies on time and space. In the recent years, large-scale involuntary migration emerged as one of the most relevant political items in the agenda of many countries, producing important shifts on national politics and international aid priorities in Europe. As a consequence, the health researchers elected health inequalities associated with race and ethnicity as a relevant investigating topic. Some findings revealed the impact of anti-immigration politics on health systems. In Barcelona in 2012, changes in the legal framework constrained immigrants to access health services (Castano 2016). In the Latin America and the Caribbean countries, race and ethnicity are relevant for observing the unfair distribution of health goods and services among national groups.

However, the migration is not frequently considered neither a prevalent nor a relevant topic on population health studies. Data availability and scientific tradition also interfere in the preference for some stratifiers. In Brazil, income is a very commonly used stratifier, while in South Africa, multivariable indicators, such as household socioeconomic status, are most popular. A common criticism related to the use of equity stratifies is the narrow representation of the social vulnerability captured by a single social dimension associated with one health indicator (WHO, 2013). In fact, vulnerable contexts (such as poverty in poor countries) encompass a large number of stratifiers which are entangled producing a complex picture of this vulnerability.

The growth in health inequalities: possible solutions

The construction of a more equitable world has been the aspiration of different political movements which understand that reducing inequalities in the various spheres of human life is essential and guarantees the existence and sustainability of human society. Health inequalities expose one of the

facets of inequalities that are prevalent among human beings, the cruel and damaging effects on one's own existence, which is reflected in the immense differences in life expectancy or the burden of disease and suffering.

The evidence of the importance of social determinants in explaining health inequalities is compelling. However, although there are clear academic and political arguments that favour the implementation of actions to redress the determinants of health inequalities, policies to mitigate these inequalities have been scarcely implemented as part of the public policies of national governments, and still less to alleviate inequalities between nations. There are several reasons for this lack of political motivation; however, some aspects have been recurrent in the literature regarding health inequalities.

One of the first aspects to consider is the consolidation of a health system based on biomedical knowledge and the resulting technologies, together with strong industrial and service sectors. These forces tend to generate and consolidate health systems that are only slightly affected in conceptual and structural terms, or favour actions directed at the social determinants of health.

It is worth noting that the existence of inequalities, and their magnitude, does not immediately result in moral imperatives or generate political action within a society. In some societies, relatively small levels of health inequalities have generated strong policy actions to reduce them (e.g. some Scandinavian countries), while others, with broad levels of health inequality, are not motivated to alleviate them as happen in a major part of the LMICs.

In poor and developing countries, where health inequalities are of the highest magnitude, there are few examples of the latter being among the priorities of public policy. For example, following the establishment by the WHO of its Commission on Social Determinants of Health, Brazil, which is a country with immense social and health inequalities, created its own national commission. However, after 2 years of work this commission produced a report which, in the main, was not assimilated into government actions (CNDSS 2008). Nevertheless, over the past two decades the implementation in many developing countries of redistributive policies such as income-transfer and micro-credits (von Gliszczynzki 2015), which are non-health policies, have had positive effects on health inequalities (Rasella et al. 2013). With regard to inequalities between countries, the proposals

and actions have been even more timid. For example, the final report of the WHO Commission on Social Determinants of Health (CSDH 2008; Landmann-Szwarcwald and Macinko 2016) places great emphasis on inequalities within a particular society and less on inequalities between nations. It has a chapter dedicated to the issue of health inequalities in the global sense, which focuses on the need to strengthen so-called global governance and explains the need for coordination among various intergovernmental agencies. Some of these ideas were subsequently deployed in policies, such as the MDGs, which focused on the eradication of extreme poverty from 2000 to 2015 and their successor, the SDGs, which include the additional aspiration of sustainable development in its three forms (economic, social, and environmental) during the period 2016–2030.

More recently, another group (*The Lancet*, the University of Oslo Commission on Global Governance for Health) (Ottersen et al. 2014) has made advances in understanding and proposing actions about global inequalities. The latest document produced by this group entitled 'The Political Origins of Health Inequity: prospects for change' is intended to 'convey a strong message to the international community and to all actors who influence global governance processes: we must no longer simply consider health to be a biomedical, technical issue; we recognize the need for multisectoral and global actions and justice to address health inequalities' (Ottersen et al. 2014, p. 631).

In conclusion, although interest in the issue of health inequalities has increased from the academic point of view, this interest has only had a limited impact on public policies aimed at improving the health of populations. Social inequalities in health are a global problem that, to a greater or lesser extent, affects all human societies. They are mainly due to the inequalities that exist between the different social groups in each society. Although the inequalities that exist between different societies and nations are relevant, and are often of a greater magnitude, they are not always considered to be unjust, and as such they are subject to political actions. The most plausible theory that has been put forward to solve the latter type of inequalities has been to improve the mechanisms of global governance, insofar as this includes an understanding of how nations were historically founded and the effect of the position of each country in the productive global system (Robinson and Acemoglu 2012).

How to reduce health inequalities in developing countries?

Combating health inequality demands an articulation of different actions that are associated with its determinants. A comprehensive approach should consider: (1) To be a distinguished goal of the national health policy; (2) To intervene on social determinants of health inequality; and (3) To improve living conditions and to overcome the exclusions mechanisms related to barriers for reaching the health systems.

It is difficult to access the national commitment on converting a public goal into effective policies. In Latin America, the concept of 'social institutionalization' has been used to access such degree of governmental commitment. According to Repetto (2004), social institutionalization refers to a set of formal and informal norms, including routine and organizational behaviours, necessary to process and to prioritize the social problems, delimiting the content and the political and administrative dynamics of social policies.

Cruces and Machinea (2010) have highlighted some institutional attributes that can be amplified to state level scale while defining a strategy for fighting health inequality: clear norms and goals, and sustainable policy. The issue of health inequality should be intrinsically part of the health policy with clear and stable norms and goals based on explicit political agreements and legal framework, with its sustainability been assured by adequate funding policy.

Martínez (2017) has performed an extensive review of the Latin American literature on social institutionalization. The following list highlights selected components based on aforementioned authors' reviews that tend to be relevant on the subject of reducing health inequality:

Consistent and sustainable long-term strategy. That indicates a resilient policy with the capacity to go beyond electoral cycles and ready to respond to economic and political shocks.
Social justice approach, as just explained.
Coordination *Fora*. Health inequality policy requires a large number of actions from different areas and levels. The coordination of different actors from different *loci* and interests requires specific spaces of horizontal and vertical coordination with mandate and capacity for easing divergences and identifying the best cost-effective solutions.
Development of technical capacities. A great limitation for implementing any action is the lack of qualified professional for implementing the adequate procedures. In developing countries, that is aggravated by the high turnover of the limited trained personnel as a consequence of precarious labour contracts and low incomes.
Social participation. That comment is strongly associated with the long-term sustainability of policy and programmes and will be further explored in the next section.

Interventions on social determinants of health inequalities are strongly associated with the reduction of morbimortality in vulnerable populations. Currently, the most effective approach to reduce poverty, extreme poverty, and social inequality is the use of articulated interventions to confront different factors associated with population vulnerabilities. Investment on education, labour policies, social protection, gender policies, and housing, compose the set of public policies that are often associated with the promotion of health equalitarianism. Ideally, a combined approach would seek to:

Promoting social development. Restoring the broader dimension of the concept of development in order to integrate an excluded population into the national and local economic and social dynamics.
Promoting equity. Using public policies to reduce unjustifiable differences and fighting discrimination. Reducing the poverty gaps faced by minority groups such as: black, indigenous, and traditional populations.
Generating opportunities. Human capital development considering gender, place of residence (urban or rural), and cultural context.
Improving quality in the delivery of public goods and services. The main deficit faced by the vulnerable populations is lack of access to quality public services.
Developing a sustainable policy. The endogenous and exogenous factors that can affect the programmes and policies sustainability should be considered while developing the model.

Box 2.5.1 presents the recent change in the Mauritian strategy to fight poverty and extreme poverty named as The Marshal Plan. An observation of its components can be useful for countries with some capacities but with huge social inequality.

Box 2.5.1 Marshall plan against poverty in Mauritius

Mauritius is implementing the Marshall Plan, aiming to fight poverty and inequality using an articulated set of 39 interventions from the areas: social protection, employment, social housing, education, health, environment, and equity in service delivery. The success of the plan lays on: (1) mobilizing all sectors, including private sector to reduce poverty and inequality; (2) produce institutional arrangements of public sector in order to avoid redundancy, overlapping responsibilities and waste of resources; (3) empowering the public institutions responsible for promoting community-based approach; (4) setting official poverty line for harmonizing the concept across public institutions; (5) improving technological capacity for accessing the needs and service performance; (6) developing creative mechanisms for funding pro-poor actions.

The first intervention is to fill the poverty gap in the whole country using conditional cash transfers. Many specific interventions on the poor areas are expected to produce positive impacts on the reduction of the health inequality, such as: dismantling of asbestos housing in 5 years; implementing community-based childhood care and education services; implementing community schools to promote emotional, physical, and creative learning.

One group of interventions was designed to reduce the barriers to access basic health services. The implementation of mobile clinics for youth-friendly and stigma-free healthcare services for providing sexual and reproductive healthcare services for youth, sex workers, people living with HIV/AIDS, people who inject drugs, and men who have sex with men.

Other community-based environmental interventions were designed to improve the sustainability in the poor areas. They comprise: increase waste recycling; producing income generating green jobs in textiles and manufacturing; developing alternative livelihoods from coastal and marine zones in management of communal forests and sustainable collection of native plant species; and to enhance ecotourism.

A third layer of policy interventions refer to improve living conditions and to overcome the exclusions mechanisms related to barriers for reaching the health systems. Living conditions encompasses a wide list of factors such as health behaviours and the hazards present in the working place. In relation to health behaviours is well known that tobacco consumption and poor diet are associated with socioeconomic status both in developed and developing countries. Thus, price interventions can reduce health inequalities, such as increasing taxation on unhealthy habits (alcohol and tobacco consumption) or providing subsidy to stimulate the healthy ones (Hill et al. 2014; McGill et al. 2015). Interventions to control diseases associated with poverty are also part of the strategy to reduce health inequality. More challenging is to promote the reduction of the barriers to access the health services. Kidd (2017) has used the concept of *social exclusion* to identify the main barriers of access to social policy services, as 'the process through individuals or categories of the population are wholly or partially excluded from full participation in the society in which they live'. He identifies three dimensions that restrict or block the access of vulnerable populations to diverse social programmes. He three dimensions require specific strategies for dealing with associated problems. The Kidd's social exclusion dimensions are described below, with focus adjusted on main problems and practical policy responses present in health sector:

Exclusionary forces. It is related to the view from those sectors that hold major share of power in the society over marginalized groups. Civil servants and institutions become embedded by discriminatory practices, institutional blindness, and no recognition of claims demanded by some groups. A large number of populations groups can be affected by discriminatory practices, such as: indigenous populations in Latin America, religious minorities in Asia, women in Arabic countries, Albinos in Africa, and disabled population in many continents. Recent interventions in Mauritius (Box 2.5.1) and Bangladesh (Box 2.5.2) have aimed to remove the barriers to health services and social protection schemes.

Structural disadvantages. They comprehend infrastructure deficits (such as roads, transportation, health facilities), lack of telecommunications systems, exposure to natural disasters (such as water crisis), and low levels of economic activity, interfering in all provision of services market. They are very associated with geographical factors and low economic dynamism of the areas located is the countryside of developing countries. Botswana (Box 2.5.3) have promoted an extensive shift on health programmes in order to improve the infrastructure of health services in poor rural areas. Brazil (Box 2.5.4) also present a relevant experience of how the interactive effects of primary healthcare and social transfers programmes can reduce morbimortality related to poverty-related diseases.

Limitations in capabilities. Some individual states or attributes may reduce the capacity to engage with public authorities and civil servants. The list of conditions is extensive, such as: disabilities, mental disorders, low education (literacy and numeracy), poor linguistic fluency, being exposed to domestic violence and abuse, low self-confidence, lack of available time, care responsibilities for children, elderlies, and ill persons.

Box 2.5.2 Reducing poverty-related diseases in Bangladesh

Bangladesh is one of the most populated, dense, and disaster-prone country in the world. With a territory of almost the size of England and Wales together, it accounted for 163 million people in 2016. In the last 20 years, it has achieved remarkable results on improving child and women health, despite of its modest public spending in health (2.6% of GDP in 2015), with meagre domestic general government health expenditure (0.39% % of GDP) in the same year according to the World Bank. The impressive health achievements at low cost turned Bangladesh in a role model for developing countries in the region (WHO 2015).

A mix of actors (public, private, and NGO) have contributed for the achievements on rapidly improvement of health indicators, such as infant mortality, under-5 mortality, and maternal mortality rates, morbidity rates from some infectious diseases including malaria, tuberculosis, diarrhoea, visceral leishmaniosis, HIV/AIDS, nutritional deficiencies, immunization coverage, and access to medicine (WHO 2015). Between 1990 to 2015, the under-5 mortality rate and maternal mortality ratio fell, respectively, to 73.8% and 69.1% (Ahmed 2016). The fall in mortality has produced significant impact in both life expectancy at birth and in fertility rate. Between 1990 and 2016, the total life expectancy at birth jumped from 58.4 to 72.49. When stratified by gender, that gains are greater for women. For men, the variation for this indicator in the period was from 58.05 to 70.89, while for women it moved from 58.78 to 74.29. Between 1990 and 2016, the total fertility rate declined from 4.49 to 2.1 (World Bank 2018). The demographic indicators suggest significant improvement of health conditions for women. That is a remarkable achievement for a patriarchal society with a tradition of gender disparities in all areas.

The results are associated with actions from different sectors, including health, education, social protection, infrastructure, and economic development. Some health programmes achieved international recognition such as: expanded immunization programme, oral rehydration, prevention, and treatment of targeted infectious diseases and undernourishment. An interdependent set of determinants are associated with these heath gains, such as the increase in female schooling, which in turn was strongly stimulated by the Female Stipend Programme designed to increase female secondary education and to delay marriage and childbearing (Hahn 2018).

Box 2.5.3 Fighting poverty as part of development strategy in Rwanda

After violent conflicts in Rwanda in the mid-90s, that included a genocide in 1994, the country experienced a rapid and sustained growth, sharp poverty reduction, and reduced inequality. According to Verpoorten (2014), among the most relevant drivers are: massive aid; increased public investment on health, education, and agriculture; the fertility transition; and the rise in international prices of the commodities.

The World Bank Open Data allows for comparison between the immediate post-conflict period with the current context. The net official development assistance received by the country increased from USD$321 million (2000) to USD$1.148 billion (2016). The secondary school enrolment (% gross) increased from 11% (2000) to 37% (2016). The health indicators presented substantive improvement, such as: the total life expectancy grew from 48 years at birth (2000) to 67 years (2016), and the total fertility rate declined from 5.6 births per woman to 3.9, respectively, in the same years of reference. Its economic prospects are marked by the real GDP growth averaged at about 8% per year (between 2001 and 2015).

The development policy of Rwanda is shaped by the Vision 2020 (long-term) and the second Economic Development and Poverty Reduction Strategy—EDPRS 2 (medium-term). They stress the combination of economic with social goals, such as: economic transformation with poverty reduction, e.g. moving the country from agriculture-based economy to knowledge-based and service-oriented-economy while reducing the percentage of population living on poverty from 30% by less than 9% by 2020 (World Bank 2017).

Rwanda also promoted a health sector reform that produced a relevant impact on maternal and infant health. According to Bucagu (2012), between 2006 and 2010, the improvements on the heath sector was particularly high in rural areas as a consequence of scaling up of facility-based childbirth policy based on: the heath workforce skills, performance-based financing, community-based health insurance, better governance.

The data compiled by the World Bank Open Data accounts that under-5 mortality rates per 1,000 live births fell from 195 in 2000 to 39 in 2016. That is strongly correlated with the rise in the coverage of the proportion of births attended by skilled health staff, which raised from 31% (2000) to 91% (2016). The DHS 210 allowed for comparison between the top and bottom quintile of wealth, indicating small difference on health indicators between the two groups, such as: antenatal care (at least one visit), DPT3 and measles immunizations, full immunization coverage (1 year), early initiation of breastfeeding, vitamin A supplementation (WHO 2013). The difference in percentage points between the two groups (top/bottom) varied from –1.6 (early initiation of breastfeeding) to 8.6 (full immunization coverage). At that time, the biggest difference was related to births attended by skilled health staff. The access of the bottom quintile was 61.2% while top quintile was covered by 85.9%. The subsequent progress on the indictor had reduced that gap.

Box 2.5.4 The interactive effects of primary healthcare and social transfers on the reduction of health inequalities in Brazil

From 1981 to 2014, Brazil presented one of the most remarkable drops in poverty in the world. According to the World Bank Data (2018), the poverty head count ratio at US$ 1.90 a day (2011 PPP) fell from 21.4% to 2.8% of the population in the period. The indicator refers to the population classified as ‘extreme poor’. From 1980 to 2014, according to the Brazilian Census Bureau (IBGE), the infant mortality rate (per 1,000 live birth) fell from 69.12 to 14.4, suggesting that the gains on poverty reduction influenced the improvement of public health indicators, as it is largely described in the scientific literature in relation to different countries. What is surprising is the fall of the GINI Index that was long stabilized from 1981 to 2003 around 58, reaching 51.3 in 2015 (World Bank 2018). Three determinants seemed to be associated with the reduction of poverty and inequality: the rise of the formal employment and the minimum wage, and the fast increase of the coverage of the social transfer programmes, which Bolsa Familia Programme is the major one in Brazil. The former was particularly relevant for protecting the extreme poor population against socioeconomic adversities, accounting for 66% in the reduction on the severity of poverty (Tesouro Nacional 2016).

The Brazilian primary healthcare programme (Family Health Strategy, FHS) is the biggest of its kind in the world. In 2018, it covered 206 million people (98.5% of the Brazilian population). More than 263,000 community health agents, organized in more than 42,000 teams, deliver a comprehensive and preventive healthcare services to the local populations. There is strong evidence of the contribution of the FHS on the reduction of infant mortality, in particular post-neonatal mortality due to diarrhoea and respiratory infections, also improving oral health, and increasing the detection of neglected disease and the reporting of vital data (Macinko and Harris 2015). Despite being a universal programme, its presence is more relevant to poor populations that use only the public health system, which contributes to the reduction of health inequalities. The FHS has contributed to the reduction of the disparities between Black/Pardo and White populations in relation to deaths due to ambulatory-care-sensitive conditions (Hone et al. 2017).

The effect of the poverty alleviation (including better nutrition) component of the Bolsa Familia with the incentive for accessing FHS is associated with the reduction of under-5 years of age mortality (Rasella et al. 2013), and the incidence of tuberculosis (Nery et al. 2017) and leprosy (De Andrade et al. 2018), among the poorer municipalities.

The interactive effects of social transfer and high coverage primary healthcare programmes are not only effective, but also very efficient. In 2017, the investment on Bolsa Family Programme

represented only 0.43% of the Brazilian GDP, while FHC responded for just 0.31 of the GDP (Senado Federal do Brasil 2018).

Social participation

Social participation is usually indicated as a condition for the sustainability of health inequality interventions in terms of funding, content, and governance. When active, it can prevent political interference on the health systems, specially under the shift of governments. It is considered a fundamental component of the institutionalization process and referred in many targets of the Agenda 2030. However, social participation refers to a wide set of engagements of representatives on policy fora, including inequalities of access to decision-making processes.

Arnstein (1969) described the participatory process as a ladder with eight steps, which social participation would vary from information to citizen control, e.g. different degrees of access to power from convening only to inform about decision taken elsewhere to effective participation.

Gaventa (2006) suggests a more complex model (cube shape), which participation can vary in three different levels (local, national, and global). The spaces for participation can be closed, invited, or claimed/created. And forms of power *per se* can be:

Visible, if the decision-making process is observable;
Hidden, if the setting of the political agenda is controlled by powerful individuals and institutions;
Invisible, if the controllers operate through psychological or ideological mechanism, preventing the appreciation of inconvenient topics.

Conclusions

In the recent decades, poverty has presented a sharp decline in all regions. However, in Africa, Asia, and in the Americas, poverty is still a vulnerability experienced by large parts of the population. That risk is exacerbated by economic crises that exterminate jobs, and shrink the fiscal space, thus compromising the countries' level of social protection and

safety nets. The face of poverty and the extreme poverty is changing, becoming more urbanized and affecting the older population, while remaining highly concentrated in rural areas and among children. If poverty is moving down, socioeconomic inequality is growing in size and complexity. Global and national wealth concentrations challenge the capacity of the states to produce equity, integration, and the feeling of living in a fair society.

Both poverty and inequality are no longer isolated national issues. The degree of economic and cultural integration at the global level shows how the prosperity of some social groups across the planet may be connected, through a complex web of economic unfair advantages, to the unfavourable condition of far larger population groups. The migration crisis faced specially in Africa and Asia, and secondarily in Europe illustrates that natural disasters, political conflicts, economic crises, religious or ethnical rivalries can drive masses of destitute people towards safer areas, expanding the effects of a context, which combines recent crisis with historical poverty to other regions.

Poverty and inequality interfere on the health dimensions: determinants of health, access to promotive and preventive actions, and health intervention. As a consequence, an inequal health patterns of diseases, deaths, and disabilities flows across the different population groups. Health inequality is present across the planet with far more intensity on the poorer areas and dwellings.

Policies for fighting poverty and inequality require local, national, and global approaches. The Agenda 2030 is a laudable initiative for presenting a multisectoral approach for confronting simultaneously health problems and their social determinants. The commitments of the countries encompass fighting poverty, hunger, health problems, and inequalities in articulation with environmental and economic interventions. However, the lack of a clear funding strategy limits its feasibility.

National and local interventions have decisively contributed to reducing health inequalities. Leveraging social participation has helped to promoting public policies for reducing health inequalities. Different health and other social policy initiatives from countries such as Bangladesh, Rwanda, Brazil, and Mauritius show alternatives ways for improving the health status of the population while reducing health inequalities.

REFERENCES

Ahmed, S.M., et al.. (2016). Cross-country analysis of strategies for achieving progress towards global goals for women's and children's health. *Bulletin of the World Health Organization*, 94(5), 351.

Arnstein, S.R. (1969). *A ladder of citizen participation. Journal of the American Institute of Planners*, 35(4), 216–24.

Bongaarts, J. (2017). Africa's unique fertility transition. *Population and Development Review*, **43**, 39–58.

Bucagu, M., Kagubare, J.M., Basinga, P., Ngabo, F., Timmons, B.K., and Lee, A.C., (2012). Impact of health systems strengthening on coverage of maternal health services in Rwanda, 2000–2010: a systematic review. *Reproductive Health Matters*, 20(39), 50–61.

Castano, J., Ospina, J.E., Caylà, J.A., and Greer, S.L. (2016). Restricting access to health care to immigrants in Barcelona: a mixed-methods study with immigrants who have experienced an infectious disease. *International Journal of Health Services*, 46(2), 241–61.

Comissão Nacional Sobre Determinantes Sociais da Saúde (CNDSS) (2008). *As Causas Sociais das Iniqüidades em Saúde no Brasil*. Editora Fiocruz. [Online] Available at: http://bvsms.saude.gov.br/bvs/publicacoes/causas_sociais_iniquidades.pdf

Commission on Social Determinants of Health (CSDH) (2008). *Closing the Gap in a Generation: Health Equity Through Action on the Social Determinants of Health: Final Report of the Commission on Social Determinants of Health.* [Online] Available at: http://apps.who.int/iris/bitstream/handle/10665/43943/9789248563706_por_contents.f?sequence=8

Cruces, G.A. and Machinea, J.L. (2010). Instituciones de la política social: objetivos, principios y atributos. In: Martínez, R. (ed.) *Institucionalidad Social en América Latina*. Santiago: CEPAL, pp. 31–54.

De Andrade, K.V.F., Nery, J.S., Penna, M.L.F., Penna, G.O., Barreto, M.L. & Pereira, S.M. (2018). Effect of Brazil's conditional cash transfer programme on the new case detection rate of leprosy in children under 15 years old. *Leprosy Review*, **89**, 13–24.

Deaton, A. (2013). *The Great Escape: Health, Wealth, and the Origins of Inequality*. Princeton, NJ: Princeton University Press.

Eckert, S. and Kohler, S. (2014). Urbanization and health in developing countries: a systematic review. *World Health & Population*, 15(1), 7–20.

Evans, T. and Brown, H. (2003). Road traffic crashes: operationalizing equity in the context of health sector reform. *Injury Control and Safety Promotion*, 10(1–2), 11–12.

Food and Agriculture Organization (FAO) (2015). *El Estado de la Inseguridad Alimentaria en El Mundo 2014. Fortalecimiento de un Entorno Favorable Para la Seguridad Alimentaria y la Nutrición.* Rome: Food and Agriculture Organization Publications.

Franco, R. and Székely Pardo, M., 2010. *Institucionalidad social en América Latina.* Chile: Documentos de Proyectos, Comisión Económica para América Latina y el Caribe, CEPAL.

Galobardes, B., Shaw, M., Lawlor, D.A., Lynch, J.W. and Smith, G.D. (2006). *Indicators of socioeconomic position (part 1). Journal of Epidemiology & Community Health*, 60(1), 7–12.

Gaventa, J. (2006). Finding the spaces for change: a power analysis. *IDS Bulletin*, 37(6), 23–33.

Guanais, F.C. 2015. The combined effects of the expansion of primary health care and conditional cash transfers on infant mortality in Brazil, 1998–2010. *American Journal of Public Health*, 105(S4), S593–9.

Hahn, Y., Islam, A., Nuzhat, K., Smyth, R., and Yang, H.S. (2018). Education, marriage, and fertility: long-term evidence from a female stipend program in Bangladesh. *Economic Development and Cultural Change*, 66(2), 383–415.

Hill, S., Amos, A., Clifford, D. and Platt, S. (2014). Impact of tobacco control interventions on socioeconomic inequalities in smoking: review of the evidence. *Tobacco Control*, 23(e2), e89–97.
Hoffman, K. and Centeno, M.A. (2003). The lopsided continent: inequality in Latin America. *Annual Review of Sociology*, 29(1), 363–90.
Hone, T., Rasella, D., Barreto, M.L., Majeed, A. and Millett, C. (2017). Association between expansion of primary healthcare and racial inequalities in mortality amenable to primary care in Brazil: a national longitudinal analysis. *PLoS Medicine*, 14(5), p.e1002306.
Hotez, P.J. (2017). Global urbanization and the neglected tropical diseases. *PLoS Neglected Tropical Diseases*, 11(2), p.e0005308.
Johnston, D.C. (ed.) (2015). *Divided: The Perils of Our Growing Inequality* (Vol. 10). New York: The New Press.
Kidd, S. (2017). Social exclusion and access to social protection schemes. *Journal of Development Effectiveness*, 9(2), 212–44.
Labonté, R., Baum, F. and Sanders, D. (2015). Poverty, justice, and health. In: Detels, R., et al. (eds.) *Oxford Textbook of Global Public Health*. Oxford: Oxford University Press, p. 89.
Landmann-Szwarcwald, C. and Macinko, J. (2016). A panorama of health inequalities in Brazil. *International Journal for Equity in Health*, 15(1), 174.
Macinko, J. and Harris, M.J. (2015). Brazil's family health strategy—delivering community-based primary care in a universal health system. *New England Journal of Medicine*, 372(23), 2177–81.
Marmot, M. (2015). The health gap: the challenge of an unequal world. *The Lancet*, 386(10011), 2442–4.
Martínez, R. (2017). Institucionalidad y desarrollo social. In: Martínez, R. (ed.) *Institucionalidad Social en América Latina y el Caribe*. Santiago: CEPAL, pp. 23–51.
McGill, R., et al. (2015). Are interventions to promote healthy eating equally effective for all? Systematic review of socioeconomic inequalities in impact. *BMC Public Health*, 15(1), 457.
Milanovic, B. (2016). *Global Inequality: A New Approach for the Age of Globalization*. Cambridge, MA: Harvard University Press.
Ministério da Saúde (2018). *Histórico de Cobertura da Saúde da Família.* Departamento de Atenção Básica. Brasília: Ministério da Saúde. [Online] Available at: http://dab.saude.gov.br/portaldab/historico_cobertura_sf.php
Nery, J.S., et al. (2017). Effect of Brazil's conditional cash transfer programme on tuberculosis incidence. *The International Journal of Tuberculosis and Lung Disease*, 21(7), 790–6.
Ottersen, O.P., et al. (2014). The political origins of health inequity: prospects for change. *The Lancet*, 383(9917), 630–67.
Rasella, D., Aquino, R., and Barreto, M.L. (2013). Impact of income inequality on life expectancy in a highly unequal developing country: the case of Brazil. *J Epidemiol Community Health*, **67**, 661–6.
Rasella, D., Aquino, R., Santos, C.A., Paes-Sousa, R., and Barreto, M.L. (2013). Effect of a conditional cash transfer programme on childhood mortality: a nationwide analysis of Brazilian municipalities. *The Lancet*, 382(9886), 57–64.
Repetto, F. (2004). *Capacidad Estatal: Requisito Para el Mejoramiento de la Política Social en América Latina*. Washington, DC: Inter-American Development Bank.
Requejo, J.H., et al. (2015). Countdown to 2015 and beyond: fulfilling the health agenda for women and children. *The Lancet*, 385(9966), 466–76.
Robinson, J.A. and Acemoglu, D. (2012). *Why Nations Fail: The Origins of Power, Prosperity and Poverty*. New York: Crown Business.
Ruger, J.P. (2004). Health and social justice. *The Lancet*, 364(9439), 1075–80.

Ruger, J.P. (2016). *Global Health Inequalities and Justice. Understanding Health Inequalities and Justice: New Conversations across the Disciplines*. Oxford: Oxford University Press, p. 64.
Senado Federal do Brasil (2018). *Orçamento Federal*. SIGA Brasil. [Online] Available at: http://www9.senado.gov.br/QvAJAXZfc/opendoc.htm?document=senado%2Fsigabrasilpainelcidadao.qvw&host=QVS%40www9&anonymous=true&Sheet=shOrcamentoVisaoGeral
Social Europe (2018). Europe's refugee crisis. Available at: https://www.socialeurope.eu/focus/europes-refugee-crisis
Stiglitz, J.E. (2012). *The Price of Inequality: How Today's Divided Society Endangers Our Future*. London: WW Norton & Company.
Tesouro Nacional (2016). Gasto social do Governo Central 2002 a 2015. [Online] Available at: http://www.tesouro.fazenda.gov.br/documents/10180/318974/Gasto+Social+Governo+Central/c4c3d5b6-8791-46fb-b5e9-57a016db24ec
UNDESA (2013). *World Population Ageing 2013*. United Nations, Population Division, Department of Economic and Social Affairs. ST/ESA/SER.A/348.
UNDESA (2014). *HLPF Issue Briefs 5: From Silos to Integrated Policy Making*. United Nations, Division for Sustainable Development. [Online] Available at: http://sustainabledevelopment.un.org/index.php?page=view&type=400&nr=1322&menu=35
UNDESA (2018). *UN Major Conferences and Summits*. [Online] Available at: http://www.un.org/en/development/desa/what-we-do/conferences.html
UN General Assembly (2000a). *The Millennium Summit*. [Online] Available at: http://www.un.org/en/events/pastevents/millennium_summit.shtml
UN General Assembly (2000b). *The Millennium Declaration*. [Online] Available at: http://www.un.org/millennium/declaration/ares552e.htm
UN General Assembly (2012). *Resolution A/RES/66/288—The Future We Want*. [Online] Available at: http://www.un.org/ga/search/view_doc.asp?symbol=A/RES/66/288&Lang=E
UN General Assembly (2015a). *Third International Conference on Financing for Development. Addis Ababa Action Agenda (AAAA)*. [Online] Available at: https://www.un.org/esa/ffd/ffd3/index.html
UN General Assembly (2015b). *Resolution A/RES/70/1. Transforming our world: the 2030 Agenda for Sustainable Development*. [Online] Available at: http://www.un.org/ga/search/view_doc.asp?symbol=A/RES/70/1&Lang=E
Verpoorten, M. (2014). *Growth, Poverty and Inequality in Rwanda: A Broad Perspective*. WIDER Working Paper 2014/138.
von Gliszczynzki, M. (2015). *Cash Transfers and Basic Social Protection—Towards a Development Revolution?* Basingstoke: Palgrave Macmillan.
World Health Organization (WHO) (2013). *Handbook on Health Inequality Monitoring: With a Special Focus on Low-and Middle-income Countries*. Geneva: World Health Organization.
World Health Organization (WHO) (2013). *Global Action Plan for the Prevention and Control of Noncommunicable Diseases 2013–2020*. Geneva: Public Health England. Interactive map—NHS health check. [Online] Available at: https://www.healthcheck.nhs.uk/interactive-map/
World Health Organization (WHO) (2014). *Global Tuberculosis Report 2014*. Geneva: WHO.
World Health Organization (WHO) (2015). *World Health Statistics 2015*. Geneva: WHO.
World Health Organization (WHO) 2015. *Bangladesh Health System Review*. Manila: WHO Regional Office for the Western Pacific.
World Bank (2018). *World Bank Database*. [Online] Available at: https://data.worldbank.org
World Bank (2018). *Poverty Headcount Ratio at $1.90 a Day (2011 PPP) (% of Population)*. [Online] Available at: https://data.worldbank.org/indicator/SI.POV.DDAY?locations=BR
World Bank (2018). *GINI Index (World Bank Estimate)*. [Online] Available at: https://data.worldbank.org/indicator/SI.POV.GINI?locations=BR

World Bank (2013). *World Bank Database: Poverty*. [Online] Available at: https://data.worldbank.org/topic/poverty
World Bank (2013). *World Bank Database: Poverty*. [Online] Available at: https://data.worldbank.org/topic/poverty
Worku, E.B. and Woldesenbet, S.A. (2015). Poverty and inequality—but of what-as social determinants of health in Africa?. *African Health Sciences*, 15(4), 1330–8.

2.6

Genomics and public health

Veron Ramsuran and Tulio de Oliveira

Introduction

In the last two decades there have been several significant advances in the field of genomics. Some of these advancements include, the completion of the Human Genome Project in 2003, use of genomic tools to diagnose patients at point-of-care facilities, and the use of gene editing technology to cure several diseases. The discipline of ‘public health genomics’ emerged as a need to provide ‘responsible and effective translation of genome-based knowledge and technologies for the benefit of the population’ (Bellagio 2005, in Bellagio 2019). It is becoming increasingly accepted that the one size fits all approach is going to be outdated soon. Targeted personalized medicines, therapies, and interventions approach will be the preferred form of treatment in the not too distant future (Molster et al. 2018).

Genomics has served as the platform for ‘precision medicine’, which utilizes the advances made by genomic technologies to provide information that is person centric. Determining the interactions of an individual’s genes, environment, and lifestyle to deliver a targeted intervention specific to that person (Khoury et al. 2017). The rapid advancement of genomic technologies has led to significant volumes of individual and population level data. Combining data generated, at the individual level on genes, environment, behaviour, and other social and economic determinants of health, provide a useful new approach of utilizing big data to guide the right intervention to the right people at the right time, has been termed ‘precision public health’ (Khoury et al. 2017).

The fields of public health genomics, precision medicine, and precision public health have proven important aspects in improving population health. However, despite the advancements made, there are still

improvements that need to be implemented to significantly improve health, especially within the developing countries. In this chapter we discuss the advancements made within the field and highlight challengers or future directions.

Genetic variation has been shown to associate with a number of clinical outcomes. These variations could be single nucleotide polymorphisms (SNPs), combinations of SNPs within a gene or in multiple genes and chromosomal aberrations (Molster et al. 2018). Genetic testing can be used as a diagnostic or a predictive tool to screen individuals who present with or are at risk for a particular disease or to determine a suitable course of treatment through pharmacogenomic studies. As a predictive tool, it allows for newborn screening of metabolic diseases such as phenylketonuria, it can be used to determine if individuals are carriers of a recessive trait which can be passed on to offspring, it allows non-invasive testing of foetuses for genetic disorders or chromosomal abnormalities and it can also be used during *in-vitro* fertilization to test if an embryo has a particular genetic or chromosomal abnormality. Genetic testing methods vary depending on the question being asked. Simple tests such as SNP genotyping and target gene capture with sequencing are used for single-gene disorders or Mendelian diseases. Chromosomal disorders on the other hand are abnormalities of the chromosome structure or number and these are detected usually through karyotype analyses which reveals minimal information about the genetic code itself. Finally, there are the polygenic diseases that are caused by a collection of mutations that may occur across a number of different genes contributing to a given phenotype (Khoury et al. 2017). For example, there are several combinations of mutations that can lead to hypertension, coronary artery disease, and neuropsychiatric disorders. Testing for these disorders can sometimes be more complicated especially when the causative genes are unknown. When the causative genes are known, targeted sequencing of specific genes can be performed. However, when they are unknown it becomes more beneficial to perform whole exome (WES) or whole genome sequencing.

The era of genome-wide association studies has allowed us to identify variants across the genome that contribute to a disease phenotype. Many of the variants identified that reach genome-wide significance are found within the non-protein coding regions of the genome. One of the major benefits of performing whole genome sequencing (WGS) is that it covers variation in

the coding and non-coding regions of the genome. Another is that WGS data from an individual does not change over time and contains a plethora of invaluable information about the health outcomes one might experience in their lifetime (Molster et al. 2018).

Public health genomic tests currently in use

The rapid proliferation of genetic tests has made it possible to detect a range of health problems at the clinic level or direct-to-consumer for personal genetic tests. According to the Genetic Testing Registry (GTR) (https://www.ncbi.nlm.nih.gov/gtr/), a central location of submitted genetic test information hosted by the National Institute of Health, has 60,027, 11,971 conditions, 18,631 genes from 526 labs across the world (last accessed 16 August 2019) (NCBI 2019). The United States still dominates the number of labs contributing to the GTR, about 50% of the information submitted originates from a United States-based laboratory. The conditions reported range from several rare diseases to a variety of common diseases. The risk assessment, predictive, diagnostic, and prognostic are described for the tests. The information provided on the GTR is voluntary based submissions and is not verified for accuracy (NCBI 2019). The application of these tests ranges from diagnostic, predictive, and reproductive genetic testing, which could be used in response to a specific clinical problem or as a screening tool.

Diagnostic genomic testing

The diagnostic genomic testing is often used to confirm a clinical diagnosis. The genetic information obtained proves valuable to relatives and for monitoring of treatment. The finding of genetic makers observed within an individual that is linked to an inherited disease should encourage their family members to perform the diagnostic test (Molster et al. 2018).

Predictive genomic testing

Genetic variants linked to a person’s risk of developing heritable disorders are used as predictive tools or biomarkers. There are a few tests that have

received significant attention due to their usefulness for screening before signs or symptoms appear. These predictive tests include screening for cancers, newborn diseases, and pathogen detection.

Hereditary breast and ovarian cancer

More than 50 hereditary cancer syndromes have been reported and genetic predictive testing is available for most of these cancers (NCI 2019). Mutations within the *BRCA1* and *BRCA2* genes have been associated with Hereditary Breast and Ovarian Cancer (HBOC). These genes have tumour suppressor abilities and serve the function of repairing damaged DNA. Therefore, mutations within these genes inhibits the overall functioning. In breast cancer mutations account for 5% of all cases, while within ovarian cancer mutations account for 10–15% of cases (Armstrong et al. 2019). Women with *BRCA1* and *BRCA2* mutations have a significantly higher lifetime risk of developing breast cancer (about 50–85%) and ovarian cancer (about 15–40%) chance. There are several thousand mutations that have been identified within the *BRCA1* and *BRCA2* genes (BRCA 2019). Many of these mutations do not contribute towards cancers, while certain mutations have the ability to cause truncated proteins when translated. Understanding the genetics basis of mutations with regards to HBOC from diverse ethnic backgrounds around the world is the aim of collaborative efforts like the BRCA Exchange programme (BRCA 2019). Resources that provide clinical significance of genetic variants specific to the country and ethnicity an individual originates from has greater relevance to the patient than general information.

Apart from *BRCA1* and *BRCA2* genes, which show the strongest lifetime risk for developing breast cancer, other human genes have also been linked to breast cancer predisposition. The genes identified and the lifetime risk for breast cancer are TP53 (50–80%), PTEN (50–85%), CDH1 (39–52%), STK11/LKB1 (35–50%), PALB (20–30%), CHEK2 (15–25%), NF1 (15–25%), and p16 (15–25%) (NCI 2019). Risk genes for ovarian cancer include BRIP1, RAD51C, and RAD51D (NCI 2019). Genomic testing is currently being offered direct to consumers or through healthcare providers and genetic counsellors.

Newborn screening

Infants that are a few days old undergo predictive testing as a form of newborn screening (NBS). A growing list of NBS tests provide predictive evidence that an infant is at risk for certain metabolic, hormone, and genetic diseases (Therrell et al. 2015; Urv and Parisi 2017). The metabolic disorders NBS examines includes; phenylketonuria (PKU), methylmalonic acidaemia, maple syrup urine disease (MSUD), tyrosinemia, citrullinemia, and medium chain acyl CoA dehydrogenase (MCAD) deficiency. Congenital hypothyroidism and congenital adrenal hyperplasia are the hormone disorders tested. Some of the other disorders tested can include sickle cell disease, haemoglobin SC disease, beta thalassemia, galactosemia, biotinidase deficiency, cystic fibrosis, severe combined immunodeficiency (SCID), Pompe disease (glycogen storage disease type II), mucopolysaccharidosis type 1, X-linked adrenoleukodystrophy and spinal muscle atrophy (SMA) (Therrell et al. 2015; Urv and Parisi 2017). Results from these tests provide information for diagnosis, treatments, follow-up, and evaluation. Most the genetic testing performed for NBS is performed on mass spectrometry, which allows multiple conditions to be screened for at once. Traditional genetic testing, such as next-generation sequencing is not recommended as the initial technique for NBS, mainly due to the time taken to receive results (Berg et al. 2017; Ceyhan-Birsoy et al. 2017). Currently, researchers are exploring alternate DNA-based genetic tests to complement existing NBS methods (Friedman 2015).

Preimplantation and prenatal genetic testing

Genetic testing that determines if an embryo possesses genetic abnormalities prior to *in-vitro* fertilization is termed preimplantation genetic testing. This screening test decreases the chance of having a child with genetic abnormalities, since only embryos ‘passing’ the test are implanted into the uterus. Whole genome and exome sequencing are used to screen the embryos.

Another test, which examines genetic changes present within genes and chromosomes from a foetus is termed prenatal genetic testing. This test allows for detection of disorders such as Down’s syndrome and neural tube defect. Results from this test provide parents with information to prepare

for the child's care after birth, some parents choose termination of pregnancy. Novel testing methods have been evolving to detect genetic abnormalities, which include whole genome or exome sequencing (Vora and Wapner 2018). Measuring the cell-free DNA (cfDNA) from the foetus, which circulate in the mother's blood, serves as a form of non-invasive prenatal testing (NIPT) (Vora and Wapner 2018). NIPT whole genome or exome sequencing have also been reported (Fan et al. 2012; Kitzman et al. 2012), however, these techniques still possess a few limitations (i.e. cost of sequencing, limited cell-free placental DNA, and requirement of paternal DNA to determine the fetal genome).

Pathogen genomics

The use of genomics within pathogen detection, specifically infectious diseases, has made the greatest advances in genomics of public health. Advances in molecular technologies and bioinformatics have allowed researchers to examine pathogen genomes more rapidly, which has led to altering both medical and public health practice (Gwinn et al. 2019). Some of the public health applications for pathogen genomics, as described by the United States Centers for Disease Control and Prevention (CDC), are diagnosing infection, investigating outbreaks, describing transmission patterns, monitoring antimicrobial resistance, and developing interventions, including vaccines (CDC 2019). The CDC has an advanced molecular detection (AMD) programme that uses the latest technology to examine the key areas of pathogen genomics and has committed to making genomic data freely available. Access to pathogen genomic data is key to understanding pathogen mutations that could threaten public health, identifying new pathogens, development of diagnostic tests and determining and monitoring outbreaks (CDC 2019). Increasing the number of countries that subscribe to the data sharing model will dramatically improve pathogen genomics.

Next-generation sequencing is currently being used for pathogen genomics which provide rapid and reliable results compared to traditional clinical diagnostics. A new sequencing method, nanopore sequencing, has the potential to be the biggest game-changer for public health genomics. The first available commercial instrument has unrivalled portability (i.e. sequencing can be performed at any location, even from the outbreak site). This was demonstrated by studies that used nanopore sequencing in a

mobile genomics laboratory that travelled through rural Brazil monitoring the genetic diversity of Zika virus (Faria et al. 2017) and yellow fever virus (Faria et al. 2018). The sequences produced from these studies were analysed in real time, which allowed for the generation of pathogen diversity, distribution, and transmission data. Apart from the device portability, nanopore sequencing has a number of other benefits such as fast sample preparation, flexibility, long sequence reads, and relatively low cost. These benefits make this technology an attractive option as a first line strategy for pathogen sequencing in clinical and public health settings (Besser et al. 2017; Quainoo et al. 2017; Gwinn et al. 2019).

Technologies aiming to be implemented within a clinical health setting need to meet certain criteria in order to be fully utilized and provide public health benefit. Some of these criteria include, ease of use, cost-effectiveness, rapid results, minimized risk to patients, extended shelf life of reagents and the ability to function within resource-limited settings (i.e. interrupted electricity and extreme temperatures). The GeneXpert MTB/RIF test is one example where the technology has tremendous potential due to the ability to detect tuberculosis (TB) and the multidrug resistant form with very high sensitivity and specificity. However, this test does not meet some of the requirements to function effectively within a resource-limited setting, which has a large proportion of the TB infected patients (Evans 2011; Van Rie et al. 2011). It is therefore suggested that new technologies targeting diseases prevalent within resource-limited settings carefully consider the criteria to achieve maximum impact on public health.

Evaluation of genetic testing

Genetic testing is a rapidly expanding area with many clinical applications. As mentioned earlier the GTR (https://www.ncbi.nlm.nih.gov/gtr/), there are serval genetic tests currently being used around the world. While the introduction of new genetic tests creates tremendous potential for improving patient care, it is essential to evaluate these tests thoroughly to ensure that they are accurate and lead to improved patient outcomes when used in clinical practice settings. Therefore, methods have been developed to guide stakeholders (patients, clinicians, healthcare system policymakers, payers, and public health officials) in the assessment of tests, including

genetic tests, in a broad array of clinical settings. The process of implementing a molecular genetic test for diagnostic use is complex and involves many levels of assessment and validation. The key components of the process, as detailed by the ACCE framework, are analytical validation, clinical validation, clinical utility, and consideration of the ethical, legal, and social implications of the test (Pitini et al. 2018). Analytical validity and clinical validity are the two main measures of accuracy that apply to genetic tests. Another measure of the quality of a genetic test is its usefulness, or clinical utility (Holtzman and Watson 1999).

Analytical validity refers to how well the test predicts the presence or absence of a particular gene or genetic change (Teutsch et al. 2009), i.e. can the test accurately detect whether a specific genetic variant is present or absent?
Clinical validity refers to how well the genetic variant being analysed is related to the presence, absence, or risk of a specific disease (Teutsch et al. 2009; Wylie Burke 2014).
Clinical utility refers to whether the test can provide information about diagnosis, treatment, management, or prevention of a disease that will be helpful to a consumer (Teutsch et al. 2009; Wylie Burke 2014).

The Agency of Healthcare Research and Quality (AHRQ) report evaluated the four commonly used methods that cover the principal domains of test evaluation used in the environment of genetic testing. The four methods include; the US Preventative Services Task force (USPSTF) method (Harris et al. 2001; Trinite et al. 2009), the Fryback Thornbury hierarchy (Fordyce and Douglas 2017), the ACCE model and the Evaluation of Genomic Applications in Practice and Prevention (EGAPP) model (Teutsch et al. 2009). Framework such as Fryback–Thornbury hierarchy provide general conceptual guidance, analytic framework such as USPSFT and EGAPP provide additional benefit for important questions regarding the relevant population. A framework of interest was the McMaster University evaluation due to its thoroughness and richness in detail for making decisions about genetic testing (Giacomini et al. 2003). Also reviewed was the Genetic Testing Evidence Tracking Tool (GETT). GETT allows (1) researchers to summarize the current evidence and to identify knowledge gaps for further research, and; (2) stakeholders to collect data related to a

given molecular test and improve their decision-making process (Rousseau et al. 2010).

In the USPSTF method, the purpose is defined as preventive interventions in primary care. It provides a formal analytic framework; evidence assessment related to key questions to establish a chain of evidence. The downfall of focusing on preventive services results in a focus on clinical utility outcomes that are not relevant to all clinical applications (Sawaya et al. 2007). In contrast, the Fryback–Thornbury hierarchy, recognizes the clinical value of diagnostic efficacy and diagnostic thinking. It provides a hierarchic representation of levels of efficacy for medical tests, which allows an evaluator to determine what evidence types need to be assessed for a given test purpose or desired outcome (Morrison and Boudreau 2012). However, this method lacks a formal evidence-based assessment procedure. ACCE is an analytic process for evaluating evidence on genetic tests. It has a standard set of 44 questions that are organized according to different evidence types (analytic validity, clinical validity, clinical utility, ELSI) (Sanderson et al. 2005). This method provides a highly granular approach to assessing different evidence types. However, ACCE does not provide details on evaluating the strength of evidence. It was developed for single-gene tests and may be difficult to extend to multigene panels or genome-scale sequencing test. EGAPP has a systematic approach to evidence-based assessment of genetic tests (Teutsch et al. 2009).

A hybrid approach using analytic framework similar to USPSTF and the evaluation of evidence types articulated by ACCE might provide an enhanced method. The flexibility of this framework allows to evaluate different topics in genetic testing including a wide array of potential outcomes of interest, and integration of formal evidence-based reviews (Sanderson et al. 2005). The downside is the flexibility to evaluate different topics in genetic testing focus on single-gene tests that may be difficult to extend to broader genomic technologies. GETT provides structure for systematic identification and organization of published evidence on genetic testing. GETT comprises 72 clearly defined items/questions, grouped into 10 categories and 26 subthemes, including an overview of disease epidemiology and genetics, the available diagnostic tools, and their analytical and clinical performances, availability of quality control programmes, laboratory, and clinical best practice guidelines, clinical

utility, impact on healthcare, and psychosocial, ethical, and legal implications. The method helps stakeholders to determine whether the knowledge base is sufficient for genetic technology assessment (Rousseau et al. 2010). The disadvantage of this method is the lack of details on evaluating the strength of evidence; developed for single-gene tests and may be difficult to extend to multigene panels or genome-scale sequencing tests. The McMaster University framework is an evaluation model to guide public coverage for new predictive genetic tests in Ontario, Canada (Giacomini et al. 2003). This method combines technology assessment with coverage decision-making from payer's perspective. It defines criteria for determining coverage, anticipates the need for payers to identify evidentiary thresholds, and considers conditional coverage scenarios. Since it is developed for the Canadian health system, it lacks a formal evidence-based assessment procedure.

In order to achieve the full potential of genetic tests and to avoid inappropriate use, genomic tests need to be broadly understood and a dialogue between different types of clinicians, researchers, and drug companies needs to be facilitated. The frameworks mentioned earlier are intended to facilitate ongoing discussion and developmental activities among manufacturers, test researchers, systematic reviewers, regulators, and other policymakers. These discussions need to then facilitate understanding among clinicians and patients. Future frameworks should be broad enough to be applied to a wide variety of genomic testing. It is appreciated that the complexities and rapid advances within genomics (next-generation sequencing) and related fields (proteomics, metabolomics) will necessitate re-evaluating and tailoring frameworks over time.

Future aspects require examining patient outcome efficacy to determine if patients benefit from the test and if patients that take the test do better than similar patients who do not. Another aspect is observing at societal level, it should be noted that genetic testing should have a favourable effect on the health and allow effective health interventions throughout the entire population (Hall and Olopade 2005; Phillips et al. 2014). It is also important to examine ethical, legal, and social implications of genomic medicine (ELSI) and the effects of applying genetic technologies on a large scale (Clayton 2003). Technical efficacy for accurate detection of genetic variants, analytic, sensitivity, and specificity also include therapeutic and management efficacy as well as diagnostic thinking efficacy. This reflects a

useful synthesis of existing frame working efforts to date, it may allow more consistency in definitions for terms and concepts going forward and provide a platform for future collaborative efforts.

Pharmacogenomic testing

The combination of pharmacology and genomics has led to the new field termed pharmacogenomics, which essentially examines the effects a person's genes have on response to drugs. Tailoring effective, safe medications and doses according to the genetic makeup is one of the major goals of precision medicine. Most of the drugs currently available and being produced are designed as 'one size fits all', however, increasing evidence has demonstrated that drugs do not have the same effect on everyone.

Pharmacogenomics involves DNA and RNA influence on drug efficacy whereas pharmacogenetics only involves DNA and is usually observed for single genetic variants rather than a combination of variant effects. Pharmacogenomics is a relatively new field of research which is gaining momentum due to the variable effectiveness of drugs between individuals and it is an area of study that shows promise for the design of safe and efficacious treatments in a dose dependent manner according to an individual's genetic makeup (Roden et al. 2019). The majority of the drugs available today have been created without taking pharmacogenomics into account. These drugs have been created with a one drug fits all approach. Furthermore, many of the drugs that have passed clinical trials have been tested in mostly American/European Caucasians, some Asians, and a few individuals of African descent, therefore, extending these treatments to individuals from developing countries does not always prove successful. One of the major problems with using these non-tailored drugs are the adverse drug reactions (ADRs). ADRs occur at alarming frequencies and are well documented in developed countries. Most clinicians today use a trial and error approach when prescribing drugs. Usually, when an individual fails treatment due to ineffectiveness or the occurrence of a negative reaction the drug dosage is altered, or the drug is changed. This process is time-consuming, costly, and the potential of untimely death. In the US alone it is estimated that more than 100,000 individuals die from serious adverse events every year (Tan et al. 2016). Within developing

countries, it is estimated that death by ADRs greatly exceeds that of developed countries. This is mainly due to the genetic variation unaccounted for during drug development. The interindividual genetic diversity within individuals of African descent is several folds higher than other ethnic groups (Masimirembwa et al. 2014), which highlights the dire need for pharmacogenomic drug development within Africa.

Two major considerations should be considered when establishing whether a drug will prove efficacious in an individual, i.e. pharmacokinetics and pharmacodynamics. Pharmacokinetics is the process of absorption, transport to target site, metabolism, and excretion (Roden et al. 2019). Pharmacodynamics is variability in drug action that is not contributed by drug concentrations, and represents changes of the interaction of active drug with its effector molecules or other mechanisms such as variability in disease mechanisms (Roden et al. 2019). A large part of pharmacogenetic traits identified to date have a pharmacokinetic controlling mechanism.

Pharmacokinetic mechanism

There are two main scenarios that demonstrate how single gene variants affect pharmacokinetics that could lead to large effects. The first example requires the use of bioactivation pathways to achieve the desired therapeutic effects. Changes caused by mutations in the drug metabolizing enzyme can decrease or block the drug effect. An example includes codeine that gets bioactivated to active metabolite morphine with the *CYP2D6* gene (Roden et al. 2019). The majority of the variants identified offer partial or complete loss of function. There are fewer variants that lead to gain of functions within the bioactivation pathways, which subsequently leads to excess drug response. An example of gain of function variants include the *CYP2C19*17* that was linked to bleeding during clopidogrel therapy (Sibbing et al. 2010).

The second scenario is demonstrated when pharmacokinetic variants apply large effects during the administration of a drug that possesses a narrow therapeutic range of activity, that leads to either optimum activity or toxic doses. This is mainly through the elimination by a single drug metabolizing system. An example of this scenario is demonstrated by the antileukaemic drug 6-mercaptopurine that gets bio inactivated by thiopurine methyltransferase (TPMT) and xanthine oxidase. Individuals that are

homozygous for the loss of function variants within TPMT exhibit life-threatening bone marrow toxicity (Relling et al. 2019).

Transporter molecules, responsible for moving drugs into and out of cells, are affected by mutations that possess the capability of altering drug levels at target effector sites. An example of an efflux transporter gene that is affected by variants is the OATP1B1 encoded by SLCO1B1, which are shown to be variable in individuals of African descent (Naidoo et al. 2018; Naidoo et al. 2019).

Pharmacodynamics mechanism

Some of the initial well-described syndromes affected by pharmacogenetics are from the pharmacodynamics area. The anticoagulant drug warfarin is a well-studied pathway that acts by inhibiting the production of vitamin K dependent clotting factors in order to treat thrombotic disorders. The drug dosage is influenced by genes encoding the cytochrome P450 family (CYP2C9, CYP4F2, CYP2C) of enzymes responsible for the metabolizing the drug as well as a VKORC1 which recycles vitamin K in the body (Belley-Cote et al. 2015; Nagai et al. 2015). Testing for variation within these genes allow clinicians to make informed decisions when administering warfarin. If the drug is administered at a standard dose without testing for variation in these genes it may result in either too much drug in the bloodstream which can lead to excessive bleeding and death or not enough drug which could lead to thrombosis.

Another example can be observed during HIV treatment, where a combination of antiretrovirals are administered. One such drug known as abacavir has shown serious adverse reactions ranging from fever, nausea, vomiting, and death. Studies have shown that administering abacavir to individuals carrying the HLA-B*57 allele results in increased hypersensitivity and therefore testing for HLA-B*57 is recommended prior to administering abacavir (Sousa-Pinto et al. 2015; Illing et al. 2017). The allele frequency of HLA-B*57 in Caucasians is approximately 6% while in African Americans it is between 2% and 3%. However, in native Africans the allele frequency is much lower. Another HLA-B allele (HLA-B*15:02) has also shown an effect on Stevens–Johnson syndrome and toxic epidermal necrolysis (SJS/TEN). HLA-B*15:02 confers risk of the disease mainly

within southeast Asian populations, in which the allele is found in relatively high frequencies (Chung et al. 2004). However, a different HLA allele (HLA-A*31:01) has been linked to the SJS/TEN disorder within individuals from European ancestry (McCormack et al. 2011). These studies reporting differing alleles contributing to SJS/TEN highlights the need to examine the genetic profiles for each population the drug is being administered to. However, there is much controversy about whether enforced HLA-B*15:02 testing is cost-effective (Rattanavipapong et al. 2013; Chong et al. 2017). The implementation of a testing campaign, within Hong Kong that examined more than 100,000 patients, showed a decrease in the prescription of carbamazepine and a decrease in related SJS/TEN (Chen et al. 2014). However, the prescription of other antiseizure medications increased in this study and examining this group showed no overall change in SJS/TEN (Chen et al. 2014). The authors report, clinicians preferring antiepileptic drug (AED) therapy that do not require genetic screening are preferred, this exemplifies the need for educational programmes to supplement the implementation of testing programmes.

An example of the type of infecting organism found within a person also plays a role on the type of drug administered. Maraviroc, an HIV drug that binds onto the HIV entry receptor, CCR5, prevents the virus entering into the cell for viral replication. However, the desired effect of maraviroc is achieved within individuals that possess only the CCR5 receptor strain of HIV. The virus also uses CXCR4 to enter into the cell. Maraviroc does not function with CXCR4 strains of HIV, therefore, health providers need to determine the receptor usage before prescribing the drug (Ryst 2005).

Pharmacogenomics and drug development

A huge amount of effort and resources goes into the development of drugs to treat a range of diseases and disorders, however, the number of drugs candidates entering clinical trials are very small in number. Increasing evidence suggests that studies examining the human genetic components of the drug have an increased likelihood of successful marketing than those drugs not testing the genetics (Sanseau et al. 2012; Nelson et al. 2015; Zhang et al. 2015). Drugs that have proved to be successful in clinical trials still have remarkable failure rates when these drugs are given to the target

audience. Since most of these trials are being performed in developed countries and do not take into account the genetic effect on individuals from underdeveloped countries not knowing the human genetics within these populations hinders the desired effect of the drug.

Ideally, when a drug is created for an individual or individuals within a population, their genetic diversity must be considered. Since the assembly of the first human genome there has been an explosion in the generation of genomic data with projects providing full length whole genome sequences that are curated and annotated. Using this data, there is evidence that the haplotype frequency and overall genetic diversity is much higher in Africans than Caucasians and Asians (Rotimi et al. 2017). Furthermore, within Africans there is an additional spread of genetic diversity which suggests that drugs need to be designed for specific populations. For example, a drug created to treat African individuals from Northern Africa should be reconstructed to treat African individuals in Southern Africa due to their genetic differences. In regions where there is minimal diversity, genetic screening can be used to test whether a drug is suitable for that candidate.

Currently, there are a number of platforms that offer the ability to test the suitability of a drug. The cheapest and smallest scale screening is the genotyping approach which works using polymerase chain reaction (PCR) technology to determine the presence of single variants. This can be used when a single SNP has an effect on drug efficacy. Another approach is the use of genome-wide association study (GWAS) to identify the variants carried by an individual. This test provides a more comprehensive picture than the single variant genotyping. Special chips can be designed to suit a specific population or trait. WGS provides the maximum genomic information for an individual. Patients who screen their whole genome can be assessed not only for genetic variants that alter drug efficacy but also a myriad of information such as their susceptibility to infectious diseases as well as the likelihood of developing heritable diseases.

Pharmacogenomics is a step in the direction towards precision medicine and may help to reduce the burden of disease by improving outcomes with specialized drugs. The epigenome has been shown to have a significant effect on gene regulation and interactions between the genome and epigenome may provide an additional layer of complexity to the field of pharmacogenomics. The use of WGS and whole genome methylation in

conjunction with pharmacology will allow us to understand much more about the landscape for treating disease.

Epigenetic therapy

The field of epigenetics has rapidly expanded in the past few decades to the extent that novel forms of therapy are emerging from this field that warrant discussion in this chapter. It is well established that epigenetic changes contribute to the development of human diseases such as cancer, autoimmune disease, as well as neurodegenerative and psychological disorders (Hahn and Weinberg 2002; Jones and Baylin 2002; Zoghbi and Beaudet 2016). Epigenetics refers to all heritable alterations in gene expression and chromatin structure due to chemical modifications that do not involve changes in the primary nucleotide sequence (Strahl and Allis 2000; Jenuwein and Allis 2001; Nightingale et al. 2006; Bernstein et al. 2007). Epigenetic mechanisms have the ability to either appropriately regulate cell growth (Portela and Esteller 2010; Moosavi and Motevalizadeh Ardekani 2016) or in contrast, dysregulation of epigenetic patterns lead to induction and propagation of disease states (Hahn and Weinberg 2002; Jones and Baylin 2002; Feinberg et al. 2006).

There are several factors that contribute to epigenetic alterations, the factors with the strongest contribution include DNA methylation, histone modifications, and non-coding RNAs. All of these are crucial mechanisms that ensure the stable propagation of gene activity from one generation of cells to the next (Jenuwein and Allis 2001; Feinberg et al. 2006). Disruption of any of these three distinct and mutually reinforcing epigenetic mechanisms leads to inappropriate gene expression, resulting in cancer development and other ‘epigenetic diseases’ (Jones and Baylin 2002; Egger et al. 2004; Feinberg et al. 2006). In some instances, pathogens have been able to take advantage of epigenetic alterations for their propagation.

In recent years, scientific interest in epigenetics has increased substantially so that this field represents an important new avenue to advance our understanding of pathogenesis and to help in the development of strategies for disease. It is hoped that by understanding epigenetic mechanisms effective therapies would soon become available. Epigenetic therapy, a relatively new outlook to treatment, involves the use of drugs or

epigenome altering tools to change the non-DNA environment to treat medical conditions. There are about 191 clinical trials around the world involving epigenetic therapy, according to the NIH clinical trials database, indicating the potential of this rapidly growing field (NIH 2019). Most of these clinical trials have targeted different forms of cancer, however there are several other diseases which also have been targeted. Drug reposition efforts have been launched to explore epigenetics therapies for non-oncology conditions (Best and Carey 2010). Some of these conditions include diabetes mellitus, depression, rheumatoid arthritis, schizophrenia, etc.

Clinical conditions treated with epigenetic therapy

DNA methyltransferase (DNMT) enzymes are responsible for the addition of methylation to DNA, these enzymes are inhibited with the drugs 5-azacytidine and 5-aza-20-deoxycytidine. Both drugs have been approved for use by the FDA for the treatment of various forms of cancer. These drugs have been shown to reactivate the cellular antitumor systems repressed by the cancer, enabling the body to weaken the tumour (Foulks et al. 2012; Vendetti and Rudin 2013; Li et al. 2014; Wells et al. 2014). Zebularine, an activator of a demethylation enzyme, has also been used with some success (Kowluru et al. 2013). The wide-ranging effects of these drug on the entire genome may result in major side effects, however, survival rates are significantly increased when the drugs are used for treatment.

The largest class of epigenetic therapies is inhibitors of histone-modifying enzymes. These include drugs that inhibit histone methyltransferases (HMTs) and histone deacetylases (HDACs) and are the most prevalent form of epigenetic therapy. In addition, there are several compounds that activate HDACs or inhibit HDMs, currently being developed. These include inhibitors of the H3K79 methyltransferase DOT1L for the treatment of Mixed-lineage leukaemia (MLL)-rearranged leukaemias (Barski et al. 2007). The treatment of non-Hodgkin lymphoma involves the inhibitors of the H3K27 methyltransferase EZH2, which form the catalytic component of the PRC2 complex (Bradley et al. 2014). Use of HDAC inhibitors, specifically HDAC6, are being explored for the treatment

of certain cancers such as multiple myelomas and a wide variety of haematologic malignancies (Li et al. 2018). Epigenetic therapy also includes the use of bromodomain and extraterminal motif (BET) bromodomain inhibitors. Bromodomains are evolutionary conserved protein-protein interaction modules that preferentially recognize acetylated chromatin (Fujisawa and Filippakopoulos 2017). The use of BET bromodomain inhibitors that possess broad binding specificity have been tested for treatment of advanced haematologic and non-haematologic malignancies (Sanchez and Zhou 2009).

Treatment of diabetic retinopathy using epigenetic therapy has also been explored. One therapeutic approach includes inhibition of methylation within manganese superoxide dismutase (SOD2) and matrix metalloproteinase (MMP-9). The DNMT inhibitors 5-azacytidine and 5-aza-20-deoxycytidine inhibit the SOD2 and MMP-9 methylation patterns and have been reported to reduce diabetic retinopathy symptoms (Kowluru et al. 2013). Limited success has also been reported for the treatment of this disorder with histone acetyltransferase (HAT) inhibitors epigallocatechin-3-gallate, vorinostat, and romidepsin (Kowluru et al. 2013).

The epigenetic mechanisms of lupus have been studied in detail, which has assisted the use of epigenetic-based therapies into clinical trials against lupus. A cytidine analogue, Cytarabine, reported to increase DNA methylation, was used for treatment of refractory cutaneous lupus in a human clinical trial (Mishra et al. 2003). Results showed relapse of disease during the fourth week in two-thirds of tested cases, however, treatment with Cytarabine also showed rapid improvement of refractory cutaneous lupus in treated patients (Yung and Richardson 1995).

There have been many experimental drugs that target epigenetic mechanisms against rheumatoid arthritis (RA). DNMT inhibitors such as 5-azacytidine was used to treat normal synovial fibroblasts and this reproduced RA synovial fibroblasts phenotypes (Kooistra and Helin 2012). Trichostatin A (TSA) and nicotinamide, both HDAC inhibitors, were separately used to reduce TNF-α which in turn reduces IL-6 expression in macrophages isolated from PBMCs of RA patients (Gillespie et al. 2012). Other studies have shown that targeting HDACs, using HDACi such as valproic acid, can mitigate the severity of murine chronic inflammatory arthritis in part via improving regulatory T-cell function (Chung et al. 2003; Saouaf et al. 2009). These show potential therapeutic value, due to their

ability to reduce or disrupt the production of inflammatory signals that fuel RA disease activity. Another promising study silenced SIRT1 and promoted apoptosis in RA synovial fibroblasts from synovial tissues and cells from RA patients (Niederer et al. 2011).

Multiple sclerosis (MS) is another clinical disorder for which investigators have considered epigenetic therapy for treatment. Histone deacetylases inhibitor, such as TSA, has reduced spinal cord inflammation, demyelination, neuronal and axonal loss in experimental autoimmune encephalomyelitis (EAE) models, and thus serves as a potential treatment option for MS (Camelo et al. 2005). However, HDAC inhibitors used at the systemic level have shown to negatively affect the generation of new myelin, since HDAC is crucial to developmental of myelination (Marin-Husstege et al. 2002; Shen et al. 2005). Chirivi et al. have proposed citrullination as a target for epigenetic intervention in MS and other inflammatory diseases (Chirivi et al. 2013). The authors also postulate that shielding citrullinated histone epitopes, which play a large role in NET formation, can help prevent the intensification of the inflammatory response in MS and other inflammatory diseases (Chirivi et al. 2013).

Advantages and disadvantages of epigenetic therapy

Unlike genetic defects, epigenetic defects are potentially reversible and are therefore more amenable to corrective therapeutic reversal. This realization has driven the rapid emergence and expansion of epigenetic drug discovery, and drug-targeting epigenetic enzymes are in the clinic for the treatment of human disorders, especially for clinical management of haematological malignancies and neurological disorders. However, treatments with epigenetic drugs still need to overcome the challenge of avoiding off-target effects. It should be considered that multiple epigenetic enzymes are contained in chromatin complexes, and as a result the inhibition of an epigenetic enzyme can result in changes in other members of the same protein complex. For example, treatment with the HDAC inhibitor sodium butyrate alters the acetylation level of histones, but also affects their methylation status. In addition, drug development based on epigenetics is difficult and expensive similar to other drug targets.

Epigenomic technologies

Similar to GWAS, epigenome-wide association studies (EWAS) could be a diagnostic tool for screening epigenetic disorders. An EWAS is an examination of a genome-wide set of quantifiable epigenetic marks, such as DNA methylation, in different individuals to derive associations between epigenetic variation and a particular identifiable phenotype/trait (Flanagan 2015). Bisulfite conversion is used for the detection of DNA methylation levels, methylated cytosines are saved from conversion to uracil and thus measured using WGS. Whole genome and RNA sequencing are valuable tools for the screening of non-coding RNAs. Continued study of genome sequences can help determine whether new genetic variations are associated with health conditions.

Epigenetic therapy in the future

Future studies may be able to identify methods for recognition response mechanisms by combining genomic sequencing and gene expression profile. In addition, DNA could also undergo histone modifications, phosphorylation, ubiquitination, sumoylation, and acetylation. Each of these mechanisms could provide novel opportunities for epigenetic therapies in the future. The use of non-coding RNA (ncRNA) is also an avenue that is currently being explored. MicroRNAs (miRNA), a class of ncRNA, that are about 19–24 nucleotides long, prevent the translation of mRNA and protein production (Nam et al. 2009). As miRNAs play critical roles in regulating functions of the cells, disruption in their structure and turnover can cause diseases. There are 651 miRNA clinical trials reported on the NIH clinical trials database (NIH 2019) highlighting the growing interest of this avenue for treatment of clinical disorders. The use of specific miRNA mimics and inhibitors that modulate miRNAs and regulate gene expression could be an attractive avenue for future research. miRNA mimics have shown therapeutic benefit in mouse models of lung cancer (Trang et al. 2011; Kasinski et al. 2015). Therapeutic delivery was demonstrated using mimics of the tumour suppressors, microRNA-34a (miR-34a) and let-7, both of are often downregulated in lung cancer. Specifically, mice treated with miR-34a displayed a 60% reduction in the tumour area compared to mice treated with a miRNA control (Trang et al. 2011). Similar results were obtained

with the let-7 mimic. As a single miRNA may be able to target more than one gene it would be interesting to combine several microRNA with different targets so that various proteins involved in different steps of a disease can be regulated, enabling a combinatorial, multitarget approach that could complement combination drug treatment.

Long non-coding RNA (lncRNAs) should be taken into account as future advances in epigenetic therapy, since lncRNAs could also be regulated by epigenetic mechanisms, including CpG methylation as well as histone marks (Kung et al. 2013). They are involved in chromatin remodelling, as well as transcriptional and post-transcriptional regulation, through a variety of chromatin-based mechanisms and via cross-talk with other RNA species (Zampetaki et al. 2018). There are currently 32 clinical trials for lncRNA according to the NIH clinical trials database (NIH 2019). The data gathered so far concerning the involvement of lncRNA in epigenetic regulation is still the tip of the iceberg of this emerging field, and further expansion of these data may reveal untapped targets and pathways amenable to epigenetic therapy. In addition, as mentioned earlier, screening for epigenetic diversity across diverse populations prior to administering epigenetic therapies will assist with the success of the treatment.

Genome editing and public health

Genome or gene editing provides the ability to edit an organism's genetic material to achieve a desired trait. Using genome editing technologies researchers can modify precisely specific additions, deletions, and alterations to specific sites within the genome. There are several tools used for genetic editing, however since the advent of clustered regularly interspaced short palindromic repeats (CRISPR) there has been much excitement. The reason for the excitement is mainly due to CRISPR being faster, cheaper, increased in accuracy and efficiency when compare to the existing genome editing methods. Some of the common methods used prior to the arrival of CRISPR include meganucleases, zinc fingers nucleases (ZFNs), and transcription activator-like effector nucleases (TALENS).

Meganucleases

These are endonuclease molecules, also called homing endonucleases, that can be grouped into five families (Pingoud and Silva 2007). The large recognition site, that is rarely found within an entire genome, allows this tool to function highly specifically for the editing of genomes. The meganucleases carry the genetic portion that influences recombination during homing (Orlowski et al. 2007). Meganucleases have been implemented in an industrial scale to target research, bio-manufacturing, bio-production, and have a potential role in human healthcare. A limitation of this tool is that small sequences that may play a role in disease outcome are not targeted.

Zinc finger nucleases (ZFNs)

ZFNs are fusion proteins that have been adapted from zinc finger containing transcription factors, which possess numerous site-specific DNA-binding domains. These fusion proteins are linked to the bacterial FokI restriction enzyme, which forms the endonuclease domain (Gupta and Musunuru 2014). A single zinc finger domain possesses the ability to recognize and bind a 3- to 4-bp DNA fragment, the combination of domains allows for the ability to detect a large target DNA sequence. Cleavage of a site in the DNA sequence is bi-directional using two flanking sequences. Once the ZNFs are bound at a specific site in the genome, the two FokI domains dimerize and cuts the DNA, which generates a double stranded break (Urnov et al. 2010; Gupta and Musunuru 2014). A major limitation of this tool has been found to be the off targets effects. For a more precise edit, more zinc finger domains are required, which increases costs of the application (Urnov et al. 2010).

Transcription activator-like effector nucleases (TALENS)

This endonuclease has a similar mechanism of action to ZFN. However, instead of ZFN binding to three nucleotides (codon), each nucleotide has its own zinc finger protein (33 to 35 amino acids), which are much larger than ZFNs (Gupta and Musunuru 2014; Miller et al. 2010). In addition to the length advantage, TALENS are also easier to design and have fewer constraints of site selection than the ZFNs. Limitations of TALENS is the large size and their highly repetitive nature (Gupta and Musunuru 2014).

CRISPR-Cas9

CRISPR is a gene editing tool based on the bacterial adaptive immune system. Bacterial cells retrieve protospacers from invading bacteriophages and incorporates them within their genomes. These protospacers are then used as short guide RNAs by a CRISPR-cas system and target any foreign DNA possessing the matching incorporated protospacer sequence. The invading DNA is cut using the cas9 enzyme, which enables cutting of the double stranded DNA. The CRISPR-Cas9 currently being used in research labs is mimicked from the bacterial system (Hsu et al. 2014). Several versions of a modified Cas9 protein exists, which enables CRISPR gene editing system to be the most versatile and precise tool on the market.

CRISPR-based clinical trials

CRISPR has the potential to be used as an editing tool for genetic diseases. Recently, a group of scientists produced CRISPR engineered mouse pluripotent stem cells, which are resistant to proinflammatory cytokines. Not only do these engineered cells prevent inflammation, they also alleviate inflammation in chronic inflammatory diseases such as arthritis (Brunger et al. 2017). In 2018, the mutated *DMD* gene, which causes Duchenne muscular dystrophy, was edited in dogs using the CRISPR-Cas9 system and delivered using an Adeno-associated virus. As a result of CRISPR-Cas9 mediated editing, dystrophin expression was restored in a canine model of Duchenne muscular dystrophy (Duan 2018).

Following the success of CRISPR as a therapy against disease in animals, many human clinical trials against various diseases have been initiated. According to the NIH clinical trials database (www.clinicaltrials.gov), a total of 28 CRISPR-based clinical trials are underway by multiple groups targeting a range of diseases including HIV-1 and other viral infections, sickle cell disease, thalassemia, Kabuki syndrome, Leber congenital amaurosis (blindness), malignant hyperthermia, various leukaemias and lymphomas, as well as various cancers (NIH 2019).

Interest was sparked in 2016, the first phase I pre-clinical trial involving CRISPR gene editing technology in cancer research in China, as well as the second CRISPR initiative in the treatment of two debilitating inherited blood disorders, beta thalassemia (TDT) and sickle cell disease (SCD), both

in the United States and by the Swiss (Baylis and McLeod 2017; Nasdaq:CRSP 2019). Currently, these two CRISPR trials are being approved to edit immune cells against tumours and to activate genes that are inactive, making these the initial two out of several human trials that are underway. These treatment approaches may become a novel and powerful therapeutic intervention against various other diseases as well.

The trial entitled CTX001-121 aims to treat various hemoglobinopathies such as severe SCD and thalassemia. The first patient to be treated with the CTX001 drug had a form of thalassemia. Following treatment, the patient reportedly did not require any blood transfusions (the usual treatment for thalassemia) for 4 months. The first patient with severe SCD was also treated with CXT001 in 2019 (NIH Clinical Trials 2019). There are also plans by the companies behind this trial to run gene therapy trials on patients suffering with Duchenne muscular dystrophy (DMD) and myotonic dystrophy type 1 (MD1) (Wang et al. 2018).

In addition to CRISPR, the FDA has approved other gene-based therapies to treat an array of diseases. These include (1) Luxturna, used for the treatment of a different form of Leber congenital amaurosis (blindness—retinal dystrophy) (Maeder et al. 2019); (2) Zolgensma, approved for the treatment of SMA, a leading cause of infant mortality (Pena et al. 2020); and iii) Kymriah, which is an approved cell based therapy for the treatment of some blood cancers such as acute lymphoblastic leukaemia (ALL) (Montano et al. 2018).

CRISPR's potential to hit drug markets as a prescribed therapy against not only genetic but also against infectious diseases is huge. HIV is one infectious disease that has been targeted by CRISPR technology. Using the Cas9/sgRNA parts of the HIV genome was targeted for cleavage in an attempt to decrease the replication of the virus, however, the virus was able to overcome CRISPR editing by escape mutations located within the cas9 cleavage site of the virus (Wang et al. 2016a, b). The therapeutic potential of CRISPR needs further exploration and it is estimated within the next few decades the genome editing tool will play a major role in personalized and public health.

Advantages and limitations of CRISPR

The important advantage of the CRISPR Cas9 system over other gene editing technologies is the simplicity, cost-effectiveness, and efficiency. In addition, the technology has the potential to be applied directly to embryos to modify target genes, which enables rapid desired affects when compared to alternative gene editing technologies that require use of embryonic stem cells. Another benefit of the CRISPR system is the improved bioinformatic tools that are available to design guide RNAs. Optimizing of experimental conditions is simple and robust, which almost guarantees introduction of the desired edit will be successful. Furthermore, CRISPR can edit virtually any genome, this provides it with the capability to treat diseases, destroy infecting microorganisms, resurrect species, and create healthier foods without the use of antibiotics (Omodamilola and Ibrahim 2018). Additionally, the CRISPR tool proves to be far superior to competitors, due to its ability to modify multiple genes in a single experiment, up to 25 targets have been reported successfully (Campa et al. 2019). CRISPR possesses yet another benefit over competing gene editing technologies with its ability to alter the epigenetic environment. This is achieved by using a deactivated form of the cas9 protein, which does not cut or edit the DNA sequence but instead changes the surrounding epigenetic environment with accessory molecules attached to deactivate cas9 (Bae et al. 2019).

One of the main limitations of the CRISPR technology is the off-target effects. The guide RNA, which is intended to bind onto a specific target may also bind on other unintended sites resulting in unplanned altering or deactivating of genes, thus producing side effects. The simplicity of the CRISPR tool may also be its downfall, questions about the tools ability to create bioweapons or designer babies are raising significant ethical concerns.

Ethics of CRISPR

The CRISPR technology received a significant amount of negative attention from November 2018, after the first human babies created from edited embryos were revealed to the world by the Chinese scientist, Dr Jianku He. This sent the scientific research community into debates regarding the ethical, legal, and social implications regarding the germline editing (Kleiderman and Ogbogu 2019). The research has violated Chinese regulations and international bioethical standards. The Chinese regulations

on gene editing seem to align with international standards (Isasi et al. 2016; Kleiderman and Ogbogu 2019). In addition to the ethics concerns of the 'CRISPR babies' experiment, researchers have also questioned the soundness of the research conducted. The gene targeted for editing, CCR5, is required for viral entry only for the CCR5 strain of the virus, however the CXCR4 viral strain can still be transmitted despite altering the CCR5 gene. There is currently a drug, maraviroc, against CCR5 that has been used in numerous HIV clinical trials. Individuals with mutated forms of CCR5 have an increased risk of symptomatic West Nile virus infection (Glass et al. 2006). These events have certainly created more focused attention on human genetic editing, despite this study not being the first to edit human genes in an embryo (Bonora et al. 2015). Researchers in 2015, reported CRISPR/Cas9-mediated editing of endogenous beta-globin gene in human embryos (Liang et al. 2015). The authors also reported unplanned off-target effects as determined by whole exome sequencing and cautioned the pressing need to further improve the fidelity and specificity of the CRISPR/Cas9 platform (Liang et al. 2015).

Any future clinical trials aiming to alter human embryos should be subject to much discussion, especially for the need of such an experiment. Due to the simplicity of the technique, potential to leap into the public platform, and the convincing potential applications, it probably wouldn't be too long before another unorthodox individual announces the results of their human embryo editing experiments.

The US National Academy of Sciences and the National Academy of Medicine constituted a multidisciplinary, international committee to review the current status and make recommendations based on the ethical concerns raised (Coller 2019). The report makes recommendations, which include experimental germline genome editing that is subject to restrictions that include research focused on preventing transmission of a serious disease or condition, modification of well-documented DNA changes associated with disease, and research conducted under a set of stringent ethical and regulatory guidelines (Coller 2019). However, with simplicity of use and easy accessibility, mail-order do-it-yourself CRISPR kits are currently available for a few hundred US dollars, experts are concerned about what amateur scientists may achieve with the powerful tool (Sneed 2017). Is the world ready to implement such strategies or should CRISPR remain in a laboratory setting until future stable gene editing techniques are developed?

There is also the growing concern of CRISPR being used as a biotechnological threat (Vogel and Ouagrham-Gormley 2018). This technology can make a treatable pathogen become deadly by altering the drug target sites.

Another major concern is, if CRISPR germline editing eventually becomes regulated and allowed in the future, who will the technology be accessible to? Will it simply be for the rich and famous? The main challenge for nations experimenting with gene editing is to come up with feasible, rational, and ethical guidelines for its practical implementation. CRISPR is still a developing technology, which provides a strong foundation for the development of new advanced tools in the field of gene editing.

Ethical implications of genomics and public health

The rapid advances in the field of genomics and public health arises with a new set of ethical problems that need to be addressed. A framework by Burke et al. highlights several important aspects of improved population health (Burke et al. 2006). These include engaging with a range of stakeholders that are within and external to the public health and scientific communities. The aim is to provide a support of core activities, some of these include; advising on public policy; design, implementation, and evaluation of preventive and clinical health services; communication and stakeholder engagement; and finally dissemination of knowledge via education and training (Burke et al. 2006). Roberts and colleagues have reviewed in detail some of the emerging issues in public health genomics (Roberts et al. 2014). We provide an updated summary of these emerging issues in this chapter.

Informed consent

Genomic testing can provide specific confirmation or diagnosis as described earlier, however in certain instances genomic testing may raise more questions, i.e. some tests may not provide meaningful information within the present setting. However, after ongoing interpretation of the results, researchers may find relevant information in the future, therefore

the informed consenting process needs to be adequate to deal with these wide range of outcomes (Horton and Lucassen 2019). Several researchers believe we are not achieving valid consent in the genomic setting, since the consent to be valid 'the patient needs to be competent; they need to be appropriately informed; and their consent needs to be voluntarily given' (Horton and Lucassen 2019). The process of informing a patient appropriately means possessing a clear understanding of the genetic results obtained, this is in most instances not the case. Even within experts in the field, as demonstrated by the APOE status of Dr James Watson, one of the first people to be sequenced. He released his sequenced genome to online public databases, except for apolipoprotein E (ApoE), which was redacted due to its linkage with Alzheimer's disease, Dr Watson preferred not to know his status (Nyholt et al. 2009). However, based on linkage information found throughout the rest of the genome, the Alzheimer's information could be ascertained. The linkage information located in other parts of the genome was only discovered later (Nyholt et al. 2009; Horton and Lucassen 2019). Does this mean the informed consent was violated? If so, what could have been done to better protect Dr Watson? These are some of the questions that should be asked in future when achieving consent in the genomic setting. Dr Watson's case is just one of many scenarios where public health genomics have left patients with undesirable outcomes. Patients who participate in genomic research should also be informed of both the risks and benefits of testing (Department of Health et al. 2014; Miller et al. 2008).

Prenatal screening

Within the United States alone, nearly all the 4 million infants born a year are screened for rare genetic diseases (Roberts et al. 2014). Genetic screening has enabled parents to make informed choices regarding the welfare of their child's health. Testing for Down's syndrome is especially common in pregnant women over 50 years of age as there is an increased risk of non-disjunction during meiosis resulting in ova that could lead to the birth of a child with Down's syndrome (Reynolds 2003; Dukhovny and Norton 2018). Some of the symptoms of Down's syndrome include learning disabilities, stunted or delayed growth as well as characteristic facial features, however, many of the more serious health problems associated

with this disease can be treated. The life expectancy of children with Down's has increased dramatically as there is now reduced stigma and better healthcare available. Genetic testing provides the ability for a positive Down's syndrome infant to be treated earlier, however, certain parents choose the option of premature pregnancy termination (Dukhovny and Norton 2018). As the number of prenatal genetic tests become available, there needs to be a balance of autonomy and distributive justice (Dukhovny and Norton 2018).

Data storage and sharing

Sharing of genomic data has become an actuality with the advent of genomic databases that allow access to thousands of whole genome sequences for the purpose of scientific advancement (Mathaiyan et al. 2013; Dijkers 2019). At the point of data collection, is the participant privy to their own information? Does that information remain within the institute or is it shared with other institutes or companies? As WGS becomes more affordable the risk of data sharing with employers and health insurances or medical aids becomes a potential problem. If an individual who works in a laborious field is predicted to develop Parkinson's, sensitive information such as this may lead the employer to retrench him/her due to this information. Additionally, information like this may cause their insurance premiums to skyrocket or medical aid be denied or delayed. Access to such sensitive data can lead to its abuse allowing government, employers, and companies to exploit individuals (Dijkers 2019).

The need for confidentiality cannot be disputed when performing genomic research. The genetic data from WGS gives important information about current or potential diseases as well as risk of disease. In the case of rare diseases, it may be easy for the identity of the patient to be leaked because of the rarity of the disease itself (Dijkers 2019). The breach of confidentiality could expose the individual to discrimination in the workplace as well as in social settings. Proper channels for blinding patient data are necessary similar to what is performed during clinical trials, however, there needs to be additional measures to ensure that there is no leak in data.

Storage of samples

Biological specimens can be frozen for many years at a time. Researchers use these samples to answer many questions and may store them for use in future studies. Indefinite storage of samples means that the research being conducted on these samples will never end (Vaz and Srinivasan 2014). Although consent may be given by the participant for a particular study or overall type of study there is no definite way for them to know what will be done on the samples in the future (McGuire and Beskow 2010). Generally, when consent is given there is very little a patient can do regarding the studies being performed on their samples. Considering the stability of DNA, these risks to privacy of genetic and personal information does not decrease with time. Considering the long-term storage and dissemination of samples, should researchers constantly request updated consent, where the patient is informed of new or ongoing studies and recompensated (Haga and Beskow 2008)?

Conclusion

The field of public health genomics has been growing rapidly over the past few years and is projected to increase steadily in the future. The significant improvement of new technologies has made genomics accessible within the public health setting. An explosion of genetic tests examining conditions that range from several rare diseases to a variety of common diseases have started to reach the market. Caution should be implored when a new genomic test reaches the market. Each test needs to meet certain diagnostic requirements within a target population before it is implemented as a genomic test within the public health sector. In addition, pharmacogenomics testing is encouraged before a new drug is administered within diverse ethnic groups.

Advancements in the field of epigenetics has led to new therapeutic treatment options against a range of conditions. Epigenetic therapy is in its early stages and has enormous potential for future applications. Despite setback within the field, gene therapy is still argued to be the treatment of the future. Significant policies and ethics guidelines are required to be reviewed and implemented for gene therapy to be utilized as a therapeutic tool in the future. Public health genomics is an exciting area that can benefit

a larger segment of the population by bringing the right interventions to the right population at the right time.

REFERENCES

Armstrong, N., et al. (2019). A systematic review of the international prevalence of BRCA mutation in breast cancer. *Clinical Epidemiology*, **11**, 543–61.

Bae, T., et al. (2019). Recent trends in CRISPR-Cas system: genome, epigenome, and transcriptome editing and CRISPR delivery systems. *Genes Genomics*, 41(8), 871–7.

Barski, A., et al. (2007). High-resolution profiling of histone methylations in the human genome. *Cell*, 129(4), 823–37.

Baylis, F. and McLeod, M. (2017). First-in-human phase 1 CRISPR gene editing cancer trials: are we ready?. *Current Gene Therapy*, 17(4), 309–19.

Bellagio (2019). Report of an expert workshop held at the Rockefeller Foundation Study and Conference Center. [Online] Available at: http://www.phgfoundation.org/documents/74_1138619841.pdf

Berg, J.S., et al. (2017). Newborn sequencing in genomic medicine and public health. *Pediatrics*, 139(2), e20162252.

Bernstein, B.E., Meissner, A., and Lander, E.S. (2007). The mammalian epigenome. *Cell*, 128(4), 669–81.

Besser, J., et al. (2017). Next-generation sequencing technologies and their application to the study and control of bacterial infections. *Clinical Microbiology and Infection*, 24(4), 335–41.

Best, J.D. and Carey, N. (2010). Epigenetic therapies for non-oncology indications. *Drug Discovery Today*, 15(23–24), 1008–14.

Bonora, M., et al. (2015). Molecular mechanisms of cell death: central implication of ATP synthase in mitochondrial permeability transition. *Oncogene*, 34(12), 1608.

Bradley, W.D., et al. (2014). EZH2 inhibitor efficacy in non-Hodgkin's lymphoma does not require suppression of H3K27 monomethylation. *Chemistry & Biology*, 21(11), 1463–75.

BRCA Exchange (2019). *BRCA Exchange*. [Online] Available at: https://brcaexchange.org/

Brunger, J.M., et al. (2017). Genome engineering of stem cells for autonomously regulated, closed-loop delivery of biologic drugs. *Stem Cell Reports*, 8(5), 1202–13.

Burke, W., et al. (2006). The path from genome-based research to population health: development of an international public health genomics network. *Genetic Medicine*, 8(7), 451–8.

Burke, W. (2014). Genetic tests: clinical validity and clinical utility. *Current Protocols in Human Genetics*, **81**, 9.15.1–9.15.8.

Camelo, S., et al. (2005). Transcriptional therapy with the histone deacetylase inhibitor trichostatin A ameliorates experimental autoimmune encephalomyelitis. *Journal of Neuroimmunology*, 164(1–2), 10–21.

Campa, C.C., et al. (2019). Multiplexed genome engineering by Cas12a and CRISPR arrays encoded on single transcripts. *Nature Methods*, 16(9), 887–93.

Centers for Disease Control and Prevention (CDC) (2019). Public Health Genomics. [Online] Available at: https://www.cdc.gov/genomics/pathogen/index.htm

Ceyhan-Birsoy, O., et al. (2017). A curated gene list for reporting results of newborn genomic sequencing. *Genetic Medicine*, 19(7), 809–18.

Chen, Z., Liew, D., and Kwan, P. (2014). Effects of a HLA-B*15:02 screening policy on antiepileptic drug use and severe skin reactions. *Neurology*, 83(22), 2077–84.

Chirivi, R., et al. (2013). Citrullination: a target for disease intervention in multiple sclerosis and other inflammatory diseases? *Journal of Clinical Cell Immunology*, **4**, 3.

Chong, H.Y., et al. (2017). Is universal HLA-B*15:02 screening a cost-effective option in an ethnically diverse population? A case study of Malaysia. *British Journal of Dermatology*, 177(4), 1102–12.
Chung, W.H., et al. (2004). Medical genetics: a marker for Stevens–Johnson syndrome. *Nature*, 428(6982), 486.
Chung, Y. L., et al. (2003). A therapeutic strategy uses histone deacetylase inhibitors to modulate the expression of genes involved in the pathogenesis of rheumatoid arthritis. *Molecular Therapy*, 8(5), 707–17.
Clayton, E.W. (2003). Ethical, legal, and social implications of genomic medicine. *The New England Journal of Medicine*, **349** (6), 562–9.
Coller, B.S. (2019). Ethics of human genome editing. *Annual Review of Medicine*, **70**, 289–305.
Department of Health, et al. (2014). The Belmont Report. Ethical principles and guidelines for the protection of human subjects of research. *Journal of the American College of Dentists*, 81(3), 4–13.
Dijkers, M.P. (2019). A beginner's guide to data stewardship and data sharing. *Spinal Cord*, 57(3), 169–82.
Duan, D. (2018). CRISPR alleviates muscular dystrophy in dogs. *Nature Biomedical Engineering*, 2(11), 795–6.
Dukhovny, S. and Norton, M.E. (2018). What are the goals of prenatal genetic testing?. *Seminal Perinatology*, 42(5), 270–4.
Egger, G., et al. (2004). Epigenetics in human disease and prospects for epigenetic therapy. *Nature*, 429(6990), 457–63.
Evans, C.A. (2011). GeneXpert—a game-changer for tuberculosis control?. *PLoS Medicine*, 8(7), e1001064.
Fan, H.C., et al. (2012). Non-invasive prenatal measurement of the fetal genome. *Nature*, 487(7407), 320–4.
Faria, N.R., et al. (2017). Establishment and cryptic transmission of Zika virus in Brazil and the Americas. *Nature*, 546(7658), 406–10.
Faria, N.R., et al. (2018). Genomic and epidemiological monitoring of yellow fever virus transmission potential. *Science*, 361(6405), 894–9.
Feinberg, A.P., Ohlsson, R., and Henikoff, S. (2006). The epigenetic progenitor origin of human cancer. *Nature Review and Genetics*, 7(1), 21–33.
Flanagan, J.M. (2015). Epigenome-wide association studies (EWAS): past, present, and future. *Methods in Molecular Biology*, **1238**, 51–63.
Fordyce, C.B. and Douglas, P.S. (2017). Outcomes-based CV imaging research endpoints and trial design: from pixels to patient satisfaction. *JACC: Cardiovascular Imaging*, 10(3), 253–63.
Foulks, J.M., et al. (2012). Epigenetic drug discovery: targeting DNA methyltransferases. *Journal of Biomolecular Screening*, 17(1), 2–17.
Friedman, E. (2015). Next generation sequencing for newborn screening: are we there yet? *Genetic Research (Cambridge)*, **97**, e17.
Fujisawa, T. and Filippakopoulos, P. (2017). Functions of bromodomain-containing proteins and their roles in homeostasis and cancer. *Nature Reviews Molecular Cell Biology*, 18(4), 246–62.
Giacomini, M., Miller, F., and Browman, G. (2003). Confronting the 'gray zones' of technology assessment: evaluating genetic testing services for public insurance coverage in Canada. *International Journal of Technology Assessment in Health Care*, 19(2), 301–16.
Gillespie, J., et al. (2012). Histone deacetylases are dysregulated in rheumatoid arthritis and a novel histone deacetylase 3-selective inhibitor reduces interleukin-6 production by peripheral blood mononuclear cells from rheumatoid arthritis patients. *Arthritis Rheumatism*, 64(2), 418–22.

Glass, W.G., et al. (2006). CCR5 deficiency increases risk of symptomatic West Nile virus infection. *Journal of Experimental Medicine*, 203(1), 35–40.
Gupta, R.M. and Musunuru, K. (2014). Expanding the genetic editing tool kit: ZFNs, TALENs, and CRISPR-Cas9. *Journal of Clinical Investment*, **124** (10), 4154–61.
Gwinn, M., MacCannell, D., and Armstrong, G. L. (2019). Next-generation sequencing of infectious pathogens. *The Journal of the American Medical Association*, 321(9), 893–4.
Haga, S.B. and Beskow, L.M. (2008). Ethical, legal, and social implications of biobanks for genetics research. *Advanced Genetics*, **60**, 505–44.
Hahn, W.C. and Weinberg, R.A. (2002). Rules for making human tumor cells. *The New England Journal of Medicine*, 347(20), 1593–603.
Hall, M. and Olopade, O.I. (2005). Confronting genetic testing disparities: knowledge is power. *The Journal of the American Medical Association*, 293(14), 1783–5.
Harris, R.P., et al. (2001). Current methods of the US Preventive Services Task Force: a review of the process. *American Journal of Preventive Medicine*, 20(3 Suppl), 21–35.
Holtzman, N.A. and Watson, M.S. (1999). Promoting safe and effective genetic testing in the United States. Final report of the Task Force on Genetic Testing. *Journal of Child and Family Nursing*, 2(5), 388–90.
Horton, R. and Lucassen, A. (2019). Consent and autonomy in the genomics era. *Current Genetic Medicine Reports*, **7** (2), 85–91.
Hsu, P.D., Lander, E.S., and Zhang, F. (2014). Development and applications of CRISPR-Cas9 for genome engineering. *Cell*, 157(6), 1262–78.
Illing, P.T., Purcell, A.W., and McCluskey, J. (2017). The role of HLA genes in pharmacogenomics: unravelling HLA associated adverse drug reactions. *Immunogenetics*, 69(8–9), 617–30.
Isasi, R., Kleiderman, E., and Knoppers, B.M. (2016). Genetic technology regulation. Editing policy to fit the genome? *Science*, 351(6271), 337–9.
Jenuwein, T. and Allis, C.D. (2001). Translating the histone code. *Science*, 293(5532), 1074–80.
Jones, P.A. and Baylin, S.B. (2002). The fundamental role of epigenetic events in cancer. *Nature Reviews Genetics*, 3(6), 415–28.
Kasinski, A.L., et al. (2015). A combinatorial microRNA therapeutics approach to suppressing non-small cell lung cancer. *Oncogene*, 34(27), 3547–55.
Khoury, M.J., et al. (2017). From public health genomics to precision public health: a 20-year journey. *Genetic Medicine*, 20(6), 574–82.
Kitzman, J.O., et al. (2012). Noninvasive whole-genome sequencing of a human fetus. *Science Translational Medicine*, 4(137), 137ra76.
Kleiderman, E. and Ogbogu, U. (2019). Realigning gene editing with clinical research ethics: what the ‘CRISPR Twins’ debacle means for Chinese and international research ethics governance. *Account Research*, 26(4), 257–64.
Kooistra, S.M. and Helin, K. (2012). Molecular mechanisms and potential functions of histone demethylases. *Nature Reviews Molecular Cell Biology*, 13(5), 297–311.
Kowluru, R.A., Santos, J.M., and Mishra, M. (2013). Epigenetic modifications and diabetic retinopathy. *Biomedical Research International*, **2013**, 635284.
Kung, J.T.Y., Colognori, D., and Lee, J.T. (2013). Long noncoding RNAs: past, present, and future. *Genetics*, 193(3), 651–69.
Li, H., et al. (2014). Immune regulation by low doses of the DNA methyltransferase inhibitor 5-azacitidine in common human epithelial cancers. *Oncotarget*, 5(3), 587–98.
Li, T., et al. (2018). Histone deacetylase 6 in cancer. *Journal of Hematology & Oncology*, 11(1), 111–11.
Liang, P., et al. (2015). CRISPR/Cas9-mediated gene editing in human tripronuclear zygotes. *Protein Cell*, 6(5), 363–72.

Maeder, M.L., et al. (2019). Development of a gene-editing approach to restore vision loss in Leber congenital amaurosis type 10. *Natural Medicine*, 25(2), 229–33.
Marin-Husstege, M., et al. (2002). Histone deacetylase activity is necessary for oligodendrocyte lineage progression. *Journal of Neuroscience*, 22(23), 10333–45.
Masimirembwa, C., Dandara, C., and Hasler, J. (2014). Population diversity and pharmacogenomics in Africa. In: Padmanabhan, S. (ed.) *Handbook of Pharmacogenomics and Stratified Medicine* (1st ed.) London: Academic, pp. 971–98.
Mathaiyan, J., Chandrasekaran, A., and Davis, S. (2013). Ethics of genomic research. *Perspectives in Clinical Research*, 4(1), 100–4.
McCormack, M., et al. (2011). HLA-A*3101 and carbamazepine-induced hypersensitivity reactions in Europeans. *The New England Journal of Medicine*, 364(12), 1134–43.
McGuire, A.L. and Beskow, L.M. (2010). Informed consent in genomics and genetic research. *Annual Review of Genomics and Human Genetics*, **11**, 361–81.
Miller, F.G., Mello, M.M., and Joffe, S. (2008). Incidental findings in human subjects research: what do investigators owe research participants? *The Journal of Law, Medicine & Ethics*, 36(2), 271–11.
Miller, J.C., et al. (2010). A TALE nuclease architecture for efficient genome editing. *Nature Biotechnology*, 29(2), 143–8.
Mishra, N., et al. (2003). Histone deacetylase inhibitors modulate renal disease in the MRL-lpr/lpr mouse. *Journal of Clinical Investigation*, 111(4), 539–52.
Molster, C.M., et al. (2018). The evolution of public health genomics: exploring its past, present, and future. *Frontiers of Public Health*, **6**, 247.
Montano, A., et al. (2018). Targeted genome editing in acute lymphoblastic leukemia: a review. *BMC Biotechnology*, 18(1), 45.
Moosavi, A. and Motevalizadeh Ardekani, A. (2016). Role of epigenetics in biology and human diseases. *Iranian Biomedical Journal*, 20(5), 246–58.
Morrison, A. and Boudreau, R. (2012). Evaluation frameworks for genetic tests. *Ottowa: Canadian Agency for Drugs and Technologies in Health Environmental Scan*, **37**, 1–9.
Naidoo, A., et al. (2018). Effect of genetic variation in UGT1A and ABCB1 on moxifloxacin pharmacokinetics in South African patients with tuberculosis. *Pharmacogenomics*, 19(1), 17–29.
Naidoo, A., et al. (2019). Effects of genetic variability on rifampicin and isoniazid pharmacokinetics in South African patients with recurrent tuberculosis. *Pharmacogenomics*, 20(4), 225–40.
Nam, S., et al. (2009). MicroRNA and mRNA integrated analysis (MMIA): a web tool for examining biological functions of microRNA expression. *Nucleic Acids Research*, **37**, W356–62.
Nasdaq:CRSP (2019). CRISPR therapeutics and vertex announce progress in clinical development programs for the investigational CRISPR/Cas9 Gene-Editing Therapy CTX001 Nasdaq:CRSP. [Online] Available at: https://globenewswire.com/news-release/2019/02/25/1741524/0/en/CRISPR-Therapeutics-and-Vertex-Announce-Progress-in-Clinical-Development-Programs-for-the-Investigational-CRISPR-Cas9-Gene-Editing-Therapy-CTX001.html
NCBI (2019). GTR:Genetic Testing Registry. [Online] Available at: https://www.ncbi.nlm.nih.gov/gtr/
NCI (2019). NIH National Cancer Institute. [Online] Available at: https://www.cancer.gov
Nelson, M.R., et al. (2015). The support of human genetic evidence for approved drug indications. *Nature Genetics*, 47(8), 856–60.
Niederer, F., et al. (2011). SIRT1 overexpression in the rheumatoid arthritis synovium contributes to proinflammatory cytokine production and apoptosis resistance. *Annals of Rheumatic*

Diseases, 70(10), 1866–73.
Nightingale, K.P., O'Neill, L.P., and Turner, B.M. (2006). Histone modifications: signalling receptors and potential elements of a heritable epigenetic code. *Current Opinion in Genetics & Development*, 16(2), 125–36.
NIH (2019). ClinicalTrials.gov. [Online] Available at: https://clinicaltrials.gov/
NIH Clinical Trials (2019). A safety and efficacy study evaluating CTX001 in subjects with severe sickle cell disease. [Online] Available at: https://clinicaltrials.gov/ct2/show/NCT03745287
Nyholt, D.R., Yu, C.E., and Visscher, P.M. (2009). On Jim Watson's APOE status: genetic information is hard to hide. *European Journal of Human Genetics*, 17(2), 147–9.
Omodamilola, O.I. and Ibrahim, A.U. (2018). CRISPR technology: advantages, limitations and future direction. *Journal of Biomedical and Pharmaceutical Sciences*, 1(2), 115.
Orlowski, J., Boniecki, M., and Bujnicki, J.M. (2007). I-Ssp6803I: the first homing endonuclease from the PD-(D/E)XK superfamily exhibits an unusual mode of DNA recognition. *Bioinformatics*, 23(5), 527–30.
Pena, S.A., et al. (2020). Gene therapy for neurological disorders: challenges and recent advancements. *Journal of Drug Targeting*, 28(2), 111–28.
Phillips, K.A., et al. (2014). The economic value of personalized medicine tests: what we know and what we need to know. *Genetic Medicine*, 16(3), 251–7.
Pingoud, A. and Silva, G.H. (2007). Precision genome surgery. *Nature Biotechnology*, 25(7), 743–4.
Pitini, E., et al. (2018). How is genetic testing evaluated? A systematic review of the literature. *European Journal of Human Genetics*, 26(5), 605–15.
Portela, A. and Esteller, M. (2010). Epigenetic modifications and human disease. *Nature Biotechnology*, **28**, 1057.
Quainoo, S., et al. (2017). Whole-genome sequencing of bacterial pathogens: the future of nosocomial outbreak analysis. *Clinical Microbiology Review*, 30(4), 1015–63.
Rattanavipapong, W., et al. (2013). Economic evaluation of HLA-B*15:02 screening for carbamazepine-induced severe adverse drug reactions in Thailand. *Epilepsia*, **54** (9), 1628–38.
Relling, M.V., et al. (2019). Clinical Pharmacogenetics Implementation Consortium Guideline for Thiopurine Dosing Based on TPMT and NUDT15 Genotypes: 2018 Update. *Clinical Pharmacology Therapy*, 105(5), 1095–105.
Reynolds, T.M. (2003). Down's syndrome screening is unethical: views of today's research ethics committees. *Journal of Clinical Pathology*, 56(4), 268.
Roberts, J.S., Dolinoy, D., and Tarini, B. (2014). Emerging issues in public health genomics. *Annual Review of Genomics and Human Genetics*, **15**, 461–80.
Roden, D.M., et al. (2019). Pharmacogenomics. *Lancet*, 394(10197), 521–32.
Rotimi, C.N., et al. (2017). The genomic landscape of African populations in health and disease. *Human Molecular Genetics*, 26(R2), R225–36.
Rousseau, F., et al. (2010). Development and description of GETT: a genetic testing evidence tracking tool. *Clinical Chemistry and Laboratory Medicine*, 48(10), 1397–407.
Ryst, E. van der (2005). TNA endorses statement on medical marijuana use. *Texas Nursing*, 79(3), 12.
Sanchez, R. and Zhou, M.-M. (2009). The role of human bromodomains in chromatin biology and gene transcription. *Current Opinion in Drug Discovery & Development*, **12** (5), 659–65.
Sanderson, S., et al. (2005). How can the evaluation of genetic tests be enhanced? Lessons learned from the ACCE framework and evaluating genetic tests in the United Kingdom. *Genetic Medicine*, 7(7), 495–500.
Sanseau, P., et al. (2012). Use of genome-wide association studies for drug repositioning. *Nature Biotechnology*, **30** (4), 317–20.

Saouaf, S.J., et al. (2009). Deacetylase inhibition increases regulatory T cell function and decreases incidence and severity of collagen-induced arthritis. *Experimental Molecular Pathology*, 87(2), 99–104.
Sawaya, G.F., et al. (2007). Update on the methods of the U.S. Preventive Services Task Force: estimating certainty and magnitude of net benefit. *Annals of Internal Medicine*, 147(12), 871–5.
Shen, S., Li, J., and Casaccia-Bonnefil, P. (2005). Histone modifications affect timing of oligodendrocyte progenitor differentiation in the developing rat brain. *Journal of Cellular Biology*, 169(4), 577–89.
Sibbing, D., et al. (2010). Cytochrome 2C19*17 allelic variant, platelet aggregation, bleeding events, and stent thrombosis in clopidogrel-treated patients with coronary stent placement. *Circulation*, 121(4), 512–18.
Sneed, A. (2017). Mail-order CRISPR kits allow absolutely anyone to hack DNA. [Online] Available at: https://www.scientificamerican.com/article/mail-order-crispr-kits-allow-absolutely-anyone-to-hack-dna/
Sousa-Pinto, B., et al. (2015). Pharmacogenetics of abacavir hypersensitivity: a systematic review and meta-analysis of the association with HLA-B*57:01. *Journal of Allergy and Clinical Immunology*, 136(4), 1092–4.e3.
Strahl, B.D. and Allis, C.D. (2000). The language of covalent histone modifications. *Nature*, 403(6765), 41–5.
Tan, Y., et al. (2016). Improving drug safety: from adverse drug reaction knowledge discovery to clinical implementation. *Methods*, **110**, 14–25.
Teutsch, S.M., et al. (2009). The evaluation of genomic applications in practice and prevention (EGAPP) Initiative: methods of the EGAPP Working Group. *Genetic Medicine*, 11(1), 3–14.
Therrell, B.L., et al. (2015). Current status of newborn screening worldwide: 2015. *Seminars in Perinatology*, 39(3), 171–87.
Trang, P., et al. (2011). Systemic delivery of tumor suppressor microRNA mimics using a neutral lipid emulsion inhibits lung tumors in mice. *Molecular Therapy*, 19(6), 1116–22.
Trinite, T., Loveland-Cherry, C., and Marion, L. (2009). The U.S. Preventive Services Task Force: an evidence-based prevention resource for nurse practitioners. *Journal American Academy of Nurse Practitioners*, 21(6), 301–6.
Urnov, F.D., et al. (2010). Genome editing with engineered zinc finger nucleases. *Nature Reviews Genetics*, 11(9), 636–46.
Urv, T.K. and Parisi, M.A. (2017). Newborn screening: beyond the spot. *Advances in Experimental Medicine and Biology*, **1031**, 323–46.
Van Rie, A., et al. (2011). Xpert((R)) MTB/RIF for point-of-care diagnosis of TB in high-HIV burden, resource-limited countries: hype or hope?. *Expert Review of Molecular Diagnostics*, 10(7), 937–46.
Vaz, M. and Srinivasan, K. (2014). Ethical challenges in biobanking: moving the agenda forward in India. *Indian Journal of Medical Ethics*, 11(2), 79–88.
Vendetti, F.P. and Rudin, C.M. (2013). Epigenetic therapy in non-small-cell lung cancer: targeting DNA methyltransferases and histone deacetylases. *Expert Opinion on Biological Therapy*, 13(9), 1273–85.
Vogel, K.M. and Ouagrham-Gormley, S.B. (2018). Anticipating emerging biotechnology threats: a case study of CRISPR. *Politics Life Science*, 37(2), 203–19.
Vora, N.L. and Wapner, R.J. (2018). Introducing new and emerging genetic tests into prenatal care. *Seminars in Perinatology*, 42(5), 283–6.
Wang, G., et al. (2016a). CRISPR-Cas9 can inhibit HIV-1 replication but NHEJ repair facilitates virus escape. *Molecular Therapy*, 24(3), 522–6.
Wang, Y., et al. (2018). Therapeutic genome editing for myotonic dystrophy type 1 using CRISPR/Cas9. *Molecular Therapy*, 26(11), 2617–30.

Wang, Z., et al. (2016b). CRISPR/Cas9-derived mutations both inhibit HIV-1 replication and accelerate viral escape. *Cell Reports*, 15(3), 481–9.

Wells, R.A., et al. (2014). Optimizing outcomes with azacitidine: recommendations from Canadian centres of excellence. *Current Oncology (Toronto, Ont.)*, **21** (1), 44–50.

Yung, R.L. and Richardson, B.C. (1995). Cytarabine therapy for refractory cutaneous lupus. *Arthritis Rheumatism*, 38(9), 1341–3.

Zampetaki, A., Albrecht, A., and Steinhofel, K. (2018). Long non-coding RNA structure and function: is there a link?. *Frontiers in Physiology*, **9**, 1201.

Zhang, J., et al. (2015). Use of genome-wide association studies for cancer research and drug repositioning. *PLoS One*, 10(3), e0116477.

Zoghbi, H.Y. and Beaudet, A.L. (2016). Epigenetics and human disease. *Cold Spring Harbour Perspectives in Biology*, 8(2), a019497.

2.7

Water and sanitation

Thomas Clasen

Introduction to water and sanitation

Background

Safe drinking water and sanitary waste disposal are among the most fundamental of public health interventions. When readers of the *British Medical Journal* were asked in 2006 to name the 'greatest medical advance' since 1840, their top choice was clean drinking water and waste disposal, beating antibiotics, anaesthesia, vaccines, and germ theory (Ferriman 2007). Deaths from diarrhoeal diseases and typhoid fever showed dramatic declines in Europe and North America when cities and towns began filtering and chlorinating their water and safely disposing of human and animal excreta (Cutler and Miller 2005). The field of epidemiology arguably has its origins in John Snow's nineteenth-century mapping of cholera cases and the eventual intervention at London's Broad Street pump that demonstrated waterborne transmission of the disease.

While diseases associated with poor water and sanitation are now comparatively unknown in higher-income countries, they still impose a heavy burden elsewhere, especially among young children, the infirm, the poor, the immunocompromised, and the displaced. Diarrhoeal diseases alone are responsible for an estimated 1.3 million deaths annually, including a half million children under 5 years (Troeger et al. 2017). Diarrhoea is the third leading cause of deaths of children <5 years in low-income countries, accounting for 11% of the overall disease burden in this population. Improvements in water, sanitation, and hygiene have the potential to reduce this disease burden by an estimated 58% (Prüss-Ustün et al. 2014). As discussed next, poor water and sanitation is also associated with a heavy disease burden from malnutrition, parasite and worm infection, trachoma, and environmental enteropathy. Water and sanitation are not only a matter of public health, but also of poverty, equity, and justice (United Nations Development Programme 2007). Because they are less likely to have access to safe water and sanitation, the poor bear most of the burden of water-related diseases, driving them further into poverty through lost productivity and expenditure on treatment

(Blakely et al. 2005). Time spent in collecting water from distant sources and the inability to procure sufficient quantities of water for irrigating crops, watering animals, and carrying out other productive activities aggravates poverty. While urban-rural disparities in water and sanitation coverage are well documented (WHO and UNICEF 2017), a recent analysis of sub-Saharan Africa also demonstrated important geographic inequalities in use of water and sanitation previously hidden within national statistics (Pullan et al. 2014). These results confirm the need for targeted policies and metrics that reach these marginalized populations (Pickering and Davis 2012). Inadequate water and sanitation are also associated with poor cognitive development and school attendance (Sclar et al. 2017). For these and other reasons, water and sanitation have been recognized as a fundamental human right (United Nations 2010).

Nevertheless, basic water security and sanitation still elude much of the world's population living in low-income countries. An estimated 582 million people still lack improved drinking water sources; hundreds of millions more rely on water that is not safely managed (WHO and UNICEF 2017). An estimated 2.3 billion people—30% of the world's population—lack access to improved sanitation; 892 million still practice open defecation. Coverage is lowest in developing regions, where people are most vulnerable to infection and disease. In sub-Saharan Africa, safely managed or basic water and sanitation coverage is just 42% and 28%, respectively. Rural areas also lag behind their urban counterparts. In 2015, 85% of urban populations had access to safely managed drinking water; whereas only 55% of rural populations had access. If current trends continue, people in rural areas will only account for one-third of those who have access to water free from contamination (WHO and UNICEF 2017).

The shortfall in water and sanitation coverage is not the result of a failure to recognize the need or declare goals at the highest international levels. The 1977 Mar del Plata Declaration by the United Nations (UN) expressed the goal of providing safe water and sanitation for all by 1990, launching the Water and Sanitation Decade (1981–1990) (Cairncross 1992). In 1990, the UN renewed the call and extended the deadline to the end of the century. The UN Millennium Development Goals (MDGs) call for halving, by 2015, the portion of the population without sustainable access to safe drinking water or basic sanitation (United Nations 2000). While the UN claimed to have meet the water target, others have shown that the indicators used fail to capture key aspects of the goal, including sustainability, accessibility and safety (Clasen 2012); at the end of 2015, more than 1.3 billion people were estimated to lack safe water (Bain et al. 2013). Among all the MDGs, the least progress was made on the sanitation target. Increases in population largely offset gains in latrine construction, leaving an

estimated 2.3 billion without improved sanitation at the end of the MDG period (WHO and UNICEF 2017).

In 2015, the UN followed the MDGs with the Sustainable Developmental Goals (SDGs) (United Nations 2015). As discussed more fully to follow, SDG Goal 6 seeks to achieve availability and sustainable management of water and sanitation for all by 2030. Achieving the goal would not only advance environmental security, but also make contributions to reducing poverty (Goal 1), ensuring good health and well-being (Goal 3), ensuring quality education (Goal 4), and promoting gender equality (Goal 5). According to the UN, achievement of the SDGs will help advance the overall goal of the ending poverty, protecting the planet, and ensuring prosperity for all (United Nations 2015).

Traditionally, much of the work in water and sanitation has been undertaken by engineers and has consisted of infrastructural improvements. Low-cost community- and household-based interventions, such as protected wells, boreholes, and communal standpipes for improved water supplies, and various types of latrines, septic tanks, and composting systems for improved sanitation, have been largely conceived by and constructed with the assistance of engineers. There are numerous books, manuals, and other resources that describe these systems in detail, including Cairncross and Feachem (1993), the UK Department of International Development (1998), Davis and Lambert (2002), the quarterly *Waterlines*, and the World Bank Water and Sanitation Programme (WSP) 'Field Notes'. Readers are encouraged to refer to such sources for details on the design, installation, and operation of such systems, technology innovations, and the programmatic challenges associated with achieving widespread use on a sustained basis.

This chapter focuses solely on the public health issues concerning water and sanitation. After introducing some basic terminology, it begins by describing the diseases associated with inadequate water and sanitation and their contribution to the overall burden of disease. It then presents evidence of the effectiveness of water and sanitation interventions to prevent such diseases, the economic implications (especially cost-effectiveness and cost–benefits) of such interventions, and some of the other non-health benefits associated therewith. Recent and emerging developments in water, sanitation, and health are then discussed, along with some issues relevant to water and sanitation interventions. The chapter closes with a discussion of some of the continuing challenges in water and sanitation.

Terminology

At the outset, it is useful to understand some of the terminology used in describing the diseases, transmission routes, and interventions associated with the water and

sanitation sectors. Water-related diseases are sometimes classified according to their disease transmission routes as *waterborne* (ingested in drinking water), *water-washed* (associated with inadequate supplies of water for proper personal hygiene), *water-based* (transmitted through an aquatic invertebrate host), or linked to a *water-related vector* (involving an insect vector breeding in or near water) (White et al. 1972). Most waterborne organisms that are human pathogens colonize the gut of humans and certain other mammals and are transmitted through the *faecal–oral* route. The transmission of common waterborne diseases can thus be interrupted by improvements in sanitation (*excreta disposal*), personal hygiene (especially *hand washing with soap*), and microbiological *water quality*, while those that are water-washed are impacted by improvements in *water supplies* (quantity and access) for personal hygiene. Improving water supplies can also help prevent water-based diseases (such as schistosomiasis and dracunculiasis) by reducing the need to enter infected water bodies.

The term ‘sanitation’ is vague and has multiple meanings. Within the public health sector two definitions are used. Under the broader definition, sanitation extends to the process whereby people demand, effect, and sustain a hygienic and healthy environment for themselves. This definition could include safe food production, solid waste management, industrial waste, animal waste, control of chemicals, environmental pollution, storm water drainage, wastewater disposal, human settlements, personal and domestic hygiene, vector and vermin control, occupational health and safety, mining and quarrying, port health, and disposal of the dead. A second definition is more specific, extending only to the process of separating humans from their excreta. This chapter uses this second, more specific definition and regards sanitation as a system in which excreta is: (1) collected safely and with dignity, (2) transported to a suitable location, (3) treated or contained to eliminate pathogenicity, and (4) reused and/or discharged to the environment.

While the MDG targets for water and sanitation were expressed in terms of *sustainable access to safe drinking water* and of *basic sanitation*, the SDGs focus on ‘*safely managed*’ water and sanitation. The WHO/UNICEF Joint Monitoring Programme for Water, Sanitation and Hygiene (JMP) is charged with monitoring progress toward international targets, including the SDGs (WHO/UNICEF 2017). It defines a safely managed drinking water service is one that meets three criteria: it should be accessible on premises, the water should be available when needed, and the water supplied should be free from contamination.

The JMP uses service ladders to benchmark and compare progress across countries (WHO/UNICEF 2017) (Fig. 2.7.1). The bottom rung of the water service ladder is ‘surface water’— drinking water directly from a river, dam, lake, pond, stream, canal, or irrigation canal. The next two rungs are ‘unimproved water

sources’ (surface water and unprotected well/spring) and ‘improved water sources’ (piped-in tap water, borehole/tubewell, protected well/spring, and packaged/delivered water). Under the SDGs, the water service ladder includes three types of improved water services, safely managed, basic, and limited. A ‘safely managed drinking water service’ is one that meets three criteria: it is accessible on premises, the water is available when needed, and the water supplied is free from contamination. If the improved source does not meet any one of these criteria, but a round trip to collect water takes 30 minutes or less, it will be classified as a ‘basic’. If water collection from an improved source exceeds 30 minutes, it is categorized as a ‘limited service’.

The JMP sanitation ladder also has multiple runs. At bottom is ‘open defecation’— disposal of human faeces in fields, forests, bushes, open bodies of water, beaches, or other open spaces, or with solid waste. Moving up the ladder is ‘unimproved sanitation’ (pit latrines without a slab or platform, hanging latrines or bucket latrines), and ‘improved sanitation’ (flush/pour flush to piped sewer systems, septic tanks, or pit latrines, ventilated improved pit (VIP) latrines, and composting toilets, or pit latrines with slabs). Similarly, improved sanitation sources are now subdivided into three groups during the SDG period (safely managed, basic, and limited). For the service level to be ‘safely managed’, the sanitation facilities must be improved and not shared with other households, and the faecal sludge must be (i) treated and disposed of in situ, (ii) stored temporarily and then emptied, transported and treated off-site, or (iii) transported through a sewer with wastewater and then treated off-site. Sanitation services where the faecal sludge is not so managed but where the facilities are otherwise improved facilities that are shared with other households are defined as ‘limited’ and those that are share are defined as ‘limited’ (WHO and UNICEF 2017).

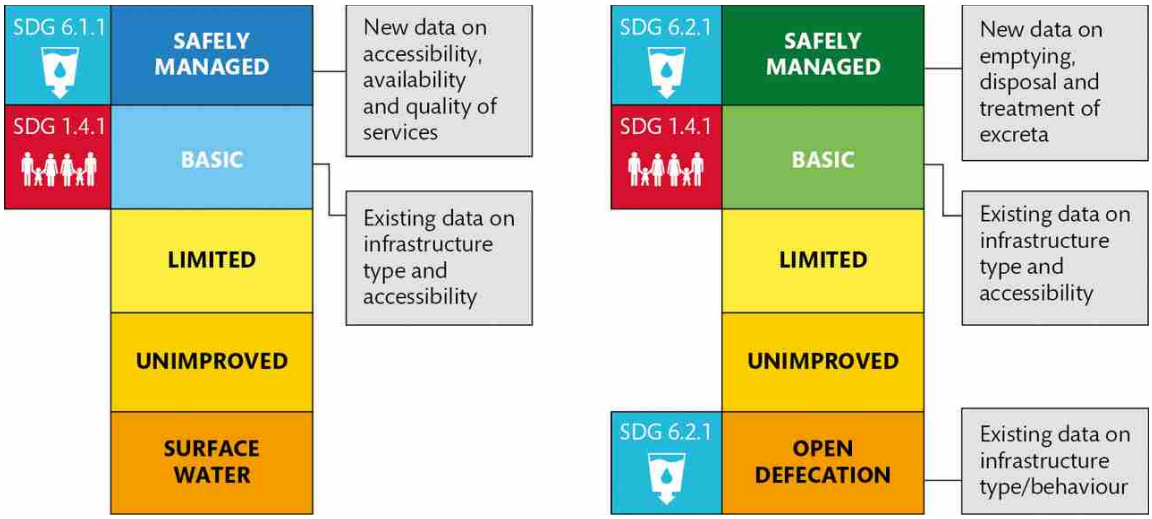


Fig. 2.7.1 Updated JMP service ladders.

Reprinted from 'Progress on Drinking Water, Sanitation and Hygiene: 2017 Update and SDG Baselines'. Geneva: World Health Organization (WHO) and the United Nations Children's Fund (UNICEF), 2017. Licence: CC BY-NC-SA 3.0 IGO.

Burden of disease

General

Poor water and sanitation are associated with a variety of infectious diseases transmitted through various pathways by helminths, protozoa, bacteria, and viruses. Table 2.7.1 summarizes the most important of these diseases, their transmission routes, aetiological agents, and epidemiological significance. Some of these diseases also contribute to malnutrition, a separate cause of substantial morbidity and mortality that is not reflected in the direct burden of disease figures cited in this chapter (Black et al. 2003).

Certain diseases associated with water are not addressed in this chapter. First, in addition to microbial agents, water is a medium for the transmission of chemical pathogens, including arsenic and other metals, fluoride, nitrates, and volatile organic compounds (including pesticides and herbicides). Accordingly, WHO guidelines and many national water standards establish maximum allowable limits for such chemicals (WHO 2012). However, except for arsenicosis and fluoridosis, which are especially serious in focal areas in Asia and parts of Africa, most of these contaminants represent hazards to health only over the longer term. Second, although improvements in water supplies (to discourage contact with water) and point-of-use water treatment (with filters) are important interventions in preventing dracunculiasis (Cairncross et al. 2002), efforts to control Guinea worm infection have been largely successful and the disease is now of public health interest in

limited areas. Finally, this chapter does not address a variety of diseases associated with waterborne pathogens, such as poliomyelitis and hepatitis A and E, which are controlled mainly by vaccines and other non-environmental measures (Leclerc et al. 2002).

Table 2.7.1 Principal infectious diseases, disease agents, transmission routes, and annual morbidity and mortality related to poor water and sanitation

Disease	Aetiological agent	Transmission	Infected[1] (millions)	Mortality[1] (thousands)
Diarrhoea (dysentery, cholera)	**Virus** *Rotavirus*	Faecal–oral	2,390 (annual morbidity)	1,233–1,391
	Bacteria *Escherichia coli (ETEC)* *Shigella* spp. *Salmonella* spp. *Vibrio* spp. *Campylobacter* sp.	Faecal–oral		
	Protozoa *Giardia lambia* *Cryptosporidium parvum* *Entamoeba histolytica*	Faecal–oral		
Schistosomiasis	*Schistosoma haematobium* *S. mansoni* *S. japonicum*	Penetration through skin exposed to contaminated freshwater	207	4.4
Ascariasis	*Ascaris lumbricoides*	Faecal–oral	807–1.200	2.7
Trichuriasis	*Trichuris trichuria*	Faecal–oral	440	3–10
Hookworm infection	*Necator americanus* *Ancylostoma duodenale*	Penetration through skin exposed to faecally contaminated soil Faecal–oral	460	3–65
Typhoid and paratyphoid fever	*Salmonella* spp.	Faecal–oral	21	177.2
Trachoma	*Chlamydia trachomatis*	Fingers Clothing Eye-seeking flies (*Musca sorbens*) Coughing/sneezing	Blind from trachoma: 1.2 Active trachoma: 21 Trichuriasis: 7.3	

Note:

[1] estimates vary according to method.

Source: data from Centers for Disease Control and Prevention (2018). Parasites—Ascariasis. Available at: https://www.cdc.gov/parasites/ascariasis/Diarrhoeal; Troeger, C., Forouzanfar, M., Rao, P.C., Khalil, I., Brown, A., Reiner Jr, R.C. et al. (2017). Estimates of global, regional, and national morbidity, mortality, and aetiologies of diarrhoeal diseases: a systematic analysis for the Global Burden of Disease Study 2015. The Lancet Infectious Diseases 17(9): 909–48. doi: 10.1016/S1473-3099(17)30276-1; World Health Organization (2017). *Guideline: preventive chemotherapy to control soil-transmitted helminth infections in at-risk population groups*. Geneva: Lustigman, S., Prichard, R.K., Gazzinelli, A., et al. (2012). A research agenda for helminth diseases of humans: the problem of helminthiases. *PLoS Neglected Tropical Diseases*, 6(4), e1582; World Health Organization (2017). Global Health Observatory. [Online] Available at: http://www.who.int/gho/neglected_diseases/trachoma/en/

Diarrhoeal diseases

Diarrhoeal diseases represent the leading cause of mortality and morbidity associated with unsafe drinking water and poor sanitation. For those infected with the human immunodeficiency virus (HIV) or who have developed acquired immune deficiency syndrome (AIDS), diarrhoea can be prolonged, severe, and life-threatening (Hayes et al. 2003). Diarrhoea is characterized by stools of decreased consistency and increased number. The clinical symptoms and course of the disease vary greatly with the age, nutritional and immune status, and the pathogen (Black and Lanata 1995). Most cases resolve within a week, though a small percentage continue for 2 weeks or more and are characterized as ‘persistent’ diarrhoea. Dysentery is a diarrhoeal disease defined by the presence of blood in the liquid stools. Though epidemic diarrhoea such as cholera and shigellosis (bacillary dysentery) are well-known risks, particularly in emergency settings, their global health significance is small compared to endemic diarrhoea (Hunter 1997). The immediate threat from diarrhoea is dehydration, a loss of fluids and electrolytes. Thus, the widespread promotion of oral rehydration therapy (ORT) has significantly reduced the case-fatality rate associated with the disease. Such improvements in case management, however, have not reduced morbidity, which is estimated at 2.4 billion cases annually (Troeger et al. 2017). And since diarrhoeal diseases inhibit normal ingestion of foods and adsorption of nutrients, continued high morbidity is an important cause of malnutrition, leading to impaired physical growth and cognitive function, reduced resistance to infection, and potentially long-term gastrointestinal disorders.

The infectious agents associated with diarrhoeal disease are transmitted chiefly through the faecal–oral route (Leclerc et al. 2002). Safe excreta disposal thus represents a primary barrier that should contribute to the prevention of indirect transmission via food, water, hands, fomites, and mechanical vectors (flies) (Fig. 2.7.2). A wide variety of bacterial, viral, and protozoan pathogens excreted in the faeces of humans and animals are known to cause diarrhoea. While the importance of individual pathogens varies among settings, seasons, and conditions, a recent

multicountry study has shed important light on the principal aetiological agents responsible for much of the disease burden in Africa and South Asia (Kotloff et al. 2013). Investigators found that most attributable cases of moderate-to-severe diarrhoea were due to four pathogens: rotavirus, *Cryptosporidium*, enterotoxigenic *Escherichia coli* producing heat-stable toxin (ST-ETEC) and *Shigella*. *Aeromonas*, *Vibrio cholerae* O1, and *Campylobacter jejuni* were also important in selected sites. They reported that the odds of dying during follow-up were 8.5 times higher in patients with moderate-to-severe diarrhoea than in controls (odds ratio (OR) 8.5; 95% confidence interval (CI) 5.8–12.5, $p < 0.0001$) and that 88% of deaths occurred during the first 2 years of life. Pathogens associated with increased risk of case death were ST-ETEC (hazard ratio (HR) 1.9; 95% CI: 0.99–3.5) and typical enteropathogenic *E. coli* (HR 2.6; 95% CI: 1.6–4.1) in infants aged 0–11 months, and *Cryptosporidium* (HR 2.3; 95% CI: 1.3–4.3) in toddlers aged 12–23 months. Although diarrhoea is also associated with the ingestion of metals, nitrates, organics, and other chemicals, the burden of disease arising from such exposure is small relative to infectious diarrhoea (Hunter 1997).

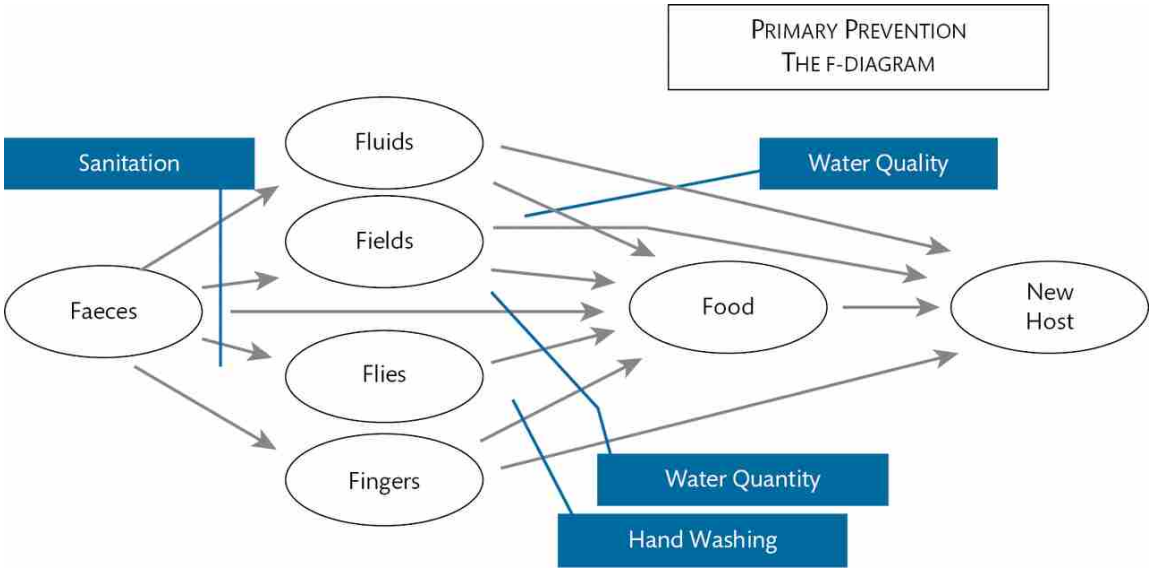


Fig. 2.7.2 The F-diagram.

Reproduced with permission from Wagner, EG and Lanois, JN, *Excreta disposal for rural areas and small communities*. WHO monograph series No.39, World Health Organization, Geneva, Switzerland, Copyright © 1958.

Schistosomiasis

Schistosomiasis affects 200 million people all over the world, with *Schistosoma haematobium* and *S. mansoni* being the most common species. The schistosomiasis life cycle involves contamination of freshwater by eggs carrying excreta and urine and an intermediate host, a freshwater snail. Larvae released in the water infect humans by penetrating through the skin. Parasites develop and migrate to the intestines and bladder where thousands of eggs are produced. Like other intestinal helminths, schistosomes are associated with impaired physical and mental development and anaemia. Furthermore, schistosomiasis may cause serious damage to the bladder and intestine walls as a result of parasite egg entrapment within tissues. Chronic infection with *S. haematobium* has been associated with increased risk of bladder cancer in adulthood (Gryssels et al. 2006).

Soil-transmitted helminth infection

Almost 2 billion of the world's population are infected with soil-transmitted helminths. A disproportionate burden of helminthiasis occurs in marginalized, low-income, and resource-constrained regions of the world, largely in developing areas of sub-Saharan Africa, Asia, and the Americas (WHO 2017). About 300 million people suffer from heavy worm load and related severe morbidity (deSilva et al. 2003). *Ascaris lumbricoides*, *Trichuris trichuria*, and hookworm (*Ancylostoma duodenale* and *Necator americanus*) are the most prevalent intestinal helminths. Transmission occurs via ingestion of eggs present in faecally contaminated soil, or via penetration of the larvae through the skin. Children are particularly vulnerable to chronic and heavy infections, which result in malnutrition, stunted growth, reduced physical fitness, and impaired cognitive development (Stephenson et al. 2000). Hookworm infection is an important cause of anaemia, not only in children, but also among women of reproductive age and pregnant women leading to premature birth and low birth weight (Hotez et al. 2004).

Typhoid and paratyphoid fevers

While enteric fevers such as typhoid and paratyphoid were leading causes of waterborne disease in previous centuries, morbidity and mortality diminished dramatically with the provisions of disinfected water supplies and improved sanitary facilities (Cutler and Miller 2005). The aetiological agents for typhoid and paratyphoid fevers are *Salmonella enterica* serovar Typhi (formerly *S. typhi*) and *S. enterica* serovar Paratyphi A and B (formerly *S. paratyphi*). There are an estimated 21 million cases of typhoid annually, causing 148,800 deaths. The milder paratyphoid accounts for an additional 5 million cases each year (Crump et al. 2004; Troeger et al. 2017).

Trachoma

Trachoma accounts for 1.4 per cent of world blindness, with 1.9 million people affected and 190.2 million people living in trachoma endemic areas that are at risk of trachoma blindness (WHO 2017). Trachoma is caused by repeated eye infection with *Chlamydia trachomatis*. Children are the main reservoir for infection, with high prevalence of active trachoma (Mabey et al. 2003). Repeated infections result in deformation of the upper eye lid, abrasion of the cornea, and progressive loss of vision in later life. *C. trachomatis* is transmitted from the discharge of an infected eye via contaminated fingers, clothing, and eye-seeking flies (*Musca sorbens*) (Hu et al. 2010). Although the role of flies in the transmission of infection remains unclear, studies have shown that *M. sorbens* breed mainly in solid human faeces present in the environment and not in latrines (Emerson et al. 2000). Thus, safe excreta disposal may play an important role in reducing trachoma transmission. Improving water supplies and sanitation is part of the WHO-backed SAFE (surgery, antibiotics, facial hygiene, environmental improvement) strategy for controlling and preventing the disease.

Malnutrition

Water and sanitation are also linked to malnutrition, a separate source of significant morbidity and mortality. An estimated 165 million children under 5 are stunted (i.e. have a height-for-age Z score of less than −2); 55 million children are wasted (i.e. have a weight-for-height Z score of less than −2), of whom 19 million have severe wasting (weight-for-height Z score of less than −3) or severe acute malnutrition (weight-for-height Z score of –3 or lower or associated oedema) (Bhutta et al. 2013). Child underweight or stunting causes about 20% of all mortality of children younger than 5 years of age; it also contributes to long-term cognitive impairment, poorer performance in school, fewer years of completed schooling, and lower adult economic productivity.

Effectiveness of water and sanitation in preventing disease

Barriers to transmission of faecal–oral diseases

As illustrated by the so-called ‘F-diagram’ (Fig. 2.7.2), the safe disposal of human faeces is the primary barrier in preventing faecal–oral transmitted diseases. Without removing excreta from potential contact with humans, animals, and insects, pathogens may be carried on unwashed hands, in contaminated water or food, or via flies and other insects on to further human hosts. Whether or not

sanitation is adequate, hands can become contaminated with faeces, especially during anal cleansing following defecation or in cleaning a child after defecation. This may result in further transmission, either directly or indirectly through food, water, or other beverages, or fomites. Accordingly, hand washing is an important secondary barrier to faecal–oral disease transmission (Curtis and Cairncross 2003). Other secondary barriers include: (1) water quality interventions (e.g. safe distribution and storage or use of boiling or other point-of-use water treatment); (2) water supply interventions to increase the quantity and availability of water for personal and domestic hygiene; (3) proper cooking and food handling; and (4) control of mechanical vectors such as flies.

Evidence of effectiveness

Scores of studies have been conducted and published on the effectiveness of water, sanitation, and hygiene interventions to prevent infection and disease. Systematic reviews (Chapter 5.15) are a means of identifying, summarizing, synthesizing, explaining, and assessing the methodological quality of evidence of the effectiveness of health interventions with a variety of studies relating to a particular health intervention. In some cases, such reviews also employ meta-analyses or other statistical methods to estimate the pooled effect of the intervention across the studies included in the review. A number of such reviews have examined the evidence of effectiveness of water, sanitation, and hygiene interventions to prevent disease and infection.

Diarrhoeal diseases

Table 2.7.2 summarizes the results of seven different systematic reviews of water and sanitation interventions to prevent diarrhoeal diseases published over the last 25 years. While there is conflicting evidence on the impact of water supply (quantity and access), there is consistent evidence that improvements in water and sanitation can make substantial contributions to the prevention of diarrhoeal diseases.

Soil-transmitted helminth infection

One review assessed not only the impact of sanitation interventions on soil-transmitted helminth infections, but also water and hygiene (Strunz et al. 2014). The review covers 94 studies, though only 5 were randomized controlled trials. Use of treated water was associated with lower odds of infection from any soil-transmitted helminth (OR 0.46, 95% CI 0.36–0.60). Piped water access was

associated with lower odds of *A. lumbricoides* (OR 0.40, 95% CI 0.39–0.41) and *T. trichiura* infection (OR 0.57, 95% CI 0.45–0.72), but not any soil-transmitted helminth (STH) infection (OR 0.93, 95% CI 0.28–3.11).

Comparatively, the most recent review by Freeman et al. (2017) examined 65 studies, including 31 from the 2014 Strunz review. The overall effect of sanitation was associated with lower odds of infection for *A. lumbricoides* (OR 0.73, 95% CI 0.61–0.86) and Hookworm (OR 0.65, 95% CI 0.53–0.78), but not *T. trichiura* (OR 0.80, 95% CI 0.63–1.03).

When examining the association between STH infections and any sanitation vs. none/non-use, a protective association was found for all STH infections with pooled ORs of 0.63 (95% CI: 0.49–0.81), 0.72 (95% CI: 0.55–0.95), and 0.61 (95% CI: 0.51–0.72), for *A. lumbricoides*, *T. trichiura*, and hookworm respectively. However, when restricted to intervention studies alone, no association was found between STH infections and sanitation: *A. lumbricoides* (OR 0.77, 95% CI: 0.49–1.21), *T. trichiura* (OR 1.00, 95% CI 0.37–2.69), and hookworm (OR 0.94, 95% CI 0.76–1.15). The researchers attribute this result to the low number of such studies available (Freeman et al. 2017).

Table 2.7.2 Estimate of effect[1], [2] (and number of studies in square brackets) of systematic reviews of water and sanitation interventions to prevent diarrhoeal diseases

Intervention (Improvement)	Esrey et al. (1985) (range)	Esrey et al. (1991)	Fewtrell et al. (2005) (95% CI) (No. studies)	Clasen et al. (2006) (95% CI) (No. studies)	Waddington et al. (2009) (95% CI) (No. studies)	Wolf et al. (2014) (95% CI)[2]	Clasen et al. (2015) (95% CI) (No. studies)	Freeman et al. (2017) (95% CI) (No. studies)
Water quantity	25% (0–100%) [17]	27% [7]						
Water quality			0.69 (0.53–0.89) [15]	0.57 (0.46–0.70) [38]	0.58 (0.50–0.67) [31]		0.59 (0.51–0.69)	
Source	16% (0–90%) [9]	17% [7]	0.89 (0.42–1.90) [3]	0.73 (0.53–1.01) [6]	0.79 (0.62–1.02) [3]	0.89 (0.78–1.01)	0.76 (0.48–1.19) [6]	
Household (point of use)			0.65 (0.48–0.88) [12]	0.53 (0.39–0.73) [32]	0.56 (0.45–0.65) [28]		0.58 (0.48–0.69) [58]	
Chlorination				0.63 (0.52–0.75) [16]		0.63 (0.55–0.72)	0.72 (0.61–0.84) [19]	
Filtration				0.37 (0.28–0.49) [6]		0.41 (0.33–0.50)	0.48 (0.38–0.59) [23]	
Solar disinfection				0.69 (0.63–0.74) [2]		0.63 (0.55–0.72)	0.62 (0.42–0.94) [4]	
Floc-disinfection				0.69 (0.58–0.82) [6]			0.48 (0.20–1.16) [7]	

Intervention (Improvement)	Esrey et al. (1985) (range)	Esrey et al. (1991)	Fewtrell et al. (2005) (95% CI) (No. studies)	Clasen et al. (2006) (95% CI) (No. studies)	Waddington et al. (2009) (95% CI) (No. studies)	Wolf et al. (2014) (95% CI)[2]	Clasen et al. (2015) (95% CI) (No. studies)	Freeman et al. (2017) (95% CI) (No. studies)
Water supply (point-of-distribution)	37% (0–82%) [8]	16% [22]	0.75 (0.62–0.91) [6]	0.73 (0.53–1.01) [6]	0.98 (0.89–1.06) [8]	0.89 (0.78–1.01)		
Water supply and sanitation/hygiene		20% [7]			0.43 (0.33–0.55) [3]			
Sanitation	22% (0–48%) [10]	22% [11]	0.68 (0.53–0.87) [2]		0.63 (0.43–0.93) [6]	0.84 (0.77–0.91)		0.88 (0.83–0.92) [27]

Note:

[1] for studies by Esrey and colleagues, estimate of effect is the median reduction in diarrhoeal disease from the reported studies; for other studies, estimate of effect is the pooled risk ratio from meta-analysis using random effects model. To compare results, the percentage reduction is 1 −RR (e.g. RR of 0.69 implies a 31% reduction in risk).

[2] Pooled estimates of effect in settings with unimproved water supply and prior to adjustment for non-blinding. For household interventions, estimates are for studies that combine treatment with safe storage.

Schistosomiasis

Esrey and colleagues (1991) reported a median reduction in the prevalence of schistosomiasis of 73% (range 59–87%) from four water and sanitation interventions; the reduction was 77% among the three studies they deemed rigorous. Piped-in water supplies and community washing and bathing facilities that reduced contact with surface waters were especially protective, leading to reductions in both prevalence and severity (Esrey et al. 1991). The reviewers noted that in Kenya, the installation of boreholes without laundry or shower facilities failed to reduce infection.

Two quasi-randomized, controlled interventional studies from China also suggest that improved excreta disposal is protective against schistosomiasis. In a 3-year quasi-randomized controlled trial (RCT), Chen et al. (2004) recorded a 43% reduction from combined water, sanitation, and hygiene interventions that also included a snail control component. Zhang et al. (2005) reported a 45% reduction in a 2-year quasi-RCT that included water, sanitation, and hygiene. Once again, these trials did not include sufficient clusters to reliably calculate confidence intervals around the point estimates of effect.

Typhoid and paratyphoid

No RCTs of water or sanitation interventions to prevent typhoid or paratyphoid have been reported. A recent review of global trends in these enteric fevers identified little direct evidence of the potential contribution of water or sanitation interventions (Crump and Mintz 2010). Nevertheless, there is evidence suggesting the effectiveness of water quality interventions. Cutler and Miller (2005) have shown the historical evidence on reductions in mortality associated with the introduction of clean water and sanitation in the United States. Case–control studies do suggest that the diseases are still associated with unsafe water and sanitation. In a study in Uzbekistan where typhoid remains endemic, cases were more likely to drink unboiled surface water outside the home (OR 3.0; 95% CI: 1.1–8.20) (Srikantiah et al. 2007). In a similar case–control study in Bangladesh, drinking unboiled water at home was a significant risk factor (OR 12.1; 95% CI: 2.2–65.6) (Ram et al. 2007). Among the risk factors for typhoid and paratyphoid in an urban setting in Indonesia were lack of a toilet in the household (OR 2.20; 95% CI: 1.06–4.55) and use of ice cubes (OR 2.27; 95% CI: 1.31–3.93). Enteric fevers are also a risk among travellers to endemic area, but effective vaccines are usually recommended against the risk of waterborne or food-borne transmission.

Trachoma

The evidence of the effectiveness of environmental interventions (including water and sanitation) alone to prevent active trachoma is not clear (Rabiu et al. 2012). Reviews of the WHO-backed SAFE strategy for trachoma control conclude that there is comparatively weak evidence of the effectiveness of the ‘F’ (facial cleanliness) and ‘E’ (environmental improvement) components that encompass improved access to water and better sanitation (Emerson et al. 2000a; Kuper et al. 2003). In a 6-month cluster RCT of 21 villages in the Gambia, Emerson and colleagues (Emerson et al. 2004) reported a reduction in fly catches among study clusters receiving latrines. However, the prevalence of active trachoma associated with the intervention was not statistically lower than among seven control clusters (relative risk (RR) 0.81; 95% CI: 0.54–1.22). The study has not been repeated. In another review, overall sanitation was associated with lower trachoma as measured by the presence of trachomatous inflammation-follicular or trachomatous inflammation-intense (TF/TI) (OR 0.70, 95% CI 0.62–0.79) and *C. trachomatis* infection (OR 0.62, 95% CI 0.44–0.87) (Freeman et al. 2017). Altogether, findings from this study provide strong evidence for the association between sanitation and active trachoma in both observational and intervention studies; however, the impact upon *C. trachomatis* is less conclusive due to lack of intervention studies conducted.

Malnutrition

Malnutrition is linked to water and sanitation, both through food production (irrigation and safe use of waste for fertilizer) and by preventing diarrhoea and enteric infections that interfere with normal adsorption of nutrients (Fewtrell et al. 2007). An association of malnutrition with diarrhoea has been well established, but the impact of interventions is less clear. In a pooled analysis of nine studies with diarrhoea and growth data for 1,393 children, the probability of stunting at 24 months of age increased by 2.5 per cent per episode of diarrhoea, and 25% of all stunting in 24-month-old children was attributable to having five or more episodes of diarrhoea in the first 2 years of life (Checkley et al. 2008). The Maternal and Child Undernutrition Series in *The Lancet* estimated that sanitation and hygiene interventions implemented with 99% coverage would reduce diarrhoea incidence by 30%, which would in turn decrease the prevalence of stunting by only 2.4 per cent (Bhutta et al. 2013). A report for the World Bank suggests that open defecation explained 54% of international variation in child height by contrast with gross domestic product, which only explained 29% (Spears 2013). However, a recent systematic review and meta-analysis of sanitation's effect on infectious disease and nutritional status found no evidence of impact on weight-for-age Z-scores, weight-for-height Z-scores, and only a borderline effect on height-for-age Z-scores in children under 5 years of age (Freeman et al. 2017).

Other benefits of water and sanitation interventions

Improving water availability (quantity and access)

As the section 'Evidence of effectiveness' suggests, improving water availability—even without a corresponding improvement in water quality—is associated with reductions of diseases transmitted through waterborne, water-washed, and water-based routes. This is partly due to the well-established relationship between the amount of water that people use and the time required to collect it (Cairncross and Feachem 1993). Figure 2.7.2 shows the L-shaped curve that characterizes water consumption patterns based on service levels. While significant quantities of water are used if delivered directly to the home, the quantity used is fairly constant when the collection time is 5–30 minutes and further diminishes for longer collection times. In lower-income settings, average daily per capita consumption is about 150 L for those with household connections, 50 L for yard taps, and just 15 L for communal sources such as stand posts, wells, and springs. Thus, assessments of water availability are expressed in terms of distance: normally 1 km or round trip

collection time (normally 30 minutes) (WHO and UNICEF 2013) but 0.5 km in disaster response (Sphere Project 2004).

A WHO report concluded that a minimum quantity of water for basic health protection is 20 L/person/day (Hutton and Bartram 2003). Of this, 7.5 L is for consumption (hydration and food preparation) and therefore must be of a quality to present minimal health risk; the balance is for basic personal and domestic hygiene. The report recognizes, however, that in addition to the direct health benefits associated with improving water supplies, there are indirect health and other benefits. Indirect health benefits may accrue from reducing the amount of time collecting water which can then be used more effectively at home caring for children or engaging in other productive activities (Cairncross 1987). Services at clinics, medical posts, and other healthcare facilities also benefit from improved water supplies. Sufficient water for irrigating gardens and crops can improve nutrition and generate income (Thompson et al. 2001). Vending water and making and selling beverages can also impact poverty. Finally, to the extent that people are paying for water, improved water supplies may result in savings that can be used for food and other necessities that may impact health outcomes.

Improved school attendance

Although weaker than often asserted, there is some evidence that water and sanitation interventions at schools can improve school attendance (Sclar et al. 2017). A recent study from India showed that a national government programme to install toilets in schools resulted in an 8% increase in enrolment among pubescent-age boys and girls and a 12% increase among younger children of both genders (Adukia 2014). In Africa alone, Hutton et al. (2007) estimated that meeting the MDGs for water and sanitation would increase school attendance by 99 million days annually; full access to basic water and sanitation for all would increase school attendance by 140 million days each year. The sanitary needs of girls and the negative impact of lack of sanitation adversely impacts their attendance levels. In a recent trial in Kenya, schools with poor water access during the dry season that received combined water supply, hygiene promotion and water treatment together with improvements in sanitation experienced an increase in average attendance equal to 26 additional pupils per school on average (Garn et al. 2013). The proportion of girls enrolled in school also increased by 4%. In Uganda, 94% of girls reported problems at school during menstruation and 61% reported staying away from school (IRC International Water and Sanitation Centre 2006). Cultural and religious constraints in many settings make menstruation a taboo. If menstruation lasts over a week, there is a tendency for girls to skip the entire school year (Bharadwaj and Patkar 2004). Sanitation can play an important role in

improving educational access for children with disabilities, through the improvement of paths, latrine floors, and installation of handrails (Bannister et al. 2005).

Security and gender equality

Improved sanitation and water supplies advance security and gender equality for women and girls (Caruso et al. 2017a). Household sanitation can increase their safety by avoiding the dangers of sexual assault and harassment faced when practising open defecation or using latrines away from their homes. Safe, private, and proximate latrines are now considered a basic human right (United Nations 2010); they are a particular issue for women in emergencies and conflicts (Sphere Project 2004). Research in Kenya revealed that women would defecate into plastic bags and throw them into streets ('flying toilets') because they feared being raped when using latrines shared with men (Maili Saba Research Report 2005). They were also afraid of being seen to be using latrines too regularly and preferred to bathe within their own homes after dark where they felt safer. Caruso and colleagues have developed a scale for assessing sanitation insecurity among women (2017). In a study in rural India, women reported unaddressed urination, defecation and menstruation concerns due to inadequate sanitation facilities and were linked with an array of adverse adaptations (suppression, withholding food and water) and consequences (scolding, shame, fear) (Caruso et al. 2017b). Young children often prefer open defecation due to fear of falling into pits in poorly-designed or unsuitably-adapted latrines.

HIV/AIDS

It is well known that access to safe drinking water and sanitation prolongs the lives of people living with HIV/AIDS (PLWHA) by reducing the risk of opportunistic infections, including diarrhoeal diseases (Peletz et al. 2013). Household-based water treatment has been shown to be an effective intervention in preventing mortality and morbidity in a population with one or more persons infected with HIV (Colford et al. 2005; Lule et al. 2005; Peletz et al. 2012). A recent systematic review by Yates and colleagues (2015) further confirmed these results, finding that household water treatment, water supply, and hygiene interventions reduced morbidity from diarrhoea and waterborne disease-causing agents in PLWHA. There is also evidence that household-based water filters and insecticide-treated bednets for malaria slow the progression of HIV (Walson et al. 2013). Water and sanitation are now recognized as particular priorities in-home-based care of

PLWHA (WHO 2008). Point-of-use water treatment products are now included in health kits for people living with HIV/AIDS (Colindres et al. 2007).

Recent and emerging developments in water and sanitation

SDGs in water and sanitation

The MDG water and sanitation targets—to halve the proportion of the 1990 population without sustainable access to safe drinking water and basic sanitation—have been a major factor in mobilizing efforts to expand water and sanitation coverage. While there is still some controversy about JMP claims on the water target and overall disappointment on progress to the sanitation target, successor goals have now been established. SDG targets 6.1 and 6.2 relate to drinking water, sanitation, and hygiene. They represent significant advancements the MDG by (i) calling for universal (as opposed to reducing by half those who were unserved) and addressing inequalities in service levels, (ii) addressing hygiene, which was not been included in the MDGs, (iii) specifically requiring that drinking water should be safe and affordable, and that sanitation should be adequate, and (iv) including explicit references to ending open defecation and to the needs of women and girls and those in vulnerable situations. As discussed just now, the JMP has developed enhanced indicators that allow for more direct monitoring of the new water and sanitation targets (WHO and UNICEF 2017).

WHO Guidelines on Sanitation

While the WHO has long established guidance on water (e.g. Guidelines for Drinking Water Quality, GDWQ) and certain aspects of excreta management. For the first time in 2018, the WHO has published a set of Guidelines on Sanitation (WHO 2018)). The guidelines are intended to set forth a comprehensive framework for health-protecting sanitation, covering policy formulation, implementation of sanitation technologies and behavioural interventions, risk-based hazard identification, and management approaches and monitoring. They also aim to provide an authoritative basis for the effective consideration of public health in setting international, regional, national, or municipal sanitation policies and actions. They are intended to encourage achievable practices and the formulation of sound regulations, applicable to low-income, middle-income, and industrialized countries alike. The guidelines provide three general recommendations that emphasize the need for universal coverage and use of sanitation, focus on the entire faecal management chain, and encourage linkages with other sectors. More specific guidance is provided on best practices. Like the

GDWQ, the Guidelines on Sanitation are seen as an evolving document that will reflect emerging research and experience in sanitation.

Systems approach to understanding enteric pathogen transmission

Much of the research described in previous sections regarding the health impact of water and sanitation interventions has focused on assessing the impact of interrupting particular transmission pathways. Increasingly, epidemiologists are recognizing the interaction and interdependencies that characterize the transmission of enteric pathogens. This is leading to calls for a systems approach that not only acknowledges the multiple transmission pathways and varied aetiological agents involved in diarrhoea and other enteric diseases but also their dynamism and the community level impact of interventions (Eisenberg et al. 2012). Examples of this community effect are the health benefits that inure to non-adopters who live adjacent to households with latrines or who practise effective household water treatment. Interventions that interrupt transmission of one or more pathways can also impact transmission through other pathways of one or more pathogens. It is important that studies endeavour to capture externalities and the wider impact of an intervention at the community level, both to understand and design interventions that that optimize this larger effect and to correctly assess the impact (and cost-effectiveness) of the interventions.

Need for consistency/adherence in safe water consumption

Epidemiological modelling using QMRA has also been used to estimate the impact of sporadic exposure to contaminated drinking water, either as a result of interruptions in water supplies (Hunter et al. 2009) or from inconsistent use of point-of-use water treatment (Brown and Clasen 2012; Enger et al. 2013). The models show that in settings with moderately contaminated water supplies, much of the health benefit from improved drinking water quality is vitiated by even occasional consumption of untreated water. These results raise questions about the health benefits that can be gained from unreliable water supplies. Moreover, insofar as point-of-use water treatment campaigns often fail to achieve high levels of adherence, these models raise questions about the potential health benefits from household-based water treatment.

Household-based water treatment

For the hundreds of millions who lack household water connections that provide drinking water on a 24/7 basis, water is often collected and stored in the home until

needed. It is well known that even water that is safe at the point of collection undergoes frequent and extensive re-contamination during collection (or compromised distribution), storage, and use in the home (Wright et al. 2004). While providing safe, piped-in, disinfected water to each household is an important goal, even the purported achievement of the MDG water target leaves more than three-quarters of a billion people without access to improved water and hundreds of millions more without safe drinking water (Clasen 2012). Accordingly, the WHO and others have called for other approaches that will accelerate the heath and economic gains associated with safe water while progress is made in improving infrastructure.

Household water treatment and safe storage (HWTS) represent one such alternative (Sobsey 2002). Boiling, filtering, and chemically disinfecting water in the home are already in use, with more than 1 billion people reportedly treating their water at home before drinking it (Rosa and Clasen 2010). Recent systematic reviews have shown household-based interventions (home-based boiling, chlorination, filtration, solar disinfection, and flocculation/disinfection) to be significantly more effective than traditional, non-reticulated source-based interventions (protected wells and springs, boreholes, communal tap stands) in improving microbiological water quality and reducing diarrhoeal disease (Fewtrell et al. 2005; Clasen et al. 2015).

Shared and public sanitation

An estimated 600 million people—including a fifth of the population of sub-Saharan Africa—rely on public or other shared sanitation facilities (WHO and UNICEF 2017). Historically, such shared facilities were excluded from the definition of ‘improved’ sanitation for purposes of monitoring progress toward the MDG sanitation target, regardless of the level of service or number of people sharing them, because they are deemed unacceptable, unhygienic, and inaccessible. As discussed previously, the JMP made a revision to this policy that allows shared latrines to be considered ‘improved’ and categorized as a ‘limited’ sanitation service if they utilize wet or dry sanitation technologies and are shared by two or more households (WHO and UNICEF 2017). However, there is little evidence to support this change. A systematic review and meta-analysis of 11 studies reporting on diarrhoea found increased odds of disease associated with reliance on shared sanitation (OR 1.39; 95% CI: 1.14–1.70) (Heijnen et al. 2014). While there was some evidence of increased risk with the number of households sharing, there was an elevated risk at any level of sharing. An analysis of JMP data also found evidence of increased risk of diarrhoea associated with sharing latrines, even after adjusting for likely confounders (Fuller et al. 2014).

Safe disposal of child faeces

While there are few published studies, the evidence suggests that in many low-income settings, nappies and potties are rarely available or used, making the hygienic collection of early child faeces difficult; if collected, such faeces are often disposed of in a manner that does not prevent further exposure (Gil et al. 2004). In fact, the unsanitary disposal of child faeces may present a greater health risk than that of adults. First, young children represent the highest incidence of enteric infections, and their faeces are most likely to contain agents. Second, young children tend to defecate in areas where susceptible children could be exposed. Third, young children who are also most at risk of mortality and the serious sequelae associated with enteric infection are most likely to be exposed to these ambient agents due to the time they spend on the ground, their tendency to put fingers and fomites in their mouths, and common behaviours such as geophagia. In a meta-analysis of ten observational studies published between 1987 and 2001, Gil et al. (2004) found that child faeces disposal behaviours considered risky (open defecation, stool disposal in the open, stools not removed from soil, stools seen in household soil, and children seen eating faeces) were significantly associated with an increased risk of diarrhoeal diseases (RR 1.23; 95% CI: 1.15–1.32); behaviours considered safe (use of latrines, nappies, potties, toilets, washing diapers) was borderline protective (RR 0.93; 95% CI: 0.85–1.00).

There is evidence that even households with latrines often fail to disposal of child faeces safely; more commonly, they are either not collected or are disposed of with the household's other solid waste (Majorin 2012; Majorin et al. 2014). By definition, this renders the practice 'open defecation' even though these households are usually counted as having 'improved sanitation' due to the presence of the latrine (WHO and UNICEF 2013). Efforts to end open defecation will not be successful unless they include interventions to ensure the safe disposal of child faeces.

Technological and programmatic innovations

Steps to improve water supplies and sanitation, particularly in rural and remote locations, have proved especially challenging. Despite concerted efforts over past decades, vast numbers in Africa and South/South East Asia still lack improved water and sanitation (WHO and UNICEF 2017). While a variety of communal and household-level options have been promoted as alternatives to customary approaches in order to improve water quality, quantity, and proximity, some of these have been found wanting in terms of technological suitability, cost, and sustainability. New challenges include natural or man-made chemical contamination, saline intrusion, increasing urbanization, falling water tables,

threats associated with climate change, and increasing demand for agricultural and industrial uses of water. High upfront costs, lack of financing, uncertain land tenure, inadequate skilled masons for construction, pit-emptying, longer-term waste disposal, and urbanization are major challenges in sanitation.

Public health professionals, donors, programme implementers, social entrepreneurs, and the private sector have responded to these challenges by developing and promoting a variety of technological and programmatic innovations for improving water supplies and sanitation, especially among low-income populations. In water, these include developments in rainwater harvesting, water locating, borehole drilling, well digging, locally fabricated pumps, and other water lifting devices, self-supply strategies, small community water treatment and/or distribution systems, water filling stations/kiosks, and a variety of point-of-use water treatment technologies. In sanitation, much of the effort in low-income settings focuses on improved on-site solutions in rural settings and alternatives to conventional sewerage in urban settings. Communal private latrines have been promoted widely in India and elsewhere. New technologies include cheaper, lightweight squatting slabs, composting toilets, digestion chemicals, multichamber pits, pit-emptying, and improved separation of liquid and solid excreta. Sludge is increasingly being viewed as a useful product rather than waste, with innovations in collection and processing. Many of these innovations are accompanied by entrepreneurial initiatives and public–private partnerships; others through micro-finance and base-of-the-pyramid marketing. While some of these innovations appear promising, lessons from the past suggest that understanding and responding to the particular circumstances present in a given setting—and especially what the target population itself wants and is willing to pay for—are especially important in achieving large-scale sustainable improvements that will also impact public health. Technical innovations in water and sanitation that do not address the need for behaviour change are also unlikely to achieve optimal uptake and impact.

Community-led total sanitation

First developed in Bangladesh in 1999 and expanding widely elsewhere, community-led total sanitation (CLTS) is an approach that empowers local communities to stop open defecation and to build and use latrines without the support of external hardware subsidies. Through the use of participatory techniques, community members analyse their own sanitation situation, including the extent of open defecation and the possibilities of faecal–oral contamination. This is designed to ignite a personal sense of disgust and shame that translates into collective action to reduce the impact of open defecation (Kar 2003). By triggering collective behaviour change, CLTS places the community, rather than the

household, at the centre of the decision-making process. Peer pressure and civic pride are important motivating factors. The particular design of a latrine is secondary to the emphasis on 100% coverage. The results can be impressive, with whole communities changing from open defecation to latrine use in a matter of months (Kar 2003). The approach has since been rolled out in Africa and Asia, and there is some evidence of its sustainability (Kar and Bongartz 2006).

Nevertheless, a recent systematic review of included 200 studies of CLTS interventions concluded that the evidence base on CLTS effectiveness is weak (Venkataramanan 2018). They found that 14 quantitative evaluations reported decreases in open defecation, but did not corroborate the widespread claims of open defecation-free (ODF) villages found in case studies; they also found that over a quarter of the literature overstated conclusions, attributing outcomes and impacts to interventions without an appropriate study design. The authors highlighted the need for more rigorous research on CLTS impacts as well as applied research initiatives that bring researchers and practitioners together to address implementation challenges to improve rural sanitation efforts.

Sanitation marketing

Sanitation marketing uses a commercial approach to the production and delivery of sanitation technologies and engages the private sector for production and delivery in a financially and institutionally sustainable manner (Jenkins and Sugden 2006). Such a marketing approach has been recommended over typical public sector promotion of sanitation since it helps ensure that people choose to receive what they want and are willing to pay for, is financially sustainable, is cost-effective, and can be taken to scale (Cairncross 2004). Sanitation marketing adopts a consumer perspective, starting with an understanding of which products and services the target population wants, will pay for, will maintain, and are appropriate to the local context. It seeks to develop a sustainable sanitation industry, which is not dependent on external donors for hardware subsidies or long-term support for its continuation (Water and Sanitation Programme n.d.). It recognizes the household as the key decision-maker regarding their defecation practice and the importance of effective public–private partnerships. The extent to which the approach is capable of reaching the base of the economic pyramid has not yet been shown.

Enteropathy

Beyond the impact of diarrhoea on malnutrition, researchers have postulated that repeated exposure to faecal pathogens may be a significant cause of environmental

(tropical) enteropathy, a subclinical condition characterized by malabsorption, villus atrophy, crypt hyperplasia, T-cell infiltration, and general inflammation of the jejunum (Humphrey 2009). Enteropathy is caused by faecal bacteria ingested in large quantities by young children living in conditions of poor sanitation and hygiene. Humphrey has suggested that that the primary causal pathway from poor sanitation and hygiene to undernutrition is tropical enteropathy and not diarrhoea. If so, current estimates may have substantially underestimated the contribution of sanitation and hygiene to growth because the effect is modelled entirely through diarrhoea. A recent study in Bangladesh found that markers of enteropathy linked to unhygienic environments (Lin et al. 2013). However, recent randomized controlled studies have shown mixed results on the efficacy and effectiveness of water and sanitation interventions to prevent stunting (Pickering and Davis 2012; Luby et al 2018, Null et al. 2018). This has raised questions about the link between enteropathy and stunting and the potential of environmental interventions to prevent it.

Challenges in water and sanitation

Failure to treat diarrhoea as a serious disease

One of the threshold constraints to scaling up water and sanitation is the belief that diarrhoea—the main health threat associated with poor water and sanitation coverage—is not a disease. Figueroa and Kincaid (2010) cite numerous studies from various countries in which participants reported diarrhoea to be a natural and even desirable condition, especially in young children, not worthy of special preventive measures. Although health benefits often lack significant motivational impetus for driving preventive measures, the fact that diarrhoea is not even considered a disease by many of the most vulnerable populations further limits this strategy. Among policymakers and health officials faced with a variety of life-threatening diseases, diarrhoea may not receive the commitment of resources that its status as the third leading cause of morbidity and mortality from infectious disease would suggest it deserves.

Lack of public sector coordination

In most countries, a variety of agencies and authorities are responsible for some part of water and sanitation. These typically include the ministries of water, health, water resources, environment, local government, rural development, and education. In many cases, there are also federal, regional, district, and local levels of government. Rarely do any of these ministries take full responsibility for all

aspects of water or of sanitation. The result is often a patchwork effort that lacks funding and coordination. There are important examples of successful coordinated public sector efforts. In South Africa, strides in sanitation are occurring because of a national decision and plan which set out targets, clear strategies, significant resources, and accountability (Muller 2002). Ethiopia has also achieved considerable success in improving water and sanitation coverage, particularly in the Southern Nations, Nationalities and Peoples Region, where a commitment at the senior levels translated into coordinated and sustained action (Bibby and Knapp 2007).

Mixed results from water and sanitation interventions

The research just described provides consistent evidence of an association between poor water and sanitation conditions and adverse health outcomes. Nevertheless, a series of recent efficacy studies of WASH interventions have raised questions about the potential of existing approaches to achieve the health outcomes have been shown in systematic reviews (Luby et al. 2018; Null et al. 2018, SHINE trial). Continued research is necessary in order to provide more guidance on the means of optimizing scalable water and sanitation interventions in low-income settings.

Conclusion

Unlike many of the other challenges in public health, the solutions for eliminating most of the disease burden associated with poor water and sanitation are well known. All but the poor have enjoyed the health, economic, and other benefits associated with safe drinking water and basic sanitation for decades. The fact that hundreds of millions still lack access to these fundamental resources is a scandal that generations have allowed to persist simply as a matter of misguided priorities. And as the ‘haves’ continue to make rapid gains, they are not only increasing the gap over the ‘have-nots’ but also using up larger amounts of the world’s limited water supplies and capacity for waste disposal, making it more difficult and costly for others to join their privileged club.

The need to extend water and sanitation coverage is acknowledged at the highest international levels, and progress is being made. Whether these efforts will be any more successful than those expressed in previous international declarations and goals is not yet clear. As the poor continue to wait for the piped-in water supplies and sanitary disposal that they deserve, however, it is incumbent on the public health community to develop, assess, and promote effective, low-cost, and sustainable alternatives and creative delivery strategies in order to accelerate access to the health gains associated with safe drinking water and basic sanitation.

Key points

While safe drinking water and sanitation are widely recognized as fundamental public health interventions, eight per cent of the world's population still lack improved water supplies and 30% lack basic sanitation.

The infectious diseases associated with unsafe drinking water and poor sanitation impose a heavy burden, especially on the poor, the very young, and the immunocompromised; they also aggravate poverty, education, and economic development.

There is strong evidence that interventions to improve water supplies or sanitation can be effective in preventing diarrhoea, soil-transmitted helminth infections, schistosomiasis, and typhoid fevers.

Water and sanitation interventions have also been shown to be cost-effective and cost-beneficial, with significant savings to the public sector from reduced healthcare costs; there is also evidence of other economic and developmental benefits from improved access to water and sanitation.

A variety of recent and emerging developments, including new methods for assessing and monitoring the risk of diseases associated with water and sanitation as well as alternative technologies, programmatic approaches, and implementation strategies, may contribute to improved targeting, coverage, uptake, and sustainability.

Nevertheless, significant political, social, economic, and developmental challenges must be addressed in order to successfully scale up some of these interventions on a sustainable basis and thus provide the most vulnerable populations with the health and other benefits of safe drinking water and sanitation.

ACKNOWLEDGEMENTS

The author acknowledges the contributions of Steven Sugden to a previous version of this chapter and of Christian Landon in support of the current version.

REFERENCES

Adukia, A. (2014). *Sanitation and Education* (PhD thesis). Cambridge, MA: Harvard University.

Allan, S.C. (2003). *The WaterAid Bangladesh/VERC 100% Sanitation Approach: Bangladesh*. IDS Working Paper 194. Brighton: Institute of Development Studies, Brighton.

American Public Health Association, American Water Works Association, and the Water Environment Federation (2012). *Standard Methods for the Examination of Water and Wastewater* (22nd ed.). Washington, DC: American Public Health Association, the American Water Works Association, and the Water Environment Federation.

Bannister, M., et al. (2005). *Water and Sanitation for All: Practical Ways to Improve Accessibility for Disabled People*. Kampala, Uganda: 31st WEDC Conference, Maximising the benefits from water and environmental sanitation.

Bharadwaj, S. and Patkar, A. (2004). *Menstrual Hygiene and Management in Developing Countries: Taking Stock*. Mumbai: Junction Social, Social Development Consultants.

Bhutta, Z.A., et al. (2013). The Lancet Nutrition Interventions Review Group, and the Maternal and Child Nutrition Study Group. Evidence-based interventions for improvement of maternal and child nutrition: what can be done and at what cost? *The Lancet*, **382**, 452–77.

Bibby, S. and Knapp, A. (2007). *From Burden to Communal Responsibility: A Sanitation Success Story from Southern Region in Ethiopia*. Field Note. Washington, DC: World Bank Water and Sanitation Programme.

Black, R.E. and Lanata, C.F. (1995). Epidemiology of diarrhoeal diseases in developing countries. In: Blaser, M.J., Smith, P.D., Ravdin, J.I., Greenberg, H.B., and Guerrant, R.L. (eds.) *Infections of the Gastrointestinal Tract*. New York: Raven Press, pp. 11–29.

Black, R.E., Morris, S.S., and Bryce, J. (2003). Where and why are 10 million children dying every year? *The Lancet*, **361**, 2226–34.

Blakely, T., Hales, S., Kieft, C., Wilson, N., and Woodward, A. (2005). The global distribution of risk factors by poverty level. *Bulletin of the World Health Organization*, 83(2), 118–26.

Boisson, S., et al. (2013). Effect of household-based drinking water chlorination on diarrhoea among children under five in Orissa, India: a double-blind randomised placebo-controlled trial. *PLoS Medicine*, 10(8), e1001497.

Brown, J. and Clasen, T. (2012). High adherence is necessary to realize health gains from water quality interventions. *PLoS One*, 7(5), e36735.

Burton, M.J. and Mabey, D.C. (2009). The global burden of trachoma: a review. *PLoS Neglected Tropical Diseases*, 3(10), e460.

Cairncross, S. (1987). The benefits of water supply. In: Pickford, J. (ed.) *Developing World Water*. London: Grosvenor Press, pp. 30–4.

Cairncross, S. (1992). *Sanitation and Water Supply: Practical Lessons from the Decade*. Washington, DC: World Bank.

Cairncross, S. (2004). *The Case for Marketing Sanitation.* Field Note. Nairobi: Water and Sanitation Programme Africa.

Cairncross, S. and Feachem, R. (1993). *Environmental Health Engineering in the Tropics: An Introductory Text* (2nd ed.). Chichester: John Wiley and Sons Ltd.

Cairncross, S., Muller, R., and Zagaria, N. (2002). Dracunculiasis (Guinea worm disease) and the eradication initiative. *Clinical Microbiology Reviews*, **15**, 223–46.

Caruso BA, et al. (2017a). Assessing women's negative sanitation experiences and concerns: the development of a novel sanitation insecurity measure. *International Journal of Environmental Research and Public Health*, 14(7), pii: E755.

Caruso BA, et al. (2017b). Women's gendered experiences of urination, defecation, and menstruation in Rural Odisha, India: understanding and defining sanitation insecurity. *BMJ Global Health*, 2(4), e000414.

Centers for Disease Control and Prevention (2018). *Parasites—Ascariasis*. Available at: https://www.cdc.gov/parasites/ascariasis/

Checkley, W., et al. (2008). The Childhood Malnutrition and Infection Network. Multi-country analysis of the effects of diarrhoea on childhood stunting. *International Journal of Epidemiology*, **37**, 816–30.

Chen, G., et al. (2004). Observation on the effect of the comprehensive measures of replacing cattle with machine and reconstructing water supply and lavatory to control the transmission of schistosomiasis. *Journal of Tropical Disease and Parasitology*, **2**, 219–22.

Clasen, T. (2009). *Scaling Up Household Water Treatment in Low-Income Settings*. Geneva: World Health Organization. [Online] Available at: http://whqlibdoc.who.int/hq/2009/WHO_HSE_WSH_09.02_eng.pdf

Clasen, T. (2012). MDG water target claim exaggerates achievement. *Tropical Medicine & International Health*, 17(10), 1178–80.

Clasen, T. F., et al. (2015). Interventions to improve water quality for preventing diarrhoea. *The Cochrane Database of Systematic Reviews*, (10), 1–201.

Clasen. T., et al. (2007). Cost-effectiveness analysis of water quality interventions for preventing diarrhoeal disease in developing countries. *Journal of Water and Health*, **5** (4), 599–608.
Colford, J.M. Jr, et al. (2005). A pilot randomized, controlled trial of an in-home drinking water intervention among HIV + persons. *Journal of Water and Health*, 3(2), 173–84.
Colindres, R., et al. (2007). Utilization of a basic care and prevention package by HIV-infected persons in Uganda. *AIDS Care*, **24**, 1–7.
Crump, J.A., Luby, S.P., and Mintz, E.D. (2004). The global burden of typhoid fever. *Bulletin of the World Health Organization*, **82**, 346–53.
Crump, J.A. and Mintz, E.D. (2010). Global trends in typhoid and paratyphoid fever. *Clinical Infectious Diseases*, **50**, 241–6.
Curtis, V. and Cairncross, S. (2003). Effect of washing hands with soap on diarrhoea risk in the community: a systematic review. *The Lancet Infectious Diseases*, **3**, 275–81.
Cutler, D. and Miller, G. (2005). The role of public health improvements in health advances: the twentieth-century United States. *Demography*, **42** (1), 1–22.
Dangour, A.D., et al. (2013). Interventions to improve water quality and supply, sanitation and hygiene practices, and their effects on the nutritional status of children. *Cochrane Database of Systematic Reviews*, **8**, CD009382.
Davis, J. and Lambert, R. (2002). *Engineering in Emergencies*. London: Intermediate Technology Publications, Ltd.
Department of International Development (1998). *Guidance Manual on Water Supply and Sanitation Programmes*. London: Department of International Development.
Desilva, N.R., et al. (2003). Soil-transmitted helminth infections: updating the global picture. *Trends in Parasitology*, **19**, 547–51.
Eisenberg, J.N., Trostel J., Sorensen R.J., and Shields K.F. (2012). Toward a systems approach to enteric pathogen transmission: from individual independence to community interdependence. *Annual Review of Public Health*, **33**, 239–57.
Emerson, P.M., Bailey, R.L., and Mahdi, O.S. (2000). Transmission ecology of the fly *Musoca sorbens*, a putative vector of trachoma. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, **94**, 1–5.
Emerson, P.M., et al. (2000a). Review of the evidence base for the ‘F’ and ‘E’ component of the SAFE strategy for trachoma control. *Tropical Medicine and International Health*, 5(8), 515–27.
Emerson, P.M., et al. (2004). Role of flies and provision of latrines in trachoma control: cluster-randomised controlled trial. *The Lancet*, **363**, 1093–98.
Enger, K.S., Nelson, K.L., Rose, J.B., and Eisenberg, J.N. (2013). The joint effects of efficacy and compliance: a study of household water treatment effectiveness against childhood diarrhea. *Water Research*, 47(3), 1181–90.
Ensink, J.H. and van der Hoek, W. (2007). New international guidelines for wastewater use in agriculture. *Tropical Medicine and International Health*, 12(5), 575–7.
Esrey, S.A., Feachem, R.G., and Hughes, J.M. (1985). Interventions for the control of diarrhoeal diseases among young children: improving water supplies and excreta disposal facilities. *Bulletin of the World Health Organization*, **63**, 757–72.
Esrey, S.A., et al. (1991). Effects of improved water supply and sanitation on ascariasis, diarrhoea, dracunculiasis, hookworm infection, schistosomiasis, and tracoma. *Bulletin of the World Health Organization*, **69**, 609–21.
Evans, B., Hutton, G., and Haller, L. (2004). *Closing the Sanitation Gap—The Case for Better Public Funding of Sanitation and Hygiene*. Oslo: Commission on Sustainable Development.
Ferriman, A. (2007). BMJ readers choose the ‘sanitary revolution’ as greatest medical advance since 1840. *British Medical Journal*, **334**, 111.
Fewtrell, L., et al. (2005). Water, sanitation, and hygiene interventions to reduce diarrhoea in less developed countries: a systematic review and meta-analysis. *The Lancet Infectious Diseases*, **5**, 42–52.
Fewtrell, L., et al. (2007). *Water, Sanitation and Hygiene—Quantifying the Health Impact at National and Local Levels in Countries with Incomplete Water Supply and Sanitation Coverage*. Environmental Burden of Disease Series No. 15. Geneva: World Health Organization.
Figueroa, M.E. and Kincaid, D.L. (2010). *Social, Cultural and Behavioral Correlates of Household Water Treatment and Storage*. Center Publication HCI 2010-1: Health Communication Insights. Baltimore,

MD: Johns Hopkins Bloomberg School of Public Health, Center for Communication Programs.
Freeman, M.C., et al. (2017). The impact of sanitation on infectious disease and nutritional status: a systematic review and meta-analysis. *International Journal of Hygiene and Environmental Health*, 220(6), 928–49.
Freeman, M.C., et al. (2012). Assessing the impact of a school-based water treatment, hygiene and sanitation programme on pupil absence in Nyanza Province, Kenya: a cluster-randomized trial. *Tropical Medicine & International Health*, 17(3), 380–91.
Fuller, J.A., Clasen, T., Heijnen, M., and Eisenberg, J.N. (2014). Shared sanitation and the prevalence of diarrhea in young children: evidence from 51 countries, 2001–2011. *American Journal of Tropical Medicine and Hygiene*, 91(1), 173–80.
Garn, J.V., et al. (2013). A cluster-randomized trial assessing the impact of school water, sanitation, and hygiene improvements on pupil enrollment and gender parity in enrollment. *Journal of Water, Sanitation and Hygiene for Development*, 3(4), 10.2166/washdev.2013.217.
Gil, A., Lanata, C., Kleinau, E., and Penny, M. (2004). *Children's Feces Disposal Practices in Developing Countries and Interventions to Prevent Diarrheal Diseases.* Environmental Health Project. New York: USAID.
Gleeson, C. and Gray, N. (1997). *The Coliform Index and Waterborne Disease*. London: E and FN Spon.
Gryssels, B., et al. (2006). Human schistosomiasis. *The Lancet*, **368**, 1106–18.
Haas, C.N., Rose, J.B., and Gerba, C.P. (1999). *Quantitative Microbial Risk Assessment*. New York: John Wiley and Sons.
Hayes, C., et al. (2003). Food and water safety for persons infected with human immunodeficiency virus. *Clinical Infectious Diseases*, 36(Suppl 2), S106–9.
Heijnen, M., et al. (2014). Shared sanitation versus individual household latrines: a systematic review of healthoutcomes. *PLoS One*, 9(4), e93300.
Hotez, P., et al. (2004). Hookworm infection. *The New England Journal of Medicine*, **351**, 799–807.
Hu, V.H., Harding-Esch, E.M., Burton, M.J., Bailey, R.L., Kadimpeul, J., and Mabey, D.C. (2010). Epidemiology and control of trachoma: systematic review. *Tropical Medicine & International Health*, 15(6), 673–91.
Humphrey, J.H. (2009). Child undernutrition, tropical enteropathy, toilets, and handwashing. *The Lancet*, **374**, 1032–5.
Hunter, P.R. (1997). *Waterborne Disease Epidemiology and Ecology*. Chichester: John Wiley and Sons.
Hunter, P.R. (2009). Household water treatment in developing countries: comparing different intervention types using meta-regression. *Environmental Science & Technology*, 43(23), 8991–7.
Hunter, P.R., Zmirou-Navier, D., and Hartemann, P. (2009). Estimating the impact on health of poor reliability of drinking water interventions in developing countries. *Science of the Total Environment*, 407(8), 2621–4.
Hutton, G. (2013). Global costs and benefits of reaching universal coverage of sanitation and drinking-water supply. *Journal of Water and Health*, 11(1), 1–12.
Hutton, G. and Bartram, J. (2003). *Domestic Water Quantity, Service Level and Health*. Geneva: World Health Organization.
Hutton, G., Haller, L., and Bartram, J. (2007). Global cost-benefit analysis of water supply and sanitation interventions. *Journal of Water and Health*, 5(4), 481–502.
Hutton, G. and Varughese, M.C. (2016). The costs of meeting the 2030 Sustainable Development Goal targets on drinking water sanitation, and hygiene. *Water and Sanitation Programme*. Washington, DC: World Bank Group.
IRC International Water and Sanitation Centre (2006). *Girl Friendly Toilets for School Girls*. [Online] Available at: http://www.washinschools.info/page/319
Jackson, B. (2005). *A Review of EcoSan Experience in Eastern and Southern Africa*. Field Note. Washington, DC: World Bank Water and Sanitation Programme.
Jenkins, M.W. and Curtis, V. (2005). Achieving the 'good life'; why some people want latrines in rural Benin. *Social Science & Medicine*, **61**, 2446–59.
Jenkins, M.W. and Scott, B. (2007). Behavioral indicators of household decision-making and demand for sanitation and potential gains from social marketing in Ghana. *Social Science & Medicine*, 64(12), 2427–42.

Jenkins, M.W. and Sugden, S. (2006). *Rethinking Sanitation—Lessons and Innovation for Sustainability and Success in the New Millennium*. UNDP HDR, Sanitation Thematic Paper. London: London School of Hygiene and Tropical Medicine.

Kar, K. (2003). *Subsidy or Self-Respect? Participatory Total Community Sanitation in Bangladesh*. IDS Working Paper 194. Brighton: Institute of Development Studies, Brighton.

Kar, K. and Bongartz, J. (2006). *Update on Some Recent Developments in Community-Led Total Sanitation*. Brighton: Institute of Development Studies, Brighton.

Kosek, M., Bern, C., and Guerrant, R.L. (2003). The global burden of diarrhoeal disease, as estimated from studies published between 1992 and 2000. *Bulletin of the World Health Organization*, **81**, 197–204.

Kotloff, K.L., et al. (2013). Burden and aetiology of diarrhoeal disease in infants and young children in developing countries (the Global Enteric Multicenter Study, GEMS): a prospective, case–control study. *The Lancet*, **382**, 209–22.

Kumaresan, J. and Mecaskey, J. (2003). The global elimination of blinding trachoma: progress and promise. *The American Journal of Tropical Medicine and Hygiene*, **69**, 24–8.

Kuper, H., et al. (2003). A critical review of the SAFE strategy for the prevention of blinding trachoma. *The Lancet Infectious Diseases*, 3(6), 372–81.

Leclerc, H., Schwartzbrod, L., and Dei-Cas, E. (2002). Microbial agents associated with waterborne diseases. *Critical Reviews in Microbiology*, 28(4), 371–409.

Lin, A., et al. (2013). Household environmental conditions are associated with enteropathy and impaired growth in rural Bangladesh. *The American Journal of Tropical Medicine and Hygiene*, 89(1), 130–7.

Luby, S.P., et al. (2018) Effects of water quality, sanitation, handwashing and nutritional interventions on diarrhoea and child growth in rural Bangladesh: a cluster randomized trial. *Lancet Global Health*, 6(3), e302–15.

Lule, J.R., et al. (2005). Effect of home-based chlorination and safe storage on diarrhea among persons with HIV in Uganda. *Tropical Medicine and International Health*, **73**, 926–33.

Lustigman, S., et al. (2012). A research agenda for helminth diseases of humans: the problem of helminthiases. *PLoS Neglected Tropical Diseases*, 6(4), e1582.

Mabey, D.C., Solomon, A.W., and Foster, A. (2003). Trachoma. *The Lancet*, **362**, 223–9.

Maili Saba Research Report (2005). *Livelihood and Gender in Sanitation and Hygiene Water Services Among Urban Poor*. London: Overseas Development Institute.

Majorin, F. (2012). *Child Faeces Disposal Practices in Rural Orissa: A Cross Sectional Study*. MSc Thesis. London: London School of Hygiene and Tropical Medicine.

Majorin, F., et al. (2014) Child feces disposal practices in rural Orissa: a cross sectional study. *PLoS One*, 9(2), e89551.

Methra, M. and Knapp, A. (2005). *The Challenge of Financing Sanitation for Meeting the Millennium Development Goals*. Commissioned by the Norwegian Ministry of the Environment for the Commission on Sustainable Development. Water and Sanitation Programme—Africa. Nairobi: World Bank.

Mukherjee, N. (2000). *Myth vs. Reality in Sanitation and Hygiene Promotion*. Field Note, Jakarta: World Bank Water and Sanitation Programme—East Asia and the Pacific.

Muller, M. (2002). *The National Water and Sanitation Programme in South Africa*. Field Note, Water and Sanitation Programme—Africa. Nairobi: World Bank.

Nauges, C. and Whittington, D. (2009). Estimation of water demand in developing countries: an overview. *The World Bank Research Observer*, **25**, 263–94.

Null C, et al. (2018) Effects of water quality, sanitation, handwashing, and nutritional. Interventions on diarrhoea and child growth in rural Kenya: a cluster-randomised controlled trial. *Lancet Global Health*, 6(3), e316–29.

Obika, A., et al. (2002). *Social Marketing for Urban Sanitation, Inception Report*. DFID KAR Project R7982. Loughborough: WEDC.

Peletz, R.L., et al. (2013). Water, sanitation and hygiene interventions to improve health among people living with HIV/AIDS: a systematic review. *AIDS*, 27(16), 2593–601.

Peletz, R., et al. (2012). Assessing water filtration and safe storage in households with young children of HIV-positive mothers: a randomized, controlled trial in Zambia. *PLoS One*, 7(10), e46548.

Pickering, A.J. and Davis, J. (2012). Freshwater availability and water fetching distance affect child health in sub-Saharan Africa. *Environmental Science & Technology*, 46(4), 2391–7.

Prüss-Ustün, A., et al. (2014). Burden of disease from inadequate water, sanitation and hygiene in low- and middle-income settings: a retrospective analysis of data from 145 countries. *Tropical Medicine & International Health*, 19(8), 894–905.

Pullan, R.L., Freeman, M.C., Gething, P.W., and Brooker, S.J. (2014). Geographical inequalities in use of improved drinking water supply and sanitation across Sub-Saharan Africa: mapping and spatial analysis of cross-sectional survey data. *PLoS Med*, 11(4), e1001626.

Rabiu, M., Alhassan, M.B., Ejere, H.O., and Evans, J.R. (2012). Environmental sanitary interventions for preventing active trachoma. *Cochrane Database of Systematic Reviews*, **2**, CD004003.

Ram, P.K., et al. (2007). Risk factors for typhoid fever in a slum in Dhaka, Bangladesh. *Epidemiology and Infection*, 135(3), 458–65.

Rosa, G. and Clasen, T. (2010). Estimating the scope of household water treatment in low- and medium-income countries. *The American Journal of Tropical Medicine and Hygiene*, 82(2), 289–300.

Sclar, G.D., et al. (2017). Effects of sanitation on cognitive development and absence: a systematic review. *International Journal of Hygiene and Environmental Health*, 220(6), 917–27.

Sobsey, M.D. (2002). *Managing Water in the Home: Accelerated Health Gains from Improved Water Supply*. WHO/SDE/WSH/02.07. Geneva: World Health Organization.

Spears, D. (2013). *How Much International Variation in Child Height can Sanitation Explain?* World Bank policy research working paper, no. WPS 6351. [Online] Available at: http://go.worldbank.org/SZE5WUJBI0

Sphere Project (2004). *Humanitarian Charter and Minimum Standards in Disaster Response*. Geneva: The Sphere Project.

Srikantiah, P., et al. (2007). Epidemiology and risk factors for endemic typhoid fever in Uzbekistan. *Tropical Medicine and International Health*, 12(7), 838–47.

Stephenson, L.S., Latham, M.S., and Ottensen, E.A. (2000). Malnutrition and parasitic helminth infections. *Parasitology*, **121**, S23–38.

Stocks, M.E., Ogden, S., Haddad, D., Addiss, D.G., McGuire, C., and Freeman, M.C. (2014). Effect of water, sanitation, and hygiene on the prevention of trachoma: a systematic review and meta-analysis. *PLoS Med*, 11(2), e1001605.

Strunz, E.C., Addiss, D.G., Stocks, M.E., Ogden, S., Utzinger, J., and Freeman, M.C. (2014). Water, sanitation, hygiene, and soil-transmitted helminth infection: a systematic review and meta-analysis. *PLoS Med*, 11(3), e1001620.

Thompson, J., et al. (2001). *Drawers of Water II: 30 Years of Change in Domestic Water Use and Environmental Health in East Africa*. London: IIED.

Troeger, C., et al. (2017). Estimates of global, regional, and national morbidity, mortality, and aetiologies of diarrhoeal diseases: a systematic analysis for the Global Burden of Disease Study 2015. *The Lancet Infectious Diseases* 17(9): 909–48.

UN Millennium Project (2005). *UN Millennium Project Task Force on Water and Sanitation—Health, Dignity and Development: What Will it Take?* London: Earthscan.

United Nations (2000). *United Nations Millennial Declaration*. General Assembly Res. 55/2, 18 September 2000. New York: United Nations.

United Nations (2010). *The Human Right to Water and Sanitation*. United Nations General Assembly Res. 64/292, 3 August 2010. New York: United Nations.

United Nations (2015). *Transforming our world: the 2030 Agenda for Sustainable Development*. United Nations General Assembly Res 70/1, 25 September 2015. New York: United Nations

United Nations Development Programme (2007). *Beyond Scarcity: Power, Poverty and the Global Water Crisis*. Human Development Report 2006. New York United Nations Development Programme.

Waddington, H., Snilstveit, B., White, H., and Fewtrell, L. (2009). *Water, Sanitation and Hygiene Interventions to Combat Childhood Diarrhoea in Developing Countries*. New Delhi: 3ie.

Wagner, E.G. and Lanois, J.N. (1958). *Excreta Disposal for Rural Areas and Small Communities*. WHO monograph series no. 39. Geneva: WHO.

Walson, J.L., et al. (2013). Evaluation of impact of long-lasting insecticide-treated bed nets and point-of-use water filters on HIV-1 disease progression in Kenya. *AIDS*, 27(9), 1493–501.

Wang, H., et al. (2016). Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980–2015: a systematic analysis for the Global Burden of Disease Study 2015. *The Lancet*, *388*(10053): 1459–544.

Water and Sanitation Programme (2002). *Selling Sanitation in Vietnam: What Works?* Jakarta: World Bank Water and Sanitation Programme—East Asia and the Pacific.
Water and Sanitation Program (n.d.). *Sanitation Marketing Toolkit*. World Bank. [Online] Available at: http://www.wsp.org/toolkit/toolkit-home
Water Resources Institute (2007). *Water Resources Institute. EarthTrends Environmental Resource Portal.* [Online] Available at: https://www.wri.org/our-work/project/earthtrends-environmental-information
White, G.F., Bradley, D.J., and White, A.U. (1972). *Drawers of Water: Domestic Water Use in East Africa*. Chicago, IL: University of Chicago Press.
Whittington, D., Jeuland, M., Barker, K., and Yuen Y (2012). Setting priorities, targeting subsidies among water, sanitation, and preventive health interventions in developing countries. *World Development*, 40(8), 1546–68.
Winblad, U. and Simpson-Hébert, M. (eds.) (2004). *Ecological Sanitation*. Stockholm: Stockholm Environmental Institute.
Wolf, J., et al. (2014). Assessing the impact of drinking water and sanitation on diarrhoeal disease in low- and middle-income settings: systematic review and metaregression. *Tropical Medicine and International Health*, 19(8), 928–42.
World Bank (n.d.). *Water and Sanitation Program*. [Online] Available at: http://www.wsp.org/
World Health Organization (WHO) (2006). *Guidelines for the Safe Use of Wastewater, Excreta and Greywater*, Vols. 1–4. Geneva: WHO.
World Health Organization (WHO) (2008). *Essential Prevention and Care Interventions for Adults and Adolescents Living with HIV in Resource-Limited Settings*. Geneva: WHO.
World Health Organization (WHO) (2012). *Guidelines for Drinking-Water Quality* (Vol. 1, 4th ed.). Geneva: WHO.
World Health Organization (WHO) (2013). *Global Health Observatory*. [Online] Available at: http://www.who.int/gho/en/
World Health Organization (WHO) (2015). *Typhoid*. [Online] Available at: http://www.who.int/immunization/diseases/typhoid/en/
World Health Organization (WHO) (2017). *Global Health Observatory*. [Online] Available at: http://www.who.int/gho/neglected_diseases/trachoma/en/
World Health Organization (WHO) (2017). *Guideline: Preventive Chemotherapy to Control Soil-Transmitted Helminth Infections in At-Risk Population Groups*. Geneva: WHO.
World Health Organization (WHO) (2018). *Guidelines on Sanitation and Health*. Geneva: WHO. [Online] Available at: https://www.who.int/water_sanitation_health/sanitation-waste/sanitation/sanitation-guidelines/en/
World Health Organization and the United Nations Children's Fund (WHO and UNICEF) (2002). *Global Water Supply and Sanitation Assessment*. Geneva: WHO and UNICEF.
World Health Organization and the United Nations Children's Fund (WHO and UNICEF) (2005). *Water for Life: Decade for Action 2005–2015*. Geneva: WHO and UNICEF.
World Health Organization and the United Nations Children's Fund (WHO and UNICEF) (2014). *Progress on Sanitation and Drinking Water: 2014 Update*. New York: WHO and UNICEF Joint Monitoring Programme on Water and Sanitation.
World Health Organization and the United Nations Children's Fund (WHO and UNICEF) (2017). *Progress on Drinking Water, Sanitation and Hygiene: 2017 Update and SDG Baselines*. Geneva: WHO and UNICEF.
Wright, J., Gundry, S., Conroy (2004). Household drinking water in developing countries: a systematic review of microbiological contamination between source and point-of-use. *Tropical Medicine & International Health*, **9** (1), 106–17.
Yates, T., et al. (2015). The impact of water, sanitation, and hygiene interventions on the health and well-being of people living with HIV: a systematic review. *Journal of Acquired Immune Deficiency Syndromes*, *68*, 318–30.
Zhang, S.-Q., et al. (2005). [Observation on comprehensive measures of safe treatment of night-soil and water supply, replacement of bovine with machine for schistosomiasis control]. *Zhongguo Xue Xi Chong Bing Fang Ji Za Zhi [Chinese Journal of Schistosomiasis Control]*, 17(6), 437–42.
Ziegelbauer, K., Speich, B., Mäusezahl, D., Bos, R., Keiser, J., and Utzinger, J. (2012). Effect of sanitation on soil-transmitted helminth infection: systematic review and meta-analysis. *PLoS Medicine*, 9(1),

e1001162.

2.8

Food and nutrition

Roger Shrimpton, David Sanders, and Anne Marie Thow

Introduction

Nutritional science concerns the study of nutrients in food, how the body uses nutrients, and the relationship between diet, health, and disease. In previous centuries the focus was more on identifying the nutrient components in the diet, and single nutrient deficiencies, such a scurvy caused by vitamin c deficiency, beriberi caused by thiamine deficiency, and goitre caused by iodine deficiency. By the end of the Second World War, virtually all the nutrients had been identified and the work was thought to have been completed (Carpenter 2003). Early nutrition programme efforts around the turn of the millennium focused on tackling child undernutrition especially, but in more recent times nutrition science has begun to study population-based evidence for the relationships between specific dietary patterns, nutrient intakes, and the non-communicable diseases of obesity, cancer, diabetes, and atherosclerosis.

Nutrition paradigms have shifted in the last few decades, and now there is an urgent need for the policy discourse and public health practice to catch up. Effective public health policy and practice responses to nutrition depend on an accurate understanding of the problem of ‘malnutrition’, its implications for human health, and its underlying causes. They also depend on the political economy of nutrition policy—including policies relating to food environments and the global food system. The aim of this chapter is to map out the public health dimensions of food and nutrition problems that increasingly assail the world as well as to provide some guidance on the population-based interventions necessary for diet-related diseases to be contained.

The double burden of malnutrition

Malnutrition refers to nutritional excesses of macronutrients and micronutrients as well as deficiencies (WHO 1995). Undernutrition is the result of insufficient intake, poor absorption, and/or poor biological use of the nutrients. This can result in impaired body functions, impaired growth, and underweight. Excess or imbalanced nutrient intakes can result in impaired body functions, as well as overweight and/or obesity. The double burden of malnutrition (DBM) was presented as a 'new paradigm', at the 1992 International Conference on Nutrition (ICN), a member state meeting jointly hosted by the United Nations Food and Agriculture Organization (FAO) and the World Health Organization (WHO). It was subsequently recognized that DBM occurs not only in the same populations, but also in the same households, and even the same individuals (Shrimpton and Rokx 2012). At the population level women are the most affected by DBM, with most low- and middle-income countries (LMICs) having problems of both overweight and underweight women (Jaacks et al. 2015). The most common form of DBM at the household level is stunting in children and overweight/obesity in women (Garrett and Ruel 2005). For individuals the most common form of DBM is overweight and iron deficiency (Zhao et al. 2015). Furthermore, new evidence emerged suggesting that maternal and foetal undernutrition increased susceptibility to overweight and obesity and diet-related non-communicable diseases (NCDs) in adulthood. Consequently, the understanding of the causal links to overweight and obesity evolved to reflect the double burden causality across the 'life course' (Darnton-Hill et al. 2004). Separating the treatment and/or prevention of nutritional deficiencies and excesses was therefore no longer considered sensible.

Most countries are now grappling with multiple burdens of malnutrition, and globally overnutrition is increasing faster than undernutrition is decreasing. Estimates are that at least 88% of countries face a serious burden of either two or three forms of malnutrition, i.e. childhood stunting, anaemia in women of reproductive age and/or overweight in adult women (Development Initiatives 2017). Two billion adults were overweight or obese globally in 2017 (a third being obese), while 815 million people were still going to bed hungry and 2 billion people lacked key micronutrients like iron and vitamin A. Furthermore 155 million children were stunted, and 52

million children were wasted, while 41 million children were overweight. Obesity prevalence more than tripled from 3.2% to 10.8% among men and more than doubled (6.4–14.9%) among women from 1975 to 2014. By comparison, the prevalence of underweight people decreased from 13.8% to 8.8% among men and 14.6% to 9.7% among women (NCD Risk Factor Collaboration 2016). Between 1975 and 2016 the prevalence of moderate and severe underweight in girls decreased from 9.2% to 8.4% and from 14.8% to 12.4% in boys (NCD Risk Factor Collaboration 2017). More than 20 million infants worldwide, representing 15.5% of all births in 2000, were born with low birthweight, 95.6% of them in LMICs (UNICEF/WHO 2004). Accurate monitoring is challenging, since nearly half of the world's infants are not weighed at birth, but this situation does not seem to have improved. Iron deficiency is the most common and widespread nutritional disorder in the world, being common in both high and low-income countries.

Measuring nutritional status

Methods of defining nutritional status and describing the size of nutritional problems have evolved considerably in the last few decades. Until the turn of the millennium the focus of nutrition programme efforts was on child undernutrition, the population estimates of which largely relied on the adequacy of weight-for-age in children under five. These efforts essentially focused on trying to understand the aetiology of the severe forms of undernutrition, kwashiorkor, and marasmus, trying to prove the hypothesis that these were protein deficiency and energy deficiency diseases (Scrimshaw and Viteri 2010). There was little awareness at that time that non-severe forms of malnutrition existed, and that their consequence for a population perspective was of immense importance. During this era, much was achieved in terms of definitions for undernutrition. Moderate malnutrition was defined as between 2 and 3 Z scores, and severe malnutrition was more than 3 Z scores, below the median of the Harvard growth standard (Jelliffe 1966). However, as noted by Waterlow (1973), a deficit in weight-for-age conceals two different conditions, a deficit in height-for-age or 'stunting' and a deficit in weight-for-height or 'wasting'. It is important that these conditions should be differentiated and assessed separately, because they reflect different underlying problems: stunting tends to reflect chronic undernutrition, and wasting, acute undernutrition.

Recognizing the needs for more representative references for child growth and nutrition, WHO developed the WHO International Growth Reference (de Onis et al. 2006), constructed by measuring the growth of cohorts of well-fed children from several nations. This was essentially to demonstrate that in the right circumstances, i.e. healthy environments, children grew similarly independent of their ethnic differences or geographical location. Indeed, this is now the growth standard children of all nations should attain.

In the years since the pioneering work on child undernutrition, it has become increasingly obvious that the burden of malnutrition is changing, with dietary risk factors for overweight and NCDs becoming more and more of a problem. In consequence increasing effort has gone into better defining and measuring overnutrition. Overweight and obesity are most commonly measured in adults with the body mass index (BMI), which is weight divided by height squared. A BMI greater than 25 is regarded as overweight and greater than 30 is obesity. Child overweight is also defined as a weight-for-height greater than +2 standard deviations of the WHO Child Growth Reference (WHO 2017a). The World Health Organization has also recommended BMI cut-offs for overweight and obesity be 23 kg/m^2 and 27 kg/m^2, respectively in Asia, lower than the cut-offs used in Caucasian populations (WHO 2004). Further measures of adiposity include waist circumference and skinfold thickness, but these are more often used for individual than population assessments.

Nutrition and health

Nutritional diseases are caused by both under and overnutrition and affect people throughout the lifecourse. Undernutrition, including foetal growth restriction, suboptimal breastfeeding, stunting, wasting, and deficiencies of vitamin A and zinc, caused 45% of child deaths, resulting in 3.1 million deaths annually in 2011 (Black et al. 2013). Breastfeeding is the most important start to life in terms of feeding patterns (Victora et al. 2016). Appropriate breastfeeding practices prevent child morbidity due to diarrhoea, respiratory infections, and otitis media. Where infectious diseases are common causes of death, breastfeeding provides major protection. Despite these benefits, in 2013, only about two-fifths of infants in LMICs were exclusively breastfed for the first 6 months of life, and three-quarters

were still being breastfed at 12–15 months (Rollins et al. 2016). Rates in high income countries (HICs) are less sure but are much lower at all ages.

At least a third of the world's population, some 2 billion people, are anaemic, many due to iron deficiency, and in resource-poor areas, this is frequently exacerbated by infectious diseases. Maternal anaemia, especially during the first trimester of pregnancy, can be considered as a risk factor for many pregnancy outcomes including low birth weight (Rahmati et al. 2017). High BMI accounted for 4 million deaths globally in 2015, nearly 40% of which occurred in persons who were not obese but overweight (The GBD 2015 Obesity Collaboration 2017). More than two-thirds of deaths related to high BMI were due to cardiovascular disease. Furthermore, Asians at any given level of adiposity have a much greater predisposition to cardiometabolic disorders than Caucasians (Haldar et al. 2015).

Increasingly, diet-related metabolic risk factors and NCDs are becoming the major driving force behind the global burden of disease. NCDs accounted for 64 % of the attributed risk factors for deaths in 2016 (GBD 2016 Risk Factor Collaborators 2017). High blood pressure was the fourth-leading risk factor for both men and women in 1990 and had risen to be the second leading risk factor for men and the leading risk factor for women by 2016. High fasting plasma glucose (FPG) was the third leading risk factor for deaths and the fourth-leading risk factor for disability-adjusted life years (DALYs). Diabetes was the largest source of DALYs attributable high FPG. The global prevalence of diabetes has nearly doubled since 1980, rising to 8.5% in the adult population, with 422 million so affected, three times more than in 1980 (WHO 2016). By increasing the risks of cardiovascular and other diseases, diabetes, and high FPG caused 3.7 million deaths, with 43% of these occurring before the age of 70 years. The percentage of deaths attributable to high FPG or diabetes that occurs prior to age 70 is higher in LMICs than in HICs.

Systematic reviews find that red meats, processed meats, sugar-sweetened beverages, high glycaemic load foods, transfats, and sodium are the principal dietary factors that increase the risk of cardiovascular disease and diabetes (Micha et al. 2017).

Changing dietary patterns

Globally, food consumption per person measured as dietary energy intake, increased by 15% in the 30 years up to 2000 (Kearney 2010). In the same period annual meat consumption in LMICs rose from 10 kg to 26 kg per person (WHO/FAO 2003). Global consumption of milk and dairy products also doubled to 45 kg, and the average global supply of fat increased by 20 g per capita per day. The type of carbohydrate in the diet also changed, with decreases in complex carbohydrates such as starchy foods (for example, whole grain cereals and legumes) and an increase in refined sugar. The average global consumption of fish increased from 13.4 kg per capita in 1990 to 16.3kg per capita in 1999 (FAO 2003). The concept of this 'consumption' is more related to food purchased than intake, as it doesn't take into consideration food waste. Food lost and wasted along the way from field to fork may be as much as a third, amounting to about 1.3 billion tonnes per year (FAO 2011).

WHO recommends that adults and children should maintain a reduced intake of free sugars over the life course, and this should be less than 10% of total energy intake (WHO 2015). Global assessments of sugar intake as a percentage of energy intake range from 20.0% to 38.4% in young children and 13.5–24.6% in adults (Newens and Walton 2016). Various systematic reviews and meta-analyses provide evidence that consumption of sugar-sweetened beverages (SSBs) (Malik et al. 2006) or free sugars (Te Morenga et al. 2012) promotes weight gain in adults and children. Per capita sugar consumption is strongly associated with the prevalence of diabetes mellitus in a 165-country analysis, with the strongest association found in the Asia region (Weeratunga et al. 2014). Another problem is that many processed foods have a high glycaemic index (Fardet et al. 2017), with blood sugar levels remaining high for a long period after the food is eaten. Consumption of higher glycaemic index foods are associated with increased risk of type 2 diabetes (T2D) (Bhupathiraju et al. 2014) and coronary heart disease (CHD) (Fan et al. 2012).

The proportion of energy contributed by dietary fats exceeds 30% in the HICs, and in nearly all other regions this share is increasing, even though the WHO recommendation is for not more than 30% (WHO/FAO 2003). Thinking has now changed however, and fat intake is regarded as less important in the genesis of obesity and cardiovascular disease with calls to revise the dietary guidance (Willett 2011; Harcombe et al. 2016). Contrary to popular commentary regarding the fat intake being bad for health, a recent

dynamic time series analysis of US and global population data sets suggests that while increases in carbohydrate consumption are associated with increasing weight, increases in fat consumption are associated with decreasing weight (Riera-Crichton and Tefft 2014). Similarly, a cohort study in 18 countries from five continents has also found that high carbohydrate intake was associated with higher risk of total mortality, whereas total fat and individual types of fat intake were related to lower total mortality (Dehghan et al. 2017). Alongside this changing understanding of the impact of fat and carbohydrate (particularly sugar) intakes on health has been increasing evidence for the role of 'ultra-processed' foods in contributing to diet-related NCD risk (Fardet 2016; Fiolet et al. 2018).

It is increasingly recognized that what is eaten is as important as how much is eaten in the causality of obesity and cardiovascular disease. Traditional thinking that 'eating too much' and 'moving too little' are the causes of the rising global obesity problem is increasingly coming under scrutiny (Shrimpton et al. 2017). Average adult weight began to increase by 1.5 kg a year in the early eighties in the United States and the main drivers were the consumption of processed foods, and especially of potato chips and SSBs (Mozaffarian et al. 2011). This suggests that relatively minor changes in diets—for example, marginal increases in energy intakes due to increased discretionary food intake—can have significant impacts on body weight and health over the long term.

The food environment

These dietary changes have occurred during a time of systemic changes to the food environments in which people make decisions about what—and how much—they eat. The food environment acts as an important interface between the wider food system and peoplés food acquisition and consumption through continuous and complex interactions between external and personal food environments. The external food environment domain relates to the world of opportunities and constraints that exists within a given context, and includes exogenous dimensions such as food availability, prices, vendor and product properties and marketing and regulation. The personal food environment domain includes a set of endogenous dimensions including food accessibility, affordability, convenience, and desirability (Turner et al. 2017).

Processed foods now account for 80% of country food sales (Mazzocchi et al. 2012), that is foods sold through retail outlets. Ultra-processed products are becoming dominant in the global food system, including in LMICs where they have become ubiquitous, especially in urban areas (Igumbor et al. 2012; Monteiro et al. 2013). These foods are formulations of several ingredients which, besides salt, sugar, oils, and fats, include food substances not used in culinary preparations in the home, such as flavours, colours, sweeteners, emulsifiers, and other additives. They contribute almost 60% of the energy and 90% of added sugars consumed by the USA population (Martínez Steele et al. 2016). Chemical food additives are used to make processed foods look, smell, and taste better, as well as to improve their shelf life. These include a large assortment of colourings, preservatives, antioxidants, emulsifiers, stabilizers, anticaking agents, and flavour enhancers. In 2000, the food industry spent US$20 billion on such additives, and the average consumption was 7 kg per person a year in industrialized countries (Millstone and Lang 2003).

Changing food systems

Changes to food environments reflect broader changes to food systems, that have revolutionized how food is produced, processed, and marketed. Global food production has outpaced rising demand based on population growth rates during the last four decades (FAO 2002). The number of farms globally is around 570 million, and of these 83% are small farms of less than two hectares (Lowder et al. 2016), which operate about 12% of the world´s land. Despite the population almost doubling, the per capita supply of energy from food for human consumption rose by just over 20% between the early 1960s and the late 1990s. The greatest growth came from increases in the production of wheat, vegetable oils, sugar, and meat.

The increase in global food production has been largely attributed to the green revolution (Pingali 2012). The production of cereal crops tripled in the last 50 years with only a 30% increase in land area cultivated. Using petroleum-based fertilizers and chemical pesticides, expanded irrigation, and genetically modified disease-resistant cultivars, cereal production (wheat, rice, maize) doubled in Asia between 1970 and 1990. This success was concentrated in areas where irrigation water, infrastructure (roads) and access to markets (input and output) as well as credit were well developed.

In areas without these conditions, such as in most of Africa, the impact of the green revolution was less impressive.

Of the 446 million small farms, the majority (87%) are in Asia, and only 8% are in Africa and 1% in the Americas. About a half of small farms are pre-commercial, i.e. not benefitting from productivity enhancing technologies, financing, and market access. While these small farms of the peasant food web provide 70% of total food eaten globally (ETC Group 2015) and harvests 60–70% of food crops from 20% to 30% of arable land, the larger farms of industrial agriculture food system provide 30% of all food consumed globally and uses 70–80% of world's arable land. Even where the green revolution was successful in raising productivity this was concentrated in the richer larger landholders, as the poorer farmers of the peasant food web could not afford to pay for these inputs. The green revolution thus gave impetus to the development of agribusiness (Engdahl 2007), designed to increase the flow of food crops from the places of production through the food system, to the place of consumption.

The number of actors providing the inputs needed to achieve increased production through the industrial agricultural system is very limited (ETC Group 2013). A few big multinational conglomerates, including Monsanto and DuPont, Syngenta and Bayer, Dow, and BASF, control 75% of all private sector plant breeding research, 60% of the commercial seed market and 76% of global agrochemical sales. About a third of global cereal production and just over 90% of soybean production is used to feed farm animals, mostly in factory farms of the industrial agriculture food system (FAO 2002).

Around 56 billion land animals are estimated to be reared and slaughtered for human consumption annually (Koneswaran and Nierenberg, 2008). These estimates vary considerably however, largely depending on terminology and definitions used. It is agreed that factory farms are the most rapidly growing system of farm animal production (FAO 2009), providing 72% of poultry, 43% of eggs, and 55% of pork globally. Furthermore, capture fisheries production increased from 69 million to 93 million tonnes during the last three decades, while world aquaculture production increased from 5 million to 63 million tonnes (World Bank 2013). Aquaculture will provide close to two-thirds of global food fish consumption by 2030 as catches from wild capture fisheries level off and demand from an emerging global middle class, especially in China, substantially increases

Food is now a major component of international trade, accounting for 10% of all global trade (FAO 2011a), and the aggregate value of agricultural exports passing through the international food system was expected to reach a record US$1.29 trillion in 2011. The amount of food calories traded on the international market is about 20% and had more than doubled between 1986 and 2009 (D'Odorico et al. 2014). Much of the increases in global food trade have been in highly processed commodities. Since the early seventies and the development of the Washington Consensus on the conditionality for structural adjustment programmes (SAPs) from the World Bank, and the 'standard prescription' by the International Monetary Fund (IMF), LMICs have been increasingly encouraged to liberalize trade and investment, privatize national industries, and reduce state support to local agricultural production (Weis 2007). The Washington Consensus also bolstered the global free trade negotiations under the General Agreement on Tariffs and Trade (GATT), which eventually resulted in the creation of the World Trade Organization (WTO) in 1994. Alongside this, multilateral, regional, and bilateral trade agreements, such as the North American Free Trade Agreement (NAFTA), have proliferated in the past 30 years.

The WTO, World Bank, and IMF are thus key institutions governing trade in the modern global food system. All these institutions have mandates that include food security to some extent—but this inclusion has been along economic lines, focused on calorie availability at the national level, and paying little attention to inequalities and sociocultural aspects of food production and distribution (Rosset 2006; Margulis 2013). Another key actor in global food governance is the Codex Alimentarius Commission, which sets food standards, including approval for use of food additives. The Codex is recognized by the WTO as an international reference point for the resolution of disputes concerning food safety and consumer protection, thus lending it greater authority (Halabi 2015). However, while Codex is a member state organization, under the WHO and UN FAO, it is also a standards-setting body with open membership. This means that industry actors can act as formal observers and has resulted in standards that often preference trade interests over public health (Downes 2014).

Growing industry concentration has led to a situation where just a few companies control large shares of the flow of food through the global food system (FAO 2002). In 2003 just four firms controlled 73% of the global grain trade (Murphy et al. 2012). Furthermore, these multinational

corporations are highly diversified and integrated both vertically and horizontally; for example, Cargill, ADM, and Bunge not only account for more than 60% of total financing of soy production in Brazil, but they also provide the seed, fertilizer, and agrochemicals to the growers, and subsequently buy the soy and store it in their own facilities (Murphy et al. 2012). In the USA Cargill is number one in terms of domestic grain handling, as well as grain and soybean exports but also the second largest compound feed producer and the number three turkey producer. Policies of trade liberalization have facilitated the rising availability and consumption of relatively cheap meat, dairy products, and processed foods in LMICs (Thow and Hawkes 2009). Imported foods tend to be cheaper because of continued state support to the industrial agriculture system in HICs, and their dumping of cheaper food commodities onto the world market.

Notwithstanding the increasing production of food globally, food prices have risen sharply since 2006, impacting most negatively on the poor. The main drivers of food price rises are the increasing use of land for biofuel production in the face of rising energy costs, and greatly increased speculative acquisition of land, as well as financial speculation on food futures (McMichael 2012). Speculative acquisition of land, sometimes termed ‘land grabbing’ authorizes large-scale removal of rural populations from (often) ancestral lands to install ‘agriculture without farmers’, i.e. agro-industrialization. It sacrifices land and its inhabitants in favour of what is a necessary global good (food yields, green fuels, and carbon offsets). For example, in Colombia between 2001 and 2005, 263,000 peasant families were expropriated from 2.6 million hectares by agrobusiness and/or paramilitaries interested primarily in palm oil development. (McMichael 2012).

Another recent and major contributor to rising food prices is financial speculation in food. The deregulation of commodities derivatives markets in 2000 as part of financial liberalization allowed large institutional investors to enter these markets. Between 2003 and 2008 commodity index holdings increased from $13 billion to $317 billion (McMichael 2012). Already in 2006, Merrill Lynch Wealth Management and Financial Services estimated that speculation was causing commodity prices to trade at 50% higher than if they were based on fundamental supply and demand (De Schutter 2010). The increase of the agricultural commodities futures between March 2003 and March 2008 was accompanied by price increases during the same period

for coffee with 167%, for soybean oil with 199% and for wheat with 314% (Kerckhoffs et al. 2010).

Global food retail sales are about US$4 trillion annually, with supermarkets accounting for the largest share of sales (Meade 2017). The urban demand for food in LMICs is expected to more than double in the next decade. It will be responsible for almost the entire growth in global food demand, spiking from about 40% to some 60% by 2050. The emergence of supermarkets in LMICS is leading to a radical change in the global food marketing and distribution system (Reardon et al. 2015). Structural changes in the food distribution system, which took 50 years to unfold in the United States, have taken place in little more than a decade in Latin America. In China, where no supermarkets existed in 1989, annual supermarket sales were US$100 billion in 2008 (Reardon and Gulati 2008). Unlike HICs, where supermarkets are often the principal, if not only, source of fruits and vegetables (Larson et al. 2009), supermarkets in LMICs are often distribution channels for cheap processed foods including less than healthy snacks and 'junk' food, as well as being venues for fast food chains.

Given the changing nature of the food system one of the greatest challenges to improving nutrition is how to deal with the burgeoning private sector. The food industry is now a massive business, with ten big food and beverage companies generating revenues of more than US$1.1billion a day (OXFAM 2013). Together they are part of an industry valued at US$7 trillion, larger than the energy sector and representing roughly ten percent of the global economy. Although spending on processed foods is still lower in LMICs (US$143 per capita per year in middle-income countries and US$63 per capita in low-income countries), it is growing at 28% and 13% a year respectively in those country groups.

The political economy of food and nutrition

The economic consequences of the DBM across the life course are massive. Estimates are that more than 200 million stunted children under 5 years of age in LMIC countries that survive growth failure in the first 1,000 days, are not developing to their full potential, and that this is associated with more than a 20% deficit in their income as adults (Grantham-McGregor et al. 2007). It is hard to put a single figure on this, but in comparison with the costs of over nutrition they are likely to be small. Current trends project 65

million more obese adults in the United States and 11 million more obese adults in the United Kingdom by 2030, consequently accruing an additional 6–8.5 million cases of diabetes, 5.7–7.3 million cases of heart disease and stroke, 492 000–669 000 additional cases of cancer, and 26–55 million quality-adjusted life years forgone for United States and UK combined (Wang et al. 2011). Combined medical costs associated with treatment of these preventable diseases are estimated to increase by $48–66 billion/year in the United States and by £1.9–2 billion/year in the UK by 2030. Over the next 20 years, the cost of treating NCDs and cumulative output losses globally are likely to be at least US$30 trillion and US$47 trillion, respectively, representing 48% and 75%, respectively, of global GDP in 2010 (Bloom et al. 2011). Of course, the great majority of these NCDs are diet-related.

At the turn of the millennium the prospects for eliminating young child malnutrition were considered grim. An assessment done by the World Bank and UNICEF (Gillespie et al. 2003) revealed that nutrition was improving slowly in some regions but was stagnant in others and was often side-lined in poverty agendas despite its potential to improve health, mental development, and productivity. Although nutrition professionals broadly agreed on key interventions and on success factors for implementation, this was not reflected in action. Few large-scale programmes were rigorously monitored and evaluated and inadequate capacity to tackle malnutrition was a major factor limiting progress.

The publication of the Lancet Nutrition Series in 2008, including an article with a package of low cost efficacious interventions for tackling maternal and child undernutrition (Bhutta et al. 2008), stimulated a tremendous increase in political commitment to the reduction of child undernutrition. In consequence most development agencies revised their strategies to address undernutrition and focused on the 1,000 days during pregnancy and the first 2 years of life. National commitment in LMICs grew, donor funding rose, and civil society and the private sector were increasingly engaged. However, despite this ‘awakening’, progress in tackling stunting was still considered insufficient 5 years later (Black et al. 2013). In 2012, the World Health Assembly endorsed a comprehensive implementation plan on maternal, infant, and young child nutrition which specified six global nutrition targets for 2025. The first target is a 40% reduction in the number of children under-5 who are stunted. The global prevalence of stunting in

children under-5 declined 36% during two decades—from an estimated 40% in 1990 to 26% in 2011 (UNICEF 2013), hence the likelihood of a further 40% reduction by 2025 seems remote.

Translating the evidence, interest, and government concern regarding nutrition into action on the ground has proved challenging—often as the result of politico-economic challenges (Balajaran and Reich 2016). One such challenge is that the nutrition system and community are fragmented from global to community level. Among the various subgroups at the country level are clinic-based medical nutritionists, academic nutritionists, development economists, and planners, social scientists, public health nutritionists, and agricultural specialists. Building coalitions across these different groups is challenging and time-consuming. Furthermore, this multiplicity of stakeholders creates multiple narratives concerning nutrition approaches, which can lead to contention and rivalry. All these challenges are compounded by a limited capacity among nutritionists, be it to engage in the broader political dynamics of managing nutrition policy or the delivery of interventions by multiple actors/sectors.

The private sector must surely play a role in controlling the malnutrition problem, but this will require balancing public health needs with commercial interests and managing conflicts of interest requires transparent systems and mechanisms (Hawkes and Buse 2011). Efforts to achieve some balance have been developed though the scaling up nutrition (SUN) initiative (Bezanson and Isenman 2010), with countries increasingly looking to the private sector to join national SUN multistakeholder committees. Recent evaluations of SUN suggest that it is still far from being successful in achieving any impact however (Mokoro 2015). The challenge facing policymakers trying to integrate business into national nutrition strategies remains and mistrust and opposition to the private sector are the hardest challenges. Until now the SUN ‘movement’ has been largely focused on maternal and child undernutrition, but whether and how it can evolve to include overnutrition and the metabolic and NCDs remains to be seen, as the potential for conflict of interest are obviously enormous.

Multinational retailers have introduced into LMICS the types of supply chain seen before in HICs, based around tight vertical coordination, centralized purchasing, and sophisticated marketing (Mazzocchi et al. 2012). They accelerate the rate at which local firms adopt these strategies and tailor them to local needs, and in doing so provide employment and generate

income locally. Most of the profits however still revert to the head office of the chains located in HICs (Hawkes 2005). These trends risk eroding the more traditional access to fruit in urban areas of LMICs, such as in South Africa for example, where it is primarily through the informal market; in particular, wet markets or stalls and street vendors. Specific policy opportunities that could increase access to fruit and vegetables in South Africa and other LMICs are: targeted support for periurban farmers to access urban markets (e.g. transport); including food production considerations in initiatives for urban greening; and, initiatives that could increase access to, or incentivize the sale of, fresh fruit and vegetables by street vendors (Thow et al. 2017)

The food industry spends at least US$ 40 billion a year globally on advertising processed foods (Millstone and Lang. 2003). The evidence indicates that unhealthy food and beverage marketing increases dietary intake and preference for unhealthy food and beverages. Marketing of unhealthy food and beverages increased dietary intake and influenced dietary preference in children during or shortly after exposure to such advertisements (Sadeghirad et al. 2016). Exposure to child-targeted fast food TV advertising is positively associated with fast food consumption among children of pre-school age (Dalton et al. 2017). Trade liberalization is considered to be driving the NCD epidemic in Asia (Baker et al. 2014). Evolving layers of the international and regional trade regimes have facilitated increased market penetration by transnational tobacco, alcohol, and ultraprocessed food corporations and thereby driven consumption of these unhealthy commodities in Asia especially.

Challenges to taking action

Nutrition has been identified as a ‘chosen issue’, meaning that champions are required to put it on the policy agenda (Garrett et al. 2011). Moreover, the implementation of multisectoral approaches for maternal and child undernutrition (MCU) improvement demands strong and sustained support from high levels of government (Levinson and Balarajan 2013), and this is no different for overnutrition. Accelerating progress in nutrition through multisectoral approaches requires building leadership at nationally and locally.

The lack of a natural institutional home in government is a challenge, as nutrition doesn't fit cleanly into any single government agency, with health or agriculture usually nominated as a 'home' agency—but both departments are usually unable to lead cross-sectoral action across government. To surmount this problem many countries have created a 'home' for nutrition in suprasectoral bodies such as the prime minister's office or the ministry of planning. Giving high profile leaders—such as Prime Ministers and Mayors—a central role in driving initiatives has been shown to be a key factor in bringing different actors together to collaborate towards NCD prevention goals (Meiro-Lorenzo et al. 2011).

Nutrition capacity is often weak in many countries and in LMICs especially. This is largely related to the absence of a common understanding of the malnutrition problem (Heaver 2005). This lack of understanding as well as the relatively recent development of public health nutrition (PHN) science have undoubtedly contributed to the shortage of professionals capable of developing population-based approaches for the prevention of all forms of malnutrition as opposed to (or in addition to) a cadre capable of treating the consequences in individuals. Indeed, in most African countries nutrition programmes and interventions are led by nurses whose nutrition knowledge is generally very limited. Typically, other development actors look to the health sector for nutrition guidance, while health sector workers' nutrition training is all too often woefully outdated and clinical.

A consensus is growing that a robust and knowledgeable workforce is essential for PHN policy implementation (Yngve et al. 2012). How to build that capacity is the question that still needs to be answered, especially in LMICs. A hybrid distance learning model for PHN workforce managers' in-service training is urgently needed, especially in LMICs (Shrimpton et al. 2016), and unless increased priority and funding is given to building capacity for SUN programmes in LMICs, MCU rates are likely to remain high and nutrition-related NCDs to escalate.

The World PHN Association has placed considerable emphasis on improved definition of the nutrition capacity and how to build it since its inception in 2008. A competency framework was first developed (Hughes et al. 2012) which described the core functions of PHN practice, and the work done at different levels of the public health system. Subsequently a nutrition capacity assessment framework was proposed (Shrimpton et al. 2013) to guide the systematic assessment of capacity development needs at system,

organizational, workforce, and community levels. While nutrition has to be everybody's business at the district level, somebody has to be in charge. That is the role proposed for the public health nutritionist.

Another serious challenge that faces nutrition is demonstrating the effectiveness of policies and programmes. Improving nutrition in practice is complicated and context specific, since it requires careful strategic planning, informed by the situational analysis and prioritization of programmatic interventions to be funded with limited resources (Balajaran and Reich 2016). There is a growing repertoire of efficacious nutrition interventions for maternal and child undernutrition but translating that knowledge into practice and demonstrating tangible results at scale is still difficult to achieve (Leroy and Menon 2008). It is easier for academic experts in nutrition to get funding for carrying out research than for programme evaluations.

While intervening in the first 1,000 days is critical, addressing the DBM requires expanding the focus of nutrition interventions to cover the rest of the life course and the food system (Fig. 2.8.1). The importance of employing multisector policies and interventions for improving nutrition outcomes has been widely agreed during this last decade (World Bank 2006). Indeed, these nutrition-specific and nutrition-sensitive interventions need to be continued across the school years into adulthood and the workplace (Shrimpton and Rokx 2012). Moreover, in order to tackle the double burden a range of policy interventions and programmes are needed to address the overall food environment (Shrimpton et al. 2016) and concerted efforts are needed from the health, education, agriculture, industry, social protection, and water and sanitation sectors as shown in Figure 2.8.1.

The nutrition-specific interventions directed at immediate levels of causality are mostly delivered through the health sector, while nutrition-sensitive interventions directed at the underlying and basic levels of causality are often delivered through other sectors such as education, agriculture, water and sanitation, and social welfare. Scaling up the priority nutrition-specific interventions through the health sector to cover 90% of the population in the 36 countries with 80% of the global stunted children will only reduce stunting by 20% (Bhutta et al. 2013), still short of the WHO CIP target for 2025. Scaling up the nutrition-sensitive interventions through other sectors is therefore required to achieve these MCU targets and perhaps as important, do so in a sustainable way. Although the evidence base doesn't yet permit precise estimation of the contribution of nutrition-sensitive

interventions to stunting reduction, early indications are that social protection, agriculture, and water and environment, and public health can all greatly contribute to the acceleration (Fanzo et al. 2013).

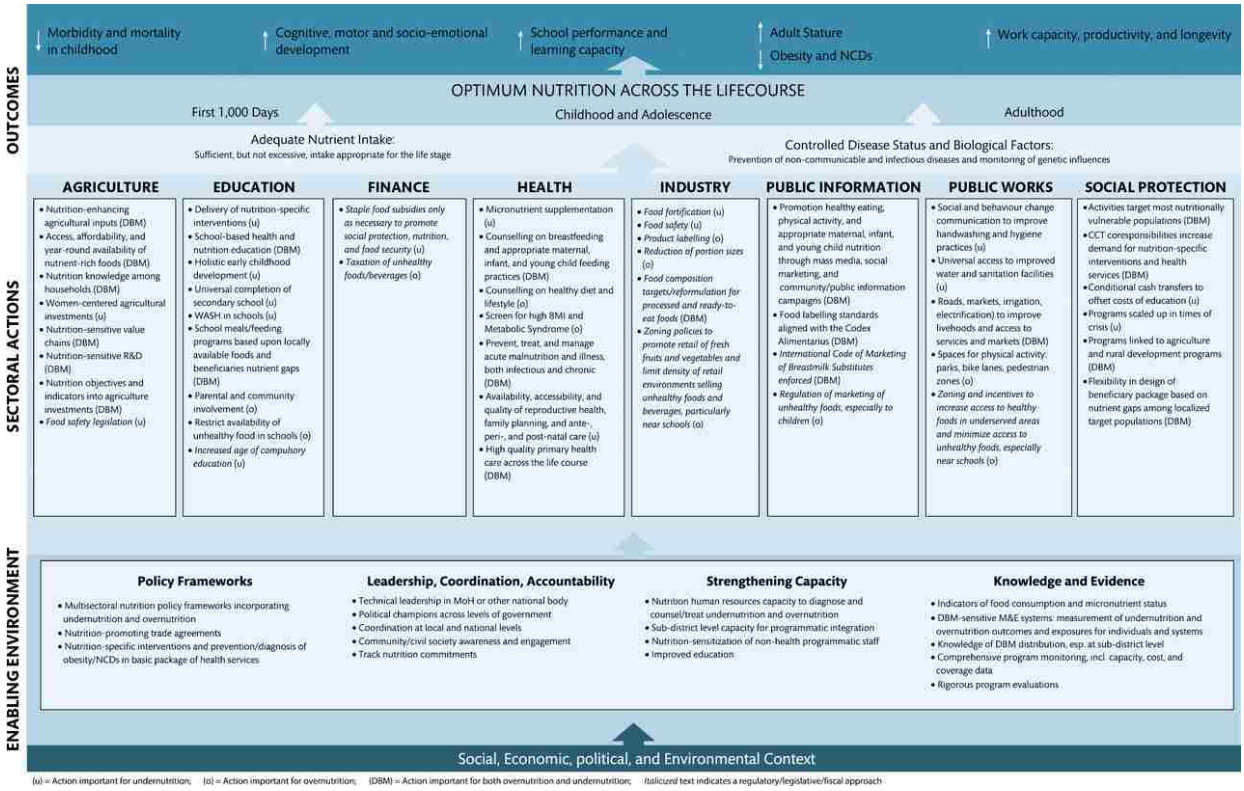


Fig. 2.8.1 Framework of sectoral actions and systems strengthening efforts needed for a multisectoral response to the double burden of malnutrition.

Reprinted from Shrimpton R, Mbuya NK, Provo AM (2016) 'The Double Burden of Malnutrition in East Asia and the Pacific: Evidence and Lessons for a Multisectoral Response'. Washington: World Bank under a CC BY 3.0 IGO license.

The estimates of impact of the various packages of nutrition interventions assumes that all are delivered together. Rather than joint programme implementation, the various sectors need to converge their programme interventions to be delivered together as a package. This process is essentially the responsibility of local government and is facilitated by decentralization (Shrimpton 2002). Community-based programmes are of importance here, especially if they have the essential elements of community ownership, adequate population coverage, targeting, and central support for supplies and training (Mason et al. 2006). The promotional activities, which

aim at ensuring optimal growth of the foetus/infant from conception through to 2 years of age, i.e. the first 1,000 days, typically include: treating maternal anaemia; treating infections such as intestinal parasites and avoiding/treating malaria; monitoring growth of the infant; promoting breast feeding and adequate complementary feeding; using oral rehydration for diarrhoea, and getting children immunized.

In a typical community-based programme, community 'mobilizers' take on the task of looking after households in their neighbourhood, allowing whole communities to be mapped and covered. A certain level of intensity of effort must be ensured, with optimal ratios of not more than 20 mobilizers per facilitator and not more than 20 households per mobilizer. If these levels of intensity of effort with a sufficiently wide coverage are not ensured, no impact will be seen in the population, and the resources invested will be largely lost. A key challenge for scaling up such community-based programmes lies in institutionalizing and mainstreaming community participation (Lehmann and Sanders 2007).

As no single sector alone can solve the problem of undernutrition, neither can a health-centric approach reverse the growing trend of overnutrition. Multiple sector contributions are needed, and the necessary actions range from national and district level policies to creating supportive community environments and supporting individuals to change their behaviours. National level political support—through taxation, regulation, and/or subsidization—is necessary to ensure social environments that support healthy decisions as the default. Interventions are needed beyond the realms of 'self-will': ensuring that physical activity is facilitated and easily available and access to quality diets is affordable, without the promotion of unhealthy products and habits through the media. The education sector has an important role to play in nurturing the development of children who graduate from school with adequate life skills and appropriate habits that help them avoid these risks in adulthood. The health sector must do more to deliver the messages and screening that can contribute to primary and secondary prevention, particularly for adults.

There is clear evidence that policy intervention that fosters a food environment in which the healthy choices are more available and affordable than the unhealthy choices, and in which marketing and point of sale incentives do not preference unhealthy foods, can support consumers to make healthier choices (Hawkes et al. 2015). Macro, or environment-

focused, policy intervention can help to correct for existing incentives in food environments which incentivize the purchase of highly processed (and highly profitable) unhealthy foods and beverages. For example, fiscal policy interventions, including taxes and subsidies, can help to incentivize consumption of healthier, rather than unhealthy foods (Thow et al. 2018). Similarly, nutrition labelling interventions are critical in supporting healthier point-of-purchase decision making, by clear external labelling detailing the internal properties of the foods (Cecchini and Warin 2016).

The WHO recommendations for childhood obesity prevention programmes include key points which have population-wide relevance (WHO 2016a). These recommendations include: use a mixture of 'top-down' and community-based actions, employ a range of policy instruments, establish cross-sectoral platforms and a multisectoral approach; promote interventions across a range of settings, including early childcare settings, schools, workplaces, and community organizations; and leverage legislative and financial tools to ensure availability and affordability of healthy foods and physical activity opportunities. Similarly, for reducing the risk of chronic NCDs, WHO has found multicomponent interventions that are adapted to the local context to be the most successful (WHO 2013a, 2013b). Interventions that used the existing social structures of a community, such as schools or the weekly meetings of older adults, reduced barriers to implementation. Effective interventions invariably involved participants in the planning and implementation stages, such as involving the workers themselves in workplace interventions, and the leaders of the range of civil society organizations in the community.

In 2015 the United Nations adopted a set of goals to be achieved by 2030 with the aim of ending poverty, protecting the planet, and ensuring prosperity for all as part of a new sustainable development agenda (UN 2015). While considerable progress has been made over the past decade across all areas of development, the pace of progress observed in recent years is insufficient to fully meet the sustainable development goals (SDGs) and targets by 2030 (UN 2017). One of the seventeen SDGs is directly related to food and nutrition, namely SDG 2 which is to end hunger, achieve food security and improved nutrition, and promote sustainable agriculture. While some progress has been made to reduce hunger and malnutrition, it is not enough to meet the goal. Failure to achieve SDG2 will also make other SDGs difficult to achieve, since the pervasive negative influence of our

current food system endangers many others (Shrimpton 2017). The SDGs provide an exciting opportunity to achieve global nutrition targets while catalysing other development goals through 'double duty' actions (WHO 2017b). These include interventions, programmes and policies that have the potential to simultaneously reduce the risk or burden of both undernutrition (including wasting, stunting and micronutrient deficiency or insufficiency) and overweight, obesity, or diet-related NCDs.

Learning from success

Examples of successful programmes are not easily found, but experiences gained in Brazil with undernutrition and France with overnutrition offer some important insights.

In Brazil the Bolsa Familia Programme (BFP) is part of the National Food and Nutrition Security System (SISAN), which has evolved over the last two decades to become a very effective large-scale multisectoral nutrition programme. The principal sectors participating in SISAN include the Ministries of Social Development, Health, and Education. At national level SISAN is overseen by the National Council on Food and Nutrition Security (CONSEA). CONSEA is a deliberative body with two-thirds of its members made up of civil society, and that is linked to the President's Office. Coordination and implementation of SISAN is carried out by an interministerial chamber on food and nutrition security (CAISAN) which is located in the Ministry of Social Development (Chmielewska and Souza 2011).

The BFP targets the mothers and children of poor families across Brazil, with some 12 million families and 48 million beneficiaries in 2010 (Lindert et al. 2007). The conditionality of BFP includes either school attendance of children aged 5–18, or regular health check-ups for mothers and their children aged under 5 years of age. At the municipal level there are councils and committees that oversee the conditionality requirements of BFP, with the creation of registries of poor families (Cadastro Unico) that are requirements for a family to become BFP beneficiaries. The delivery platforms at the municipal level are the family grant that is dispensed through the banking system based on the family being in the 'Cadastro Unico'. The conditionality of the BFP through the health system includes regular pre- and postnatal care, growth monitoring, immunization, and participation in nutrition

education seminars. These services are largely provided through the Family Health Programme (Programa de Saude da Famila or PSF) of the Ministry of Health. Together, a doctor, nurse, nurse auxiliary, and four to six community health workers (CHW) comprise a multidisciplinary family health support team working in health units in defined geographic areas covering no more than 5,000 residents. (Harris and Haines 2010). There are some 250,000 CHWs. These CHWs must live in the area they work, are responsible for up to 120 families in a defined area and perform home visits to each family once a month.

The PSF provides primary healthcare services in 95% of all 5,564 municipalities, covering 55% of the population or some 85 million people. Coverage of PSF is not universal as it is targeted at poorer areas. Since 2008 the PSF has included a nutritionist in the family health support teams to strengthen nutrition capacity (Jaime et al. 2011). The SISAN also includes the school meal programme (Programa Nacional de Alimentaçâo Escolar or PNAE) through the Ministry of Education, and the Food Acquisition Programme (FAP) through the Ministry of Social Development. There is a requirement that 30% of food used in schools be locally produced, and the Ministry of Agriculture oversees the production, with funds transferred to municipal governments. The FAP primary focus is on small farm families and gives special attention to native fruits and places a high premium on diversified production and consumption by small producer families. School meals are required to be healthy, using fruits and vegetables produced by local farmers and minimizing the use of ultra-processed foods.

The interventions delivered through the various delivery platforms of SISAN including the BFP, PSF, FAP, and PNAE are largely targeted at mothers and their children in poor families. The BFP monthly cash transfers range from $US7 to US$45 per family depending on eligibility (very poor and moderately poor). The health conditionality means that BFP beneficiaries are most likely going to be PSF clients. The PSF CHW’s primary concern is maternal and child health: they provide counselling on maternal and infant and young child feeding, ensure adherence to micronutrient supplementation and immunization, as well as provide guidance on oral rehydration in the treatment of diarrhoea as appropriate. Increasingly CHWs are involved in curative care, triage, and referral to the health unit and health promotion for chronic diseases.

The impact of the various component parts of the SISAN has been quite remarkable in the last decade. The BFP has contributed to reduction of poverty with those on less than US$ 1.25 a day falling from 25.6% in 1990 to 4.8 percent in 2008. Estimates are that the BFP contributed to about 35% of the reduction of the extreme poverty gap. Although there is limited rigorous evidence that the BFP resulted in improved nutrition outcomes (Martins et al. 2013), there is considerable ecological evidence that it contributed to national reductions.

National stunting rates in young children fell from 37.1% in 1974–5 to 7.1% in 2006–7, accompanied by a narrowing of the difference in stunting rates between poor and rich quintiles. The steepest decline was in the last 10 years of that period (Monteiro et al. 2010). In the Northeast of Brazil, the prevalence of child undernutrition fell by one-third between 1986 and 1996 and by three-quarters between 1996 and 2006 (de Lima et al. 2010). The improvements in the first period were mostly associated with improvements in maternal schooling and in water and sanitation coverage, whereas in the second period increasing purchasing power of the poorer families as well as maternal schooling were the strongest associations. Quite remarkably the median duration of breastfeeding in Brazil increased from 2.5 months in 1970 to 14 months in 2006–7 (Victora et al. 2011). Furthermore, it seems likely that PSF has also contributed to the dramatic decreases in infant mortality seen in Brazil in the last few decades (Aquino et al. 2009)

In France, the ‘Ensemble Prévenons l’Obésité Des Enfants’ (EPODE) was first developed in 2003, with the aim to reduce the prevalence of obesity in children under 12 years of age. The institutional arrangements include a central coordination in the Ministry of Health in Paris which liaises with the office of the Mayor in the cities that are involved, where there is a local project manager that works with a local steering committee. EPODE is a large-scale, coordinated, capacity-building approach for communities to implement effective and sustainable strategies to prevent childhood obesity (Borys et al. 2011). EPODE is designed to involve all relevant local stakeholders in an integrated and concrete prevention programme aimed at facilitating the adoption of healthier lifestyles in everyday life (Van Koperen et al. 2013). The delivery platforms include local crèches, health clinics, sports facilities, cultural associations, supermarkets, restaurants, school catering services, school recreation facilities, local parks, and gymnasiums. The activities developed and supported include advocacy and social

mobilization through local media campaigns, together with capacity building and coaching of local actors, as well as the monitoring and evaluation of these activities.

The initial evaluations of the first ten EPODE cities gave encouraging results with ten to 20% reductions in obesity prevalence (Romon et al. 2009). EPODE now extended to nearly 1.8 million inhabitants in 167 French cities, 20 cities in Spain, and 8 cities in Belgium. In June 2008 the first EPODE Mayors' Club European congress was held. Success to date is measured by a large field mobilization in the pilot cities and by the encouraging reduction of the BMI of children in France within the pilot cities (European Public Health Alliance 2008).

The sale of sugary soft drinks and vending machines was also banned in all schools across France. A tax on sugary soft drinks was introduced in 2012, which has led to companies making demands at the EU for evidence of the impact of such taxes. The resulting study found that food taxes achieve a reduction in the consumption of the taxed products and, in some cases, product reformulation aimed at reducing the sugar, salt, and fat levels of the product. It is also found that product substitution takes place, both through an increase in the consumption of taxed products from cheaper brands and non-taxed or less-taxed product substitutes. To what extent such changes in consumption lead to public health improvements was still considered inconclusive however (Ecorys 2014).

Conclusion

Shifting the paradigm for addressing the DBM will require attention to the common drivers of malnutrition. In particular, translating the science into interventions that support and empower people to consume diets of high nutrition quality. The DBM is not two separate problems, but the product of inadequate *access to affordable, quality* sufficient foods. However, transitioning to a food system that delivers foods of high nutritional quality across the population will require addressing the politico-economic factors that shape food production, processing, distribution, and retail. It will also require re-visioning nutrition interventions to consider implications for undernutrition as well as risk factors for overweight and NCDs. For example, programmes like Bolsa Familia and the Mexican Opportunidades have been associated with weight gain and consumption of sugar and SSBs

(de Bem Lignani et al. 2011). To address these global challenges adequately, a key enabler will be strategic investment in PHN capacity at the local, national, and global level. Among the key aims should be promoting exclusive breastfeeding and adequate complementary feeding in children up to 2 years of age, as well as promoting the local production of fruits and vegetables and increasing their availability in local markets and reducing the supply of ultra-processed foods.

REFERENCES

Aquino, R., de Oliveira, N.F., and Barreto, M.L. (2009). Impact of the family health program on infant mortality in Brazilian municipalities. *American Journal of Public Health*, 99(1), 87–93.

Baker, P., Kay, A., and Walls, H. (2014). Trade and investment liberalization and Asia's noncommunicable disease epidemic: a synthesis of data and existing literature. *Globalization and Health*, **10**, 66.

Balajaran, Y. and Reich, M.R. (2016). Political economy challenges in nutrition. *Globalization and Health*, **12**, 70.

Bezanson, K. and Isenman, P. (2010). Scaling up nutrition: a framework for action. *Food Nutrition Bulletin*, 31(1), 178–86.

Bhupathiraju, S.N., et al. (2014). Glycaemic index, glycaemic load, and risk of type 2 diabetes: results from 3 large US cohorts and an updated meta-analysis. *American Journal of Clinical Nutrition*, **100**, 218–32.

Bhutta, Z.A., et al. for the Maternal and Child Undernutrition Group (2008). What works? Interventions for maternal and child undernutrition and survival. *The Lancet*, **371**, 417–40.

Bhutta, Z., et al. for the Lancet Nutrition Interventions Review Group, and the Maternal and Child Nutrition Study Group. (2013). Evidence-based interventions for improvement of maternal and child nutrition: what can be done and at what cost? *The Lancet*, **382**, 452–77.

Black, R.E. et al. and the Maternal and Child Nutrition Study Group (2013). Maternal and child undernutrition and overweight in low-income and middle-income countries. *The Lancet*, **382**, 427–51.

Bloom, D.E., et al. (2011). *The Global Economic Burden of Noncommunicable Diseases*. Geneva: World Economic Forum.

Borys, J.M., et al.; EEN Study Group (2011). EPODE approach for childhood obesity prevention: methods, progress and international development. *Obesity Reviews*, **13**, 299–315.

Carpenter, K.J. (2003). A short history of nutritional science: part 4 (1945–1985). *Journal of Nutrition*, **133**, 3331–42.

Cecchini, M. and Warin, L. (2016). Impact of food labelling systems on food choices and eating behaviours: a systematic review and meta-analysis of randomized studies. *Obesity Reviews*, 17(3), 201–10.

Chmielewska, D. and Souza, D. (2011). *The Food Security Policy Context in Brazil*. IPCIP Country study No 22. Brasilia: UNDP.

Dalton, M.A., et al. (2017). Child-targeted fast-food television advertising exposure is linked with fast-food intake among pre-school children. *Public Health Nutrition*, 20(9),1548–56.

Darnton-Hill I., Nishida, C., and James, W.P.T. (2004). A life course approach to diet, nutrition and the prevention of chronic diseases. *Public Health Nutrition*, **7** (1A), 101–21.

de Bem Lignani, J., Sichieri, R., Burlandy, L., and Salles-Costa, R. (2011). Changes in food consumption among the Programa Bolsa Família participant families in Brazil. *Public Health*

Nutrition, 14(05), 785–92.
de Lima, A.L.L., et al. (2010). Causes of the accelerated decline in child undernutrition in North-eastern Brazil (1986–1996–2006) *Rev Saúde Pública*, 44(1), 1–10.
de Onis, M., Garza, C., Onyango, A.W., and Martorell, R. (2006). *WHO Child Growth Standards*. *Acta Paediatrica Supplementum*, **450**, 1–101.
de Onis, M. and Branca, F. (2016). Childhood stunting: a global perspective. *Maternal & Child Nutrition*, 12(Suppl. 1), 12–26.
Dehghan, M., et al. on behalf of the Prospective Urban Rural Epidemiology (PURE) study investigators. 2017. Associations of fats and carbohydrate intake with cardiovascular disease and mortality in 18 countries from five continents: a prospective cohort study. *The Lancet*, **390**, 2050–62.
De Schutter, O. (2010). 'Food Commodities Speculation and Food Price Crisis', United Nations Special Rapporteur on The Right to Food, Briefing Note 2, September 2010. World Economy, Ecology and Development (WEED) (2008).
Development Initiatives (2017). *Global Nutrition Report 2017: Nourishing the SDGs*. Bristol, UK: Development Initiatives.
D'Odorico, P., Carr, J.A., Laio, F., Ridolfi, F.L., and Vandoni, S. (2014). Feeding humanity through global food trade. *Earth's Future*, **2**, 458–69.
Downes, C. (2014). Is Codex Alimentarius all talk? The importance of standards in transnational food governance. In: Downes, C. (ed.) *The Impact of WTO SPS Law on EU Food Regulations*. Switzerland: Springer International Publishing, pp. 205–43.
Ecorys (2014). Food taxes and their impact on competitiveness in the agri-food sector. Rotterdam: European Competitiveness and Sustainable Industrial Policy Consortium, ECSIP Consortium.
Engdahl, F.W. (2007). *Seeds of Destruction: The Hidden Agenda of Genetic Manipulation*. Montreal: Centre for Research on Globalization.
ETC Group (2013). *Putting the Cartel Before the Horse … and Farm, Seeds, Soil and Peasants, etc: Who Will Control Agricultural Inputs? Communiqué 111. The State of Corporate Concentration*. Ottawa: Action Group on Erosion, Technology and Concentration.
ETC Group (2015). *With Climate Chaos…Who Will Feed Us? The Industrial Food Chain or The Peasant Food Web*. Ottawa: Action Group on Erosion, Technology and Concentration.
European Public Health Alliance (2008). *EPODE—Together Let's Prevent Childhood Obesity*. [Online] Available at: https://epha.org/epode-together-lets-prevent-childhood-obesity/
Fan, J., Song, Y., Wang, Y., Hui, R., and Zhang, W. (2012). Dietary glycaemic index, glycaemic load, and risk of coronary heart disease, stroke, and stroke mortality: a systematic review with meta-analysis. *PLoS One*, 7(12), e52182.
Fanzo, J., et al. (2013). *The Integration of Nutrition into Extension and Advisory Services: A Synthesis of Experiences, Lessons, and Recommendations*. Lindau, Switzerland: Global Forum for Rural Advisory Services.
Food and Agriculture Organization of the United Nations (FAO) (2002). *World Agriculture: Towards 2015/2030*. Rome: Food and Agriculture Organization of the United Nations.
Food and Agriculture Organization of the United Nations (FAO) (2003). *World Agriculture: Towards 2015/2030: An FAO Perspective*. London: Earthscan Publications Ltd.
Food and Agriculture Organization of the United Nations (FAO) (2009). *The State of Food and Agriculture: Livestock in the Balance*. Rome: Food and Agriculture Organization.
Food and Agriculture Organization of the United Nations (FAO) (2011). *Global Food Losses and Food Waste—Extent, Causes and Prevention*. Rome: Food and Agriculture Organization of the United Nations.
Food and Agriculture Organization of the United Nations (FAO) (2011a). *Global Food Price Monitor*. Rome: Food and Agriculture Organization.
Fardet, A. (2016). Minimally processed foods are more satiating and less hyperglycemic than ultra-processed foods: a preliminary study with 98 ready-to-eat foods. *Food & Function*, 7(5), 2338–

46.
Fardet, A., Méjean, C., Labouré, H., Andreeva, V.A., and Feron G. (2017). The degree of processing of foods which are most widely consumed by the French elderly population is associated with satiety and glycemic potentials and nutrient profiles. *Food Function*, 8(2), 651–8.
Fiolet, T., et al. (2018). Consumption of ultra-processed foods and cancer risk: results from NutriNet-Santé prospective cohort. *British Medical Journal*, **360**, k322.
Garrett, J.L. and Ruel, M.T. (2005). Stunted child-overweight mother pairs: prevalence and association with economic development and urbanization. *Food Nutrition Bulletin*, 26(2), 209–21.
Garrett, J., Bassett, L., and Levinson, F.J. (2011). Principles and a conceptual model for working multisectorally. In: Garret, J.L. and Natalicchio, M. (eds.). *Working Multisectorally in Nutrition: Principles, Practices and Case Studies*. Washington: IFPRI, Chapter 3.
GBD 2016 Risk Factor Collaborators (2017). Global, regional, and national comparative risk assessment of 84 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *The Lancet*, **390**, 1345–422.
Gillespie, S., McLachlan, M., and Shrimpton, R. (eds.). (2003). *Malnutrition: Time to Act. The World Bank-UNICEF Nutrition Assessment*. Washington, DC: The World Bank.
Grantham-McGregor S, et al. and the International Child Development Steering Group (2007). Developmental potential in the first 5 years for children in developing countries. *The Lancet*, **369**, 60–70.
Halabi, S.F. (2015). The Codex Alimentarius commission, corporate influence, and international trade: a perspective on FDA's global role. *American Journal of Law & Medicine*, 41(2–3), 406–21.
Haldar, S., Chia, S.C., and Henry, C.J. (2015). Body composition in Asians and Caucasians: comparative analyses and influences on cardiometabolic outcomes. *Food Nutrition Research*, **75**, 97–154.
Harcombe, Z., Baker, J.S., DiNicolantonio, J.J., Grace, F., and Davies, B. (2016). Evidence from randomised controlled trials does not support current dietary fat guidelines: a systematic review and meta-analysis. *Open Heart*, 3(2), e000409.
Harris, M. and Haines, A. (2010). Brazil´s family health program. *British Medical Journal*, **341**, c4945.
Hawkes, C. (2005). The role of foreign direct investment in the nutrition transition. *Public Health Nutrition*, 8(4), 357–65.
Hawkes, C. and Buse, K. (2011). Public-private engagement for diet and health: addressing the governance gap. *SCN News*, **39**, 6–10.
Hawkes, C., et al. (2015). Smart food policies for obesity prevention. *The Lancet*, 385(9985), 2410–21.
Heaver, R. (2005). *Strengthening Country Commitment to Human Development: Lessons from Nutrition*. Washington, DC: World Bank.
Igumbor, E., et al. (2012).'Big Food,' the Consumer Food Environment, Health, and the Policy Response in South Africa. *PLoS Medicine*, 9(7), e1001253.
Jaacks, L.M., Slining, M.M., and Popkin, B.M. (2015). Recent underweight and overweight trends by rural-urban residence among women in low- and middle-income countries. *Journal of Nutrition*, 145(2), 352–57.
Jaime, P.C., Silva, A.C.F., Lima, A.M.C., and Bortolini, G.A. (2011). Ações de alimentação e nutrição na atenção básica: a experiencia de organização no governo Brasileiro. *Nutrition Reviews*, 24(6), 809–24.
Jelliffe, D.B. (1966). *The Assessment of the Nutritional Status of the Community*. World Health Organization Monograph Series No 55. Geneva: World Health Organization.

Kearney J. (2010). Food consumption trends and drivers. *Philosophical Transactions of the Royal Society of London, Series Biological Sciences*, 365(1554), 2793–807.
Kerckhoffs, T., van O, R., and Vander Stichele, M. (2010). *Financing Food. Financialization and Financial Actors in Agricultural Commodity Markets*. Amsterdam: Stichting Onderzoek Multinationale Ondernemingen.
Koneswaran, G. and Nierenberg, D. (2008). Global farm animal production and global warming: Impacting and mitigating climate change. *Environmental Health Perspectives*, 116(5), 578–82.
Lehmann, U. and Sanders, D. (2007). *Report on Community Health Workers: What Do We Know About Them? Evidence and Information for Policy, Department of Human Resources for Health*. Geneva: World Health Organization.
Leroy, J.L. and Menon, P. (2008). From efficacy to public health impact: a call for research on program delivery and utilization in nutrition. *Journal of Nutrition*, 138(3), 628–9.
Levinson, F.J. and Balarajan, Y. (2013). *Addressing Malnutrition Multi-Sectorally: What Have We Learned from Recent International Experience?* New York: UNICEF.
Lindert, K., Linder, A., Hobbs, J., and de la Briere, B. (2007). *The Nuts and Bolts of Brazil's Bolsa Familia Program: Implementing Conditional Cash Transfers in a Decentralized Context. Social Protection Discussion Paper No 0709*. Washington, DC: The World Bank.
Lowder, S.K., Skoet, J., and Raney, T. (2016). The number, size, and distribution of farms, smallholder farms, and family farms worldwide. *World Development*, **87**, 16–29.
Malik, V.S., Schulze, B., and Hu, F.B. (2006). Intake of sugar sweetened beverages and weight gain: a systematic review. *American Journal of Clinical Nutrition*, **84**, 274–88.
Margulis, M.E. (2013). The regime complex for food security: implications for the global hunger challenge. *Global Governance: A Review of Multilateralism and International Organizations*, 19(1), 53–67.
Martínez Steele, E., et al. (2016). Ultra-processed foods and added sugars in the US diet: evidence from a nationally representative cross-sectional study. *BMJ Open*, **6**, e009892.
Martins, A.P.B., Canella, D.S., Baraldi, L.G., and Monteiro, C.A. (2013). Cash transfers in Brazil and nutritional outcomes: a systematic review. *Review Saude Publica*, 47(6), 1159–71.
Mason, J.B., Sanders, D., Musgrove, P., Soekirman, and Galloway R. (2006). Community health and nutrition programmes. In: Jamison, D.T. et el. (eds.) *Disease Control Priorities in Developing Countries* (2nd ed.). Washington, DC: World Bank, Chapter 56.
Mazzocchi, M., Shankar, B., and Traill, B. (2012). *The Development of Global Diets Since ICN 1992: Influences of Agri-Food Sector Trends and Policies*. FAO Commodity and Trade Policy Research Working Paper No. 34. Rome: Food and Agriculture Organization of the United Nations.
McMichael, P. (2012). The land grab and corporate food regime restructuring. *The Journal of Peasant Studies*, 39(3–4), 681–701,
Meade, B. (2017). Global Food Industry. United States Department of Agriculture. [Online] Available at: https://www.ers.usda.gov/topics/international-markets-trade/global-foodmarkets/global-food-industry/
Meiro-Lorenzo, M., Villafana, T.L., and Harrit, M.N. (2011). *Effective Responses to Non-Communicable Diseases: Embracing Action Beyond the Health Sector*. HNP Discussion Paper. Washington, DC: The World Bank.
Micha, R., et al. (2017). Etiologic effects and optimal intakes of foods and nutrients for risk of cardiovascular diseases and diabetes: systematic reviews and meta-analyses from the Nutrition and Chronic Diseases Expert Group (NutriCoDE). *PLoS One*, 12(4), e0175149.
Millstone E. and Lang, T. (2003). *The Atlas of Food. Who Eats What Where and Why*. London: Earthscan Publications.
Mokoro (2015). *Independent Comprehensive Evaluation of the Scaling Up Nutrition Movement: Final Report—Main Report and Annexes*. Oxford: Mokoro Ltd.

Monteiro, C.A, et al. (2010). Narrowing socioeconomic inequality in child stunting: the Brazilian experience, 1974–2007. *Bulletin of the World Health Organization*, 88(4), 305–11.

Monteiro, C.A., Moubarac, J.C., Cannon, G., Ng, S.W., and Popkin, B. (2013). Ultra-processed products are becoming dominant in the global food system. *Obesity Reviews*, **14** (Suppl. 2), 21–8.

Mozaffarian, D., Hao, T., Rimm, E.B., Willett, W.C., and Hu, F.B. (2011). Changes in diet and lifestyle and long-term weight gain in women and men. *The New England Journal of Medicine*, **364**, 2392–404.

Murphy, S., Burch, D., and Clapp, J. (2012). *Cereal Secrets: The World's Largest Commodity Traders and Global Trends in Agriculture*. Oxfam Research Reports. Oxford: Oxfam International.

NCD Risk Factor Collaboration (2016). Trends in adult body-mass index in 200 countries from 1975 to 2014: a pooled analysis of 1698 population-based measurement studies with 19.2 million participants. *The Lancet*, **387**, 1377–96

NCD Risk Factor Collaboration (2017). Worldwide trends in body-mass index, underweight, overweight, and obesity from 1975 to 2016: a pooled analysis of 2416 population-based measurement studies in 128.9 million children, adolescents, and adults. *The Lancet*, **390**, 2627–42.

Newens, K.J. and Walton, J. (2016). A review of sugar consumption from nationally representative dietary surveys across the world. *Journal of Human Nutrition and Dietetics*, 29(2), 225–40.

OXFAM (2013). *Behind the Brands. OXFAM Briefing Paper 166*. Oxford: OXFAM International.

Pingali, P.L. (2012). Green revolution: impacts, limits, and the path ahead. *PNAS*, 109(31), 12302–308.

Rahmati, S., Delpishe, A., Azami, M., Ahmadi, M.R.H., and Sayehmiri, K. (2017). Maternal anemia during pregnancy and infant low birth weight: a systematic review and meta-analysis. *International Journal of Reproduction and Biomedicine*, 15(3), 125–34.

Reardon, T., et al. (2015). Urbanization, diet change, and transformation of the downstream and midstream of the agrifood system: effects on the poor in Africa and Asia. *Faith and Economics*, **66**, 43–63.

Reardon, T. and Gulati, A. (2008). *The Rise of Supermarkets and Their Development Implications: International Experience Relevant for India.* IFPRI Discussion Paper 00752. Washington, CD: International Food Policy Research Institute.

Riera-Crichton, D., and Tefft, N. (2014). Macronutrients and obesity: revisiting the calories in, calories out framework. *Economics and Human Biology*, **14**, 33–49.

Rollins, N.C., et al. (2016). Why invest, and what it will take to improve breastfeeding practices? *The Lancet*, **387**, 491–504.

Romon, M., et al. (2009). Downward trends in the prevalence of childhood overweight in the setting of 12-year school- and community-based programs. *Public Health Nutrition*, 12(10), 1735–42.

Rosset, P.M. (2006). *Food is Different: Why We Must Get the WTO Out of Agriculture*. London: Zed Books.

Sadeghirad, B., Duhaney, T., Motaghipisheh, S., Campbell, N.R., and Johnston, B.C. (2016). Influence of unhealthy food and beverage marketing on children's dietary intake and preference: a systematic review and meta-analysis of randomized trials. *Obesity Review*, 17(10), 945–59.

Scrimshaw, N.S. and Viteri, F.E. (2010). INCAP studies of kwashiorkor and marasmus. *Food Nutrition Bulletin*, 31(1), 34–41.

Shrimpton, R., Victora, C.G., de Onis, M., Lima, R.C., Blössner, M., and Clugston, G. (2001). Worldwide timing of growth faltering: implications for nutritional interventions. *Pediatrics*, 107(5), E75.

Shrimpton, R. (2002). Nutrition and communities. In: UN ACC/SCN (ed.) *Nutrition: A Foundation for Development*. Geneva: UN ACC/SCN.

Shrimpton, R. and Rokx, C. (2012). *The Double Burden of Malnutrition: A Review of Global Evidence*. Washington, DC: The World Bank.
Shrimpton, R., et al. (2013). Nutrition capacity development: a practice framework. *Public Health Nutrition*, 17(3), 682–8.
Shrimpton, R., et al. on behalf of the WPHNA Capacity Building Task Force (2016). Public health nutrition capacity: assuring the quality of workforce preparation for scaling up nutrition programmes. *Public Health Nutrition*, 19(11), 2090–100.
Shrimpton, R., Mbuya, N.K., and Provo, A.M. (2016). *The Double Burden of Malnutrition in East Asia and the Pacific: Evidence and Lessons for a Multisectoral Response*. Washington, DC: World Bank.
Shrimpton, R., Bazzano, A., and Mason, J. (2017). Getting fat: 'what' is eaten is as important as 'how much' is eaten. *Journal of Obesity and Weight Loss Therapy*, 7(4), 344.
Shrimpton, R. (2017). Fixing our food system: an imperative for achieving sustainable development. *World Nutrition*, 8(2), 207–31.
Te Morenga, L., Mallard, S., and Mann, J. (2012). Dietary sugars and body weight: systematic review and meta-analyses of randomised controlled trials and cohort studies. *British Medical Journal*, **345**, e7492.
Turner, C., et al. (2017). *Concepts and Methods for Food Environment Research in Low and Middle Income Countries*. Agriculture, Nutrition and Health Academy Food Environments Working Group (ANH-FEWG). London: Innovative Methods and Metrics for Agriculture and Nutrition Actions (IMMANA) Programme.
The GBD 2015 Obesity Collaboration 2017. Health effects of overweight and obesity in 195 countries over 25 years. *The New England Journal of Medicine*, 377(1), 13–27.
Thow, A.M. and Hawkes, C. (2009). The implications of trade liberalization for diet and health: a case study from Central America. *Global Health*, 28(5), 1–15.
Thow, A.M., Schönfeldt, H., Viljoen, A., Gericke, G., and Negin, J. (2017). Policy for the complex burden of malnutrition in Africa: a research agenda to bring consumers and supply chains together. *Public Health Nutrition*, 20(6), 1135–9.
Thow, A.M., Downs, S.M., Mayes, C., Trevena, H., Waqanivalu, T., and Cawley, J. (2018). Fiscal policy to improve diets and prevent noncommunicable diseases: from recommendations to action. *Bulletin of the World Health Organization*, **96**, 201–10.
United Nations (UN) (2015). *Transforming Our World: The 2030 Agenda for Sustainable Development*. New York: The United Nations General Assembly.
United Nations (UN) (2017). *The Sustainable Development Goals Report 2017*. New York: The United Nations.
UNICEF (2013). *Improving Child Nutrition: The Achievable Imperative for Global Progress*. New York: United Nations Children's Fund.
UNICEF/WHO (2004). *Low Birthweight: Country, Regional and Global Estimates*. New York: United Nations Children's Fund.
Van Koperen, T.M., et al. (2013). Obesity prevention: characterizing the EPODE logic model: unravelling the past and informing the future. *Obesity Reviews*, **14**, 162–70.
Victora, C.G., Aquino, E.M.L., Leal, M.C., Monteiro, C.A., Barros, F.C., and Szwarcwald, C.L. (2011). Maternal and child health in Brazil: progress and challenges. *The Lancet*, **377**, 1863–76.
Victora, C.G., et al. for The Lancet Breastfeeding Series Group (2016). Breastfeeding in the 21st century: epidemiology, mechanisms, and lifelong effect. *The Lancet*, **387**, 475–90.
Wang, C.Y., McPherson, K., Marsh, T., Gortmaker, S.L., and Brown, M. (2011). Health and economic burden of the projected obesity trends in the USA and the UK. *The Lancet*, **378**, 815–25.
Waterlow, J.C. (1973). Note on the assessment and classification of protein-energy malnutrition in children. *The Lancet*, 2(7820), 87–9.

Weeratunga, P., Jayasinghe, S., Perera, Y., Jayasena, G., and Jayasinghe, S. (2014). Per capita sugar consumption and prevalence of diabetes mellitus—global and regional associations. *BMC Public Health*, **14**, 186.

Weis, T. (2007). *The Global Food Economy: The Battle for the Future of Farming*. London: Zed Books.

World Health Organization (WHO) (1995). *Physical Status: Uses and Interpretation of Anthropometry*. WHO Technical Report Series, Report No. 854. Geneva: World Health Organization.

World Health Organization and the Food and Agriculture Organization of the United Nations (WHO/FAO) (2003). *Diet, Nutrition and the Prevention of Chronic Diseases WHO Technical Report No 916*. Report of a joint WHO/FAO expert consultation. Geneva: World Health Organization.

World Health Organization (WHO) (2004). Appropriate body-mass index for Asian populations and its implications for policy and intervention strategies. *The Lancet*, 363(9403), 157–63.

World Health Organization (WHO) (2013a). *Global Nutrition Policy Review: What Does it Take to Scale Up Nutrition Action?* Geneva: World Health Organization.

World Health Organization (WHO) (2013b). *Global Action Plan for the Prevention and Control of Noncommunicable Diseases 2013–2020*. Geneva: World Health Organization

World Health Organization (WHO) (2015). *Guideline: Sugars Intake for Adults and Children*. Geneva: World Health Organization.

World Health Organization (WHO) (2016). *Global Report on Diabetes*. Geneva: World Health Organization.

World Health Organization (WHO) (2016a). *Report of the Commission on Ending Childhood Obesity*. Geneva: World Health Organization.

World Health Organization (WHO) (2017a). *Guideline: Assessing and Managing Children at Primary Health-Care Facilities to Prevent Overweight and Obesity in the Context of the Double Burden of Malnutrition. Updates for the Integrated Management of Childhood Illness (IMCI)*. Geneva: World Health Organization.

World Health Organization (WHO) (2017b). *Double-Duty Actions.* Policy brief. Geneva: World Health Organization.

World Bank (2006). *Repositioning Nutrition as Central to Development: A Strategy for Large Scale Action*. Washington, DC: The World Bank.

World Bank (2013). *Fish to 2030: Prospects for Fisheries and Aquaculture. World Bank Report No. 83177-GLB*. Washington, DC: The World Bank.

Willett, W.C. (2011). The great fat debate: total fat and health. *Journal of American Dietetic Association*, 111(5), 660–2.

Yngve, A., Tseng, M., Haapala, I., and Hodge, A. (2012). A robust and knowledgeable workforce is essential for public health nutrition policy implementation. *Public Health Nutrition*, **15**, 1979–80.

Zhao, L., Zhang, X., Shen, Y., Fang, X., Wang, Y., and Wang, F. (2015). Obesity and iron deficiency: a quantitative meta-analysis. *Obesity Review*, 16(12), 1081–93.

2.9

Climate change and human health

Alistair Woodward and Alex Macmillan

Introduction to the environment and climate change

This chapter aims to introduce readers to the basic science underlying climate change, to explore the links between climate change and health, and examine the implications of this knowledge for public health policy and practice.

As a start, readers are invited to check their understanding of climate science against the multiple-choice questions in Box 2.9.1. (The answers are shown at the end of the chapter.)

The Earth has an average surface temperature suitable for life that is the product of a number of natural influences on the balance of radiation (radiative 'forcing' at the top of the atmosphere). These include solar activity (the amount of energy produced by the sun), the tilt and orbit of the planet, volcanic activity that injects gases and ash into the upper atmosphere, other dusts and aerosols that block incoming solar radiation, and the amount of reflection ('albedo') by clouds, ice, and snow. Also important is the concentration of so-called 'greenhouse' gases. These gases are transparent to relatively high-frequency incoming solar radiation in the visible spectrum. However, when this radiation is absorbed on the surface of the Earth it is re-emitted in the infra-red range, which is trapped by the greenhouse gases. This action is similar to the effect of a pane of glass in a greenhouse (or the window of a car sitting in the sun). Without the blanket for the planet that is provided by the greenhouse gases, the Earth would be approximately 30°C colder than it is at present. Carbon dioxide (CO_2) is the most abundant of these heat-trapping gases. Other important greenhouse agents include methane, nitrous oxide, black carbon, and manufactured halocarbons (e.g. chlorofluorocarbons or CFCs).

CO_2 levels in the atmosphere vary over long timescales due to transfers from the atmosphere to the biosphere and the oceans, and from the oceans carbon is eventually sequestered in sedimentary rocks. In the very long term, carbon from the Earth's crust is ejected by volcanic eruptions into the atmosphere, and the cycle begins again.

What has changed in the last 200 years or so is the large-scale emission of CO_2 (and other heat-trapping gases) as a result of human activities, such as clearing of forests and burning of fossil fuels. This process began in pre-industrial times as a relatively minor addition of carbon to the atmosphere but has escalated rapidly in the last 50 years. In 2017, about 37 giga-tonnes (Gt) of CO_2 were released into the atmosphere each year, roughly 150 times more than the amount emitted in the middle of the nineteenth century. CO_2 is a long-lived greenhouse gas: about half of the carbon that is emitted into the atmosphere is taken up by natural sinks within 30 years, but about 30% will persist in the atmosphere for centuries. Nitrous oxide and methane are much more potent heat-trapping gases, but are emitted in lesser quantities. Nitrous oxide remains in the atmosphere for about a hundred years; methane is short-lived, and is mostly removed within about 12 years.

The concentration of CO_2 in the atmosphere rose steadily from about 280 parts per million (ppm) in pre-industrial times, to 315 ppm in 1960 when direct measurements were first made. Since then the level has increased 2–3% per year, reaching a global average concentration of 403 ppm in 2017. Indirect measures, using ice cores and other indications of historic conditions, tell us that present levels are higher than at any time in the last 5 million years (IPCC 2013).

Consequences of these changes in atmospheric chemistry are evident in the recent temperature records at or close to the Earth's surface, sea level rise, reduced snow and ice cover, and increased ocean acidity. The year-to-year variability in average global temperatures is dominated by natural factors such as the El Niño Southern Oscillation (ENSO), so the effect of persistent, external forcing such as steadily rising greenhouse concentrations is seen most clearly on longer timescales. Indeed each of the last three decades has been warmer than all previous decades in the last 150 years and in early 2018 NOAA reported the 5 warmest years on record have occurred in the last 10 years (Climate Central 2018). This rise in global average temperatures has occurred in every region, over both land and ocean, and cannot be explained by time trends in solar activity and other natural forces.

On the other hand, climate simulations that include anthropogenic contributions (i.e. the increase in CO_2 and other greenhouse gases) fit well with the last 40 years of temperature records (Fig. 2.9.1).

Box 2.9.1 Carbon IQ quiz

Indicate which of the following, A to E, is the best answer for each question.

1. If the planet's atmosphere consisted of oxygen and nitrogen, the global average temperature would change by:

   A. −30 degrees C
   B. −10 degrees C
   C. +10 degrees C
   D. +30 degrees C
   E. 0 degrees C

2. Climate is 'average weather'. Over what period?

   A. At least a week
   B. >2 years
   C. >20 years
   D. A century at least
   E. Millennia

3. One-third of the CO_2 emitted today will still be in the atmosphere in:

   A. 1 year
   B. 10 years
   C. 100 years
   D. 1,000 years
   E. 10,000 years

4. Combustion of biomass releases black carbon, other particulates, and aerosols. The effect is to:

   A. Warm
   B. Cool
   C. Affect regional climate, mostly
   D. Accelerate ice melt
   E. All of the above

5. In the last 50 years, what proportion of the warming due to greenhouse emissions has been absorbed by the ocean?

   A. 5%
   B. 80%
   C. 0%
   D. 20%
   E. 100%

6. Up to 2018, the warmest year on record was:

   A. 2010
   B. 1998
   C. 2016
   D. 2017
   E. 1260

7. How many return flights, Singapore to Rome, are required to generate the annual average emissions of CO_2 per capita for Laos?

A. 200
B. 20
C. 1
D. 0.2
E. 0.02

8. If there was a food with the energy intensity of oil, how many litres would be required to cycle 100 km?

A. 200
B. 20
C. 1
D. 0.2
E. 0.02

9. Which US President gave a commitment to translate the UN Framework Convention on Climate Change into ‘concrete action to protect the planet’?

A. George H.W. Bush
B. Bill Clinton
C. George W. Bush
D. Jimmy Carter
E. Richard Nixon

10. Since 1960, the rise in temperatures in the lower atmosphere (the troposphere) has been accompanied by:

A. Stratospheric warming
B. Cooling in the Arctic
C. An overall reduction in precipitation, worldwide
D. Stratospheric cooling
E. Dilution of the oceans making them less acidic

Other evidence of global warming includes the predominance of highest ever temperatures in the record of temperature extremes (in the western United States, record high temperatures now outnumber record low measurements each year by about three to one) (Meehl et al. 2009), and the unique combination of warming in the troposphere (the lower atmosphere) and cooling at higher levels (due to trapping of heat in the greenhouse layer, and less heat radiating into space). Summertime Arctic sea ice is reducing year on year, and global average sea level is rising at about 3.0 mm/year (twice the rate observed before 1990) (Australian Academy of Sciences 2010).

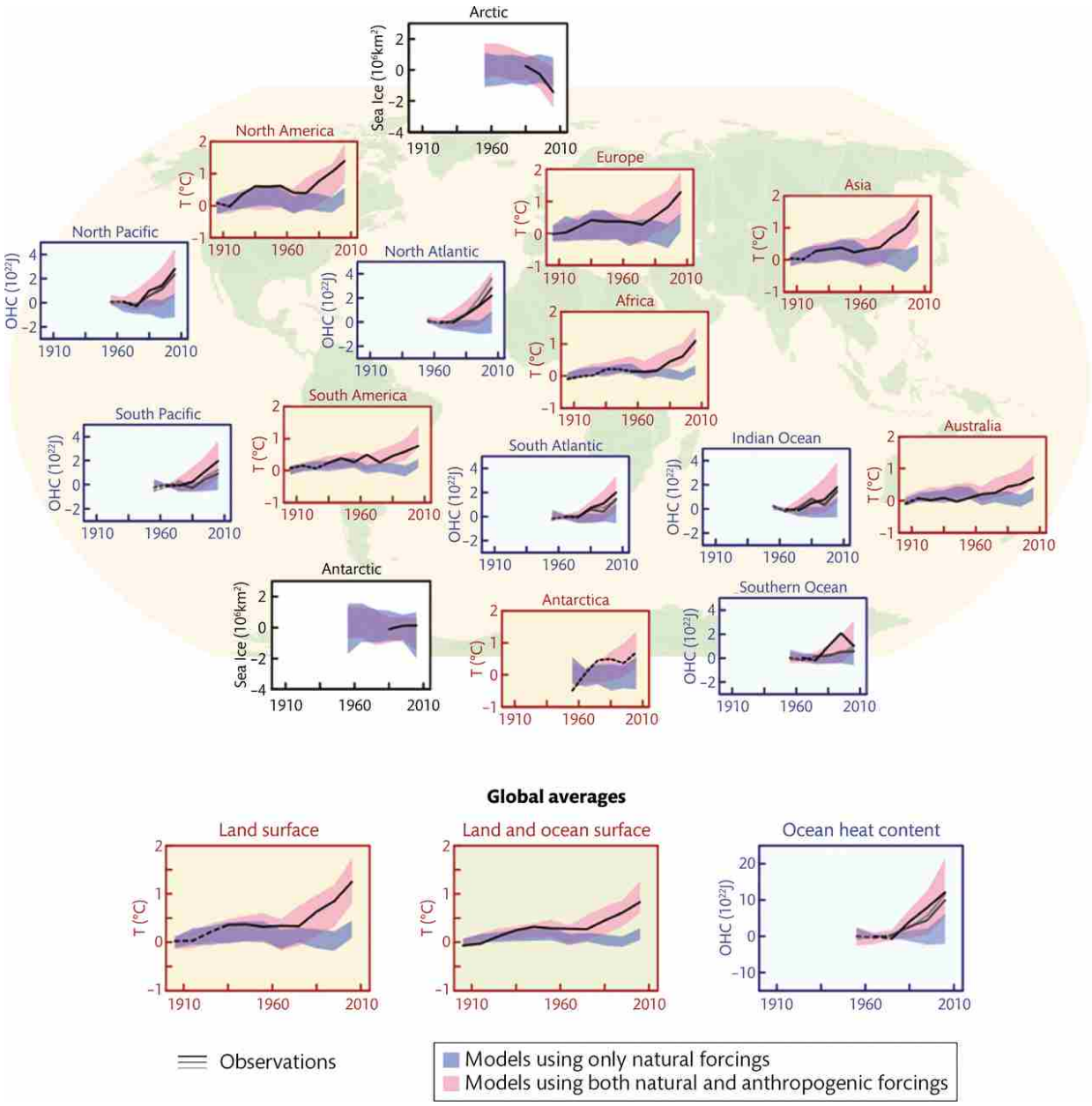


Fig. 2.9.1 Observed and simulated climate change.

Reproduced with permission from Climate Change 2013: The Physical Science Basis. Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, Figure SPM.6. Cambridge University Press.

Since the 1960s, emissions of greenhouse gases have tracked upwards, the rise halted only briefly by major economic disturbances, such as the Asian financial crisis of the late 1990s (Peters et al. 2011). Emissions in the first 15 years of the twenty-first century have matched, or exceeded, the figures

forecast by the most extreme of the climate scenarios reported in the IPCC's Third Assessment Report in 2001.

What about the future? The 5th Assessment Report of the IPCC forecast that if there was essentially no restriction on the burning of fossil fuels or harmful land use changes (the Representative Concentration Pathway 8.5), atmospheric CO_2 would double pre-industrial levels sometime between 2050 and 2100, and global average temperatures would rise as a result by 3.2–5.4°C relative to 1850–1900 by 2100 (IPCC 2013).

The increase in average temperatures will be accompanied by a marked increase in the frequency of very high temperatures. It has been projected that 'mega-heatwaves' of the kind that caused more than 50,000 excess deaths in Russia in 2010 may increase by a factor of 5–10 in Europe, within the next 40 years (Barriopedro et al. 2011). As average temperatures rise, the incidence of cold events is likely to fall, all else being equal, although there may be unexpected effects in some regions due to feedback loops and critical tipping points. In one example, melting of the Arctic ice has disturbed the high-level jet stream over Europe and North America, re-distributing very cold air masses and leading to the apparently paradoxical phenomenon of warm polar seasons and extreme cold events further south (Liu et al. 2012).

The effects of warming and changes in rainfall will be unevenly distributed, geographically. For instance, a global average of 4°C of warming may be associated with an increase in temperatures in the Arctic of 10°C or more, a decline in rainfall by 40–50% in North Africa, droughts threatening the viability of rain-fed agriculture in many parts of sub-Saharan Africa by the end of the century, and much of south-eastern Australia experiencing extreme fire risks as often as every second summer (4°C and Beyond 2009).

In summary, the greenhouse effect is long known and well described. The rate at which greenhouse gas levels in the atmosphere are increasing is unprecedented, and absolute concentrations are presently higher than at any time since the Pleiocene. In association with these changes, the world is warming, in a manner that carries a discernible human 'fingerprint'. For instance, the mix of carbon isotopes in the atmosphere has changed with the rising contribution of CO_2 from burning of fossil fuels (in which the lighter isotopes especially ^{12}C predominate) (IPCC 2013).

Why is this a public health issue? It is important to recognize that for human health, variability and frequency of extreme events are often more important than average conditions. On exposure to heat or cold, it is deviations from usual or most commonly experienced temperatures that carry the greatest risk.

While humans may be able to cope with modest increases in baseline temperatures, there are physical and ecological systems of importance to health that are exposed to environmental change and may be exquisitely sensitive to temperature rise. Agriculture generally operates as an open system, directly exposed to the weather, with very few buffers against extreme events or long-term shifts in ambient conditions. Warming of 2–4°C in the global average would lead to substantial falls in productivity in some parts of the world. In much of sub-Saharan and southern Africa, it has been projected that an increase of 5°C in the global average would reduce the growing period for some staple crops by more than 20%, threatening the coping capacity and resilience of hundreds of millions of people (Thornton et al. 2011). A small degree of ocean warming above certain threshold limits may lead to death of coral and increased rates of ciguatera (fish poisoning) (Tester et al. 2010). Mosquitoes and other disease vectors also respond acutely to small shifts in temperature and rainfall, and increases in average temperatures will increase the potential for disease transmission in some populations (Wandiga et al. 2010). Conversely, disease risks may be reduced in areas where conditions are presently close to the upper limits of vector survival (Gething et al. 2010). Perhaps as important, historical improvements in health have been built on stable governance, liberty, and relative protection from extreme violence.

Another mechanism by which climate change may affect human health is through conflict, disorder, and displacement of populations. There are many examples in history of social disruption caused by rapid fluctuations in temperature and extremes in precipitation (McMichael 2017). Although the sudden variations in climate in the past resulted from natural causes, such as volcanic eruptions and long-term climate cycles, the consequences resemble what may occur due to anthropogenic climate change. Indeed, the risks may be even higher now, given the scale and speed of change that is projected.

Observational studies of weather, climate, and health

A great deal has been written about variations in weather (day-to-day conditions) and climate (average weather over a period of decades), and their relationship with health outcomes. One of the exposures most closely studied is temperature, which may be expressed in many ways, commonly as the daily average or maximum temperature, with or without an adjustment for humidity. Figure 2.9.2 shows a pattern that is frequently observed—a J-shaped relation between temperature and health outcomes (in this case, mortality excluding accidental deaths). The optimal temperature (the point at which mortality is lowest) differs from one setting to another. In the case of Haerbin, a city located in the far north of China, the number of deaths is at its lowest when the average temperature is close to 0°C (Yang et al. 2007). In cities located closer to the equator, the optimal temperature tends to be higher (Fig. 2.9.2). The overall shape of the curve relating temperature to mortality remains similar, but the point of inflection on the curve moves towards the right in cities where high temperatures are more commonly experienced (McMichael et al. 2008). This is a reminder that populations do adapt to their ecological context, physiologically and socially. What is less clear is how quickly this adaptation can occur, if people move, or the climate changes. Another important question is: what are the limits of adaptation? how much heat is tolerable (Box 2.9.2)?

Box 2.9.2 Too big? Or too big to ignore? The challenge of scale

Climate change is fundamentally different from most other environmental problems that public health has faced. It is a matter of scale. Climate change is a global problem, not only because the effects are widely distributed, like acid rain, or influenza epidemics, but because the underlying cause is a breakdown of Earth's homeostatic systems. Other examples of environmental changes that threaten a 'safe operating space' planet-wide include the depletion of high-level atmospheric ozone by chlorofluorocarbons and loss of biodiversity (Steffen et al. 2015). Climate change moves on prolonged timescales. For instance, the full effects of a spike in greenhouse emissions on warming of the deep ocean and associated sea level rise may not be apparent for 5,000 years or more (Stouffer 2012). This provides a serious challenge to public health scientists. In the last hundred years, with advances in exposure assessment and better knowledge of cellular and molecular mechanisms, the focus has shifted to finer-grained explanations of disease. However the contexts in which chemical reactions occur, microbes multiply, and individuals make their decisions is key to understanding both risk and resilience. For example, mortality in the 1995 Chicago heatwave was a function of both individual-level variables (such as age and income), and the qualities of neighbourhoods. The poor and the elderly were at much greater risk during the heatwave if they lived in areas with rundown businesses, abandoned housing, and degraded public spaces (Klinenberg 2002). Climate change represents a particularly stark example of a big-picture problem, featuring extended timescales, ubiquitous exposures, and the confounding effects of human adaptation. In response, public health

must move upstream, beyond studies of the variations between individuals to investigations of causes on a larger scale (Pearce 2011), and pay greater attention to the ecologic perspective on human disease and injury (Krieger 2015). This will very likely require refinement of established methods of public health inquiry (such as time series analysis, and population-level research designs); an understanding of climate science and meteorology; the creative use of analogues; and the application of relevant modelling methods from other disciplines (e.g. system dynamics modelling) (Rutter et al. 2017).

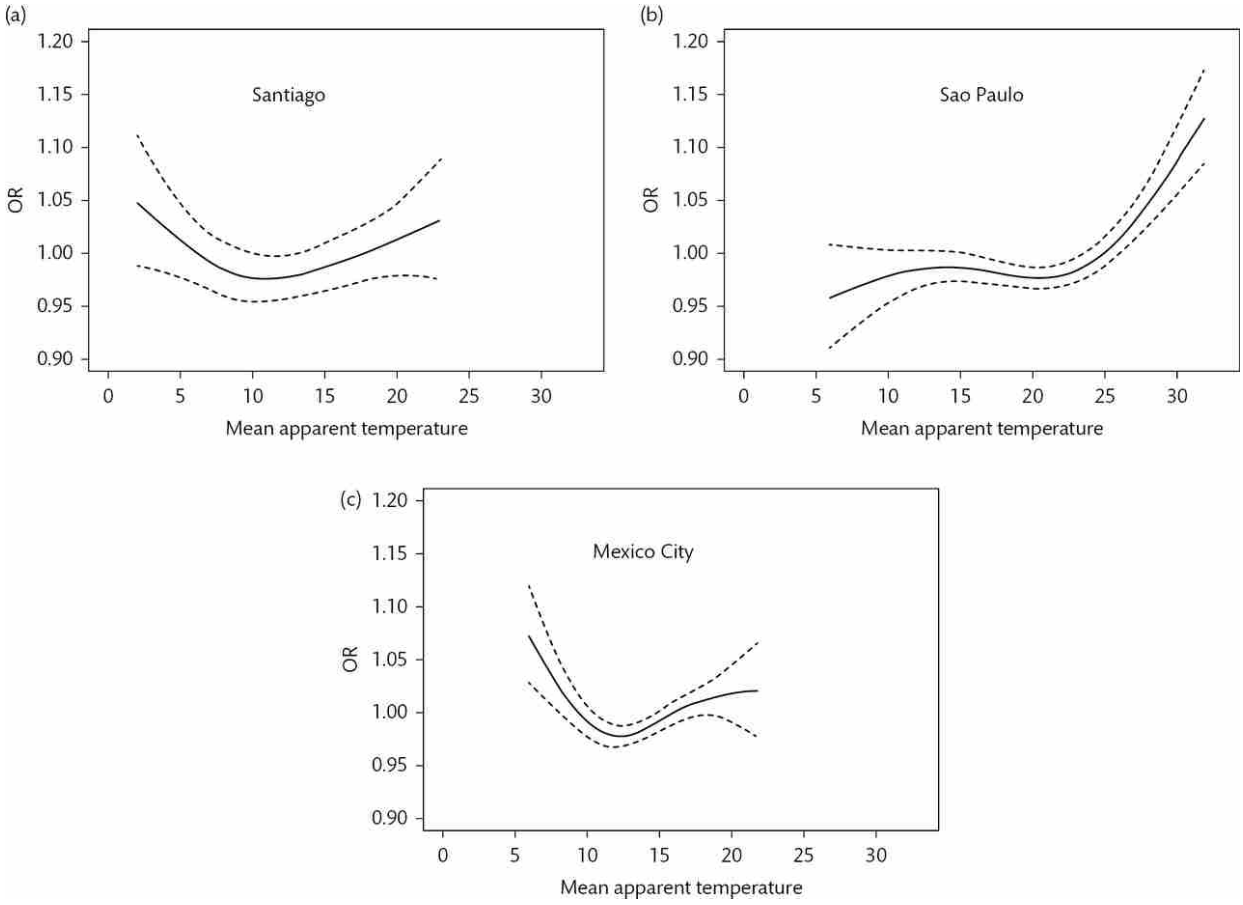


Fig. 2.9.2 Average daily apparent temperature and total mortality: Santiago (33 degrees S), São Paulo (24 degrees S), and Mexico City (19 degrees N). Solid lines show central estimate, dashed lines the 95% confidence intervals. Adjusted for particulate pollution and ozone levels.

Reproduced from Bell ML et al., Vulnerability to heat-related mortality in Latin America: a case-crossover study in São Paulo, Brazil, Santiago, Chile and Mexico City, Mexico, *International Journal of Epidemiology*, Volume 37, Issue 4, pp. 796–804, Copyright © The Author 2008, with permission from Oxford University Press.

Higher temperatures have been linked with a range of health outcomes including deaths and hospital admissions for respiratory and cardiovascular

diseases. In heatwaves, which may be defined as prolonged exposure to temperatures in the top 5–10% of the range for that location, the effects appear to be greater than would be expected by simply extrapolating the temperature–health outcome relationship (Anderson and Bell 2011; Rocklöv et al. 2012). Duration of heat, frequency of heatwaves, and timing (early or late in the season) all appear to modify the effects of high temperatures. High levels of air pollution amplify the damaging effects of heat (Qin et al. 2017).

There is considerable variation within populations in vulnerability to heat. Older people (70 years and over) and children are generally at greater risk. There is no consistent pattern with gender. Men were found to be at greater risk in São Paulo and Santiago (Bell et al. 2008); whereas a study in Brisbane, Australia reported a more marked increase in mortality risk for women than men (Huang et al. 2012). In many countries, low socioeconomic status (SES) has been associated with increased risk of heat- and cold-related poor health (Qin et al. 2017), but again, this is not universal. There is, for example a greater variation by SES in heat mortality in the United States than in England, perhaps because the use of air conditioning is distributed differentially in the United States but not so much in the United Kingdom (where few houses have artificial cooling and there is much less social variation) (Armstrong et al. 2011). Vulnerability is also expressed spatially (Uejio et al. 2011) reflecting the geographical distribution of individual and household characteristics (such as prevalence of chronic diseases, housing design, and poverty) and also district-level variables (urban form, quality of services).

It is not just the maximum or average temperatures that affect health. Variability is important as well. In 135 United States cities, for example, the standard deviation of summer temperatures was associated with mortality of older people suffering from chronic disease, after adjustment for a wide range of risk factors including average temperature (Zanobetti et al. 2012).

As with many climate change impacts, more is known about the effects of high temperatures in high-income countries than in the developing world, although the problem is likely to be more severe in the latter, given the high temperatures that already apply in many low-latitude countries, the very large populations, often with multiple health problems, that are exposed, the large number of outdoor workers and the low quality of housing and urban environments.

The effects of heat on the health of workers has emerged as an important issue, given the large number of people who may be affected, and the multiple adverse consequences of reduced productivity and loss of livelihoods (Lucas et al. 2014). This is especially significant in equatorial countries where high temperatures already place severe restraints on the duration and intensity of outdoor work (Kjellstrom et al. 2013).

The effects of climate on infectious disease have long been known (see Box 2.9.3). The first textbook of tropical medicine, by Patrick Manson, was published in 1900, and was called *A Manual of the Diseases of Warm Climates* (Manson 1900). Manson wrote principally of infectious diseases but he recognized that few conditions were peculiar to the tropics. In other words, he understood that climate variables such as temperature and precipitation play an important part in, but are seldom the entire explanation for, the distribution and occurrence of disease.

The mosquito-borne diseases are good examples. In warmer conditions, vectors such as *Aedes aegypti* (responsible most commonly for the transmission of dengue fever) develop more quickly from the immature stage (Tun-Lin et al. 2000). In tropical countries in which temperatures vary little through the year, the transition from a time of low rainfall to the rainy season may act as a powerful climatic trigger for outbreaks of dengue (Schreiber 2001). But other factors are important also. Urbanization and the quality of housing, for instance, modify the effect of climate on dengue risk (Wu et al. 2009). Overall, it is projected that climate change will extend the present geographic range of dengue and threaten populations which do not currently experience local transmission of the disease (Ebi and Nealon 2016).

Box 2.9.3 Plague and climate

Plague is caused by infection of humans with *Yersinia pestis*. The organism is established in rodent populations in parts of Asia, Africa, and the Americas, and periodically causes outbreaks of disease in local populations. The prevalence of *Yersinia* in animal hosts is climate sensitive. For example, a study of gerbils in Kazakhstan found the risk of disease was increased by warmer springs and wetter summers (Stenseth et al. 2006). An increase of 1°C in spring was associated with an increase of more than 50% in the prevalence of animals infected with *Yersinia*. Climate also affects the abundance of rodent hosts and flea vectors, as well as the chances of humans coming into contact with the pathogen. Consequently, it is not surprising that many historical studies have shown a link between climate and plague epidemics (McMichael 2017). As one example: a study of plague in China between 1850 and 1960 found the number of cases per year was positively related to rainfall,

although extremely high precipitation in the wetter parts of the country led to lower than expected plague intensity (Xu et al. 2011).

There are many ways by which physical and mental health may be affected by variations in climate or weather. In the south-western part of Australia, the prevalence of depression has been related to the severity of long-term drought (measured as dry-land salinity) (Speldewinde et al. 2009). The physical degradation of environments following climate change has been linked to loss, grief, separation, and self-questioning in many environments (Berry et al. 2010; Cunsolo and Ellis 2018). Non-communicable diseases may be affected: in Bangladesh the dry season is marked by increasing salinity and rising temperatures of the ground water, and it is thought these changes are linked to marked seasonal variations in the incidence of high blood pressure and eclampsia in pregnancy (precipitated by increases in sodium loading) (Xun et al. 2010). Cholera is another marker in Bangladesh of warming and salt intrusion, but indeed all the vibrio pathogens (including *vulnificus* and *parahaemolyticus* as well as *cholerae*) are temperature sensitive and have been described as 'microbial barometers of climate change' (Baker-Austin et al. 2017).

In summary, humans have adapted to local conditions during a period of relatively stable climate (the Holocene epoch),but are susceptible to ill-health caused by rapid changes in temperature and precipitation. The connection between extremes of cold and heat on mortality is well established, and many infectious diseases also vary with climate.

Attribution—present impacts of climate change

In 2013 the Fifth Assessment Report of the IPCC concluded that 'It is extremely likely more than half of the observed increase in global average surface temperature from 1951 to 2010 was caused by the anthropogenic increase in greenhouse gas concentrations and other anthropogenic forcings' (IPCC 2013, p. 17). In the IPCC lexicon, 'extremely likely' means the probability is more than 95%.

The IPCC assessment reports have described many changes in physical and biological systems that are consistent with global warming. For example, during the last one hundred years the distribution of North Sea cod has changed dramatically, with fish moving northwards and into deeper

waters, largely as a consequence of ocean warming (the other influence being the pressure of commercial fishing) (Engelhard et al. 2014). In the Alsace region of France, in association with higher summer temperatures, the potential alcohol content of Riesling grapes, at harvest, increased by almost 1% per year (Duchêne and Schneider 2005). It is not difficult to attribute changes such as these to climate change, as there are few alternative plausible explanations.

The story of tick-borne diseases, including Lyme disease and tick-borne encephalitis (TBE) illustrates well the complexities of attribution. In the last decade of the twentieth century the vector for TBE moved northward across the Baltic and Scandinavia, in association with warming that has been most marked in winter and early spring (Lindgren and Gustafson 2001). The incidence of TBE surged in the early 1990s in the Baltic states, but this did not fit closely with climate. There were considerable variations within the Baltic in TBE rates, and although warming continued after 2000, TBE rates fell (Sumilo et al. 2007). Clearly there are other factors that are important besides the arrival of a climate that was more favourable for transmission of the TBE virus. When the Soviet Union broke up in 1989, many collective farms ceased to operate, areas that had been used for pasture and agriculture were left untended, and these provided new habitat for rodents and deer. At the same time, during a period of economic disruption and transition, people were more dependent than before on wild food sources, including mushrooms and small game, and as a result spent more time in the countryside. In combination these changes, climatic and non-climatic, may have increased tick abundance and boosted, for a short time, the exposure of humans to tick-borne diseases (Sumilo et al. 2007).

Lyme disease emerged in the United States in the 1970s, most likely due to socioeconomic changes and transition of farmland in the north-east to forests (Ebi et al. 2017). The disease was slow to get established in Canada, although the local habitat was suitable and there was movement of natural hosts, vector, and pathogen (*Borrelia burgdorferi*) across the border from the United States. It is likely that temperature was the limiting factor. Since 2000, in the same period as lower Canada has warmed, local populations of the tick vector have established, and the number of cases of Lyme disease has increased. The links between environmental change, the shift of vectors, and the onset of disease suggest that the spread of Lyme disease, in this region, is due in part to climate change (Ebi et al. 2017).

In public health, attribution involves probabilities, based on the risk of an event occurring following a particular exposure, compared with the probability of the same event taking place without the exposure. The inability to attribute a particular instance of lung cancer to cigarette smoking is quite consistent with a firm conviction that smoking causes the majority of such cancers in populations in which tobacco use is common. Such an approach is applied also in climate sciences to make judgements about the causes of extreme events. For instance, a study of cyclones in the north Pacific in 2015 concluded the extreme energy levels that were observed were highly likely to be due to human-induced climate change (Woodward and Samet 2018). A study of the 2003 European heatwave concluded it was very likely (probability greater than 90%) that human influences on climate had more than doubled the risk of such an extreme event (Stott et al. 2004). Note that a relative risk of more than 2.0 means it is more likely than not that the outcome is attributable to the exposure.

Forecasting

To attribute past and present morbidity and mortality to climate change is not straightforward, but future attribution presents even greater difficulties. There is no certainty about how much the climate will change in the future, or what the consequences will be, but it is clear that the most important risks lie ahead of us. Present-day effects of climate change on health may be miniature versions of future effects, or perhaps not: future challenges may be threshold-crossing novelties. For this reason, a great deal of effort has been applied to attempts to project future climates and the consequent impacts through modelling studies. One approach is to attempt to isolate the effect of an altered climate on specific diseases by holding all other independent variables constant. We will look at examples of this approach using several diseases known to be highly climate sensitive. Since climate change acts as a multiplier of disadvantage, ideally studies would go further and ask: how are the ‘baseline’ risks likely to change in the short to medium term? This is a difficult undertaking but could, potentially, provide a more accurate description of future impacts of climate change on human health. Modelling studies that have taken this approach are less common and require methods not so familiar to epidemiologists—methods that can deal with significant

levels of complexity, non-linear interactions, and feedback dynamically over time.

A study of future impacts of climate change on the 'potential transmission' of malaria in Zimbabwe attempted to isolate the effects of climate between 2000 and 2050, by assuming there would be no other changes (Ebi et al. 2005). The suitability of the climate for stable transmission of *Plasmodium falciparum* malaria was related to mean temperatures, winter minimum temperatures, and monthly rainfall, based on outputs from the MARA (Mapping Malaria Risk in Africa) project. On this basis the researchers reported that within 50 years malaria could be established in the densely populated and presently malaria-free highlands of Zimbabwe. The model used in this study was risk based (suitability varied from 0 to 1), was explicitly projecting the potential for disease, rather than predictions of disease occurrence, and set aside the question of how other influences on malaria, apart from climate, might differ in 2050 from the present day.

In a less common non-communicable disease example, a study in the United States modelled the current relation between mean annual temperature and hospital admissions for stones in the urinary tract in the United States. In general, nephrolithiasis is more common in populations living in warm climates, and this trend is apparent in the United States. The authors applied projections of mean annual temperatures in the future and calculated the potential increase in hospital admissions and associated costs (Brikowski et al. 2008). Since both linear and non-linear relationships between temperature and risk fitted the present data, both forms were modelled. Projections were based on one of the IPCC's intermediate severity climate change scenarios (SRES A1B).

Even more challenging is to model the future burden of disease due to climate change impacts on the upstream determinants of health. Crucially, the effects of climate change on the global food system may have far-reaching effects on health. To investigate this issue, researchers linked models that projected food supply under different climate scenarios with estimates of the effects of changes in fruit and vegetable consumption, red meat intake and bodyweight (Springmann et al. 2016) (Fig. 2.9.3). The study was the first to assess the effects of changes in diet composition (e.g. intake of fruit and vegetables), and found that this was more important in terms of health outcomes than simple loss of calories. It was assumed that recent

improvements in most parts of the world in food security and nutrition-related health outcomes would continue, so the main effect of climate change would be to act as a brake on progress. Limitations in this assessment include the relatively short time-frame (up to 2050 only), inability in a global assessment to accurately capture local conditions and the particular difficulties faced by some populations (e.g. those on low-lying Pacific islands), and the possibility that wide-ranging mitigation policies might themselves threaten the production and distribution of food (Woodward and Porter 2016).

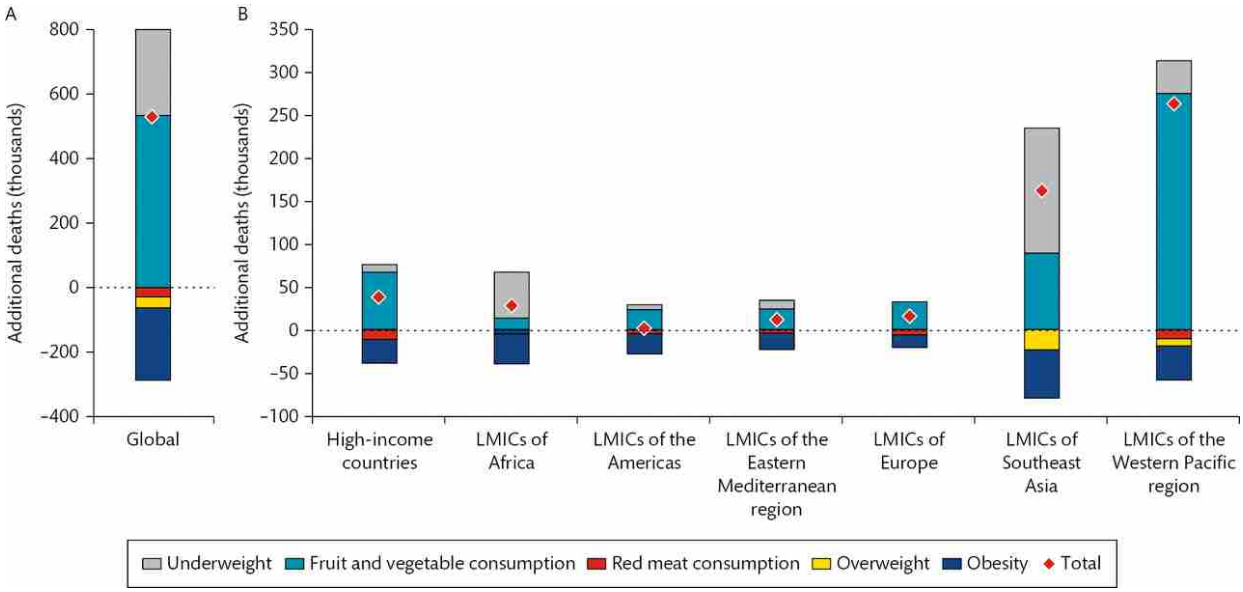


Fig. 2.9.3 Climate-related deaths (in thousands) in 2050
Reproduced from Springmann et al. (2016) 'Global and regional health effects of future food production under climate change: a modelling study', *Lancet*, with permission from Elsevier.

Demographic projections are important, since climate change impacts will depend on the size, structure, and health status of world populations. The mid-range estimate is for the world population to increase from about 7.6 billion presently to around 9.7 billion by 2050, with most of the growth occurring in developing countries (United Nations 2015b). The proportion of the population in older age groups is likely to increase, almost everywhere, although the average years of life remaining will change much less (Lutz et al. 2008). Mortality is projected to decline, in most countries, leading to increases in life expectancy at birth of around 5–6 years by the mid-century

(World Health Organization 2011). Assuming that present trends of social and economic development continue, illness rates will also decline, and it has been estimated that the overall burden of disease (measured in disability-adjusted life years) will be reduced by about 30% by 2030 (World Health Organization 2008). Child mortality is falling, worldwide, with no sign of a reduction in the rate of improvement (You et al. 2015). However, there are still very large differences between countries, and some, especially in sub-Saharan Africa, are improving slowly, if at all (World Health Organization 2011).

How could it be that health statistics are improving around the world while pressures mount on the global environment, from climate change and other big-picture problems (e.g. loss of biodiversity, overloading waterways and soils with nitrogen and phosphorus)? We have no clear reference point of course: mortality may have fallen more rapidly, child stunting may have diminished sooner, if climate change was not acting. But the more important point relates to timing. There is good reason to believe the full costs of large-scale human exploitation of the planet have not yet become apparent (McMichael 2017): human populations have been ‘eating the future’ by headlong clearing of land, burning of fossil fuels, and drawing down on sources of fresh water.

Whether social and economic development continues on the present trajectories is of course unknown, but is fundamental to projections both of climate change and vulnerability to its effects. In turn, uncertainty about the pace of social and economic development stems, in part, from the possible effects of climate change on societies and economies, and on the pathways we take in responding. Furthermore, health is sensitive to the manner in which economic growth occurs. If development is not inclusive, then vulnerability to the effects of climate change may not improve. India, for instance, has taken great strides, economically, in the last 20 years, but the health problems of poverty persist for many people. The prevalence of child underweight and stunting fell modestly between 1992 and 2005 (from 49% to 40% underweight, for instance), but there was no relation, at the state level, with income growth (Subramanyam et al. 2011). Child health measured in this way advanced no faster in the richest states than the poorest.

These examples demonstrate the intertwined nature of climate change and development and serve as an entry point to understanding that climate

change and our actions to address it are issues of social and environmental justice, equity, and human rights. The highest income countries have achieved substantial improvements in life expectancy, health, and quality of life, in significant part due to fossil-fuel dependent economic development. Up to 2010, the most developed countries had contributed more than half of all anthropogenic emissions, despite being home to only 17% of the global population (den Elzen et al. 2013). On the other hand, we have already demonstrated that the impacts of climate change are being, and will continue to be borne, by the lowest income countries and populations within countries who are already experiencing socioeconomic inequities. The relationship between climate change and human rights is discussed further in Chapter 2.9. Indigenous peoples are both particularly vulnerable to the health impacts of climate change and have specific resilience capabilities and knowledges for mitigation and adaptation (Ford 2012). These are discussed in detail in Chapter 10.6.

Will there be positive effects of climate change, and if so, how will they balance the negative impacts? This is not a straightforward question to answer. It is likely there will be some positive effects in the short term, but it is difficult to judge their magnitude. For instance, it is likely that modest warming will boost agricultural yields in temperate parts of the world, increasing food production in these regions. Moreover, food crops such as wheat, rice, barley, and oats, tend to grow more quickly in higher levels of CO_2. Together, these effects may reduce the risk of food shortages and undernutrition caused by lower production in hot, low-latitude countries (Gornall et al. 2010). The balance of gains and losses will depend on whether populations at risk of food shortages can in fact get access to the climate dividend, and the relief may be temporary: the increase in crop yields in temperate zones will almost certainly diminish as warming continues (Teixeira et al. 2011). In tropical developing countries, where food insecurity is greatest, the impact will be most severe. In sub-Saharan Africa, where the caloric intake of 60% of the population in the late 2000s was insufficient for an active life, 5 degrees of warming (global average) is projected to cut maize yields by 19% and beans by 68% (Thornton et al. 2011).

Where disease vectors and pathogens are already close to the upper temperature limit for reproduction and transmission, further warming is likely to reduce the risk of disease. In parts of equatorial Africa, for example,

projected temperatures will exceed the optimal settings for both the vector and the parasite, and the potential for malaria transmission will rapidly diminish (Ebi et al. 2017).

The effects of cold and heat on mortality have been closely studied, and shed some light on the matter of balancing gains and losses with global warming (Ebi and Mills 2013). The J-shaped relation between ambient temperature and mortality suggests that warming will lead to fewer cold-related deaths, and greater heat-related mortality. This curve is rarely symmetrical—in cooler locations, the right-hand arm of the curve (associated with heat) tends to rise more rapidly, and the opposite is true for populations living in relatively warm conditions. This means that if there is a trade-off of gains and losses with warming, the balance will depend on the initial conditions. But it is more complicated still—it is not clear whether cool-location populations will retain their (relatively greater) tolerance for cold as temperatures rise, or whether the shape of the temperature–mortality curve will change. Also, it is uncertain how many of the deaths that occur at colder times of the year are in fact due to low temperatures, rather than other phenomena that vary with season (such as infectious disease epidemics). Many populations in colder parts of the world, where the most affluent and best buffered populations in the world are located, have adapted to a large extent to cold, and the impacts of both cold spells and ambient low temperatures have much reduced over time. Consequently, the added benefits of global warming in these locations may be modest, and outweighed by the additional burden of heat-related ill-health in the much larger and more vulnerable populations living in low latitudes (Kinney et al. 2015).

It is important to identify the most important sources of uncertainty in projections. A study of diarrhoeal disease attempted to distinguish the effects of uncertainties in temperature changes and the exposure–response function (change in disease incidence per unit warming) (Kolstad and Johansson 2011). The focus in future-oriented studies tends to be on the climate modelling, given the many assumptions that have to be incorporated. However in this case it was evident that uncertainty associated with the risk coefficient (which varied almost fourfold between studies) caused more variation in the outcome than the choice of climate change model. Another study in the United States examined the effects of modelling choices on national mortality attributable to low-level ozone, in the presence of climate

change (Post et al. 2012). Results varied from about 600 deaths per year avoided as a result of climate change to 2,500 deaths attributable to climate change (although most models projected increases in mortality). The most influential analytic steps were the choices of climate change scenario and air quality model. It should be noted also that the ozone studies, and many other investigations of climate impacts, have tended to project effects based on mid-range climate scenarios, when the globe is in fact tracking on a steeper trajectory for emissions, one that fits more closely with the high-range models. This highlights the issue of thresholds and tipping points—as we move further away from familiar environments and encounter more extreme exposures, it becomes less certain that present exposure–response relations will apply in the future.

There are two ways of responding to climate change: adapt to the new environment (analogous to secondary prevention) and mitigate (primary prevention—take whatever steps are needed to avoid further, increasingly dangerous climate change occurring). Primary prevention is necessary from a public health perspective, but there are strong imperatives for now also investing in adaptation: in particular, the world is committed to warming as a result of greenhouse gases already emitted, due to the long time-course of global carbon cycles and heat transfers. This is regardless of future emission trajectories. Both adaptation and mitigation have implications for health, and we deal with each in turn.

Adaptation to expected climate change

The importance of adaptation is proportional to the extent and momentum of climate change—when there are very long lag times in the climate system then interventions to reduce emissions will take a very long time to have their full effect, and in the meantime, steps must be taken to cope with changes that are, in effect, unavoidable. The IPCC's Fifth Assessment Report estimated the Earth has already warmed by 0.85°C on average, compared with the late 1800s, with a resulting decrease in cold days and nights, a likely increase in both heat waves and heavy rainfall events (particularly in North America and Europe), and sea level rise of about 20 cm (IPCC 2013). It has been estimated that we are already committed to further warming of at least 0.7°C over the next 30 years because of the greenhouses gases that are now in the atmosphere. Also in the atmosphere

are airborne pollutants, many of which have a cooling effect, and are tending to dampen down the global warming signal. It is difficult to determine the size of this effect, but it may be as much as 0.2°C of latent warming that will be expressed when local air pollution is better controlled and fossil fuel use declines (IPCC 2013).

Then there is the question of future emissions. In 2015, the United Nations adopted the Paris Agreement, since signed by all member states and ratified by 175 of 197 members, including the United States and China (the two single largest emitting countries, responsible for 45% of global emissions). The Agreement commits parties to increasing cuts to emissions to keep average global warming to below 2°C (the globally agreed threshold for 'dangerous' climate change), while aiming for no more than 1.5 degrees to protect the most vulnerable low-lying island states from devastating sea level rise. While not legally binding, the Paris Agreement is the first truly global instrument to address climate change, and recognizes both the need to protect health, and the opportunity for health co-benefits in well-designed climate action (United Nations 2015a). At present, global emissions are close to 50 Gt CO_2 equivalent per year, about 14% higher than the level required to hold warming to around 2°C. Following decades of rapid growth, the years 2014–2016 saw global emissions plateau for 3 years but in 2017 emissions rose again, to an historic high. This resulted from increasing energy demand and a slowdown in energy efficiency improvements (International Energy Agency 2018). There is a mixed picture emerging internationally. Some countries (including the United States and members of the European Union) are tracking steadily downwards, while emissions continue to rise in large, rapidly-developing countries such as China. The year of peak global emissions is critical. The sooner greenhouse emissions worldwide begin to fall the less painful the pathway to climate targets. For example, if global emissions did not peak until 2025 and then reduced by 3% per year, global average surface temperatures would rise by as much as 4 degrees C by the end of this century. More severe reductions, about 8% per year, would be needed to stay within 2°C (Hausfather and Pearce 2017).

On the one hand, there are strong forces driving emissions upwards. These include population pressures (a 50% increase projected worldwide in the next 50 years) and economic development 'catch up': 90% of the growth in emissions this century will come from states that are presently in the low- and middle-income categories. On the other hand, there is some evidence

that economic development and emissions are beginning to 'decouple', especially in South America (Hausfather and Pearce 2017). Further, there is increasing global political leadership from many of the high-income countries responsible for the majority of emissions, as evidenced by the Paris Agreement. Even in countries where there is a lapse of national-level leadership (such as the US president's threats to leave the Paris Agreement), local and state-level governance continue to be committed to reducing emissions in line with obligations. There are also some helpful feedbacks at play. For example, as alternatives to fossil fuel energy generation such as solar panels and wind turbines become more popular, the price of these alternatives fall, making them even more attractive. Both onshore wind and photovoltaic solar generation now have cheaper lifecycle costs than the marginal cost of continuing to generate electricity by burning coal in existing power plants (International Renewable Energy Agency 2018).

Locked-in climate change and the risk of mitigation failure are both compelling reasons to find options for coping with climate change impacts, particularly adaptation to increasing average and extreme temperatures, as well as resilience to storms, flooding, sea level rise, and disruption to agricultural production and distribution. Here we will examine some of the opportunities for adaptation, ranging from individual behaviours to neighbourhood interventions, to sector-wide changes. Some of the available options for adapting to climate change can also assist with climate change mitigation, or have other co-benefits for health and equity, while others may hamper efforts in these areas. We cover some examples relating to heat and storm events.

At an individual level, improving the body's physiological response to temperature extremes can assist with adaptation. Increasing physical activity, reducing obesity, and improving diabetes control have many benefits to health, including enhanced capacity to deal with heat (Kenny and Jay 2013). Clinical trials have shown that short-term endurance training increases sweat rate and peripheral blood flow and leads to more effective control of body temperature during exercise (Ichinose et al. 2009). Overweight military recruits in the United States were found to be seven times more likely to suffer heat illnesses during training than their colleagues (Bedno et al. 2010). Work of this kind indicates that public health programmes to promote physical activity, prevent unhealthy weight gain, and reduce the prevalence of obesity are likely to reduce vulnerability to heat.

Community-oriented interventions have promise, since risk factors for heat-attributable deaths tend to cluster spatially. For instance, studies in New York City found that poverty, ethnicity, poor housing, exposure to greenspace, and the prevalence of hypertension were all associated with heat wave mortality, and vulnerable census tracts may be identified through a comprehensive neighbourhood heat vulnerability index (Madrigano et al. 2015).

Although air conditioning has been advocated by some as an important measure for coping with high temperatures (Ostro et al. 2010), it brings vulnerability, as well as consuming large amounts of (often fossil fuel-based) electricity, and transferring further heat outdoors. Artificial cooling relies on a consistent power supply, and heatwaves may trigger electricity outages due to air conditioning demand overload (Anderson and Bell 2011). In these circumstances, populations accustomed to living with air conditioning may be at increased risk of ill-health due to high temperatures. Buildings designed for air conditioning are often poorly designed in other ways to deal with heat (with large, exposed surfaces of glass, and minimal natural shade, for instance). Also, physiological adaptation requires some exposure to heat strain, that is, an increase in body temperature (Maloney and Forbes 2011). Populations with around-the-clock access to air conditioning are less likely to develop the capacity to respond to extreme temperatures, when and if this occurs. Since most deaths during heat waves occur inside buildings, a focus on passive building and surrounding measures that don't rely on energy use will be vital for equitable heat wave survival ('passive survivability') (Baniassadi and Sailor 2018). These include window design, balancing insulation for cold and heat, and balancing tensions between well-designed ventilation and shading, and energy efficiency through all seasons.

Cities are significantly warmer than their surrounding rural areas (the urban heat island effect) due to darkly paved surfaces and dense anthropogenic heat generation (e.g. from vehicles and buildings). City-level interventions to reduce the health risks due to high temperatures will therefore be important (Phelan et al. 2015). Designing cities to reduce the heat island effect can also make them safer during times of high temperature. A study in Phoenix, Arizona estimated the reduction that could be made in heat-related emergency calls by an urban re-design package (Silva et al. 2010). This included changes in building and road materials to increase heat flows, more urban green spaces, rooftop treatments, and widespread

application of reflective paints. In combination, the authors estimate, these interventions could reduce inner-city temperatures sufficiently to almost halve the number of annual heat-related emergency calls.

Measures at a city-level to manage high temperatures when they do occur include heat-health early warning systems, better cooling of at-risk environments (such as residential care facilities for the elderly), identifying the isolated and most vulnerable citizens, and improving the preparedness of the healthcare sector. In France, many such changes to heatwave responsiveness were made following the severe heatwave in 2003. When another heatwave occurred in July 2006, the question asked was: did the country cope better this time than it did in 2003? In July 2003, the increase in daily temperatures was theoretically sufficient to increase daily mortality by about 230%, according to the long-run (1975–1999) relationship between temperatures and daily deaths (Fouillet et al. 2008). But a 290% increase was observed, an excess of more than 15,000 deaths. In 2006, the heatwave temperatures were not as high as 3 years earlier, and it was theoretically projected to raise mortality by about 60%. In fact, the death rate increased by only 14%. One possible explanation is that the 2003 disaster brought forward many deaths in relatively frail individuals, greatly reducing the pool of susceptibles in 2006. However, the mortality trends indicate the 4,600 fewer-than-expected deaths in July 2006 were indeed evidence of more effective adaptation.

The contrast between more aggressive air conditioning and greater provision of green space in cities is an example of how ‘business as usual’ approaches can miss important opportunities (Hess et al. 2012). On the one hand, enhancing air conditioning helps to protect some of the most vulnerable individuals from high temperatures, though it carries a heavy cost in power consumption. On the other hand, a radical expansion in urban green spaces and street trees may bring a range of benefits (in addition to cooling), including carbon sequestration and health benefits due to reduced air pollution, better mental health, and more opportunities for physical activity (Salmond et al. 2016). But these ‘win-win’ interventions are seldom straightforward: green spaces on their own may not be sufficient to achieve health gains. A study of the 49 largest cities in the United States (Richardson et al. 2012) compared the extent of green space with all-cause mortality, adjusting for important social and demographic differences, but not for measures of urban density. They found, against expectations, that health

appeared to be worse in greener cities, and this was explained by the negative effects of factors associated in the United States with green suburbs, mainly sprawling land use and high car dependence.

Another important area for adaptation is in preparedness and resilience to severe storms and flooding. Most climate change models project an increase in the intensity of tropical storms, though it is less clear whether the frequency of storms will change. These extreme events already cause substantial loss of life and economic damage. For example, in Burma in 2008, Cyclone Nargis caused approximately 138,000 deaths and US $10 billion damage (Fritz et al. 2009); in the United States, Hurricane Katrina in 2005 was responsible for about 1,200 deaths (Blake and Gibney 2011) and direct losses of over US $130 billion (Field et al. 2012). Yet, in almost all regions more recent storms of similar or greater severity have seen lower event mortality rates, despite increasing populations living in vulnerable coastal zones globally. Southeast Asia is an important exception to this, due to frequent cyclones, very dense coastal cities and variable preparation and warning systems (Bouwer and Jonkman 2018). There are examples of countries with an improved record for managing the health risks of storms and cyclones. Cuba was struck by four destructive hurricanes (category 2 or greater) between 1998 and 2002. While tens of thousands of homes were damaged and hundreds of thousands of people were evacuated, there were very few deaths (12 in all) (Thompson and Gaviria 2004). While Hurricane Katrina killed many in the United States, no deaths resulted in Cuba from the same storm, despite 20,000 homes being destroyed. The large number of deaths associated with Hurricane Katrina in the United States was a result of serious deficiencies in the public health system, and lack of protection for the poorest and most marginalized social groups in the city (Brunkard et al. 2008).

In contrast, Cuba invests heavily in traditional civil defence activities, including stockpiles of equipment, clothing, fuel, and food. There are effective storm warning systems, and good means of communicating information to the public. Other factors that contribute to resilience are high levels of literacy, provision of basic facilities in all rural areas, and a comprehensive and accessible primary healthcare system. An Oxfam analysis of disaster preparedness in Cuba found there was strong community engagement, so that populations that are threatened can be mobilized rapidly, and protection extends to all social groups, not just those with personal

resources (Thompson and Gaviria 2004). These more immediate disaster preparedness activities are complemented by longer term shifts in the built environment, involving housing construction, land use, and urban planning (Lizarralde et al. 2015). The Cuban example is salutary for climate change adaptation for public health. Building resilience to protect health from climate change will require co-ordinated actions by central and local government, health, and other sectors, as well as communities.

Table 2.9.1 The five main pillars of disaster resilience in Cuba's built environment

Pillar	Scale	Stages of intervention	Types of intervention
Construction industry development	Country	Prevention, reconstruction, long-term development	Project[a], programme[b], national policy
Housing development	Country	Prevention, rehabilitation, reconstruction, long-development	Project, programme, national policy
Systematic urban and regional planning	Community, city, region	Prevention, reconstruction, long-term development	Plan[c], urban policy
Early warning systems	Country	Prevention, emergency	National policy and campaigns
Evacuation plans	Individual, community	Prevention, emergency	National policy and campaigns

Notes:

[a] Often undertaken by one institution with specific objectives and timeframes.

[b] Often undertaken by one institution over a long period of time.

[c] Involves more than one institution and is revised or updated periodically.

Source: data from Lizarralde, G., Valladares, A., Olivera, A., Bornstein, L., Gould, K., & Barenstein, J. D. (2015). A systems approach to resilience in the built environment: the case of Cuba. *Disasters, 39*(s1), s76–s95. doi: 10.1111/disa.12109

There are other opportunities to learn from the variability in past experience of coping with climate-related health problems. For instance, enteric infections such as that caused by *Salmonella* tend to be more

common at higher temperatures, but the size of the increase in disease rates for a given amount of warming varies from one site to another. A study of the five largest Australian cities found that all showed an increase in *Salmonella* notifications with monthly mean temperature, but one city stood out as following a much steeper trajectory than the others (D'souza et al. 2004). Perhaps aspects of the food safety programmes present elsewhere are missing in this apparently particularly temperature-sensitive city.

There is no doubt that climate change acts principally as a risk multiplier. The greatest impact will occur in populations that, for other reasons, already have the worst health statistics, while patterns of economic development now also shape future climate change vulnerability. The World Bank in its landmark 2010 report on climate change and development concluded that 'baseline health status is the single most important determinant of both future losses, and the cost of adaptation' (World Bank 2009). It follows that adaptation must include improvements in the social determinants of health, including social safety nets, as well as essential public health services. But is it true that this is all that is needed? In other words, is 'business as usual' development, the package of improvements that we would want to be making anyway, sufficient preparation for climate change?

Together with leading economists (Fankhauser and Stern 2016), we argue that 'business as usual' patterns of economic development will not be sufficient. Climate change is unique in geographic scale, long time lags, and momentum. Impacts accumulate and progress with such long lag times (sea level rise is a good example) that the process must be regarded as irreversible over our usual public health timescales. This is not a one-off event, but a new environmental dynamic, with enormous momentum. We are likely to be pushed to the edge of what is familiar, and beyond, leading potentially to abrupt and radical changes in Earth systems (Lenton and Ciscar 2013). In epidemiological terms, even if the trends in exposures (rising temperatures, changing patterns of precipitation) remain the same, it may be difficult to anticipate the dose–response relationships that apply in the future, due to unforeseen tipping points, thresholds, and non-linearities. The much greater than expected damage caused by unprecedentedly intense bushfires in Victoria, Australia in 2009 is a good example (Teague et al. 2010).

Climate change does not *only* multiply existing risks. It may contribute to new risks where ecosystems are pressed beyond significant thresholds. This

may happen on the edges of areas affected by climate-sensitive diseases. One example is the acute illness that occurs after eating sea-foods contaminated with *Vibrio parahemolyticus*. The organism is temperature sensitive, and until 2004 no instances of this particular food poisoning had been reported in North America above the US/Canada border. In that year there was an outbreak of cases on a cruise ship in Alaska, linked to shell-fish from a local oyster farm in Prince William Sound. On investigation, it was found that water temperatures at the farm (and the surrounding ocean) had been rising steadily for a decade, and presumably had crossed a critical boundary beyond which the organism could be sustained in local shell-fish (McLaughlin et al. 2005).

Altogether, these unique aspects of climate change require special responses to meet health in development goals. In particular, agricultural investments and rapid urban infrastructure development will need to simultaneously be resilient to climate change impacts, low-carbon, and health-promoting, requiring unprecedented partnerships across sectors—signalled by SDG 17.

Finally, it is crucial to note the limits to adaptation (Dow et al. 2013), emphasizing the need to address the root cause of climate change (anthropogenic greenhouse gas emissions). These limits include human physiological tolerance to extreme heat; limits to the geographical and physiological adaptation of other species on which we rely (such as food plants and pollinators); financial limits to disaster protection, response and recovery; and socio-political limits to governance under multiple stressors, leading to conflict and mass migration.

In summary, given the extent of committed warming, and the difficulty of achieving substantial reductions in emissions, the world almost certainly faces large changes in the next half century. Improving baseline health status is fundamental to adaptation, but not sufficient—there are distinctive features of climate change that have to be taken into account. Discussions about adaptation also help us to understand the limits to adaptation and the importance of urgent action to reduce the future extent of climate change (Evans et al. 2014) (Box 2.9.4).

Box 2.9.4 Health sector action on climate change

The health sector itself will need to take both adaptation and mitigation actions in direct response to climate change, over and above the general strengthening of public health activities and improving access to healthcare. In addition, the sector will need to adapt to the necessary reduction in the availability, and increase in price, of fossil fuels (and oil-based products such as plastics) that must occur in order to reduce greenhouse emissions substantially.

Adaptation actions will need to be shaped locally or regionally to respond to contextual variations in climate-related disease risk, available resources, and uncertainty (Confalonieri et al. 2015). For example, the US Resilient Hospital Dashboard (Four Twenty Seven 2016) uses localized climate and healthcare data to inform adaptation and resilience decisions.

However, some generalizable areas for consideration have been recently described in the World Bank's 2017 report on Climate Smart Healthcare. These encompass infrastructure, material resources, and supply chains, workforce planning, and resource allocation to particular services.

Decisions on location and design of new hospitals and healthcare facilities must incorporate local information about climate impacts, including flood risk, wind, and sea level rise. The Climate Smart Healthcare report includes detailed considerations for climate resilient design of healthcare facilities. Existing healthcare buildings and infrastructure (e.g. water and energy supplies) need to be assessed for their vulnerability to extreme weather and sea level rise, with responses ranging from alterations, adaptive usage plans, to planning for retreat and replacement (Wardekker et al. 2012). For example, as part of the PAHO/WHO Smart Hospitals initiative, Georgetown Hospital in St Vincent and the Grenadines has had roof repairs, solar panels installed, upgraded plumbing, water storage, and energy systems to improve its resilience to hurricanes in the Caribbean that are becoming increasingly severe, while at the same time, reducing energy consumption. In 2013, the newly refurbished hospital (Fig. 2.9.4) remained open when a severe storm closed most other health services. Georgetown Hospital continued to function, while also supplying clean rainwater to local community members (Pan American Health Organization & Regional Office for the Americas of the World Health Organization, n.d.).

Fig. 2.9.4 Georgetown Hospital, St Vincent, and the Grenadines (PAHO/WHO).

The need for increased service investment in public health early warning systems for heatwaves and other climate-related hazards has already been highlighted. Context-specific responses specifically related to infectious disease have also been developed (Confalonieri et al. 2015). These include enhanced disease surveillance, epidemic early warning systems, and new wider control efforts such as mosquito surveillance and standing water management which perhaps have not previously been needed. Changing patterns of disease may also require re-balancing of the healthcare workforce, as well as workforce re-training about new diseases.

Greater attention has been paid to reducing the health sector's own substantial greenhouse gas emissions. Although much of the focus has been in high-income countries, the Climate Smart Healthcare report points out the contribution that low-carbon changes can make to improving access to healthcare in low-income, energy-insecure settings. England, the United States, and Australia have all estimated the total emissions footprint of their healthcare sectors, with findings ranging from 3% to 10% of total national emissions (Malik et al. 2018). While the scope of these assessments varies, emissions come mainly from hospital energy use, the procurement of

material resources (particularly pharmaceuticals), and healthcare-related transport. Hospitals and health boards are often the largest employers and procurers in a region, and can therefore influence local government policy, large numbers of employees, and wider supply chains, putting them in positions of significant regional leadership.

In England, the National Health Service has a dedicated Sustainable Development Unit, whose task it is to reduce the overall environmental pollution from healthcare, with a focus on climate pollution. By taking a systematic approach to measuring emissions, setting long-term and intermediate targets, and taking action on energy, travel, reducing waste in procurement and influencing supply chains, emissions reductions have been achieved, coupled with financial savings of about GBP30 million per year (NHS Sustainable Development Unit 2016). Despite an 18% increase in inpatient admissions between 2007 and 2015, NHS England reduced its greenhouse gas emissions by 11%. Their ‘wedges’ approach demonstrates how within-health sector measures will need to be coupled with partnerships beyond the health sector to achieve required reductions (Fig. 2.9.5). The NHS England example also demonstrates the importance of the national regulatory context—in this case the UK Climate Change Act 2008 has set a consistent regulatory framework requiring all sectors to reduce their emissions by 80% by 2050 (Climate Change Act 2008 (c.27) 2008).

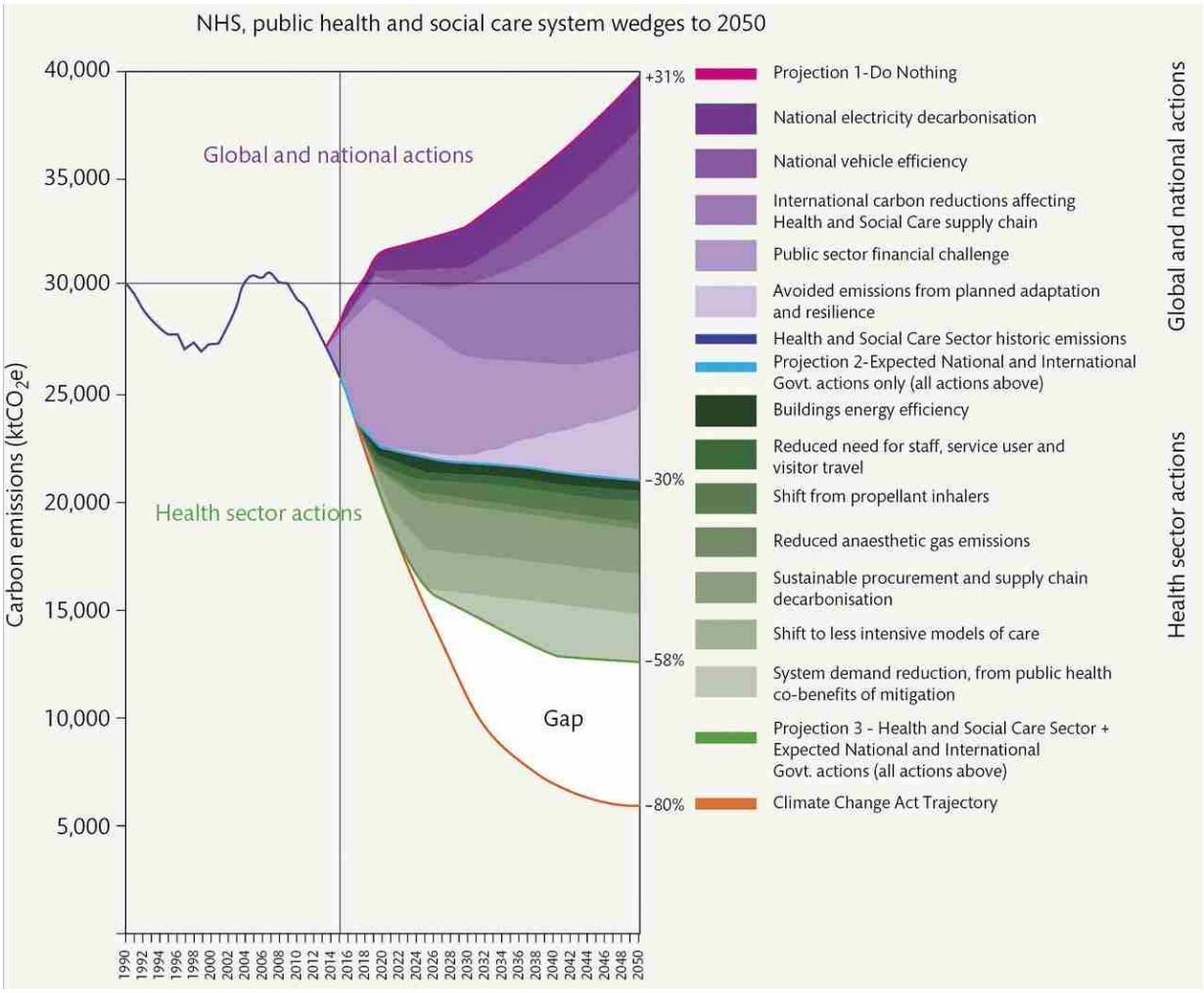


Fig. 2.9.5 NHS England, public health and social care system ‘wedges’ to 2050.
Reprinted from NHS Sustainable Development Unit. (2016). *Sustainable development in the health and care system. Health check 2016*, under the Open Government License v3.0.

Mitigation: scale, possible co-benefits, and unintended harms

To avoid dangerous climate change, greenhouse gas emissions must fall to net zero in the coming decades, requiring an industrial and energy revolution. In the past, energy revolutions have been disruptive. They have caused serious environmental damage, increased social inequalities, and shortened life expectancy. Are there paths to a low-carbon future that can achieve the opposite result—simultaneous improvements in the quality of the environment, public health, and health equity? We will review the

evidence for common solutions and discuss ways in which ‘co-benefits’ might influence policy settings for climate change.

If warming by the end of the century is to be contained to below 2°C, then CO_2 emissions must be reduced to around 14 Gt per year in 2050, less than a quarter of ‘business as usual’ projections. This is a tall order, and if reductions of such magnitude are going to be achieved, there will need to be radical changes in the way humans use energy across all sectors, including housing, transport, industry, and food and agriculture. It will require deeper social changes than merely identifying and bringing forward renewable technologies. Despite rapid growth in alternative energy sources, over the last 50 years each added unit of energy from wind, hydro, and other renewables has displaced only a quarter of a unit of energy obtained from fossil fuels and has had an even smaller impact on electricity generation (York 2012). In order to meet greenhouse targets, there must be serious deterrents to fossil fuel use, including prices that reflect the full environmental and social cost.

The 2010 World Development Report talked about changes that are so large they constitute ‘an energy revolution’. Revolutions may re-orient society in a way that is beneficial, in the long run, but they tend to be disruptive and harmful in the short term. The last energy revolution began in England in the nineteenth century and ushered in what has been in many respects a golden age of industrialization and improvements in life expectancy. But England, and every other country that went through the process, also experienced the ‘four Ds’—disruption, deprivation, disease, and death (Szreter 2004). In the large English cities in the 1840s, crowding, pollution, and social upheaval affected everyone, but particularly those at the bottom of the social hierarchy. Life expectancy at birth fell to medieval levels, less than 30 years among the working classes in many cities (Szreter and Mooney 1998).

The challenge with climate change mitigation will be to disrupt present patterns of energy use, while avoiding the very high and inequitable social and health costs that were carried by nineteenth-century cities. Early experience with alternative energy sources has not been altogether reassuring. One stark example is the encouragement of biofuels, which exemplifies the potential for health, equity, and environmental co-harms in a poorly designed low-carbon transition. There have been large-scale moves to use food and vegetable oil crops for bioethanol and biodiesel, often with

substantial government subsidies: in 2009 about a quarter of all US corn was diverted to biofuel production as a result of federal intervention. Their effectiveness for reducing greenhouse gas emissions is questionable, and their encouragement has increased competition for food staples such as soybeans, maize, and cereals—increasing biofuel demand accounted for between a quarter and a third of the rise in global grain prices in the early 2000s, and played a significant role in the 2006–2008 food price crisis (Malins 2017). It remains to be seen whether second-generation biofuels, extracted from non-edible plants such as jatropha that are grown on marginal lands and don't compete directly with food production, will bring the same energy benefits without the health risks.

It is possible to identify low-risk transitions to zero carbon that can enhance rather than worsen health and equity. One easy example is provided by the link between short-lived climate pollutants (SLCPs) and air quality. Black carbon (soot) particles, ozone, and methane have stronger heat-trapping properties than CO_2 on a weight-for-weight basis, but have much shorter atmospheric life spans (from days to decades) (Kerr 2013). SLCPs, especially black carbon and ozone, are also responsible for a large proportion of the 7 million deaths globally due to air pollution.

The United Nations Environment Programme estimated what could be achieved by widespread application of measures that are proven to reduce black carbon and methane (United Nations Environment Programme 2011). These measures include particle filters for disease engines, clean burning biomass cook stoves, and bans on the open field burning of agricultural waste. Assuming full implementation by 2030, the effects compared with 'business as usual' emissions amounted to saving 0.5°C of warming in 2050, while the reduction in local pollution would save about 50 million tonnes per year of lost crop yields, and avoid about 2.5 million premature deaths worldwide. The World Health Organization built on this work to identify priority policies with the best co-benefits for health and climate change mitigation through reductions in SLCPs. They identified four mitigation strategies with the greatest potential to effectively reduce SLCPs and achieve large health benefits: supporting active travel including through rapid mass public transport; promoting healthy plant-based diets; low-emission cook stoves and reducing household solid fuel use; and stricter vehicle emissions and efficiency standards, including for freight vehicles (World Health Organization 2015).

More broadly than just SLCPs and air pollution, there are potential health and equity co-benefits from well-designed climate action across all the major emissions sectors. Transport and housing are two sectors where health co-benefits have been well studied, but where there is also a risk of adverse consequences for health and/or health equity if interventions are not carefully designed. One example is the health and energy efficiency benefits of home insulation and efficient heating. In New Zealand, although the climate is moderate, many homes are relatively cold and damp owing to a lack of insulation, ineffective heating, and poor building design. A randomized community trial tested the effect of home insulation on indoor temperatures, energy consumption, and respiratory health (Howden-Chapman et al. 2007). Insulated homes were found to be warmer, drier, used 20% less electricity than the control homes, and children in them were sick less frequently (Table 2.9.1). The findings led subsequently to a national roll-out of insulation subsidies and installation.

However, alterations to improve the energy efficiency of housing may increase a range of other health risks. For example, there is a balance in houses between airtightness and ventilation that provides for efficient warming while allowing for adequate air flow to reduce indoor air pollution and prevent moisture from reaching levels conducive to mould growth (Shrubsole et al. 2014). The complex connections between housing, health, health equity, and energy efficiency require methods that allow for policies to be tested against their integrated impacts over time, as has been recently successfully tested for UK policies about housing (Macmillan et al. 2016) (Table 2.9.2).

Table 2.9.2 Outcomes of a randomized trial of insulation in 1350 New Zealand households. All households included one or more children with asthma

	Outcome
Average bedroom temperature in winter	Warmer by 0.5 degrees
Relative humidity in bedroom	Reduced by 2.3%
Energy consumption	81% of controls
Self-reported wheeze	OR 0.57 (0.47–0.70)
Days off school	OR 0.49 (0.31–0.80)
Visits to general practitioner	OR 0.73 (0.52–0.87)

OR, odds ratio.

Source: data from Howden-Chapman P et al., Effect of insulating existing houses on health inequality: cluster randomised study in the community, *British Medical Journal*, Volume 3, Issue 334, pp. 460, Copyright © 2007, DOI: 10.1136/bmj.39070.573032.80.

In many countries, emissions from the transport sector are growing more rapidly than those from any other sector. Yet there are many opportunities in transport policy to simultaneously cut carbon and improve health. This is particularly so in cities that are currently highly car-dependent. For example, in New Zealand about 80% of travel is by private car, and yet many vehicle trips, in the city, cover very short distances. In the New Zealand Household Travel Survey 2003–2006, it was reported that about three-quarters of car trips in urban areas were 7 km or less, and almost a third (31%) were 2 km or less. A system dynamics modelling study assessed the potential health and climate change co-benefits for investing in cycling infrastructure in New Zealand's largest city for a range of outcomes over a 40-year period (Macmillan et al. 2014). Even modest improvements in the provision for cycling in the city was estimated to have a cumulative net benefit of hundreds of millions of NZ dollars across air pollution, car crash injuries, physical inactivity, fuel costs to households, and greenhouse gas emissions, with the greatest savings through reductions in mortality due to physical inactivity. These were only marginally offset by an increase in cyclist injuries. With widespread best practice cycling infrastructure, much greater benefits were estimated, in the order of ten billion NZ dollars, through a

more substantial mode shift from car use to cycling, and greater improvements in the safety of cycling. Mortality benefits in this scenario outweighed injury deaths by more than 15 to 1.

A similar modelling exercise was undertaken in Delhi, India, comparing the health effects of different mitigation strategies (Woodcock et al. 2009). The city has set targets for greenhouse gas emission reduction, and this could be achieved by increasing the numbers of low-carbon vehicles, encouraging more active transport (walking and cycling), or a combination. The low-carbon vehicle option would avoid premature deaths and illness, mainly as a result of reduced air pollution in the city, but it was estimated these co-benefits were roughly ten times less than what could be achieved by the 'active transport' or 'combined' strategies.

In summary, global greenhouse gas emissions need to be reduced to about zero by 2050 to provide a reasonable chance of holding warming by the end of the century to 2°C or less. Radical mitigation of this kind might be seriously health damaging. But there are opportunities for co-benefits: well-chosen interventions could produce health gains (early), as well as long-term climate security. Although we have dealt with mitigation and adaptation separately in this chapter, it is clear that activities to adapt and reduce emissions need to be integrated, so that one does not undermine the other, and the potential for multisolving can be maximized. Despite our well-honed knowledge about climate science, substantial evidence for health impacts and potential for health and equity gains in action, there continues to be substantial inertia by governments, institutions, and organizations who are required to make urgent change. In the final section, we use a public health lens to reflect on obstacles to action on climate change, and why placing health at the centre of the discourse may hasten progress.

Answers to the carbon IQ quiz

Without greenhouse gases, the temperature at the Earth's surface would be below freezing (about 15°C). The global average is instead about +15°C. The answer is A.

Climate is long-range weather, over a timescale of decades. The British Meteorological Office, for example, uses 30 years as the averaging period, though in some recent assessments 20 years of weather has been considered long enough to define a climate. The answer is C.

C is closest to the mark: today's choices about the level of emissions will have long-term consequences (Stouffer 2012). AR4 says 'about 50% of a CO_2 increase will be removed from the atmosphere within 30 years, and a further 30% will be removed within a few centuries. The remaining 20% may stay in the atmosphere for many thousands of years' (IPCC 2007). There are many caveats, because the various processes that soak up carbon work at different speeds, and on a large scale, the lifetime of CO_2 depends on the amount emitted, and the rate of emission. But as a rough guide, about 40% is likely to be still in the atmosphere after 100 years, and just over 20% after 1,000 years.

Combustion of biomass releases a complex soup of pollutants, and reactions in the atmosphere generate 'secondary' agents such as ozone and other photochemical oxidants. The effects on climate are complex. Black carbon absorbs heat; sulphur aerosols have a reflective (cooling) effect. Because these agents are short-lived (remaining in the atmosphere for days or weeks), the impact is local/regional. Deposition of soot on ice causes increased absorption of solar energy, and accelerated melt. The answer is E.

The oceans absorb 80–90% of the solar energy that is trapped by greenhouse gases. Additional amounts of heat go towards melting glaciers and sea ice, as well as warming the land and parts of the atmosphere. Only a tiny fraction warms the air at the planet's surface.

As of early 2018, 2016 has the title of warmest year on record. Some have argued that the answer is 1260 (the period of the medieval warming) but this is not true.

Laos emits about 0.2 tonnes CO_2 per head per year. A 747 jet emits about 100 g of CO_2 per passenger kilometre (Carbon Independent 2015). Singapore to Rome return is roughly 10,000 km, which amounts to about 1 tonne CO_2 per passenger, based solely on the fuel consumed during the flight. Note that this does not allow for the greater warming effect of CO_2

released high in the atmosphere, nor the embedded fossil fuel costs of aviation, which include those incurred in building jets, running airports, and training staff. But on this basis the solution is 0.2 and the best answer is D.
A cyclist would travel 570 km on a litre of comparably energy dense food, which is 0.2 L/100 km (Wilson 2004). But this refers to a racing cyclist moving at 30 kph. Note that wind resistance falls away rapidly as speed drops—less than half the energy would be required for travel at an average of 20 kph (which means no more than 100 ml of oil-equivalent per 100 km).
George H.W. Bush made this commitment at the Rio Summit in 1992, after signing the UN Framework Convention on Climate Change. In the subsequent 20 years Republican leaders have shifted their ground, and increasingly they have denied that climate change is a serious threat (Oreskes and Conway 2010).
The Arctic is not cooling—temperatures are rising in this part of the world more rapidly than anywhere. The effects of climate change on rainfall are mixed—in some places conditions will be wetter than before, elsewhere precipitation will reduce. But overall the world will be a wetter place, with more intense rainfall than before, as a consequence of a warmer atmosphere carrying more moisture. The ocean will not become less acidic—increased uptake of CO_2 is expected to double the concentration of H+ ions by early in the twenty-first century. And the stratosphere is cooling, not warming, as a result of the thickening greenhouse blanket. The answer is D.

Impediments to action

More than 30 years have passed since scientists first pointed out the heating effects of rising greenhouse gas concentrations in the lower atmosphere. The United Nations Framework Convention on Climate Change was signed in 1992. Since then evidence of disruption and risk has accumulated, and the IPCC has reported, in detail, every 6 years or so on the impacts of climate change and the most effective means of adaptation and mitigation. Yet some governments appear to be more firmly opposed to taking action than their predecessors 25 years ago, and emissions continue to rise, worldwide. Why?

We suggest it is important to bear in mind the nature of the problem. Climate change results from two things, the success of a culture that equates progress with consumption and an economic model that ignores

environmental costs. The implications of climate science therefore are deeply threatening to political and commercial interests that are invested in the status quo. It is scarcely surprising that resistance has hardened over 30 years, as our knowledge about causes and consequences has increased. It is possible to tweak the present machinery to limit environmental damage without too much political disturbance (replacing petrol cars with electric cars or installing efficient lighting, for instance). But deep changes, of the kind needed to restrict warming to what has been agreed internationally, must be transformational. This is a much bigger challenge, technically, socially, and politically. There are of course other factors that make it difficult to progress more quickly in sustainable, low-carbon directions. These include development 'overhang' with a large fraction of the world's population requiring rapid access to energy and services; psychological responses to distant risks; doubt inflamed by powerful vested interests via media and politics; governance structures not well suited to long-term challenge; and modes of scientific research that privilege reductionist thinking.

But there are some grounds for optimism. The Paris Agreement was a milestone, internationally. Although it includes no legal sanctions to enforce emission reduction, and the independent nationally determined contributions (INDCs) that were promised by signatories are not enough on their own to hold warming to 2 degrees, the 2015 agreement is a necessary step in the right direction. There was for the first time international consensus on actions that need to be taken, and agreement to a process of regular review and reporting. And there are signs of effective action at local level also. For instance, cities, leaders in the public health response to the last energy revolution—the arrival of steam power and the Industrial Revolution, are once more taking a major role in innovation, collaboration, and the implantation of climate policy.

There is evidence that re-framing climate change as a health issue, as opposed to an 'environmental' issue, may increase societal willingness to act (Myers et al. 2012). Health professionals hold positions of privilege and influence, and in the past, when faced with big, difficult, threatening issues, have been thought-leaders and advocates. Consider tobacco, nuclear weapons, and the social determinants of health. It is time, we conclude, to add climate change to this list.

Conclusion

Climate change is a new kind of environmental problem—the scale of the phenomenon dwarfs the issues that public health workers have encountered previously. Its effects on health will result, mainly, from exacerbating disease and injury from familiar causes. Understanding how multiple variables will interact in the future to diminish or magnify the effects of climate change on health requires methods not currently familiar in public health. Complex models are needed that can incorporate non-linear interactions, feedbacks, and tipping points dynamically over time.

Climate change is undoubtedly an issue of health and social equity—those at greatest risk already carry a heavy load of disadvantage, and have contributed little to the CO_2 that now sits in the atmosphere. This means that many of the impacts can be avoided, or reduced in magnitude, by providing basic health improvements to the world's poor populations.

Targeted adaptation efforts, such as early warning systems and good disaster management, are necessary also: there is plenty of evidence that high-income countries are not immune from the effects on health of climate variability. But the world is not making rapid progress on either poverty reduction or targeted adaptation, so we expect the public health dimensions of climate change to become more prominent in the next few decades. In the medium to longer term (2050 and beyond), unless there are radical reductions in emissions beforehand, it is possible the average temperature of the globe will rise by more than 4°C over the pre-industrial baseline. If this happens, impacts of climate change may overwhelm all efforts at adaptation. For this reason, the response to climate change must be two-pronged—coping with adverse effects of rising temperatures and changing patterns of rainfall, and at the same time, finding ways of slashing emissions and reducing levels of climate active pollutants in the atmosphere. On the positive side, well-chosen adaptation and mitigation measures may achieve substantial health gains. Strengthening the place of public health in multidisciplinary decision-making about climate change can assist with identifying low-risk climate paths that enhance rather than worsen health and equity.

REFERENCES

4 Degrees and Beyond (2009). *Implications of a Global Climate Change of 4+ Degrees for People, Ecosystems and the Earth-System: abstract book.* Paper presented at the 4 Degrees and Beyond International Climate Conference, Oxford, UK.

Anderson, G.B. and Bell, M.L. (2011). Heat waves in the United States: mortality risk during heat waves and effect modification by heat wave characteristics in 43 U.S. communities. *Environmental Health Perspectives*, 119(2), 210–18.

Armstrong, B., et al. (2011). Association of mortality with high temperatures in a temperate climate: England and Wales. *Journal of Epidemiology and Community Health*, 65(4), 340–5.

Australian Academy of Sciences (2010). *The Science of Climate Change: Questions and Answers.* [Online] Available at: https://www.science.org.au/education/immunisation-climate-change-genetic-modification/science-climate-change

Baker-Austin, C., Trinanes, J., Gonzalez-Escalona, N., and Martinez-Urtaza, J. (2017). Non-cholera vibrios: the microbial barometer of climate change. *Trends in Microbiology*, 25(1), 76–84.

Baniassadi, A. and Sailor, D.J. (2018). Synergies and trade-offs between energy efficiency and resiliency to extreme heat—a case study. *Building and Environment*, **132**, 263–72.

Barriopedro, D., Fischer, E.M., Luterbacher, J., Trigo, R.M., and Garcia-Herrera, R. (2011). The hot summer of 2010: redrawing the temperature record map of Europe. *Science*, 332(6026), 220–4.

Bedno, S.A., et al. (2010). Exertional heat illness among overweight U.S. Army recruits in basic training. *Aviation, Space, and Environmental Medicine*, 81(2), 107–11.

Bell, M.L., et al. (2008). Vulnerability to heat-related mortality in Latin America: a case-crossover study in São Paulo, Brazil, Santiago, Chile and Mexico City, Mexico. *International Journal of Epidemiology*, 37(4), 796–804.

Berry, H., Bowen, K., and Kjellstrom, T. (2010). Climate change and mental health: a causal pathways framework. *International Journal of Public Health*, 55(2), 123–32.

Blake, E.S. and Gibney, E.J. (2011). The deadliest, costliest, and most intense United States tropical cyclones from 1851 to 2010 (NOAA Technical Memorandum NWS NHC-6). [Online] Available at: https://www.nhc.noaa.gov/pdf/nws-nhc-6.pdf

Bouwer, L.M. and Jonkman, S.N. (2018). Global mortality from storm surges is decreasing. *Environmental Research Letters*, 13(1), 014008.

Brikowski, T.H., Lotan, Y., and Pearle, M.S. (2008). Climate-related increase in the prevalence of urolithiasis in the United States. *Proceedings of the National Academy of Sciences of the United States*, 105(28), 9841–46.

Brunkard, J., Namulanda, G., and Ratard, R. (2008). Hurricane Katrina deaths, Louisiana, 2005. *Disaster Medicine and Public Health Preparedness*, 2(4), 215–23.

Carbon Independent (2015). Aviation sources. [Online] Available at: http://www.carbonindependent.org/sources_aviation.htm

Climate Central (2018). The 10 hottest global years on record. [Online] Available at: http://www.climatecentral.org/gallery/graphics/the-10-hottest-global-years-on-record

Climate Change Act (2008). (c.27), § S 1(1) (HM Government 2008).

Confalonieri, U.E.C., Menezes, J.A., and Souza, C.M.D. (2015). Climate change and adaptation of the health sector: the case of infectious diseases. *Virulence*, 6(6), 554–7.

Cunsolo, A. and Ellis, N. (2018). Ecological grief as a mental health response to climate change-related loss. *Nature Climate Change*, **8**, 275–81.

D'souza, M.R., Becker, G.N., Hall, B.A.G., and Moodie, B.A.K. (2004). Does ambient temperature affect foodborne disease? *Epidemiology*, 15(1), 86–92.

den Elzen, M.G.J., Olivier, J.G.J., Höhne, N., and Janssens-Maenhout, G. (2013). Countries' contributions to climate change: effect of accounting for all greenhouse gases, recent trends, basic needs and technological progress. *Climatic Change*, 121(2), 397–412.

Dow, K., Berkhout, F., Preston, B.L., Klein, R.J.T., Midgley, G., and Shaw, M.R. (2013). Limits to adaptation. *Nature Climate Change*, **3**, 305–7.
Duchêne, E. and Schneider, C. (2005). Grapevine and climatic changes: a glance at the situation in Alsace. *Agronomy for Sustainable Development*, 25(1), 93–9.
Ebi, K., Hartman, J., Chan, N., McConnell, J., Schlesinger, M., and Weyant, J. (2005). Climate Suitability for Stable Malaria Transmission in Zimbabwe Under Different Climate Change Scenarios. *Climatic Change*, 73(3), 375–93.
Ebi, K.L. and Mills, D. (2013). Winter mortality in a warming climate: a reassessment. *WIREs Climate Change*, 4(3), 203–212.
Ebi, K.L. and Nealon, J. (2016). Dengue in a changing climate. *Environmental Research*, **151**, 115–23.
Ebi, K.L., Ogden, N.H., Semenza, J.C., and Woodward, A. (2017). Detecting and attributing health burdens to climate change. *Environmental Health Perspectives*, 125(8), 85004.
Engelhard, G.H., Righton, D.A., and Pinnegar, J.K. (2014). Climate change and fishing: a century of shifting distribution in North Sea cod. *Global Change Biology*, 20(8), 2473–83.
Evans, L., Milfont, T.L., and Lawrence, J. (2014). Considering local adaptation increases willingness to mitigate. *Global Environmental Change*, 25(1), 69–75.
Fankhauser, S. and Stern, N. (2016). *Climate Change, Development, Poverty and Economics* (Centre for Climate Change Economics and Policy Working Paper No. 284). [Online] Available at: http://www.lse.ac.uk/GranthamInstitute/wp-content/uploads/2016/10/Working-Paper-253-Fankhauser-and-Stern.pdf
Field, C.B., Barros, V., Stocker, T.F., and Dahe, Q. (2012). *Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation: Special Report of the Intergovernmental Panel on Climate Change*. Cambridge, UK: Cambridge University Press.
Ford, J.D. (2012). Indigenous health and climate change. *American Journal of Public Health*, 102(7), 1260–6.
Fouillet, A., et al. (2008). Has the impact of heat waves on mortality changed in France since the European heat wave of summer 2003? A study of the 2006 heat wave. *International Journal of Epidemiology*, 37(2), 309–17.
Four Twenty Seven (2016). (10 February 2016). Resilient hospitals: using climate data for better healthcare planning. [Online] Available at: http://427mt.com/2016/02/10/resilient-hospitals-dashboard/
Fritz, H.M., Blount, C.D., Thwin, S., Thu, M.K., and Chan, N. (2009). Cyclone Nargis storm surge in Myanmar. *Nature Geoscience*, 2(7), 448–9.
Gething, P.W., Smith, D.L., Tatem, A.J., Snow, R.W., and Hay, S. (2010). Climate change and the global malaria recession. *Nature*, 465(7296), 342–5.
Gornall, J., et al. (2010). Implications of climate change for agricultural productivity in the early twenty-first century. *Philosophical Transactions of the Royal Society B*, 365(1554), 2973–89.
Hausfather, Z. and Pearce, R. (2017). Analysis: WRI data suggests emissions have already 'peaked' in 49 countries. [Online] Available at: https://www.carbonbrief.org/analysis-wri-data-suggests-emissions-have-already-peaked-49-countries
Hess, J., McDowell, J., and Luber, G. (2012). Integrating climate change adaptation into public health practice: using adaptive management to increase adaptive capacity and build resilience. *Environmental Health Perspectives*, 120(2), 171–9.
Howden-Chapman, P., et al. (2007). Effect of insulating existing houses on health inequality: cluster randomised study in the community. *British Medical Journal*, 334(7591), 460–4.
Huang, C., Barnett, A.G., Wang, X., and Tong, S. (2012). The impact of temperature on years of life lost in Brisbane, Australia. *Nature Climate Change*, 2(4), 265–70.
Ichinose, T.K., Inoue, Y., Hirata, M., Shamsuddin, A.K.M., and Kondo, N. (2009). Enhanced heat loss responses induced by short-term endurance training in exercising women. *Experimental Physiology*, 94(1), 90–102.

International Energy Agency (2018). *Global Energy & CO_2 Status Report 2017*. [Online] Available at: http://www.iea.org/publications/freepublications/publication/GECO2017.pdf

International Renewable Energy Agency (2018). *Renewable Power Generation Costs in 2017*. [Online] Available at: https://www.irena.org/-/media/Files/IRENA/Agency/Publication/2018/Jan/IRENA_2017_Power_Costs_2018_summary.pdf

IPCC (2007). *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*. [Online] Available at: https://www.ipcc.ch/report/ar4/wg1/

IPCC (2013). *Climate Change 2013 The Physical Science Basis: Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. [Online] Available at: https://www.ipcc.ch/report/ar5/wg1/

Kenny, G.P. and Jay, O. (2013). Thermometry, calorimetry, and mean body temperature during heat stress. *Comprehensive Physiology*, 3(4), 1689–719.

Kerr, R.A. (2013). Global warming: soot is warming the world even more than thought. *Science*, 339(6118), 382.

Kinney, P., et al. (2015). Winter season mortality: will climate warming bring benefits? *Environmental Research Letters*, 10(6), 064016.

Kjellstrom, T., Lemke, B., and Otto, M. (2013). Mapping occupational heat exposure and effects in South-East Asia: ongoing time trends 1980 & 2011 and future estimates to 2050. *Industrial Health*, 51(1), 56–67.

Klinenberg, E. (2002). *Heat Wave: A Social Autopsy of Disaster in Chicago*. Chicago: University of Chicago Press.

Kolstad, E.W., and Johansson, K.A. (2011). Uncertainties associated with quantifying climate change impacts on human health: a case study for diarrhea. *Environmental Health Perspectives*, 119(3), 299–305.

Krieger, N. (2015). The real ecological fallacy: epidemiology and global climate change. *Journal of Epidemiology & Community Health*, *69*, 803–804.

Lenton, T., and Ciscar, J.-C. (2013). Integrating tipping points into climate impact assessments. *Climatic Change*, 117(3), 585–97.

Lindgren, E., and Gustafson, R. (2001). Tick-borne encephalitis in Sweden and climate change. *The Lancet*, 358(9275), 16–18.

Liu, J., Curry, J.A., Wang, H., Song, M., and Horton, R.M. (2012). Impact of declining Arctic sea ice on winter snowfall. *Proceedings of the National Academy of Sciences of the United States of America*, 109(11), 4074–79.

Lizarralde, G., et al. (2015). A systems approach to resilience in the built environment: the case of Cuba. *Disasters*, 39(s1), s76–s95.

Lucas, R.A.I., Epstein, Y., and Kjellstrom, T. (2014). Excessive occupational heat exposure: a significant ergonomic challenge and health risk for current and future workers. *Extreme Physiology & Medicine*, 3(14), 1–8.

Lutz, W., Sanderson, W., and Scherbov, S. (2008). The coming acceleration of global population ageing. *Nature*, 451(7179), 716–19.

Macmillan, A., Connor, J., Witten, K., Kearns, R., Rees, D., and Woodward, A. (2014). The societal costs and benefits of commuter bicycling: simulating the effects of specific policies using system dynamics modeling. *Environmental Health Perspectives*, 122(4), 335–44.

Macmillan, A., et al. (2016). Integrated decision-making about housing, energy and wellbeing: a qualitative system dynamics model. *Environmental Health: A Global Access Science Source*, 15(43), S37.

Madrigano, J., Ito, K., Johnson, S., Kinney, P.L., and Matte, T. (2015). A case-only study of vulnerability to heat wave-related mortality in New York City (2000–2011). *Environmental*

Health Perspectives, 123(7), 672–8.
Malik, A., Lenzen, M., McAlister, S., and McGain, F. (2018). The carbon footprint of Australian health care. *The Lancet Planetary Health*, 2(1), e27–e35.
Malins, C. (2017). *Thought for Food: A Review of the Interaction Between Biofuel Consumption and Food Markets*. [Online] Available at: https://www.transportenvironment.org/sites/te/files/publications/Cerulogy_Thought-for-food_September2017.pdf
Maloney, S., and Forbes, C. (2011). What effect will a few degrees of climate change have on human heat balance? Implications for human activity. *International Journal of Biometeorology*, 55(2), 147–60.
Manson, P. (1900). *Tropical Diseases. A Manual of the Diseases of Warm Climates*. London: Cassell.
McLaughlin, J.B., et al. (2005). Outbreak of Vibrio parahaemolyticus Gastroenteritis Associated with Alaskan Oysters. *The New England Journal of Medicine*, 353(14), 1463–70.
McMichael, A.J. (2017). *Climate Change and the Health of Nations: Famines, Fevers, and the Fate of Populations*. New York: Oxford University Press.
McMichael, A.J., et al. (2008). International study of temperature, heat and urban mortality: the 'ISOTHURM' project. *International Journal of Epidemiology*, 37(5), 1121–31.
Meehl, G.A., Tebaldi, C., Walton, G., Easterling, D., and McDaniel, L. (2009). Relative increase of record high maximum temperatures compared to record low minimum temperatures in the U.S. *Geophysical Research Letters*, 36(23), L23701.
Myers, T.A., Nisbet, M.C., Maibach, E.W., and Leiserowitz, A.A. (2012). A public health frame arouses hopeful emotions about climate change. *Climatic Change*, *113*(3–4), 1105–12.
NHS Sustainable Development Unit (2016). *Sustainable Development in the Health and Care System. Health Check 2016*. [Online] Available at: https://www.sduhealth.org.uk/search/resources.aspx?q=health+check+2016&zoom_query=health+check+2016
Oreskes, N., and Conway, E.M. (2010). *Merchants of Doubt*. New York: Bloomsbury Press.
Ostro, B., Rauch, S., Green, R., Malig, B., and Basu, R. (2010). The effects of temperature and use of air conditioning on hospitalizations. *American Journal of Epidemiology*, 172(9), 1053–61.
Pan American Health Organization, & Regional Office for the Americas of the World Health Organization (n.d.). Transforming health systems to tackle climate change. Climate change and health: Regional Office for the Americas of the World Health Organization. [Online] Available at: https://www.paho.org/salud-en-las-americas-2017/?p=53
Pearce, N. (2011). Epidemiology in a changing world: variation, causation and ubiquitous risk factors. *International Journal of Epidemiology*, 40(2), 503–12.
Peters, G.P., et al. (2011). Rapid growth in CO_2 emissions after the 2008–2009 global financial crisis. *Nature Climate Change*, 2(1), 2–4.
Phelan, P.E., et al. (2015). Urban heat island: mechanisms, implications, and possible remedies. *Annual Review of Environment and Resources*, 40(1), 285–307.
Post, E.S., et al. (2012). Variation in estimated ozone-related health impacts of climate change due to modeling choices and assumptions. *Environmental Health Perspectives*, 120(11), 1559–64.
Qin, R.X., et al. (2017). The interactive effects between high temperature and air pollution on mortality: a time-series analysis in Hefei, China. *Science of the Total Environment, 575*, 1530–7.
Richardson, E.A., et al. (2012) Green cities and health: a question of scale? *JECH, 66*, 160–5.
Rocklöv, J., Barnett, A.G., and Woodward, A. (2012). On the estimation of heat-intensity and heat-duration effects in time series models of temperature-related mortality in Stockholm, Sweden. *Environmental Health: A Global Access Science Source*, 11(1), 1–12.
Rutter, H., et al. (2017). The need for a complex systems model of evidence for public health. *The Lancet*, 390(10112), 2602–4.

Salmond, J.A., et al. (2016). Health and climate related ecosystem services provided by street trees in the urban environment. *Environmental Health: A Global Access Science Source*, 15(43), S36.
Schreiber, K.V. (2001). An investigation of relationships between climate and dengue using a water budgeting technique. *International Journal of Biometeorology*, 45(2), 81–9.
Shrubsole, C., Macmillan, A., Davies, M., May, N., and Ucci, M. (2014). 100 Unintended consequences of policies to improve the energy efficiency of the UK housing stock. *Indoor and Built Environment*, 23(3), 340–52.
Silva, H., Phelan, P., and Golden, J. (2010). Modeling effects of urban heat island mitigation strategies on heat-related morbidity: a case study for Phoenix, Arizona, USA. *International Journal of Biometeorology*, 54(1), 13–22.
Speldewinde, P.C., Cook, A., Davies, P., and Weinstein, P. (2009). A relationship between environmental degradation and mental health in rural Western Australia. *Health and Place*, 15(3), 880–7.
Springmann, M., et al. (2016). Global and regional health effects of future food production under climate change: a modelling study. *The Lancet*, 387(10031), 1937–46.
Steffen, W., et al. (2015). Sustainability. Planetary boundaries: guiding human development on a changing planet. *Science*, *347*(6223), 1259855.
Stenseth, N.C., et al. (2006). Plague dynamics are driven by climate variation. *Proceedings of the National Academy of Sciences of the United States*, 103(35), 13110–15.
Stott, P.A., Stone, D.A., and Allen, M.R. (2004). Human contribution to the European heatwave of 2003. *Nature*, 432(7017), 610–14.
Stouffer, R.J. (2012). Oceanography: future impact of today's choices. *Nature Climate Change*, 2(6), 397–8.
Subramanyam, M.A., Kawachi, I., Berkman, L.F., and Subramanian, S.V.V. (2011). Is economic growth associated with reduction in child undernutrition in India? *PloS Medicine*, 8(3), e1000424.
Sumilo, D., Asokliene, L., Bormane, A., Vasilenko, V., Golovljova, I., and Randolph, S. E. (2007). Climate change cannot explain the upsurge of tick-borne encephalitis in the baltics (tick-borne disease upsurge). *PLoS One*, 2(6), e500.
Szreter, S. (2004). Industrialization and health. *British Medical Bulletin*, 69(1), 75–86.
Szreter, S. and Mooney, G. (1998). Urbanization, mortality, and the standard of living debate: new estimates of the expectation of life at birth in nineteenth-century British cities. *Economic History Review*, 51(1), 84–112.
Teague, B., McLeod, R., and Pascoe, S. (2010). *Victorian Bushfires Royal Commission. Parliament of Victoria*. [Online] Available at: http://royalcommission.vic.gov.au/Commission-Reports/Final-Report.html
Teixeira, E.I., Fischer, G., van Velthuizen, H., Walter, C., and Ewert, F. (2011). Global hot-spots of heat stress on agricultural crops due to climate change. *Agricultural and Forest Meteorology*, **170**, 206–15.
Tester, P.A., Feldman, R.L., Nau, A.W., Kibler, S.R., and Wayne Litaker, R. (2010). Ciguatera fish poisoning and sea surface temperatures in the Caribbean Sea and the West Indies. *Toxicon*, 56(5), 698–710.
Thompson, M. and Gaviria, I. (2004). *Weathering the Storm: Lessons in Risk Reduction from Cuba. An Oxfam America Report*. [Online] Available at: http://www.eird.org/isdr-biblio/PDF/Cuba%20Weathering.pdf
Thornton, P.K., Jones, P.G., Ericksen, P.J., and Challinor, A.J. (2011). Agriculture and food systems in sub-Saharan Africa in a 4°C+ world. *Philosophical Transactions of the Royal Society A*, 369(1934), 117–36.
Tun-Lin, W., Burkot, T.R., and Kay, B.H. (2000). Effects of temperature and larval diet on development rates and survival of the dengue vector Aedes aegypti in north Queensland, Australia. *Medical and Veterinary Entomology*, 14(1), 31–7.

Uejio, C.K., Wilhelmi, O.V., Golden, J.S., Mills, D.M., Gulino, S.P., and Samenow, J.P. (2011). Intra-urban societal vulnerability to extreme heat: the role of heat exposure and the built environment, socioeconomics, and neighborhood stability. *Health and Place*, 17(2), 498–507.

United Nations (2015a). *Paris Agreement*. [Online] Available at: https://unfccc.int/process-and-meetings/the-paris-agreement/the-paris-agreement

United Nations (2015b). World population projected to reach 9.7 billion by 2050. [Online] Available at: http://www.un.org/en/development/desa/news/population/2015-report.html

United Nations Environment Programme (2011). *Integrated Assessment of Black Carbon and Tropospheric Ozone. Summary for Decision Makers*. [Online] Available at: https://wedocs.unep.org/rest/bitstreams/12809/retrieve

Wandiga, S., et al. (2010). Vulnerability to epidemic malaria in the highlands of Lake Victoria basin: the role of climate change/variability, hydrology and socio-economic factors. *Climatic Change*, 99(3), 473–97.

Wardekker, J.A., de Jong, A., van Bree, L., Turkenburg, W.C., and van der Sluijs, J.P. (2012). Health risks of climate change: an assessment of uncertainties and its implications for adaptation policies. *Environmental Health*, 11(1), 67.

Wilson, D.G. (2004). *Bicycling Science*. Cambridge, MA: MIT Press.

Woodcock, J., et al. (2009). Public health benefits of strategies to reduce greenhouse-gas emissions: urban land transport. *The Lancet*, 374(9705), 1930–43.

Woodward, A., and Porter, J.R. (2016). Food, hunger, health, and climate change. *The Lancet*, 387(10031), 1886–7.

Woodward, A.J., and Samet, J.M. (2018). Climate change, hurricanes, and health. *American Journal of Public Health*, 108(1), 33–5.

World Bank (2009). *World Development Report 2010. Development and Climate Change*. [Online] Available at: https://openknowledge.worldbank.org/handle/10986/4387?locale-attribute=fr

World Health Organization (2008). *The Global Burden of Disease*. [Online] Available at: https://www.who.int/healthinfo/global_burden_disease/en/

World Health Organization (2011). *World Health Statistics 2011*. [Online] Available at: https://www.who.int/gho/publications/world_health_statistics/WHS2011_IndicatorCompendium.pdf?ua=1

World Health Organization (2015). *Reducing Global Health Risks Through Mitigation of Short-Lived Climate Pollutants: Scoping Report for Policy-makers*. [Online] Available at: http://apps.who.int/iris/bitstream/handle/10665/189524/9789241565080_eng.pdf?sequence=1

Wu, P.-C., Lay, J.-G., Guo, H.-R., Lin, C.-Y., Lung, S.-C., and Su, H.-J. (2009). Higher temperature and urbanization affect the spatial patterns of dengue fever transmission in subtropical Taiwan. *Science of the Total Environment*, 407(7), 2224–33.

Xu, L., et al. (2011). Nonlinear effect of climate on plague during the third pandemic in China. *Proceedings of the National Academy of Sciences of the United States*, 108(25), 10214–19.

Xun, W., Khan, A., Michael, E., and Vineis, P. (2010). Climate change epidemiology: methodological challenges. *International Journal of Public Health*, 55(2), 85–96.

Yang, C., Cai, T., and Zhou, H. (2007). Impact of low temperature on population death in a typical city of North China. *Chinese Journal of Public Health Engineering*, **2**, 77–9.

York, R. (2012). Do alternative energy sources displace fossil fuels? *Nature Climate Change*, 2(6), 441–3.

You, D., et al. (2015). Global, regional, and national levels and trends in under-5 mortality between 1990 and 2015, with scenario-based projections to 2030: a systematic analysis by the UN Inter-agency Group for Child Mortality Estimation. *The Lancet*, 386(10010), 2275–86.

Zanobetti, A., Neill, M.S., Gronlund, C.J., and Schwartz, J.D. (2012). Summer temperature variability and long-term survival among elderly people with chronic disease. *Proceedings of the National Academy of Sciences of the United States of America*, 109(17), 6608–13.

2.10

Behavioural determinants of health and disease

Lawrence W. Green, Kristin S. Hoeft, and Robert A. Hiatt

Scope of behavioural determinants of health and disease

Behaviour, and its more robust expressions, ‘lifestyle’, cultural and social ‘norms’ are widely implicated in disease incidence and prevalence. That behaviour is associated with health and disease has never been in doubt. Indeed, the tendency to blame sinful, negligent, indulgent, ignorant, reckless, slothful, or selfish behaviour for health problems has too often placed undue emphasis on individual responsibility and culpability when the solution to health problems of populations demanded attention to the social and physical environment. But no matter how behaviour is framed or moralized in relation to its causes, it remains an inescapable variable in the pathway between ultimate upstream determinants and the incidence or prevalence of most diseases and health conditions downstream. Approaches to public health have sought throughout the history of civilization to: (1) control or cajole the health-related behaviour of individuals; (2) protect individuals from the behaviour of others; and (3) mobilize the behaviour of groups to influence a population’s health-related understanding, beliefs, attitudes, values, policies, organizations, and social or physical environments.

This chapter reviews ways in which behaviour relates to the spectrum of health and disease determinants, from environmental to genetic, in shaping health outcomes. It builds on the previous chapters in recognizing the powerful influence of socioeconomic and cultural factors, especially poverty and discrimination, in influencing both behaviour and health. Many commentaries in the past four decades have attempted to correct the overemphasis on individual behavioural determinants of health by discounting and sometimes disparaging any focus on individual behaviour or personal responsibility in disease prevention and health promotion, often

referring to it as ‘victim blaming’ (e.g. Steele et al. 2014), and on another side, ‘scapegoating’ people with health problems as the cause of rising costs of healthcare (e.g. Minkler 1983). This chapter seeks a middle ground, building on the growing understanding of the ecological and cultural context of the behaviour–health relationship. It seeks to integrate that knowledge in an approach to public health that acknowledges the reciprocal determinism of behavioural, environmental, and biological determinants rather than minimizing the importance of behaviour in these complex interactions. This is consistent with a concept of ‘behavioural justice’ that Adler and Stewart (2009, p. 49) refer to as a ‘principle that individuals are responsible for engaging in health-promoting behaviours but should be held accountable only when they have adequate resources to do so. This perspective maintains both individuals’ control and accountability for behaviours and society’s responsibility to provide health-promoting environments’.

The role of behaviour as a determinant of health and disease: from discrete to complex, from individual to society

Behaviour is defined as actions by individuals or groups, and occur everywhere throughout human society—from an individual action that affects only oneself, to behaviours carried out by healthcare providers, to behaviours and decisions that are enacted at social, environment and policy levels that have widespread impact on health directly and indirectly through shaping individual behaviour and policies. While recognizing that socioeconomic status is a very strong influence on health behaviour, we first we discuss health behaviour more broadly, and then discuss socioeconomic status in relation to health behaviour in detail in the later sections of the chapter.

Some health behaviours are single, discrete actions that can have lasting effects for health (e.g. getting a vaccine, or a mosquito net to reduce malaria exposure) or injury (e.g. texting while driving). However much of today’s growing chronic disease burden relates more to complex repeated behaviours, sometimes referred to as ‘habit’ or ‘lifestyle’. We use the term ‘complex behaviour’ to refer to combinations of interrelated practices and their environmental contexts, reflecting patterns of living influenced by the family and social history of individuals and communities, their

environmental and socioeconomic circumstances, and their exposure to cultures and communications.

Public health practitioners have to think about behavioural determinants of health and disease while also considering how to encourage behaviours that improve health and discourage those that threaten it, through public health programmes, policies, or interventions. We know that discrete behaviours can be more easily influenced directly by health education targeted at individuals and groups. Complex behaviour, however, changes more slowly and usually requires some combination of educational, organizational, economic, and environmental interventions in support of changes in both behaviour and conditions of living. This combination of strategies has defined health promotion and public health programmes addressing complex behaviour change through combined efforts to predispose, enable, and reinforce the behaviours (Swan et al. 2012; Green et al. 2019). The predisposing interventions are directed particularly at the individual, but the enabling and reinforcing interventions are directed at policies in support of, and environments in which, predisposed individuals would seek to enact and maintain the behaviour.

HIV/AIDS is a disease that relies for prevention on complex behaviours, while awaiting discrete solutions (e.g. vaccine). Virtually every public health breakthrough has had a behavioural change process that served the public until the technology was at hand. Then, behavioural change processes were needed to diffuse, adapt, and apply the new technology to varying cultural and social circumstances. Until an HIV vaccine is developed, society must depend on behavioural preventive measures to curb its spread. These include, of course, policies, environmental changes, and health educational programmes that support behavioural changes (Krishnaratne et al. 2016).

Much of the early success in controlling HIV infections through change in sexual practices (especially use of condoms) among men in urban gay communities appears to have been in response to health education programmes (Petrow 1990). Reviews also show increases in the use of clean needles for at least 15 years among intravenous drug users (e.g. Wodak 2006), which has required a combination of policy and educational interventions to make clean needles accessible and more acceptable than the culture of needle sharing. An HIV outbreak in rural Indiana attributed to needle sharing among injecting drug users underscored the importance of continued reinforcement of known best practices (Conrad et al. 2015).

Medical advances such as antiretroviral therapy (ART) and pre-exposure prophylaxis (PrEP) for HIV prevention still involve complex behaviours that require daily regimens for optimum efficacy (Wilton et al. 2015; Krishnaratne et al. 2016). Attempts to minimize or circumvent the need for behaviour change, as with adult male circumcision in Africa, present additional challenges of bush surgery complications, and false assumptions that one may now ignore the admonitions about safe sex (Green et al. 2010).

Specific behaviours and health—the causal links

When identifying behavioural risk and protective factors for a given health condition, one must evaluate the scientific evidence supporting causal links between behaviours and health outcomes. The shorter and more direct (proximal) the link is, the easier it is to identify. Some behaviour clearly increases the risk of developing disease and can be considered a proximal cause of disease, such as unhygienic practices that expose one to infections. Examining the correlations between specific behavioural and social patterns and indicators of health and disease status provides the foundation for assessing behavioural factors as population health determinants (Evans et al. 1994). Examining the covariation of these relationships with other characteristics of the populations and their environmental circumstances helps put behaviour into its ecological, social, and cultural context (Institute of Medicine 2012).

The causal link is relatively easy to establish for single-agent communicable or infectious diseases. Evidence from observational epidemiological studies, human experimental trials, and animal models, together with clear mechanisms of biological action, lead one to conclude that many behaviours are, in fact, contributing causes (causal risk factors) of specific diseases. Again, the easiest examples of clear causal linkages are those established for single-action, discrete behaviours such as ingesting a contaminated food, getting an immunization that confers lifetime immunity, or taking a one-dose medication that leads to rapidly improved symptoms. As the number of booster shots required for immunization or doses required for cure increases, for example, the behaviour becomes more complex, requiring repetition or timing, and the causal linkage more difficult to unravel among the biomedical agent, the behaviour, and all the other events

and circumstances that might have intervened and influenced the person's exposure, susceptibility, and outcome along the way.

Complex behaviours and health—the synergistic causal pathways

Other, less discrete, behaviours correlate with and precede better health, increased longevity, and decreased disease risk, but their causal link is more tenuous, variable, delayed, cumulative, or indirect, warranting their inclusion with more distal determinants. Examples include the development of dietary and physical activity patterns. More difficult still are causal pathways that are long and indirect, such as long-term behavioural patterns that are not under medical supervision and relate to multiple-cause chronic diseases and conditions (Krieger 1984; Glass and McAtee 2006). We present two examples of evidence supporting causal links between complex behaviours and coronary heart disease: smoking and physical inactivity. These examples illustrate that even in the absence of direct experimental evidence in human beings, strong evidence of other types can be combined for the steps in a causal chain from behaviour through physiological effects to disease or health. In addition, the synergistic effects of two or more behaviours on health outcomes have been established.

A plausible biological model of the relationship between smoking behaviour and coronary heart disease has been available for decades (Dawber 1960). Observational epidemiological studies—including within- and between-population designs, case–control, and prospective designs—produced strong and consistent measures of association, the hypothesized temporal sequence, and dose–response relationships in subsequent decades (Stamler 2005). Additional randomized trials have included smoking cessation programmes that provide experimental evidence for smoking as a cause of coronary heart disease. Although the overall results of the Multiple Risk Factor Intervention Trial (MRFIT) in the United States were disappointing, in both MRFIT (Ockene et al. 1990) and the British trial on the effect of smoking reduction (Rose and Colwell 1992), cessation interventions showed decreases in coronary heart disease mortality respectively of 13% after 20 years and 12% after 10.5 years. In addition, when smokers at baseline from the experimental and control groups were pooled, quitters had a significant decrease in their risk of mortality from

coronary heart disease compared with non-quitters (Ockene et al. 1990). As the smoking landscape shifts with development of electronic cigarettes and heat-not-burn tobacco products, researchers and regulators are striving to determine the individual and public health risks associated with the new products (Levy et al. 2017).

Physical inactivity as a risk factor for coronary heart disease is the second example. Evidence that physical inactivity contributes to coronary heart disease came from studies of the biological effects of exercise on cardiovascular physiology, observational epidemiological studies, and randomized controlled trials of physical activity and physiological coronary heart disease risk factors, such as high blood pressure, obesity, and diabetes. The biological effects of exercise training to enhance cardiovascular health were well-established by the 1980s (McArdle et al. 1986). Epidemiological evidence continues to show consistent and relatively strong associations, the hypothesized temporal sequence, and a dose–response relationship between physical activity and coronary heart disease (Ahmed et al. 2012). Observational epidemiological studies and randomized controlled trials demonstrated the beneficial effects of physical activity on blood pressure (Arroll and Beaglehole 1992) and on blood lipids and lipoproteins (Lokey and Tran 1989), which have, in turn, been causally linked with subsequent coronary heart disease. The combination of these sources of evidence for each of the steps in a causal chain provides a plausible model for the sequence.

Some of the immediate causal risk factors are not themselves behaviours, but have determinants that are behaviours. In these cases, the behavioural determinants can be regarded as indirect risk factors that act earlier in the causal pathway. For example, a combination of overeating and sedentary behaviour produces high caloric intake and low energy output, which together mediate the behavioural determinants of obesity. Obesity, in turn, is a physiological risk factor for cardiovascular diseases (CVD) and type 2 diabetes (Institute of Medicine 2012).

Behaviour also contributes to the prognosis of diseases at each stage where the screening, diagnosis, or the compliance with prescribed regimens of treatment or self-care affects outcomes. For example, the prognosis of breast cancer depends on the stage of disease at which the person obtains screening, diagnosis, and medical care. The prognosis for type 1 (insulin-dependent) diabetes depends on the patient’s compliance with his or her

insulin prescriptions. Because behaviour is so central to disease outcomes, a large literature on patient education and patient compliance with medical regimens has been catalogued and subjected to meta-analyses with regularity (e.g. US Preventive Services Task Force 1989 and continuously updated on https://www.uspreventiveservicestaskforce.org/BrowseRec/Index/browse-recommendations).

Behavioural risk factors in population health

The causal pathways by which behaviour can influence health (or its negative manifestations) can be broadly classified as direct (as shown by arrow 1 in Fig. 2.10.1) and indirect, where the indirect pathways (arrows 2 and 3 in Fig. 2.10.1) are largely mediated through the environment or through healthcare organization and personnel. These three determinants of health—behaviour, environment, and healthcare organization—in addition to human biology, were cast as the 'Health Field Concept' as part of the Lalonde Report on the Health of Canadians (Canada 1974), which many credit with having launched or at least facilitated the health promotion, health determinants and population health movements in public health (cf. Green and Allegrante 2011). The third indirect pathway could be drawn through genetics, the main pathway by which individuals can make some reproductive decisions partially independent of the medical care system, but such non-medically mediated decisions are largely mediated by the social environment, and increasingly involve the medical care system through genetic counselling. Figure 2.10.1 is useful guide, however it is hardly a full representation of the three determinants much less the genetic interactions, insofar as there is much reciprocal determinism between and among them. These environmental influences on behaviour will be examined in the next section, and behaviour's influence on the environment in a later section. Genetics are addressed in Chapter 8.19.

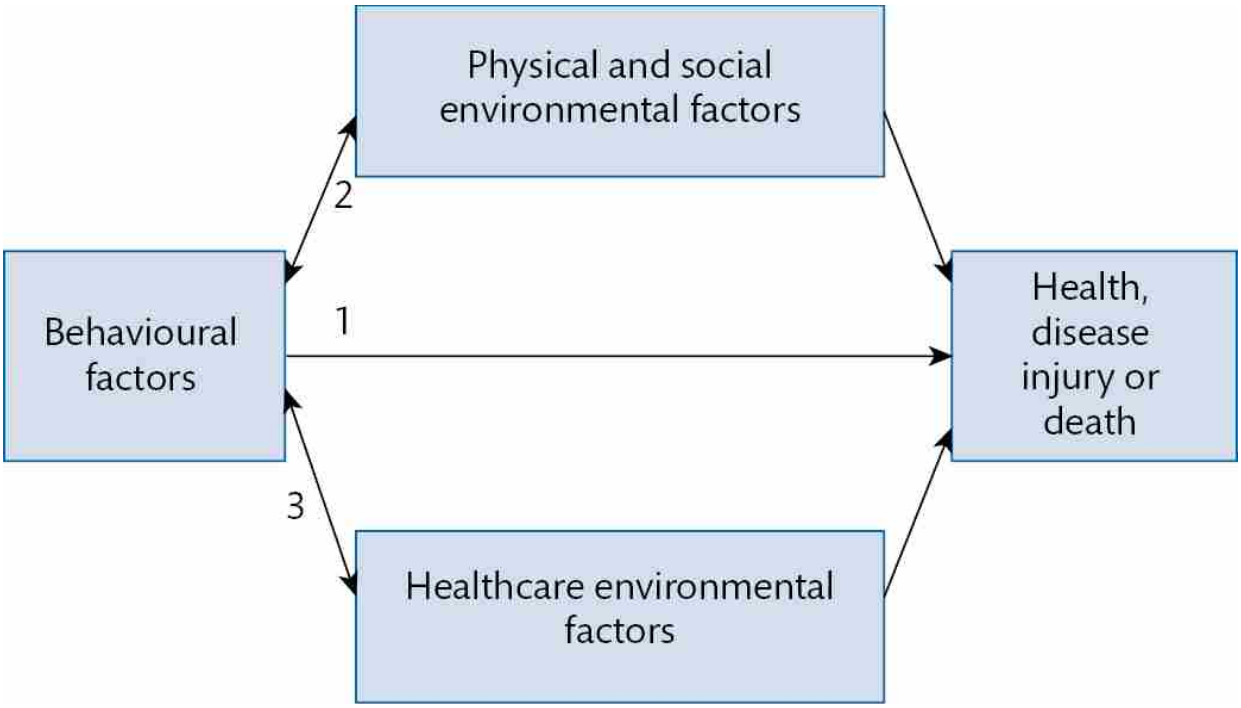


Fig. 2.10.1 Three pathways for behavioural influence on health, disease, injury, or death.

Behaviour itself as a risk factor for disease

The direct pathway suggested by arrow 1 in Fig. 2.10.1 includes the broad array of actions people take, consciously or unconsciously, that can have an immediate or cumulative effect on their health status. The effect on health may be intended (health-directed) or unintended (health-related), but the behaviour is nevertheless direct in its effect. The most dramatic of these are the violent injury-causing actions people may take behind the wheel of an automobile, with weapons, or unintentionally with the careless use of tools or toxic substances or merely walking absent-mindedly on a slippery or cluttered surface or while reading their cell phone screen. Less dramatic, but no less lethal, are the cumulative little actions people take each time they light a cigarette, imbibe or inject an addictive or mind-altering substance, or neglect physical activity or healthful foods.

Table 2.10.1 lists the five leading causes of death in most of the more developed countries, as reflected here by the United States, and relates those to the risk factors for each in column 2, which McGinnis and Foege (1993, 2004) and Mokdad et al. (2004) refer to as the ‘actual causes of death’. Colditz (2014) has noted that more than half of all cancer is a consequence

of behaviour, much of which is embedded in culture. Most of the risk factors in Table 2.10.1 are themselves behaviours. Those that are not behaviours themselves—such as high serum cholesterol, obesity, hypertension, and diabetes mellitus—are the physiological consequence, in large part, of behaviours. The behavioural determinants of risk factors in column 3 present some of the most challenging targets for public health intervention, because they require decreasing behaviours that are perceived as pleasurable (e.g. types of food, sedentary entertainment, casual or spontaneous sexual encounters) and increasing behaviours perceived as boring, painful, or inconvenient (e.g. less sugary or salty foods, more strenuous or frequent physical activity, use of condoms). Their challenges are compounded by the fact that most of the instances of these behavioural determinants of risk can be carried out very privately with limited opportunity for monitoring, enforcement, or social influence.

Tobacco consumption alone accounts for a large proportion of deaths, implicated as a direct risk factor in four of the five of the leading causes of death shown in Table 2.10.1. At least one of the three main behavioural risk factors—smoking, dietary practices, and alcohol use—is causally related to each of the leading causes of death shown in Table 2.10.1. Active smoking has been established as a risk factor for lung and many other types of cancer, coronary heart disease, diabetes, stroke, and adverse pregnancy outcomes, such as low birth weight, premature rupture of membranes, and abruption placentae (US Department of Health and Human Services 2004). As examples of how the behaviour of some individuals can influence the health of others, passive smoking (i.e. exposure to environmental tobacco smoke) has been related to lung and breast cancer and other respiratory diseases, as an independent risk factor for coronary heart disease, and exposure to environmental tobacco smoke in the home has been associated with asthma, other respiratory conditions, and with ear infections in infants and children (US Department of Health and Human Services 2006; Johnson et al. 2014).

Table 2.10.1 Five leading causes of death (United States 2015), their generally accepted risk factors, and their behavioural determinants

Causes of death (age-adjusted death rates in the United States per 100,000 population)	Selected risk factors	Behavioural determinants of risk factors
Diseases of the heart (168.5)	Tobacco smoke	Smoking
	Sedentary living	Physical inactivity
	High serum cholesterol	High-fat diet, medication
	Obesity, diabetes mellitus	Diet, physical inactivity
	Hypertension	High-salt diet, medication
Cancer (158.5)	Smoking	Smoking
	Low-fibre diet	Low-fibre diet
	Physical inactivity	Physical inactivity
	Sexually transmitted infections	Sexual behaviours, HPV vaccination compliance
	Obesity	Diet, physical activity
	Hepatitis B	Vaccination compliance
	Hepatitis C	Safe sexual practices, avoid needle sharing
Chronic lower respiratory diseases (41.6)	Hypertension	High-salt diet, medication
	Atherosclerosis	Diet
	Smoking	Smoking
Unintentional injuries (e.g. motor-vehicle related deaths, falls and drug poisonings) (43.2)	Alcohol abuse	Alcohol use before or while driving
	No seat belt engaged	Seat belt non-use

Causes of death (age-adjusted death rates in the United States per 100,000 population)	Selected risk factors	Behavioural determinants of risk factors
	Unsafe driving	Behavioural distractions (e.g. smoking, texting) while driving; speeding
	Drug abuse	Drug use before or while driving
	Falls	
	Drug poisoning	
Stroke (37.6)	Atherosclerosis	Diet, aspirin
	Cholesterol management	Diet, medication

Source: data from National Center for Health Statistics, *Health, United States, 2016 with Chartbook on Long-term Trends in Health*, p. 2 and p. 18, 2016, available from https://www.cdc.gov/nchs/data/hus/hus16.pdf Risk factor data from Johnson et al. 2014.

As noted in previous chapters, the causes of death in developing countries differ from the patterns reflected in Table 2.10.1, although the demographic profiles are converging. World Health Organization (WHO 2017) estimates of the five leading causes of death for low-income countries have been respiratory diseases, diarrhoeal diseases, stroke, CVD, and HIV/AIDS. As tobacco use has increased and problems of overconsumption of sugars, starch, and fat have replaced some of the malnutrition problems in those countries, other chronic diseases, in addition to CVD, have increased. According to WHO (2005) projections, over 8 billion deaths will be due to tobacco in 2030 (Mathers and Loncar 2006). Because it is a known probable determinant of at least 25 diseases, and the most important determinant of some of the leading causes of death, tobacco use will cause 10% of all deaths globally. This alone warrants the concerted global attention to this behaviour that has been proposed by the Framework Convention on Tobacco Control (2003).

Behaviour as a determinant of other risk factors

Besides the cumulative effect of behaviours on physiological risk factors, such as the energy balance between calorie intake and physical activity producing weight gain, obesity, and hypertension, many of the health

consequences of behaviour are secondary to their impact on the immediate environment. Individuals are not merely the passive victims of the environments they inhabit or traverse. They are agents of change in those environments, and their ability to alter or control environmental threats to their own health and that of others increases with technological innovations, such as machinery, sanitation, and food chains. The growing capacity of individuals and groups to alter their environments through technological means, such as transport and roads that decrease risk of injury, while transport also produces negative consequences for their health when combined with alcohol or texting while driving, or reducing walking as a physical activity. Hence, health promotion and health protection have emphasized mobilizing individuals and groups to undertake personal conservation behaviours and collective actions to support policy changes and regulatory initiatives in support of more healthful environments and health behaviour in response to environmental circumstances.

Among the most striking differences in health between the high- and low-income countries are the perinatal–juvenile mortality rates which have decreased dramatically in high income countries, but are decreasing slower in low-income countries. Childhood mortality has long been associated with maternal education, poverty, technology change, and sanitation, thus the major public health goals in low-income nations have related to the provision of immunizations, access to a sufficient supply of clean water, the installation of proper sanitation facilities, increasing access to HIV treatment, and improving childhood and maternal mortality (WHO 2015). The United Nations has undertaken a resolution to think more broadly and holistically to improve child health first with its Millennium Development Goals (MDG), and now, ‘Transforming our world: the 2030 Agenda for Sustainable Development’ which aims to eliminate hunger and poverty by the year 2030. These environmental and age-specific or disease-specific health measures, however, achieve their intended health goals only to the extent that an informed population accepts and uses them properly. A report by the World Bank (1993) suggested that the single most important public health policy for developing countries lies in the improvement of the education of young girls. Better educated women have fewer children, who tend to be healthier and, in turn, better educated and able to respond to the technological advances offered by such environmental measures. They also become a better informed electorate to demand and support healthful

policies. However, globally, education of girls still falls short of their male counterparts, and advocacy around this issue continues, earning Malala Yousafzai a Nobel Peace Prize in 2014. In short, behaviour remains a critical mediator of the relationships between environmental measures and health outcomes, as well as the relationship between the health needs and the political actions to create the environmental changes.

The behaviour–policy link in the causal chains becomes more important as the chronic diseases become more prevalent in middle- and low-income countries. Inspection of the recent trends in the richest of the developing countries provides evidence that improvement of the socioeconomic condition is accompanied by a shift in mortality towards the chronic diseases and their behavioural risk factors reviewed earlier. In Mexico, for example, as early as 1991, infectious and parasitic diseases were only the fourth leading cause of death, following diseases of the heart, malignant neoplasms, and unintentional injuries. Low and middle income countries are now the primary target for market expansion for the multinational tobacco and convenience food companies (WHO 2008; Monteiro et al. 2013). Behaviour will be an issue both in personally resisting the temptations offered by these industries and in collective action to restrain their advertising, promotion, ingredients, and location.

Behaviour as a consequence of cognitions, environments, and genetics

Notwithstanding the implied simplicity of identifying a few behaviours that account for the majority of deaths in developed countries, those and other behaviours are highly complex, value-laden, and overdetermined. Most behavioural risk factors and healthcare behaviours as well, are the product of a variety of component behaviours, tasks, skills, or actions. For example, food consumption confronts most people with a chain of related behaviours: procuring and selecting food, planning menus or selecting from a menu, preparing or ordering foods, and eating with literally hundreds of food-related choices, including where to shop or eat, what to purchase or prepare, how to season food, and with whom to eat (Simons-Morton et al. 1986). One can identify similar chains of component behavioural choices for each of the other health behaviours identified in Table 2.10.1.

Not only are the behaviours complex, but each behaviour has numerous influences or determinants. Factors that influence behaviours can be grouped into three major categories (Green and Kreuter 2005; with adaptation from Andersen 1969): predisposing, enabling, and reinforcing, as shown in Fig. 2.10.2. Both positive and negative behaviours are predisposed, enabled, and reinforced by forces in the culture and the environment. This broad categorization has proved useful in public health programme planning (with more than 1,000 published applications, see http://lgreen.net/bibliog.htm for bibliography) because it groups the determinants of behaviour according to the major strategies used in public health to influence them. *Direct communications* through mass media, schools, worksites, other organizations, and through patient counselling in health clinics are used to influence the predisposing factors. *Indirect communications* through parents, teachers, clergy, community leaders, employers, peers, and others are used to strengthen the reinforcing factors with organizational rewards or social-normative influence. *Community organization*, *political activation*, and *training* strengthen the enabling factors by mobilizing and moving resources, policies, and building skills and capacities.

Predisposing factors

Predisposing factors reside in the individual and include attitudes, values, beliefs, and perceptions of need, but these are shaped over time by cultural and social exposures, which produce reinforcing factors (see later). Predisposing factors are those antecedents to behaviour that provide the rationale, motivation, or drive for an individual's or group's behaviour. They mostly fall in the psychological domain of determinants, though genetics and environment shape them across the lifespan. They include the cognitive and affective dimensions of knowing, feeling, believing, valuing, and having self-confidence or a sense of self-efficacy.

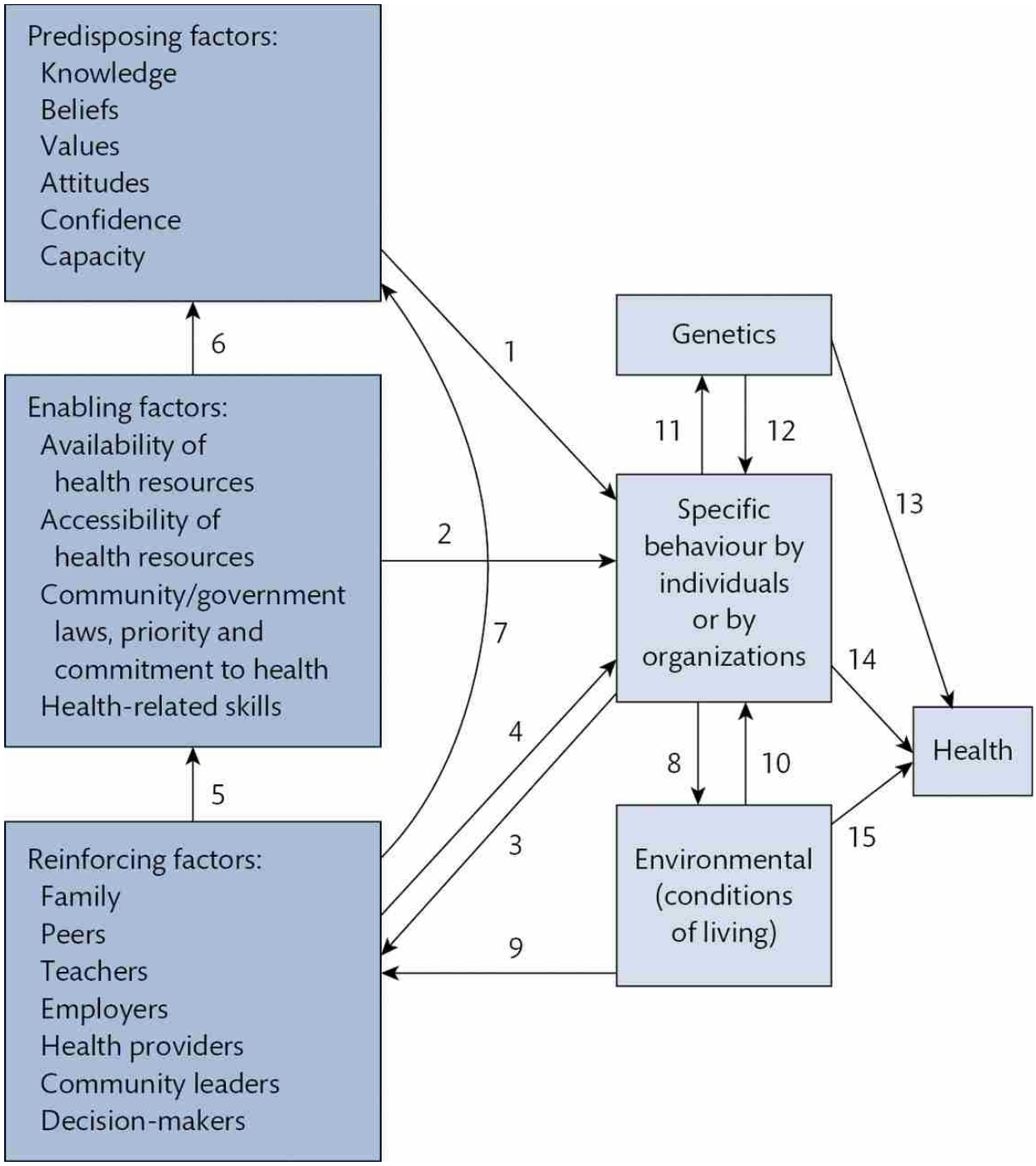


Fig. 2.10.2 This portion of the Precede–Proceed Model includes additional lines and arrows to outline a theory of causal relationships and order of causation and feedback loops for the three sets of factors influencing behaviour. In addition to the lines shown, an arrow from 'enabling factors' to 'environment' would elaborate the ecological aspect of factors that influence behaviour indirectly through changes in the environment.

From Green, L.W., Kreuter, M., Peterson, D., Geilen A., Eds. (2019). Health Program Planning: An Educational and Ecological Approach. Baltimore: Johns Hopkins University Press [in press, with permission of the author].

Because these determinants of behaviour reside in individuals, public health must seek to influence behaviour by assessing the prevalence and distribution of key predisposing factors and look for opportunities to communicate differentially with various segments of the populations, according to an educational diagnosis or 'social marketing' assessment of their knowledge, attitudes, values, beliefs, and perceptions. The behavioural science literature, especially in health psychology, is replete with competing theories of the relative importance of particular predisposing factors and how they interact with each other (e.g. Glanz et al. 2008). Some of these theories, such as the Health Belief Model (Becker 1974), Social Cognitive Theory (Bandura 2004), and the Transtheoretical Model of Stages of Change (Prochaska et al. 2002), have become mainstays of the intervention literature on health education and behaviour change. These frequent uses of selected models have produced sufficient numbers of comparable studies or programme evaluations to permit meta-analyses of their applications across different health behaviours (e.g. Spencer et al. 2005; Tuah et al. 2011). These applications of theory, in turn, have led to enough consistency in the construction of interventions and programmes that these have lent themselves to systematic reviews, notably by national panels such as the Task Force on Community Preventive Services (www.thecommunityguide.gov) and the UK's National Institute for Health and Care Excellence (http://www.nice.org.uk).

Enabling factors

Underplayed in most psychological studies of health behaviour, but critical to the role of public health as a complement to the roles of parents, schools, and mass media in the development of a healthy population, are the enabling factors influencing behaviour. These are often conditions of the environment that facilitate (or impede) the performance of a predisposition or motivated action by individuals or groups. Included are the availability, accessibility, and affordability of healthcare and community resources. Also included are conditions of living that act as facilitators or barriers to action, such as availability of transportation or childcare to release a parent from that responsibility long enough to participate in a health programme, clinic, or service. It also includes features of the built environment, such as bicycle lanes, sidewalks, proximity of housing to workplaces, and other physical

conditions of the environment that make physical activity more or less convenient and inviting (e.g. Taylor et al. 2007). Enabling factors also include new skills that a person needs—or new capacities that a group, organization, or community needs—to be able to accomplish health goals.

Public health addresses the enabling factors for behavioural change through policy changes and resource allocations at the national, regional, and local level, community organization for such policy and resource allocation at the local level when government programmes and services are insufficient or inappropriate to local perceptions of need. Training in the skills needed to take certain complicated actions is the other behavioural intervention used by public health to address the enabling factors.

In an era of constrained public health resources, many communities have resorted to more intensive community organization to build coalitions of multiple agencies including government and non-governmental organizations, and intersectoral collaboration between the health agencies and the community's schools, worksites, transportation, and other such sectors. Public–private collaboration on issues of enabling a more healthful food supply in school lunches, for example, or less accessibility of children to vending machines with cigarettes or 'junk food', have become increasingly common strategies for addressing the behaviour–environment interaction issues at the heart of most enabling factors (e.g. Institute of Medicine 2012).

Reinforcing factors

Before chronic diseases became the prevailing concern of public health in developed countries, behaviour was perceived as less important than today. The types of behaviour required to implement many of the most important public health interventions such as immunization and fluoridation of water supplies were less complex. Immunizations, for example, could confer long-term immunity for an individual with a single act. Fluoridation or chlorination of water supplies could be implemented in many communities with no engagement of the public, although when water fluoridation did become a local political issue requiring a local vote, the behaviour in question for most people was, again, a single action. As chronic diseases became more prevalent, the behaviours in question were largely ones that had to be repeated frequently, some for a lifetime, such as dietary practices,

hypertension medication regimens, smoking cessation, and physical activity. For these behaviours, the predisposing and enabling factors remain important, but reinforcing factors take on increased importance because maintaining a behaviour requires strong or repeated extrinsic reinforcement in the form of rewards or incentives, or highly internalized intrinsic reinforcements in the form of values and personal commitments.

Reinforcing factors are those consequences of action that determine whether the actor receives positive (or negative) feedback and is supported socially or financially after it occurs. Reinforcing factors thus include social support, peer influences, and advice and feedback by healthcare providers, as well as a sense that the benefits of the action outweigh the costs. In consideration of the benefits (and costs), they also include physical consequences of behaviour, which may be separate from the social context. Examples include the feeling of well-being (or pain) caused by physical exercise and the alleviation of respiratory symptoms (or the experience of side effects) following the use of asthma medication, or gaining weight while quitting cigarettes.

Public health uses reinforcing factors such as denormalizing smoking in public places, social reinforcement with encouraging words in personal counselling, small-group health education sessions in which behaviour is publicly endorsed and praised, or mass media images of attractive role models with whom people wish to be associated by their own behaviour. No matter how effective these extrinsic reinforcements might be in the short term in strengthening a behavioural tendency, they must be internalized over time to become intrinsic motivation. Token rewards lead to token behaviour if the individuals do not replace them over time with the belief that the behaviour is intrinsically valuable because it accords with their own personal values (Green et al. 1986).

Table 2.10.2 gives examples of the more commonly identified influences on each of four most important behavioural risk factors. These behaviours interact with each other, such that changes in one influence predisposing, enabling, and reinforcing factors substitutes influences on others.

Table 2.10.2 Some known and suspected predisposing (a), enabling (b), and reinforcing (c) influences on four major behavioural risk factors

Cigarette smoking	Dietary practices	Alcohol use/abuse	Physical inactivity
• Knowledge of adverse health effects of smoking (a) • Attitudes about smoking (a) • Skills in smoking cessation/prevention (b) • Cigarette cost (b and c) • Availability of cigarettes (b) • Cigarette advertising (a) • Peer influences to smoke (c) • Social support for non-smoking (c) • Smoke-free policies in worksite (b and c) • Access to cessation services (b) • Attitude towards tobacco industry practices (a)	• Personal food preferences (a) • Cultural food preferences (a) • Perceived social acceptance of foods (a and c) • Social context of eating (b and c) • Availability and convenience of healthful foods (b) • Skills in menu planning (b) • Skills in food preparation (b) • Skills in food selection (b) • Food advertising (a and c) • Healthful options in vending machines (b) • Belief in personal susceptibility to risk factors (a)	• Expectations of alcohol effects (a) • Child of alcoholic (a) • Alternatives to alcohol (a and b) • Psychological stress (a, b, and c) • Low self-esteem (a) • Early drinking experience (c) • Heavy social drinking (c) • Parent and peer influences (c) • Alcohol advertising (a, c) • Cost of alcohol (b) • Availability of alcohol (b)	• Beliefs in physical activity benefits (a) • Attitudes towards physical activity (a) • Self-efficacy (a) • Resources, facilities for exercise (b) • Cost of exercise facilities (b) • Skills in relapse prevention (b) • Skills in goal setting (b) • Enjoyability of physical activity (c) • Family support (c) • Design of the built environment (b and c) • Belief in benefits of physical activity

Public health strategies to influence determinants of behaviour

Another chapter will deal with the full range of health promotion and health education strategies (see Chapter 6.4). The focus of this section is specifically on how the aspects of the environment discussed in the previous sections that shape health-related behaviour can be targeted for strategic intervention for public health purposes. Three types of strategies are used in public health to accomplish health promotion and disease prevention goals through behaviour change:

Educational strategies inform and educate the public about issues of concern, such as the dangers of drug misuse, the benefits of automobile restraints, or the relationship of maternal alcohol consumption to fetal alcohol syndrome.
Automatic-protective strategies are directed at controlling environmental variables, that minimize the need for individual decisions in structuring each behaviour, such as public health measures providing for milk pasteurization, water fluoridation, infant immunizations, and the burning or chemical killing of marijuana crops, but these often involve individual and group decisions and actions about which policies to support, since they limit degrees of freedom in choice of behavioural options.
Coercive strategies employ legal and other formal sanctions to control individual behaviour, such as required immunizations for school entry, mandatory tuberculosis testing of hospital employees, compulsory use of automobile restraints, and arrests for drug possession or use or for drunk driving.

Illustrative of successful public health measures reflecting this range of strategies directed at influencing health-related behaviours are the efforts to control tobacco consumption. Influencing tobacco-related knowledge and attitudes has been declared one of the great public health achievements of the twentieth century, at least in the United States (Centers for Disease Control 1999) and several other countries. Considering that these successes, as measured by the reductions in tobacco consumption behaviour, were achieved largely in the last third of the century, they represent both a remarkable turn-around of an epidemic of smoking behaviour that had increased inexorably through the first two-thirds of the century, and an inspiration for public health approaches to other health-related behaviours that now show similar epidemic trends. The application of the lessons from tobacco control to reversing, for example, the obesity epidemic through similar influences on dietary and physical activity behaviours has become a point of major public health debate in the early years of the new millennium (Mercer et al. 2005; Green et al. 2006; Eriksen et al. 2007; Institute of Medicine 2012).

Influences on smoking initiation and cessation are numerous, and many of them are also intertwined with diet, alcohol, and physical activity. Predisposing factors include attitudes about smoking and beliefs about and knowledge of the health effects of smoking. Enabling factors of access to

tobacco products and price are influential. The price elasticity has been documented at a 3 per cent decrease in tobacco consumption for every 10% increase in the price, and a greater relative decrease in youth, for whom disposable income is less (Ranson et al. 2002; Wilson et al. 2012). The influence of raising taxes on cigarettes, then, has been a major public health strategy with known effects on consumption. With the dedication of part of the tax revenues to the funding of comprehensive tobacco control programmes, some jurisdictions have achieved a doubling and even tripling of the rate of decline in per capita consumption compared with other jurisdictions in the same country (Mercer et al. 2005). But more powerful still are the cumulative reinforcing factors of growing social support and pressure for smoke-free environments as smoking becomes denormalized with the decline in prevalence rates and the legal restrictions on smoking in public places (Eriksen et al. 2007; Wilson et al. 2012).

Cigarette advertising and promotions by the tobacco industry have proved to be powerful negative predisposing and reinforcing influences on smoking behaviour that are highly adaptable to changing regulatory attempts to control their content and channels. Initial restrictions on US tobacco advertising in the broadcast media in the mid-1960s, for example, resulted in the tobacco industry voluntarily withdrawing from radio and television advertising, but adroitly using those resources from mass media to expand its advertising vastly on more targeted print media in magazines, billboards, and youth-oriented outlets. Sponsorship of sports and arts events and clubs, for example, provided a more targeted venue for reaching youth and other susceptible markets. Response of some jurisdictions in restricting these sponsorships has produced reductions in youth uptake of tobacco products (Hagmann 2002; Perez et al. 2012). Diversification of advertising to youth-oriented media and point-of-purchase settings such as stores near schools has resulted in increased restrictions in many jurisdictions. The restrictions have been largely disappointing, which is attributed to the agility of the industry in finding loopholes and ways to circumvent the new legal restrictions.

Attempts to control the determinants of tobacco consumption also illustrate an essential public health lesson in the importance of comprehensive programmes to affect the complex behaviours associated with chronic diseases (Centers for Disease Control and Prevention 1999, 2007). Comprehensive approaches have proved critical because:

No single intervention reaches all segments of susceptible people in a population (Green and Glasgow 2006).
No single intervention reaches different segments with the same degree of effectiveness (Warner 2011; Green 2014).
Different interventions are differentially effective in the different phases of change (Prochaska et al. 2002).
Different interventions variously influence the predisposing, enabling, and reinforcing determinants of behaviour in a population (Green et al. 2019).
This comprehensiveness lesson has been carried forward to public health efforts in diet, alcohol, and physical activity insofar as they have increasingly engaged multiple organizations in community coalitions and multiple sectors and national programmes to address the various determinants that extend beyond the purview of the medical and public health sectors at the local level. The complexity of these organizational and ecological layers and intersections of influence on both physical activity and food consumption have driven a concerted effort in the rush to find remedies for the obesity epidemic to include ecological models and systems models (Institute of Medicine 2012).

The promotion of physical activity or active living illustrates the lessons of comprehensiveness and multisector involvement in public health strategies to influence a complex behaviour. Increasing attention has been given in recent years to the built environment, including the density of housing, mixed-use neighbourhoods, and sidewalks that encourage walking to and from home for shopping, work, and recreation; provision of mass transit that also supports walking the distance to and from transit stops rather than driving to work and other destinations (Frank 2000; Sallis et al. 2012).

The recent growth of the active living field is an extension and integration of traditional public health approaches with more collaborative approaches involving sectors such as housing, parks and recreation, and transportation. It responds to the increasing recognition of the complexity of health risk behaviours related to chronic diseases and their numerous determinants. The performance of each behaviour is interwoven with or ecologically embedded in other behaviours, organizations, and ecologies (Ottoson et al. 2009).

Behaviour as determinant of environmental and genetic predispositions

Population behavioural and educational diagnoses enable public health to intervene strategically on the behaviour of populations. But health problems have other determinants in the environment and in genetics. Behaviour also can play a role in influencing those determinants.

The reciprocal influence of individual behaviour on environments

A fundamental precept of human ecology is the reciprocal determinism of behaviour and the environment (Green et al. 1996; Sallis et al. 2012). The literature on health promotion took a sharp turn away from behaviour in the 1980s to give attention to the policy and environmental determinants of health. This was partly in response to a period in which public health was perceived to have taken too much of a psychological approach to the determinants of health (Green 2006). As early as 1968, Edward S. Rogers had appealed to sociology for the assistance public health needed from the social sciences to address the ecological issues (Rogers 1968). Psychologists, however, were more available, at least in the United States, to step into the perceived social science void of public health and to take up the new professorships in public health. They brought an emphasis on testing theories of individual behavioural change. As the ecological imperative of multilevel interventions (individual, family, organizational, community, regional, national, and global) gained growing emphasis in public health (e.g. Kickbusch 1989; Green et al. 1996), a gradual turn to the study of behaviour-in-context gave reciprocal determinism a new lease on life in public health (Stoto et al. 1997; Institute of Medicine 2001, 2003) in the decade bridging the millennium. One theory of behavioural change that gained particular public health prominence in this era was Albert Bandura's social cognitive theory, with its emphasis on self-efficacy and individual agency in changing one's environment at the same time that it gave prominence to the social environment in shaping behaviour (Bandura 2004).

Some of the environmental determinants of health beyond the behavioural control of individuals nevertheless lend themselves to group political behaviour or collective action through community intervention. Communities, neighbourhoods, or special-interest (self-help) groups can organize, vote, lobby, boycott, and otherwise support or prevent some environmental changes. To varying degrees, individuals can avoid or limit

their exposure to environmental risks such as solar radiation, lead paint, and ambient smoke. In short, individual behaviour can be mobilized to influence the environment, so that individuals need not be seen only as passive objects of environmental influence.

The projected but still limited mobilization of behaviour to influence genetics

The Human Genome Project and the explosion of research on human genetics have raised very hopeful scenarios of ‘precision medicine’, in which individuals could know more precisely their genetic risk of certain diseases, causes of death, and most effective treatments (National Research Council 2011). The usual assumptions are that such information could be made readily available to individuals, and that having such information would be considerably more compelling in motivating behaviour than the usual statistical risk of groups of people without the personalized association with the individual in question.

The first assumption remains to be supported by true evidence of effectiveness. For example, in about 30% of women with breast cancer, overexpression of a protein called HER2/neu is associated with a worse prognosis and in such women the drug trastuzumab is especially effective. The US Food and Drug Administration approved the drug and a test for HER2/neu for use in women with metastatic breast cancer in 1998 and more recently for women with early stage breast cancer (Hortobagyi 2005; Braga et al. 2006). The co-development and approval of the drug and test is considered one of the real successes of the application of genetics to modern medical practice and among the best examples of ‘precision medicine’. One might think the story is complete. However, although this information about the efficacy and availability of the test and the drug was published and promulgated by commercial backers, the behaviour of clinicians and their patients has been a more complicated matter. There is little known about how many women have access to testing and treatment. The costs of both are high and cost-effectiveness issues remain a challenge. Also how best to administer the test has not been resolved (e.g. timing) and there are reports of an increased risk of heart failure, the effect of which on acceptance is not known.

The second assumption, that having such information would motivate more concerted effort to change one's behaviour, is only partially supported by: (1) the logic and evidence from other areas of health counselling and communications that more personally tailored health information based on the individual's own family history or biological risk information adds motivational value to the experience with health advice that is based on more generic information (Kreuter et al. 2005); (2) limited direct evidence from the few instances in which such genetic information has been used to counsel individual behavioural choices of action. The latter is illustrated by the examples in the previous paragraph and those from prenatal testing and counselling for birth defects. However, a meta-analysis of 21 controlled trials showed that while genetic counselling improved knowledge of cancer genetics, it did not increase the level of perceived risk and few studies examined cancer surveillance behaviours (Braithwaite et al. 2006).

Public health faces two major limiting factors in pursuing this pathway of behaviour influencing health through genetic determinants: (1) the limited influence of the genes so far implicated in specific mortality or morbidity outcomes, and their interactions with the environment, and (2) the ethics of offering such information to the individual with anything more than a cautionary note of possible relevance to their reproductive decisions or their behavioural choices. The first is a limit that could be partially overcome with further breakthroughs in the human genomics research. But apart from some prenatal tests for genetic defects in the foetus, most of the other genetic markers associated with predispositions to illnesses or premature mortality are highly interactive with other genes and the environment (Kahan and Cheskin 2014, p. 237). Therefore, any course of action recommended to the individual remains probabilistic in its assurance of a health benefit. In combination with the other risk factors that can be more readily identified, the addition of personalized genetic information might raise the probabilities sufficiently to help the individual reach a tipping point in motivation to take action on complex health-related behaviours. But whether it really adds motivational value to what could be similarly known from a good family history has yet to be demonstrated.

The other limiting factors for this behaviour–genetic pathway as a public health consideration are the ethical complications that arise with the technology and the information. The concern, as in other screening technologies, with false-positive results can be multiplied in their ethical

considerations for genetic information on individuals. The usual issues of protecting the privacy of such information and the potential discrimination in hiring, placement, retention, and promotion of individuals with known genetic predispositions will continue to be debated before the 'precision medicine' potential of expanding the behaviour–genetic pathway can be pursued as public health policy. Meanwhile, private medicine and industry (e.g. 23andMe) is opening opportunities for individuals to explore this option in structuring their behavioural response to personal genetic information, as research on the behavioural implications mounts with three priorities: '(1) improving the public's genetic literacy in order to enhance consumer skills; (2) gauging whether genomic information improves risk communication and adoption of healthier behaviours more than current approaches; and (3) exploring whether genomic discovery in concert with emerging technologies can elucidate new behavioural intervention targets' (McBride et al. 2010). Chapter 8.19 addresses human genetics more thoroughly.

The interaction of socioeconomic status, environments, and behaviour

Of all the interactions in the association of behaviour and health, none is more pervasive, consistent, and robust than that of socioeconomic status (SES). The relationship between SES and measures of health or mortality is shown in other chapters to be a gradient rather than a threshold effect, though threshold effects (or points of diminishing returns) are sometimes found beyond which income or other SES indicators have no further beneficial effects (e.g. Finch 2003). Those at the top of the SES hierarchy usually have better health and lower mortality rates than those just below them who are themselves better off than the others below them, and so on down to those in poverty at the very bottom. The gradient adheres whether the SES measure is education, income, occupational status, or place of residence. The gradient globally is anchored at the lowest end by the poorest developing countries. These historical associations between health behaviours and SES have been confounded in the more recent tracking of obesity, physical activity, and food consumption. The more sedentary work and modes of travel of a majority of white collar workers blur the

educational, income, and occupational correlations with physical activity (e.g. Canadian Institute for Health Information 2006). In Canada, for example, 'of the demographic variables examined, income, occupation, and employment status were unrelated to obesity, while education was negatively associated with the prevalence of obesity' (Raine 2004, p. 6).

Some 767 million people globally in 2013 lived in extreme poverty on an income of just US $1.90 a day, of whom half live in sub-Saharan Africa (World Bank 2018). Many of the poor lack access to basic health services including sanitation and clean water and are at exaggerated risks for many of the leading causes of death in those countries. The gradients are further complicated by the varied statistical interactions of gender, racial and ethnic discrimination with income, education, and occupation—the usual metrics of SES—in different countries.

While SES has been found to directly affect health through mechanisms such as material well-being and psychosocial stress, this chapter focuses on the role of SES on behaviour (Lynch and Kaplan 1997; Wilkinson and Marmot 1998; Seeman et al. 2010). The question here is how the mortality and morbidity gradients with socioeconomic status might operate through health-related behaviours, cultures, and environments, whether operating through increased access to resources and material wealth or through lower psychosocial stress. Earlier discussion in this chapter suggest mechanisms by which the pervasive SES gradient influences health, and ways in which the behavioural pathway can be made more protective or preventive.

SES as a predisposing determinant of behaviour

The ecological perspective on SES as a determinant of behaviour related to health would suggest first that environments shape behaviour from early childhood onward. Shaping behaviour in the first instance (rather than enabling it in the second or reinforcing it in the third) qualifies this environmental influence as a predisposing factor. Homes, neighbourhoods, towns, cities, regions, and whole countries with their variable physical, social, economic, and cultural environments differ in relation to SES measured at individual, family, neighbourhood, and other geographic levels. Once the measures of SES and health or mortality are aggregated above the individual level, their relationships constitute ecological correlations. Studies have examined the ecological relationship between mortality rates and

various indicators of social inequalities in geographical areas varying in size from metropolitan areas (Lynch et al. 1998) to whole countries (Wilkinson and Marmot 1998). These studies showed that those areas where inequalities between those at the top of the hierarchy and those at the bottom were the largest were also those in which the mortality gradient was the strongest (Wilkinson 1996). Similar findings were found with other indicators of social inequities such as differences in social capital (Kawachi et al. 1997). Health disparities between socioeconomic groups appear from Canadian data to increase with age across the lifespan, which supports the notion of a cumulative effect over time of the health- and mortality–SES gradient (e.g. Prus 2007).

One implication of this relative deprivation dimension of the SES–health gradient is that perception of one's status relative to others de-motivates or discourages one's efforts to take greater control over the behavioural and environmental determinants of one's health. Whether consciously discouraged or unconsciously conditioned by repeated confrontations with inequalities and inequities that conspire against one's efforts, the hypothesis is that disparities make those exposed to relative deprivation less predisposed or motivated to take preventive and healthcare actions. Two specific mechanisms have been suggested for this de-motivation or lack of predisposition to undertake behaviour. One is a chronic pessimism that has been found in adolescent children of lower SES parents, and that pessimism was associated with stress (Finkelstein et al. 2007). The other possibly related mechanism is the theoretical construct of 'self-efficacy' and diminished stress, widely measured in association with social cognitive theory applications in health behaviour research (Bandura 2004; Jang and Yoo 2012).

SES as an enabling determinant of behaviour

Socioeconomic standing also confers capabilities and resources that enable the predisposed behaviours to be carried out, for better or for worse. With higher standing come more resources and the associated education and training that endow individuals, families, groups, and communities with enabling judgements, resources, and skills. No matter how motivated people may be by their predisposing factors, they may not have the income and other resources, including accessible and affordable services within reach of

their residences or workplaces, to be able to carry out the behaviour without great sacrifice and inconvenience. But environmental variables of accessible fruit and vegetable outlets in local neighbourhoods, for example, only partially explain or mediate the relationship between SES and fruit or vegetable consumption (Ball et al. 2006). The ‘food deserts’ of poor access to fruits and vegetables in neighbourhood stores of poor areas are compounded by the high accessibility of the ‘food swamps’ of fast food restaurants with their calorie-dense menus (Fielding and Simon 2011).

The educational enabling influence of SES on behaviour

Among the SES indicators, education has for decades demonstrated the strongest association with most health behaviour measures (Green 1970; Metcalf et al. 2007), and this association has been found to have a major part of its essence in health literacy (e.g. Curtis et al. 2012). It also stands out among the indicators of inequities in confirming the relative deprivation hypothesis (Kunst and Mackenbach 1994). Education can be viewed as a proxy for a variety of predisposing and enabling factors in explaining the causes of behavioural determinants as mediators of at least part of the SES–health gradient. Prominent among these are optimism (a predisposing factor) and education as an enabling or coping resource (Finkelstein et al. 2007), with the knowledge, attitudes, and skills that come with years of schooling.

The proportion of people in Canada who are smokers is double in people with elementary or lower education, compared with people with university degrees (Adams 1993; Health and Welfare Canada et al. 1993). This spread in proportions of smokers was greater for the highest and lowest education categories than for the highest and lowest income or occupational status categories (Pederson 1993). Data from the major US community trials in cardiovascular disease prevention showed that the dramatic drop in smoking prevalence over the 1980s was more pronounced for people with higher education compared with people with less education (Winkleby et al. 1992a; Luepker et al. 1993). The same trend was observed in Canada for the period 1985–1991 (Millar and Stephens 1993).

SES as a reinforcing determinant of behaviour

A behaviour that is predisposed and enabled might still fail to persist beyond a trial stage of acting if it fails to produce satisfying results. Satisfaction comes from various sources that can have the effect of reinforcing behaviour. SES can contribute to the availability of reinforcements by putting people into association with other people and environments that are more likely to produce satisfaction with behaviours. Two examples follow.

The ‘status identity factor’ and social norms

The social-normative theory underlying the notion of SES functioning like a reinforcing factor would predict that people will identify with a place in the social hierarchy that they can justify on the basis of their highest achievement. Unlike the usual measure of SES produced by averaging standardized or weighted measures of education, income, and occupation, one hypothesis derived from social reinforcement theory was that people will aspire to and adhere to that norm of a particular behaviour associated with their highest measure on the gradient of SES. For example, a person who is relatively high on education, but of moderate income and occupational status, would tend to adhere to the behavioural norms of the highest, rather than the average, of the person’s three status scores. The hypothesis was tested in a California state-wide sample of mothers of children under 5 years, demonstrating that the immunization status of the children and five other measures of early childhood care followed a gradient with SES, but that the best predictor of the mother’s behaviour on these five measures was her highest standardized SES among income, education, and occupation, not the average. This inferred ‘status identity’ of the mothers provided a basis for linking the psychological phenomena of identification and role modelling with the sociological concept of normative influence (Green 1970).

Denormalizing behaviour

In retrospect, one of the most important elements of the tobacco control success of the last third of the twentieth century in the United States and Canada was the ‘denormalization’ of smoking behaviour in public places (Eriksen et al. 2007; Rogers 2010; Livingood et al. 2011). What had been a normative behaviour of smoking in the workplace and in restaurants, meeting rooms, and other public places became increasingly unacceptable,

first by legal restrictions, then by social norms, by which the growing majority of non-smokers expected and even insisted on smoke-free environments. The combination of new smoke-free or ‘clean air’ ordinances and by-laws with mass media emphasizing the carcinogenic properties of second-hand smoke and the rights of non-smokers to be spared the exposure to this carcinogen resulted in a dramatic drop in this public behaviour. It was during this period of the 30-year decline in smoking that the rate of decline accelerated most dramatically. One reason for the importance of this element of the tobacco control campaigns was that the passage and enforcement of policies restricting a personal behaviour faced strong opposition as long as it was perceived by the public to be a matter of individual rights and the threat to be only to the person’s own health. When the threat is seen to be to other people’s health, especially to the health of children, the support of passage and enforcement of the laws and regulations grows. Other public health campaigns are attempting to model this experience, which builds on the social responsibility notions associated with communicable diseases of the past, but with most of the chronic disease-related behaviours, such as overeating and sedentary living, it has been more complicated to relate the normative behaviour of individuals to the health of others.

The interaction of socioeconomic status, gender, and behaviour

Men’s and women’s relative risk of disease or death in relation to specific behavioural risk factors such as smoking are generally similar (e.g. Oliveira et al. 2007), but their experiences with health differ markedly (Chapman Walsh et al. 1995). These differences cannot be attributed solely to biological determinants related to sexual differentiation (Krieger et al. 1993). The social construct of gender, as opposed to the biological categories of sex, was conceptualized to refer to cultural and social conventions, roles, and behaviours assigned to men and women (Krieger 1996). These in turn shape the social, political, cultural, and economic circumstances experienced by men and women. Gender thus attempts to capture this differential experience that men and women have with their environment and the possibilities and constraints associated with these differences (Potvin and Frohlich 1998). Some of these constraints and possibilities are interacting with the living conditions associated with SES to shape the health of people (Wilkinson and Pickett 2008).

The correlation of SES with health appears stronger for men than for women; the SES–health gradient is steeper for men (Mustard and Etches 2003). The gender interactions with SES and health have been variously attributed to differential occupational experiences (e.g. Ross and Bird 1994), marital experiences (Koskinen and Martelin 1994), and degree of emancipation of women (Kawachi et al. 1999). These and other possible explanations generally require assumptions of behavioural mediators. Such mediators are as likely to be 'health-related' behaviour (e.g. sedentary living or food consumption patterns) as 'health-directed' behaviour (e.g. exercise or high-fibre diet). This distinction (Green and Kreuter 2005) recognizes the centrality of behaviour in the causal chains even when (perhaps especially when) it is not consciously health-*directed* behaviour.

Relationships among health-related behaviours

The interplay among habitual behavioural patterns and the socioeconomic and cultural conditions reviewed in the last half of this chapter leads one to put into a broader context the reductionist examinations, presented in the first half, of specific behaviours as they relate to health, disease, and mortality. The dynamic relationships among the specific measures create a complex system of social, economic, cultural, and behavioural factors, interwoven with disease risk factors and health status, and influenced by the healthcare and physical environments. There is little evidence supporting a one-dimensional concept of health-related behaviours (Calnan 1994) or a uniform 'healthy lifestyle'. It supports, instead, the reciprocal and ecological system of relationships laid out in the various sections of this paper, showing behaviour as both a product and an active ingredient in the mix of cultural, socioeconomic, and environmental conditions affecting individual and population health.

Conclusion

Behaviour is an inescapable link in the chain of causation between most environmental and genetic determinants and the health outcomes in which they are implicated. Some toxins and infectious environmental agents can affect health directly without behaviour as a mediator, but even these *can* be

mediated by individual action to avoid exposures, and collective behaviour of groups or communities to protect themselves.

The *social* environment presents a further complexity in the mediating and moderating of behaviour and environment in their determination of population health. Most health-related behaviour occurs in the context of the social environment, so it involves the behaviour of other people as well as that of the person whose health is in question. The individuals are acting upon, and in reaction to, each other as their health outcomes are being shaped by their actions. This reciprocal determinism of behaviour and the social environment applies as well to the physical environment and the genetic determinants of health. These interactions make up the ecology of health, and call upon public health to take an ecological approach to the management of population health.

One way to structure the ecological approach to the planning of public health programmes in which behaviour change has a role is to examine the factors that influence behaviour in three categories of determinants: predisposing factors, enabling factors, and reinforcing factors. These roughly correspond to strategies, respectively, that would use: (1) direct communications to influence the knowledge, attitudes, beliefs, and perceptions of the population concerning the behaviour–health relationship; (2) legal, engineering, financial, organizational levers, and resource development that would enable, restrain, or prohibit the behaviour; and (3) indirect communications through social organizations, parents, peers, employers, and others who control rewards and approval that would reinforce behaviour. By combining public health strategies directed at these three categories of determinants (predisposing, enabling, and reinforcing factors) the strategies will be comprehensive in their coverage and impact.

All of what has been understood as determinants of the health behaviour of individuals applies with some variation to the behaviour of health professionals, other practitioners and policymakers who could serve as channels through which to reach individuals, groups, and whole populations to influence their health behaviour. The actions of all of these categories of individuals, as well as their organizations, can be analysed in relation to the factors that predispose, enable, and reinforce their actions, and these analyses can point, in turn, to the development of strategies to change those behaviours.

ACKNOWLEDGEMENTS

We are indebted to Denise Simons-Morton, MD, PhD, and Louise Potvin, PhD, co-authors of this chapter respectively in its first and second editions, for remnants of their earlier contributions that may remain in this fourth revision.

REFERENCES

Adams, O. (1993). Health status. In: Stephens, T. and Fowler Graham, D. (eds.) *Health and Welfare Canada. Canada's Health Promotion Survey 1990: Technical Report*, p. 23. Ottawa: Ministry of Supply and Services.

Adler, N.E., and Stewart, J. (2009). Reducing obesity: motivating action while not blaming the victim. *Milbank Memorial Fund Quarterly*, 87(1), 49–70.

Ahmed, H.M., et al. (2012). Effects of physical activity on cardiovascular disease. *American Journal of Cardiology*, **109**, 288–95.

Andersen, R.M. (1969). *A Behavioral Model of Families' Use of Health Services*. University of Chicago Center for Health Administration Studies, Research Series No. 25. Chicago, IL: University of Chicago Press.

Arroll, B. and Beaglehole, R. (1992). Does physical activity lower blood pressure: a critical review of the clinical trials. *Journal of Clinical Epidemiology*, **45**, 439–47.

Ball, K., Crawford, D., and Mishra, G. (2006). Socio-economic inequalities in women's fruit and vegetable intakes: a multilevel study of individual, social and environmental mediators. *Public Health Nutrition*, **9**, 623–30.

Bandura, A. (2004). Health promotion by social cognitive means. *Health Education and Behavior*, **31**, 143.

Becker, M.H. (1974). The Health Belief Model and personal health behaviour. *Health Education Monographs*, **2**, 324.

Braga, S., et al. (2006). Use of trastuzumab for the treatment of early stage breast cancer. *Expert Review of Anticancer Therapy*, **6**, 1153–64.

Braithwaite, D., et al. (2006). Psychological impact of genetic counseling for familial cancer: a systematic review and meta-analysis. *Familial Cancer*, **5**, 66–75.

Calnan, M. (1994). Lifestyle and its social meaning. *Advances in Medical Sociology*, **4**, 69.

Canada (1974). *A New Perspective on the Health of Canadians (Lalonde Report)*. Ottawa: Department of National Health and Welfare.

Canadian Institute for Health Information (2006). *Improving the Health of Canadians: Promoting Healthy Weights*. Ottawa: Canadian Institute for Health Information.

Centers for Disease Control and Prevention (1999). Achievements in public health 1900–1999: tobacco use—United States, 1900–1999. *MMWR Morbidity and Mortality Weekly Reports*, **48**, 986–93.

Centers for Disease Control and Prevention (2007). *Best Practices for Comprehensive Tobacco Control Programs* (2nd ed.). Atlanta, GA: Centers for Disease Control and Prevention, Office on Smoking and Health.

Chapman Walsh, D., Sorensen, G., and Leanord, L. (1995). Gender, health and cigarette smoking. In: Amick III, B.C., Levine, S., Tarlov, A.R., and Chapman Walsh, D. (eds.) *Society and Health*. New York: Oxford University Press, pp. 131–71.

Colditz, G.A. (2014). Carpe diem: time to seize the opportunity for cancer prevention. *American Society of Clinical Oncology—Educational Book*, **34**, 8–12.
Curtis, L.M., Wolf, M.S., Weiss, K.B., and Grammer, L.C. (2012). The impact of health literacy and socioeconomic status on asthma disparities. *Journal of Asthma*, 49(2), 178–83.
Dawber, T.R. (1960). Summary of recent literature regarding cigarette smoking and coronary heart disease. *Circulation*, **22**, 164.
Eriksen, M.P., et al. (2007). Thank you for not smoking: the public response to tobacco-related mortality in the United States. In: Ward, J.W. and Warren, C. (eds.) *Silent Victories: The History and Practice of Public Health in Twentieth-Century America*. New York: Oxford University Press, pp. 423–36.
Evans, R.G., Barer, M.L., and Marmor, T.R. (eds.) (1994). *Why Are Some People Healthy and Others Not? The Determinants of Health of Populations.* New York: Aldine de Gruyter.
Fielding, J.E. and Simon, P.A. (2011). Food deserts or food swamps?: comment on 'Fast food restaurants and food stores'. *Archives of Internal Medicine*, 171(13), 1171–2.
Finch, B.K. (2003). Socioeconomic gradients and low birth-weight: empirical and policy considerations. *Health Services Research*, **38**, 1819.
Finkelstein, D.M., et al. (2007). Socioeconomic differences in adolescent stress: the role of psychological resources. *Journal of Adolescent Health*, **40**, 127–34.
Frank, L.D. (2000). Land use and transportation interaction: implications on the public health and quality of life. *Journal of Planning Education and Research*, **20**, 6–22.
Glanz, K., Rimer, B.K., and Viswanath, K. (2008). *Health Behavior and Health Education: Theory, Research, and Practice* (4th ed.). San Francisco, CA: Jossey-Bass.
Glass, T.A. and McAtee, M.J. (2006). Behavioral science at the crossroads in public health: extending horizons, envisioning the future. *Social Science & Medicine*, **62**, 1650.
Green, L.W. (1970). Manual for scoring socioeconomic status for research on health behavior. *Public Health Reports*, **85**, 185.
Green, L.W. (2006). Public health asks of systems science: to advance our evidence-based practice, can you help us get more practice-based evidence? *American Journal of Public Health*, **96**, 406.
Green, L.W. and Allegrante J.P. (2011). Healthy people 1980–2020: raising the ante decennially, or just the name from health education to health promotion to social determinants? *Health Education and Behavior*, 38(6), 558.
Green, L.W. and Glasgow, R. (2006). Evaluating the relevance, generalization, and applicability of research: issues in external validation and translation methodology. *Evaluation and the Health Professions*, **29**, 126.
Green, L.W., Kreuter, M., Peterson, D., Geilen A. (eds.) (2019). *Health Program Planning: An Educational and Ecological Approach*. Baltimore, MD: Johns Hopkins University Press.
Green, L.W., et al. (2006). Inferring strategies for disseminating physical activity policies, programs, and practices from the successes of tobacco control. *American Journal of Preventive Medicine*, 31(Suppl. 4), S66.
Green, L.W., Richard, L., and Potvin, L. (1996). Ecological foundations of health promotion. *American Journal of Health Promotion*, **10**, 270–81.
Green, L.W., et al. (2010). Male circumcision and HIV prevention: insufficient evidence and neglected external validity. *American Journal of Preventive Medicine.* 39(5), 479–82.
Green, L.W., Wilson, A., and Lovato, C.Y. (1986). What changes can health promotion produce and how long will they last? Trade-offs between expediency and durability. *Preventive Medicine*, **15**, 508–21.
Green, L.W. (2014). Closing the chasm between research and practice—evidence of and for change. *Health Promotion Journal Australia*, 25(1), 25–9.
Hagmann, M. (2002). WHO attacks tobacco sponsorship of sports. *Bulletin of the World Health Organization*, **80**, 80.

Health and Welfare Canada, Stephens, T., and Fowler Graham, D. (eds.) (1993). *Canada's Health Promotion Survey 1990: Technical Report*. Ottawa: Minister of Supply and Services.
Hortobagyi, G.N. (2005). Trastuzumab in the treatment of breast cancer. *The New England Journal of Medicine*, **353**, 1734.
Institute of Medicine (2001). Health and behavior: the interplay of biological, behavioral, and societal influences. Washington, DC: National Academies Press.
Institute of Medicine (2003). *The Future of Public Health in the 21st Century*. Washington, DC: National Academies Press.
Institute of Medicine (2012). *Accelerating Progress in Obesity Prevention: Solving the Weight of the Nation*. Washington, DC: National Academies Press.
Jang, Y. and Yoo, H. (2012). Self-management programs based on the social cognitive theory for Koreans with chronic disease: a systematic review. *Contemporary Nurse*, 40(2), 147–59.
Johnson, N.B., et al. (2014). CDC National Health Report: leading causes of morbidity and mortality and associated behavioral risk and protective factors—United States, 2005–2013. *MMWR Supplement*, 63(4), 3–27.
Kawachi, I., et al. (1999). Women's status and the health of women and men: a view from the States. *Social Science & Medicine*, **48**, 21–32.
Kawachi, I., et al. (1997). Social capital, income inequality, and mortality. *American Journal of Public Health*, **87**, 1491–8.
Kickbusch, I. (1989). Approaches to an ecological base for public health. *Health Promotion*, **4**, 265–8.
Koskinen, S. and Martelin, T. (1994). Why are socioeconomic mortality differences smaller among women than among men? *Social Science & Medicine*, **38**, 1385–96.
Kreuter, M.W., et al. (2005). Cultural tailoring for mammography and fruit and vegetable intake among low-income African-American women in urban public health centers. *Preventive Medicine*, **41**, 53–62.
Krieger, N. (1984). Epidemiology and the web of causation. Has anyone seen the spider? *Social Science & Medicine*, **39**, 887–903.
Krieger, N. (1996). Inequality, diversity, and health: thoughts on 'race/ethnicity' and 'gender'. *Journal of American Medical Women's Association*, **51**, 133–6.
Krieger, N., et al. (1993). Racism, sexism, and social class: implications for study of health, disease, and well being. *American Journal of Preventive Medicine*, **9** (Suppl.), 82–122.
Kunst, A.E. and Mackenbach, J.P. (1994). The size of mortality differences associated with educational level in nine industrialized countries. *American Journal of Public Health*, **84**, 932–7.
Levy, D.T., et al. (2017). A Framework for evaluating the public health impact of e-cigarettes and other vaporized nicotine products. *Addiction*, 112(1), 8–17.
Livingood, W.C., et al. (2011). Applied social and behavioral science to address complex health problems. *American Journal of Preventive Medicine*, 41(5), 525–31.
Lokey, E.A. and Tran, Z.V. (1989). Effects of exercise training on serum lipid and lipoprotein concentrations in women: a meta-analysis. *International Journal of Sports Medicine*, **10**, 424–9.
Luepker, R.V., et al. (1993). Socioeconomic status and coronary heart disease risk factor trends: the Minnesota Heart Health Survey. *Circulation*, **88**, 2172–9.
Lynch, J.W. and Kaplan, G.A. (1997). Understanding how inequality in the distribution of income affects health. *Journal of Health Psychology*, 2(3), 297–314.
Lynch, J.W., et al. (1998). Income inequality and mortality in metropolitan areas of the United States. *American Journal of Public Health*, **88**, 1074–80.
McArdle, W.D., Katch, F.L., and Katch, V.L. (1986). *Exercise Physiology: Energy, Nutrition, and Human Performance* (2nd ed.). Philadephia, PA: Lea and Febiger.
McBride, C.M., et al. (2010). Future health applications of genomics: priorities for communication, behavioral, and social sciences research. *American Journal of Preventive Medicine*, 38(5), 556–65.

McGinnis, J.M. and Foege, W.H. (1993). Actual causes of death in the United States. *Journal of the American Medical Association*, **270**, 2207–12.
McGinnis, J.M. and Foege, W. (2004). The immediate vs the important. *Journal of the American Medical Association*, 291(10), 1263–4.
Mathers, C.D. and Loncar, D. (2006). Projections of global mortality and burden of disease from 2002 to 2030. *PLoS Medicine*, 3(11), e442.
Mercer, S.L., et al. (2005). Drawing possible lessons for obesity prevention and control from the tobacco control experience. In: Crawford, D. and Jeffrey, D. (eds.) *Obesity Prevention and Public Health*. New York: Oxford University Press, pp. 231–64.
Metcalf, P., Scragg, R., and Davis, P. (2007). Relationship of different measures of socioeconomic status with cardiovascular disease risk factors and lifestyle in a New Zealand workforce survey. *New Zealand Medical Journal*, 120(1248), U2392.
Millar, W.T. and Stephens, T. (1993). Social status and health risk in Canadian adults: 1985–1991. *Health Reports*, **5**, 143–56.
Minkler, M. (1983). Blaming the aged victim: the politics of scapegoating in times of fiscal conservatism. *International Journal of Health Services*, 13(1), 155–68.
Monteiro, C.A., Moubarac, J.C., Cannon, G., Ng, S.W., and Popkin, B. (2013). Ultra-processed products are becoming dominant in the global food system. *Obesity Review*, 14 Suppl 2, 21–8.
Mokdad, A.H., et al. (2004). Actual causes of death in the United States, 2000. *Journal of the American Medical Association*, **291**, 1238–45. Erratum, *Journal of the American Medical Association*, **293**, 293.
Mustard, C. and Etches, J. (2003). Gender differences in socioeconomic inequality in mortality. *J Epidemiol Community Health*, 57(12), 974–80.
National Center for Health Statistics (2006). *Health, United States, 2006*. Hiattsville, MD: US Department of Health and Human Services, Centers for Disease Control.
National Research Council (2011). *Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease.* Washington, DC: The National Academies Press.
Ockene, J.K., Kuller, L.H., Svendsen, K.H., and Meilahn, E. (1990). The relationship of smoking cessation to coronary heart disease and lung cancer in the Multiple Risk Factor Intervention Trial (MRFIT). *American Journal of Public Health*, **80**, 954–8.
Oliveira, A., et al. (2007). Tobacco smoking and acute myocardial infarction in young adults: a population-based case-control study. *Preventive Medicine*, **44**, 311–16.
Ottoson, J.M., et al. (2009). Policy-contribution assessment and field-building analysis of the Robert Wood Johnson Foundation's Active Living Research Program. *American Journal of Preventive Medicine*, 36(2, Suppl. 1), S34–S43.
Pederson, L. (1993). Tobacco use. In: Health and Welfare Canada, Stephens, T., and Fowler Graham, D. (eds.) *Canada's Health Promotion Survey 1990: Technical Report*. Ottawa: Minister of Supply and Services Canada, pp. 97–108.
Perez, D.A., et al. (2012). Tobacco promotion 'below-the-line': exposure among adolescents and young adults in NSW, Australia. *BMC Public Health*, 12(12), 429.
Petrow, S. (1990). *Ending the HIV Epidemic*. Santa Cruz, CA: Network Publications.
Potvin, L. and Frohlich, K.L. (1998). L'utilité de la notion de genre pour comprendre les inégalités de santé entre les homes et les femmes. *Ruptures, Revue Transdisciplinaire en Santé*, **5**, 142.
Prochaska, J.O., Redding, C.A., and Evers, K.E. (2002). The Transtheoretical model and stages of change. In: Glanz, K., Rimer, B.K., and Lewis, F.M. (eds.) *Health Behavior and Health Education: Theory, Research, and Practice* (3rd ed.). San Francisco, CA: Jossey-Bass, pp. 99–120.
Prus, S.G. (2007). Age, SES, and health: a population level analysis of health inequalities over the life course. *Sociology, Health and Illness*, **29**, 275–96.

Raine, K.D. (2004). *Overweight and Obesity in Canada: A Population Health Perspective*. Ottawa: Canadian Institute for Health Information.

Ranson, M.K., et al. (2002). Global and regional estimates of the effectiveness and cost-effectiveness of price increases and other tobacco control policies. *Nicotine and Tobacco Research*, **4**, 311–19.

Rogers, E.S. (1968). Public health asks of sociology … Can the health sciences resolve society's problems in the absence of a science of human values and goals. *Science*, **159**, 506–8.

Rogers, T. (2010). The California Tobacco Control Program: introduction to the 20-year retrospective. *Tobacco Control*, 19(2), i1–i2.

Rose, G. and Colwell, L. (1992). Randomized controlled trial of anti-smoking advice: final (20-year) results. *Journal of Epidemiology and Community Health*, **46**, 75–7.

Ross, C.E. and Bird, C.E. (1994). Sex stratification and health lifestyle: consequences for men's and women's perceived health. *Journal of Health and Social Behavior*, **35**, 161–78.

Sallis, J.F., Floyd, M.F., Rodríguez, D.A., and Saelens, B.E. (2012). Role of built environments in physical activity, obesity, and cardiovascular disease. *Circulation*, 125(5), 729–37.

Seeman, T., Epel, E., Gruenewald, T., Karlamangla, A., and McEwen, B.S. (2010). Socio-economic differentials in peripheral biology: cumulative allostatic load. *Annals of the New York Academy of Sciences*, **1186**, 223–39.

Simons-Morton, B.G., O'Hara, N.M., and Simons-Morton, D.G. (1986). Promoting healthful diet and exercise behaviors in communities, schools, and families. *Family and Community Health*, **9**, 1–13.

Spencer, L., Pagell, F., and Adams, T. (2005). Applying the transtheoretical model to cancer screening behavior. *American Journal of Health Behavior*, **29**, 36–56.

Stamler, J. (2005). Established major coronary risk factors: historical overview. In: Marmot, M. and Elliott, P. (eds.) *Coronary Heart Disease Epidemiology* (2nd ed.). Oxford: Oxford University Press, pp. 18–31.

Steele, S., Stuckler, D., McKee, M., and Pollock, A.M. (2014). The Immigration Bill: extending charging regimes and scapegoating the vulnerable will pose risks to public health. *Journal of the Royal Society of Medicine*, 107(4), 132–3.

Stoto, M.A., Green, L.W., and Bailey, L.A. (eds.) (1997). *Linking Research and Public Health Practice: A Review of CDC's Program of Centers for Research and Demonstration of Health Promotion and Disease Prevention*. Washington, DC: National Academy Press.

Swan E., Bouwman L., de Roos, N., and Koelen, M. (2012). How science thinks and practice acts: bridging the gap in weight management interventions for adolescents. *Family Practice*, 29(Suppl.), i117–25.

Taylor, W.C., et al. (2007). Changing social and built environments to promote physical activity: recommendations from low income, urban women. *Journal of Physical Activity and Health*, **4**, 54–65.

Tuah, N.A., et al. (2011). Transtheoretical model for dietary and physical exercise modification in weight loss management for overweight and obese adults. *Cochrane Database of Systematic Review*, 5(10), CD008066.

US Department of Health and Human Services (2004). *The Health Consequences of Smoking: A Report of the Surgeon General*. Washington, DC: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, Coordinating Center for Health Promotion, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health, Atlanta.

US Department of Health and Human Services (2006). *The Health Consequences of Involuntary Exposure to Tobacco Smoke: A Report of the Surgeon General—Executive Summary*. Atlanta, GA: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, Coordinating Center for Health Promotion, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health.

US Preventive Services Task Force (1989). *Guide to Clinical Preventive Services: An Assessment of the Effectiveness of 169 Interventions*. Baltimore, MD: Williams & Wilkins.
Warner, K.E. (2011). Disparities in smoking are complicated and consequential. What to do about them? *American Journal of Health Promotion*, **25**, S5–S7.
Wilkinson, R.G. (1996). *Unhealthy Society. The Afflictions of Inequality*. London: Routledge.
Wilkinson, R.G. and Marmot, M. (1998). *The Solid Facts*. Geneva: World Health Organization.
Wilkinson, R.G. and Pickett, K.E. (2008). Income inequality and socioeconomic gradients in mortality. *American Journal of Public Health*, **98**, 699–704.
Wilson, L.M., et al. (2012). Impact of tobacco control interventions on smoking initiation, cessation, and prevalence: a systematic review. *Journal of Environmental and Public Health*, **2012**, 961724
Wilton J., et al. (2015). Pre-exposure prophylaxis for sexually-acquired HIV risk management: a review. *HIV/AIDS*, 28(7), 125–36.
Winkleby, M.A., Fortman, S.P., and Rockhill, B. (1992a). Trends in cardiovascular risk factors by educational level: the Stanford Five City Project. *Preventive Medicine*, **21**, 592–601.
Wodak, A. (2006). Controlling HIV among injecting drug users: the current status of harm reduction. *HIV AIDS Policy and Law Review*, **11**, 77–80.
World Bank (1993). *World Development Report 1993. Investing in Health*. New York: Oxford University Press.
World Bank (2018). *Poverty*. [Online] Available at: http://www.worldbank.org/en/topic/poverty/overview
World Health Organization (2000). *Obesity: Preventing and Managing the Global Epidemic*. WHO technical report series no. 894. Geneva: WHO.
World Health Organization (2005). WHO Framework Convention on Tobacco Control. [Online] Available at: http://apps.who.int/iris/bitstream/handle/10665/42811/9241591013.pdf?sequence=1
World Health Organization (2008). *WHO Report on the Global Tobacco Epidemic, 2008: The MPOWER Package*. Geneva: World Health Organization.
World Health Organization (2015). World Health Statistics reports on global health goals for 194 countries. [Online] Available at: http://www.who.int/mediacentre/news/releases/2015/world-health-statistics-2015/en/
World Health Organization (2017). [Online] Available at: The top 10 causes of death. http://www.who.int/en/news-room/fact-sheets/detail/the-top-10-causes-of-death

2.11

How access to healthcare affects population health

Yvonne Inall, Rachel Lamdin Hunter, Stephen Leeder, and Angela Beaton

Introduction

This chapter explores access to healthcare—the timely availability of professional health services to prevent, diagnose, and treat illnesses and to preserve or improve the health of individuals—and its consequences. The interplay between access to care and population health has been seen dramatically in the Covid-19 pandemic. Systems of care have been stressed to breaking point with loss of life and failure of confidence in the public health system.

Access varies among and within countries. The challenge of providing access in relation to both communicable and non-communicable diseases is formidable (The Economist Intelligence Unit 2017).

Providing access is complex. Which groups or organizations are responsible for provision (the enablers), and how might access be financed? More people are living longer, with long-term comorbidities. The disease burden in developing countries is shifting. By 2020, non-communicable diseases are likely to account for approximately 70% of preventable, premature deaths (WHO 2018). How much care should be provided, and at what level? Is access confined to a minimum set of services or does it include more expensive and innovative treatments?

Countries and their governments vary in their political and social attitudes to access. Those with a strong social welfare agenda might focus attention and budgets on social factors such as education and social welfare; these not only determine health, but also access to care. Others, regarding healthcare as the individual's concern, might give these factors less weight. These variations in response have been apparent in the Covid-19 pandemic.

Despite this complexity and much national variation, access depends, in general, most often on good primary healthcare, public investment, and political will, availability of data to guide resourcing decisions, and the development of a well-trained, integrated workforce with appropriate supporting infrastructure.

Definitions

Primary healthcare

Our working definition of primary healthcare is ‘the first point of contact an individual has with the healthcare system’. It might be provided by a physician, community nurse, or by other clinical or non-clinical allied health professionals.

Effective primary healthcare provides accessible, comprehensive, culturally acceptable, community-based services to meet the health needs of individuals throughout their life-course—from family planning and infant immunizations to management of the comorbidities of the elderly and to palliative care. It also serves as the entry point for care in manging endemic and epidemic infectious disease, such as Covid-19, including vaccination.

The watershed WHO Conference on Primary Health Care held at Alma-Ata in the then USSR in 1978 defined primary health as the first level of contact with a healthcare system:

> Essential healthcare based on practical, scientifically sound and socially acceptable methods and technology made universally accessible to individuals and families in the community through their full participation and at a cost that the community and country can afford to maintain at every stage of their development in the spirit of self-reliance and self-determination. It forms an integral part both of the country’s health system, of which it is the central function and main focus, and of the overall social and economic development of the community. It is the first level of contact of individuals, the family and community with the national health system bringing health care as close as possible to where people live and work, and constitutes the first element of a continuing health care process. (WHO 1978)

The breadth of services and professions which fall within the definition means that the nature of the services might vary substantially between and within countries. This variability creates major challenges for data collection and assessment, financing, policy development, and decision-making.

Guiding principles in primary healthcare

Convened by Dr Halfdan Mahler, then Director-General of the World Health Organization, the Alma-Ata conference was a response to the social target set the previous year at the Thirteenth World Health Assembly, calling for ‘the attainment of all citizens of the world of a level of health that will permit them to lead a socially and economically productive life’ by the year 2000 (WHO 1988). Representatives from 134 UN member states and 67 UN organizations attended this seminal meeting. The objective was to agree on a set of principles for primary healthcare, to be endorsed by WHO and UNICEF. The principles agreed at Alma-Ata included:

accessibility and fairness;
relevance of the service to the needs of the population;
cost effective use of technology;
integration between primary, secondary, and tertiary care;
encouragement of self-care;
community involvement in the planning and provision of health services;
collaboration between the health and other sectors in addressing underlying causes of ill-health; and
redistribution of healthcare resources to primary healthcare, in rural and economically disadvantaged areas.

These aspirations were widely endorsed (Bhutta et al. 2018), but 40 years after Alma-Ata, primary healthcare remains inaccessible to at least 400 million people. The reasons include:

a lack (or misdirection) of public investment;
provision of inappropriate service models; and
disproportionate concentration of workers and resources in urban areas.

Secondary healthcare

We define secondary healthcare as care by specialists to whom patients are referred by a primary health carer. Depending on the field of specialist practice, it might take place in a hospital, a stand-alone purpose-built, or modified facility, or the specialist’s consulting rooms. Obtaining secondary

care can be challenging, depending on geography, transport, and funding (Araujo et al. 2018).

The relationship between primary and secondary care providers operates both ways. Patients might not only be referred to secondary care providers by primary health carers, but secondary healthcare providers might refer patients to (or expect them to utilize) primary healthcare services following discharge from hospital, or as part of a coordinated care plan, drawing upon both primary and secondary services.

A study of the relationship between primary and secondary care provision in Norway showed that primary healthcare workers felt that hospital-based carers 'do not speak [our] language' (Araujo et al. 2018). Breakdowns in communication were common. Primary and secondary carers had different ways of understanding patients' needs. Efforts to improve coordination revealed problems relating to information technology, municipal, hospital and fiscal policy, and resistance to cultural change, all of which adversely affected attempts at harmonization.

Equity

Equity in healthcare can be evaluated in several ways (Table 2.11.1). An important distinction, attributed to Aristotle, is between horizontal and vertical (Gillon 2005). Horizontal equity requires that equals should be treated in proportion to their equality, while vertical equity requires that unequals should be treated in proportion to their inequality.

In vertical equity, those with the greatest needs receive the greatest attention. Horizontal inequality, like inequality in income or health, is often measured using Gini-like metrics known as concentration indices (Mackenbach and Kunst 1997). In reality, different groups in a population have different needs and require appropriately differentiated services; for example, the healthcare needs of indigenous Australians, whose life expectancy at birth is some 10.6 years less than that of the general population (AIHW 2018). Mooney (1996) has asked about a fair distribution of health resources, given this large difference in health status?

There is little consensus on the vertical dimension of equity. Marginalized groups, such as indigenous populations, new migrants, immigrants who have not mastered the language of the host country, poorly educated people, the disabled, homeless, or prisoners might be regarded as

‘hard to reach’ or at least ‘underserved’ by standard services. In such cases, there might be a case for targeted services. There is, however, a concern that such targeted services can be seen as ‘poor services for poor people’, with the attached stigma. More generally, there is an increasing emphasis on designing local services to meet locally expressed needs, leading to greater variation in organization and delivery, contrasting with the uniform approach implied by the pursuit of horizontal equity.

In the United Kingdom, the tension between these two dimensions of equity was addressed through the introduction of national level service specifications, which accommodated local discretion in implementation. However, the National Health Service (NHS) has come under increasing pressure resulting from austerity economic cuts to funding, staffing shortages, an ageing UK population and increasing social inequality (Iacobucci 2019).

Ivan Illich, a radical critic of Medicine, claimed, ‘a vast amount of contemporary care is incidental to the curing of disease, but the damage done by medicine to the health of individuals and populations is very significant’ (Illich 1976). He acknowledged that Medicine, using vaccinations and antimicrobial drugs, aids in preventing and curing infectious diseases, but he argued that big reductions in mortality from infectious diseases in high-income countries had already occurred before these technologies became available. In his opinion, treatment of non-communicable diseases, such as cancer and cardiovascular disease, was of negligible benefit and indeed might cause considerable harm.

Richard Smith, former editor of *The British Medical Journal*, an iconoclast and a fan of Illich, put it this way:

> Health, argues Illich, is the capacity to cope with the human reality of death, pain, and sickness. Technology can help, but modern medicine has gone too far—launching into a godlike battle to eradicate death, pain, and sickness. In doing so, it turns people into consumers or objects, destroying their capacity for health.
>
> Limits to medicine. Medical nemesis: the expropriation of health. R Smith http://dx.doi.org/10.1136/jech.57.12.928

Illich’s opinion was supported by the controversial English physician, historian, and epidemiologist Thomas McKeown, in his analysis of historical trends in mortality in Britain (McKeown 1979). This analysis identified the limited role of Medicine as a past determinant of health. For example, relative reductions in mortality between 1881 and 1950 ranged

from 30-fold for tuberculosis, 20-fold for digestive diseases, and 15-fold for respiratory diseases (ONS 1998). Most of these reductions occurred before the widespread use, after 1945, of disease-specific antimicrobial treatment.

Archie Cochrane, famous today for his advocacy of the use of the randomized trial (RCT) in assessing the effectiveness of therapy, developed a somewhat similar argument in *Effectiveness and Efficiency* (Cochrane 1972). Cochrane advocated the now generally accepted view that, if at all possible, all care interventions should be tested in RCTs, and that the results of all such trials should be systematically collected, analysed, and implemented in clinical practice. Only recently, with the advent of genetic-based 'precision medicine', has a serious advance been made over RCTs in assessing the utility of proposed new therapies. Precision medicine, unlike RCTs, is specific to individuals.

Table 2.11.1 Aspects of equity in health and healthcare

Term	Meaning
Equity	Fairness, or justice, in respect of treatment of different individuals or groups (Last 2007)
Horizontal equity	The extent to which equals are treated in proportion to their equality
Vertical equity	The extent to which unequals are treated in proportion to their inequality
Equity in financial contribution	The extent to which individual or household payments for care or health insurance are consistent with their capacity to pay
Equity in access to healthcare	The extent to which there is a fair distribution of access to healthcare in relation to need
Equity in health	The extent to which there is a fair or just distribution of opportunities for health among individuals and groups in a population
Effective coverage	The proportion of the population in need of an intervention that has received it (Shengelia et al. 2005)

Access to primary healthcare

In general terms, ‘access to healthcare’ is said to exist when individuals or families can physically get to the services they need to treat illness or to maintain health (Gulliford et al. 2002). An ongoing debate seeks to distinguish access from use of services. At the simplest level, having access might be assessed in terms of the ready availability of services (Table 2.11.2).

Obstacles to use include cost; physical barriers, including distance or difficulties of travel; personal barriers, when services are not felt culturally appropriate, socially acceptable, or not consistent with personal beliefs; or organizational barriers, such as difficulties obtaining a consultation, or delays because of waiting lists or the limited capacity of services (Aday and Anderson 1981; Pechansky and Thomas 1981). Utilization offers benefit only when effective care is relevant to needs (Table 2.11.2).

‘Access’ has recently been reappraised. It is proposed that access be understood in terms of individuals’ ‘candidacy’, namely their eligibility for healthcare (Dixon-Woods et al. 2006). Candidacy is defined both by individuals’ perceptions of their own concerns, as well as by health providers’ views of the appropriateness of their services to meet those concerns. In short, ‘access’ is meaningful when a person perceives a need and care is available.

Table 2.11.2 Dimensions of access to healthcare

Dimension of access	Meaning	Potential indicators
Availability	Whether or not there is an adequate supply of health services in an area	• Number of physicians or nurses per 1,000 population • Number of hospital beds per 1,000 population • Distance to nearest primary healthcare provider • Distance to nearest hospital
Utilization	Whether or not health services are utilized	• Primary healthcare consultations per l,000 population • Hospital admissions per 1,000 population
Relevance to need	Whether or not appropriate services are received by the people who need them	• Proportion of births attended by a trained healthcare professional • Proportion of individuals with known hypertension who are being treated
Outcomes	Whether achievable health outcomes are realized	• Maternal mortality rates are improved • Mortality rate from appendicitis is improved

Access to secondary healthcare

No matter how enthusiastically primary healthcare is developed and made accessible, the range of care provided stops short of major surgery, complex obstetrics, intensive medical care, sophisticated rehabilitation services, and long-term accommodation-based care.

In many low-resource settings, where community primary healthcare is rudimentary and where access to quality care is limited, the emergency departments of hospitals overflow with patients who might otherwise be cared for in primary healthcare.

At any of these institutions in resource-poor countries, we find mothers with sick (or very sick) children. They might have travelled days from remote villages to urban hospitals. The hospital becomes a *de facto* primary health carer. Nor is this phenomenon unique to low income countries. In resource-rich countries, care given to patients with serious and continuing illnesses is being reviewed to see if it could be given more efficiently through primary healthcare, especially when the services in both settings are linked and managed as a unit. Treatment for such patients is judged to be inappropriate when they receive expensive hospital care as a substitute for primary healthcare. There is no simple answer: primary healthcare services are increasingly limited to daytime working hours. At other times, the emergency department can, by default, offer primary healthcare.

Changing global and community disease burdens

The health gain of resource-rich countries in the twentieth century is well documented. A snapshot of the interval between the Second World War and now would also show improvements in health outcomes in developing countries, in parallel with increased education, sanitation initiatives, and improved local oversight of public health (Bollyky 2018). These improvements were closely related to rising prosperity and urban development. With money to spend, many such governments invested in potable water and in sewerage disposal. In turn, a healthier workforce contributed to prosperity.

A new set of problems has come with these advances. Chronic illnesses, once the scourge of advanced nations, are now prevalent in low- and middle-income countries. This reflects changes in food supply (including a greater proportion of highly processed foods), increasing availability of tobacco, and housing shortages. Many less economically developed nations face a loss of productive workforce in middle age due to chronic non-communicable diseases such as diabetes and cardiovascular diseases. It has been projected that, by 2040, 80% of premature deaths and disability in developing countries will be caused by heart disease, cancer, and other non-communicable diseases (Bollyky 2018).

In low- and middle-income economies, healthcare expenditure per capita remains relatively low and much chronic illness is untreated. Many less

developed nations spend about US$30 per person annually on healthcare, contrasted with US$5,400 in the United States and US$3,800 in the United Kingdom (Bollyky 2018). The opportunities for primary and secondary prevention through primary healthcare are often lost.

This change in the pattern of illness has slowly been recognized. The shift in funding away from preventing and treating communicable diseases to treating and managing the growing burden of non-communicable diseases (NCDs) is also noteworthy. Initiatives, like the Global Alliance Against Chronic Disease, are addressing the threats posed by NCDs. This body has 'arranged three of its four funding rounds by disease' (Fairall et al. 2018). But, as has been seen with efforts to control HIV, the difficulties imposed by health systems which do not integrate care for the whole individual can be almost as troublesome as the underlying disease. In the case of NCDs, the usual pattern is of an individual having multiple disorders (diabetes, heart, and lung disease, for example) which need linked care. Ideally, services should respond to multimorbidity rather than to individual diseases.

Citizens' 'right to healthcare'

Does any individual (or community) have a 'right to healthcare' or a right to access to care? Answers to this question are generally approached from either a utilitarian or human rights-based viewpoint. Another, objectivist view, sees healthcare as a commodity, to be bought or sold in an open market. This section will consider these different perspectives in the context of access to healthcare.

Human-rights-based approaches to a right to healthcare draw on two core documents: the Universal Declaration of Human Rights (UDHR) and the International Covenant on Economic, Social, and Cultural Rights (ICESCR). Rights to medical care are recognized under Article 25 of the UDHR:

> Everyone has the right to a standard of living adequate for the health and well-being of himself and of his family, including food, clothing, housing and medical care and necessary social services, and the right to security in the event of unemployment, sickness, disability, widowhood, old age or other lack of livelihood in circumstances beyond his control. (UN 2007)

The UDHR was adopted by the UN General Assembly in 1948, in the wake of the atrocities perpetrated during the Second World War. It is not a legally binding document, although it is generally accepted as holding status as customary international law, which is, regrettably, hard, if not impossible, to enforce. Enforcement is generally limited to violations linked to war and armed conflict (Gunning 1991; Henckaerts 2005).

Article 25 ties health and medical care directly to living standards and social services, welfare, and security, but without explicitly stating which governing entity should be responsible for funding and provision.

The International Covenant on Economic, Social, and Cultural Rights (ICESCR adopted by the UN General Assembly in 1966, and in force since 1976) expands upon the UDHR, requiring party states to 'recognize the right of everyone to the enjoyment of the highest attainable standard of physical and mental health' and to create 'conditions which would assure all medical services and medical attention in the event of sickness' (Office of the High Commissioner for Human Rights 2007); see also Chapter 3.3).

Unlike the UDHR, the ICESCR is a 'binding treaty', which, as of December 2018, has been ratified by 169 member nations (United Nations Human Rights, Office of the High Commissioner n.d.). Four nations, including the United States of America, have signed (but not ratified) the treaty. States party to the treaty have agreed to submit a compliance report to the ICESCR 5-yearly. Alleged violations might be referred to the ICESCR committee for investigation. However, agreement on how violations should be recognized or quantified is lacking, and enforcement is non-existent (Kalantry et al. 2010).

Taken together, the UDHR and ICECSR do not guarantee access to necessary services for every individual but, as statements of political and social *intent*, they offer an important degree of hoped-for protection for marginalized and vulnerable groups and individuals. There can, of course, be no guarantee of limitless care for any individual, nor does the resolution overcome the tyranny of distance where rural dwellers cannot access health services in the big cities. But the ICECSR places responsibility on governments to facilitate the healthcare of their citizens (discussed further in Section E). Ultimately, however, human rights, including the 'right to healthcare', are weakly protected at the international level.

Human rights-based approaches to access to healthcare link to concepts of equity and draw upon social contract theory and theories of ethics. These

approaches favour a just or equitable distribution of healthcare over the maximization of potential health gains for a favoured minority.

American moral and political philosopher John Rawls's theory of justice and fairness, proposes that a just society will be arranged so as to achieve fair equality of *opportunity* (Rawls 2007) not equality of *achievement* (for example, equal health outcomes from a particular treatment, which can never be guaranteed). Although saying little about health, Rawls envisioned his theory of justice as an extension of the social contract, as understood by Locke, Rousseau, and Kant (Rawls 2007).

Norman Daniels, an eminent American political philosopher and philosopher of science, has been more explicitly concerned with health than was Rawls (Daniels 2008). He argues that health is an important determinant of the extent of an individual's opportunities and that it therefore requires careful consideration if one wishes to establish a fair society. From Rawls and Daniels, equity stands out as a major 'ethical value', one which healthcare services should strive to promote (Braveman et al. 2001).

Cruess and Cruess, an orthopaedic surgeon and an endocrinologist at Canada's McGill University, respectively, have written extensively on medical professionalism, acknowledge that the relationship between the medical profession and the social contract remains poorly defined, emphasizing that 'the social contract between medicine and society differ between countries, being influenced by cultural, economic, and political factors' (Cruess and Cruess 2008).

In direct contrast to rights-based and social contract theory approaches to healthcare stands the objectivist position in which healthcare is not a 'right', but rather a 'service' offered by trained professionals. Rather than focussing on the 'rights' of the recipient of care, this approach considers the 'rights' of those who provide care. In this approach, healthcare is a commodity provided by trained persons who can expect remuneration and respect for their services, and who have the right to practise their profession where and when it is legal to do so. They also are entitled to withhold those services. They have the same entrepreneurial legal rights as any non-medical professional.

This objectivist approach to healthcare resonates especially in the United States, where it was clearly articulated by Robert Sade, an American surgeon and bioethicist, in the *New England Journal of Medicine* in 1971

(Sade 1971). Sade argued ‘the concept of medical care as a right of the patient entails the use or threat of violence against physicians [by the state]’. In Sade’s view, the rights of the physician to remuneration for service, and freedom to choose to whom to provide it, exceed the individual’s right to care. Gaffney describes this contrast as the ‘rights-commodity dialectic’, drawing on Gøsta Esping-Andersen’s arguments around the de-commodification of welfare services, facilitating subsistence without reliance on markets (Esping-Andersen 1990; Gaffney 2018). Sade came to describe this same contrast as one between ‘central planning’ and the ‘free market’, driven respectively by beliefs that ‘healthcare is an entitlement that must be provided by society and that obtaining healthcare is the responsibility of individuals’ (Sade 2012).

Adam Gaffney, an instructor in Medicine at Harvard Medical School and a pulmonary and critical care doctor at the Cambridge Health Alliance, is president of the advocacy organization Physicians for a National Health Program. Writing about the history of healthcare as a right, he saw these two opposing positions as mutually exclusive (Gaffney 2018). Similarly, Sade’s recent work posits that government regulation of healthcare and ‘everything else that contributes to good health’ is ‘inconsistent with the notion of a free society’, arguing that regulations and the funding of services via taxation impinge on the freedoms of individuals (Sade 2012).

Much of the debate around the rights and provision of healthcare, *vis-a-vis* the rights of the provider, centre specifically on the role and remuneration of physicians, as clearly encapsulated in Sade’s position (Sade 1971). Allied health professionals are largely absent from the debate. Cruess and Cruess are properly concerned that the medical profession is viewed, almost in parody, as a homogenous monolith, rather than a complex, heterogeneous matrix of interlinked professionals, whose needs, desires, and expectations might not be unified (Cruess and Cruess 2008).

Gaffney notes a further complication in the discussion about the rights of the recipient of care and the provider. Whereas Sade ‘envisions healthcare sold by the individual physician to the individual patient’ (that is, an autonomous, sole-trader provider), the reality is that healthcare is more likely today to be provided by a state or a corporate entity, which employs providers, whose work may be constrained by corporate/government policy.

Similarly, healthcare providers are accountable, not only to their patients, but also to regulatory bodies, government and corporate agencies, including

insurers (Cruess and Cruess 2008). Sade's articulated right to withhold care might also collide with antidiscrimination legislation, if withholding care could be considered illegal because it was judged to be discriminatory. Ultimately, Sade's hypothesized example does not readily align with health systems, nor with the provision of healthcare as it is practised across the developed world. Sade himself acknowledged in a 2012 article, following the introduction of the Affordable Care Act (informally known as 'Obamacare') in the United States, that government had increasingly taken control of healthcare, policy development, and regulation of the health system, although he continues to argue that deregulation will facilitate innovative free market solutions 'guided by the ethical principle of non-interference' (Sade 2012). This libertarian perspective continues to appeal, particularly in the United States, although many such arguments fail to recognize that many corporations now rival nations in terms of their power, influence, and interference in the lives of private citizens (Babic et al. 2017). Indeed, a 2012 article in the *British Medical Journal* introduced the term 'commercial determinants of health' into the lexicon, sitting alongside the social determinants, in recognition of the 'increasingly unhealthy dominance of corporate marketing' (Hastings 2012). Each side of the rights-commodity divide sees fundamental flaws in the opposing position's arguments—a gulf which is unlikely to be reconciled in the foreseeable future, and a debate with which each community and nation must continue to struggle.

Utilitarian perspectives on access to care contend that health services should maximize the health gains obtained by all individuals. This idea is summed up by the slogan 'the greatest good for the greatest number', sometimes regarded as 'the guiding principle for many of the decisions and actions of public health professionals' (Last 2007). The objective of maximizing the health gains to be achieved from available resources is closely related to the economic principle of efficiency, which is the focus of Chapter 6.6 of the current volume. Resources for healthcare will be used most efficiently when health gain is maximized. Equity and efficiency are often viewed as conflicting objectives. For example, delivering equitable access to specialist services in sparsely populated rural areas might compromise efficiency.

The utilitarian approach raises problems because it requires only that the sum of health gains in a population should be maximized. The individual is

not of ultimate concern to the utilitarian. This could result in inefficient allocation of healthcare to individuals with greater needs; the lesser needs of many might be met instead.

By contrast, approaches based on concepts of human rights and social justice view each individual as having a right to healthcare—those needing more will have greater access to care. Proportionality of access was a key concern of the late health economist Gavin Mooney for whom ‘equal access to equal care for equal need’ served as his definition of equity, a definition with which we agree. It was not a guarantee of equal outcomes, biologically impossible to promise, but of equal access.

Responsibility and role of governments

Under the terms of the ICECSR treaty, participating governments bear some responsibility for providing a measure of healthcare to their citizens. However, the treaty does not prescribe how this should be organized, or to what extent governments should fund health systems. The United Nations’ resolution on 12 December 2012 on Global Health and Foreign Policy encourages member states to promote universal health coverage in foreign policy development. The resolution highlights, in particular, the impact of the high cost of care as a barrier to access (UN 2012). Ultimately, each sovereign government must determine how, and to what extent, it is willing or capable of organizing and funding healthcare. The result, globally, is a mix of private, public, and co-payment systems, some tightly controlled by governments and others with minimal regulation or government direction.

The United Nations Sustainable Development Summit, held in New York in 2015, ushered in a series of Sustainable Development Goals (SDG) (UN 2015), (the Millennium Development Goals (MDGs) (UN n.d.) having run their course), which addressed a raft of global challenges to be met by 2030. Target 3. aims to:

> Achieve universal health coverage, including financial risk protection, access to quality essential health-care services and access to safe, effective, quality and affordable essential medicines and vaccines for all. (United Nations (UN), Sustainable Development Goals)

The SDGs call for greater national investment in health and in improvement and maintenance of the health workforce. Success will depend on the extent

to which countries ratify the goals, invest in them, and implement strategies to achieve them.

John McDermott, an expert on public health, agriculture, nutrition and health in Africa, wrote in 2018 ‘universal healthcare is both desirable and possible, even in low income countries’ (McDermott 2018). Stigler et al. propose that universal healthcare is not possible without adequate primary healthcare, which they see as ‘the backbone of an effective health-care system’ (Stigler et al. 2016).

The Economist Intelligence Unit (EIU) report, *Global Access to Healthcare: Building Sustainable Health Systems* (2017) mentions primary healthcare as the ideal basis for a country’s health system—a point where governments can control funding (The Economist Intelligence Unit 2017). Starfield et al. note that health systems which emphasize primary healthcare are more efficient, widen access, reduce inequality, and lower the overall costs of care, while producing better population health and reduced rates of mortality—findings applicable in both national and cross-national analyses (Starfield et al. 2005).

Atun et al. charted the path to universal healthcare in Latin America from the 1980s, through to the early twenty-first century (Atun et al. 2015). Health reforms in Latin America were heavily influenced by the Alma-Ata declaration and therefore placed a strong emphasis on the role of primary healthcare (WHO 1978; Atun et al. 2015). The design and implementation of these universal healthcare systems (UHCs) and the means by which they were funded varied significantly.

With a significant surplus of physicians, the Cuban National Health Service, unlike the healthcare services in Venezuela (see next) has been particularly successful. Following the 1959 revolution, the right to healthcare was enshrined in Cuban law, and universal health services, financed via taxation, began to be developed in the 1970s. Although private healthcare is illegal in Cuba, private spending makes up approximately 10% of all healthcare expenditure. Comprehensive primary healthcare is the lynchpin of the system, with 80% of health problems being managed through healthcare networks centred on polyclinics (Atun et al. 2015).

Extended value of primary health services

Broad definitions of primary healthcare include ‘all the activities whose primary purpose is to improve or maintain health’ (Murray and Frenk 2000), including sanitation, road safety initiatives, or the education of girls. Such multisector interventions, while generally considered to contribute to a community’s health, do not contribute to the care of individuals.

Primary healthcare yields benefits beyond individual recipients. The treatment of pulmonary tuberculosis, for example, reduces the spread of the disease. Such positive ‘externalities’ (outcomes resulting from an intervention in one area but which affect other areas), might be more beneficial for communities than for individuals. Countries might facilitate access to these benefits as ‘merit goods’, underproduced if individuals valued only their own personal benefits.

Extending this argument, it might be desirable to organize care so that it benefits all. For example, the eradication of infectious diseases, or the control of antimicrobial drug resistance, offer benefits extending beyond national boundaries (Smith et al. 2003).

Importance of policy

Section 2 of the current volume examines themes surrounding the determinants of health and disease. Chapter 2.1 examines, in depth, the social determinants of health and Chapter 2.2 explores the relationship between politics and public health. Health inequalities, in both the developed and the developing world, were scrutinized in Chapters 2.4 and 2.5. Health policy has a role to play in each area. Chapters 3.5 and 3.6 examine health policy in greater detail. In this section, we examine the role of health policy as it relates to access to care.

Policies should aim at overcoming barriers to access, many of which intersect with the broader social determinants of health. Poverty, social inequality, location—especially rurality and environment—all affect access.

Low- and middle-income countries wishing to achieve SDG 3.8—the provision of universal healthcare (UHC)—will need policies to lead them towards that goal. Reich et al. examined 11 countries’ pathways towards UHC. These offer lessons for future transitions (Reich et al. 2016). The policy challenges are substantial and wide-ranging. Reich et al. underscore the need for policymakers’ ‘vision and commitment’ to ensure the

successful and sustainable introduction of UHC (Reich et al. 2016). Policies need to address medical and allied health worker training, accreditation, and regulation; service provision; taxation; building the public's trust; and commitment from the public. Starfield et al., also in a comparison of 11 nations, found that 'adequate delivery of primary healthcare services was associated with supportive governmental policies' (Starfield et al. 2005).

To give an indication of the vital importance of policy, we briefly examine the case of Venezuela. During the 1990s, Venezuela began to reform its health system. In 1999, the Venezuelan Constitution recognized healthcare as a fundamental human right (Atun et al. 2015). Funding policies drew on a number of sources, pooled to fund healthcare for the poor, the unemployed, and the rural population. The system was a mix of public and private funding, but with out-of-pocket expenses accounting for more than 50% of health expenditure (Atun et al. 2015). From 2003, a primary healthcare model, *Barrio Adentro*, was introduced, which, for a time, expanded access and resulted in better health outcomes for Venezuela's poorest citizens (Atun et al. 2015). Life expectancy rose and infant mortality dropped dramatically.

By 2013, however, the system was faltering, and by 2016 it was, and still remains, in crisis (Editorial 2018). Economic collapse followed the decline in world oil prices from 2008. For this and other reasons, government spending on public health dropped from 9.1% in 2010 to 5.8% by 2014 (Editorial 2018). The result has been a devastating increase in out-of-pocket payments—a barrier to access for poorer citizens. Maternal and infant mortality have increased. Access to medicines, even for the wealthiest citizens, has diminished due to significant shortages of medications (Fraser and Willer 2016; Editorial 2018). Rather than acknowledge the crisis, the government has ceased gathering data; nor has it, since 2016, reported on the state of the system. Offers of international aid have been declined. The wilful refusal to acknowledge or address the crisis exemplifies a spectacular policy failure. Costing lives it serves as a reminder of the importance and impact which policy implementation can have on access to even the most basic healthcare services.

Financing primary healthcare

Where governments assume a funding role for primary healthcare, policy becomes extremely important. Sustainability is a recurrent challenge. Reich et al. observed that sustaining coverage was a financing and policy issue for each of the 11 nations included in their study (Reich et al. 2016). Decisions undertaken during the development of Ghana's National Health Insurance System resulted in a fee-for-service model, which has proven unsustainable. Ten years after the adoption of the system, the government had to reassess a financially unsustainable system, which did not cover 64% of the population (Reich et al. 2016). Spain implemented a tax-based funding system during the 1980s and 1990s to both expand primary healthcare and ensure the sustainability of the system as a whole (Starfield et al. 2005). Cuba and Brazil similarly fund their universal health system via taxation (Atun et al. 2015).

Cuba has supplemented their taxation funding for healthcare through the provision of healthcare workers to Brazil and Venezuela in exchange for contracted fees paid to the Cuban government (Atun et al. 2015).

The distribution and organization of the healthcare workforce, both state-certified and non-regulated

The Institute for Health Improvement's Triple Aim (Institute of Healthcare Improvement 2009) supports, across differently-resourced regions and populations, the judicious use of 'regulated' (state-certified) and 'non-regulated' health workers. The latter include pharmacy technicians, community health staff, and volunteers.

Throughout the twentieth century, health services have largely been provided by regulated practitioners, such as nurses, midwives, and doctors. But, with need outstripping supply, this workforce is now inadequate, overtrained, and expensive to educate and remunerate. Consequently, workforce development strategies have been developed by the World Health Organization (WHO) to include unlicensed healthcare workers, trained to perform set tasks under the direction and supervision of regulated practitioners. For example, regulated Nigerian health workers make use of churches, mosques, and the premises of their civic organizations as their clinics. Traditional birth attendants assist at childbirth. Community support

is strengthened by recruiting community leaders and prominent citizens as their advocates (Gyuse et al. 2018).

In New Zealand, regulated practitioners, including nurses, midwives, and doctors, may be scarce in rural and remote areas, so the Ministry of Health's strategies make extensive use of non-medical personnel. For example, 'Quitline' advisors available by telephone, text and online, provide smoking cessation advice; and over 50,000 community-based *Kaiāwhina*, or non-regulated care workers and assistants (mainly women), support people with disabilities or chronic illnesses (Health Workforce New Zealand 2018).

Similarly, sudden unexpected death in infancy (SUDI) has been a leading cause of preventable death, the number of SUDI deaths reduced from 250 per year (over 4.4 per 1,000 live births) in the 1980s, to 44 deaths in 2015 (0.92 per 1,000 live births). Although SUDI rates have reduced significantly, they did not fall equitably across the population. The SUDI risk for Maori and Pasifika babies has remained consistently higher than for non-Maori, non-Pasifika babies. Identifying and promoting more effective, culturally appropriate ways of engaging, and staying engaged, with women most at risk of SUDI early in their pregnancies is crucial. A focus on this has resulted in the integration of Kaiiiwhina as community educators into Ministry of Health-directed policy and practice (Health Quality and Safety Commission New Zealand 2017). Likewise, initiatives, including the construction of *wahakura* (woven baskets) for infants to sleep in beside their mothers or other carers, have strengthened community networks for improving health (Abel and Tipene-Leach 2013). Ideally, non-regulated workers belong to the communities of the people they serve, thereby avoiding cultural gaps.

With adequate remuneration, training, and supervision, sufficient resourcing, defined agreed-upon tasks and the support of the community and colleagues, non-regulated workers can successfully manage outbreaks of illnesses such as malaria, and provide preventive care, including vaccination, perinatal care and health promotion (Haines et al. 2007). In a poor rural community in Pakistan, with little access to mental healthcare, integration of a cognitive behaviour therapy-based intervention into the routine work of community health workers more than halved the rate of depression in prenatally depressed women compared with those receiving enhanced 'usual' care (Rahman et al. 2008). These types of initiatives

demonstrate the potential to integrate similar roles for community health workers into other health systems, particularly in resource-poor settings.

In 2010, WHO called for the more effective use of non-regulated healthcare workers. It stressed an urgent, global need to strengthen primary healthcare, to integrate interprofessional education and collaborative practice into service, education, and health policy. That need has not abated.

In the light of budgetary constraints, primary healthcare remains one of the best investments governments can make (The Economist Intelligence Unit 2017; Mossman et al. 2017). For example, initiatives to tackle the dearth of medical practitioners in rural areas can focus on non-medical providers (Sanne et al. 2010; Fairall et al. 2012), such as nurse practitioners, to provide certain types and levels of care. New Zealand, with a population of nearly five million, including 15,000 doctors, employs more than 300 nurse practitioners throughout the country. Nurses qualify as practitioners upon gaining a clinical master's degree in addition to the three-year bachelor's degree. Those in primary healthcare can diagnose, treat, and prescribe and, like general medical practitioners, might be the lead healthcare provider. Each costs a fraction of the cost of educating and maintaining a medical practitioner. These nurses are often residents of the rural communities and small towns to which, following education and registration, they return.

Technology can support access to primary healthcare. For example, the *Mohalla* clinics in Delhi operate low-cost kits for common medical tests (Lahariya 2017), reducing the costs of screening and treatment for diabetes, dengue fever, and hepatitis. In Rwanda and many other countries, telehealth is being trialled to carry medical advice to rural and remote patients (The Economist Intelligence Unit 2017). Like the Royal Flying Doctor Service of Australia, one of the largest and most comprehensive aeromedical organizations, telehealth is an important mechanism for enhancing access, team-based care, and quality of care for patients in rural and remote Australia (Bradford et al. 2016).

Services such as the North Star Alliance, based in sub-Saharan Africa, support low-level continuing care, delivered via programmes which align with patients' experiences in a culturally responsive way, providing local support, which requires less travel, resulting in shorter waiting times. The North Star Alliance also uses clever branding and marketing to reach patients and the communities they interact with. Their not-for-profit

network includes more than 30 roadside health clinics at disease hotspots on major transport routes used by mobile workers, travellers, and displaced persons who frequently cross borders, thereby increasing the risk of disease spreading across borders (North Star Alliance 2018). Such services offer the convenience of enabling the patient's daily life to continue.

The Le-Nest initiative in Shanghai offers an inexpensive service for patients whose diabetes and hypertension can be managed in primary healthcare (International Partnership for Innovative Healthcare Delivery 2014). Such services are readily accessible for either an acute, minor problem, or a flare-up of a long-standing illness. Similarly, the Practical Approach to Care Kit (PACK) programme, offered through *The British Medical Journal*, offers non-physician clinicians in low- and middle-income countries a suite of checklists and algorithms for screening, diagnosis, and treatment for a range of commonly presented symptoms, acute and chronic conditions that can be managed within the community (Fairall et al. 2018).

The role of data in planning, monitoring, and evaluating services

To try to assure the quality of primary healthcare, health key performance indicator (KPI) data are used internationally. They allow comparison of performance among different locations as well as system-wide improvements and reforms. They can establish goals for access to care and specify the extent, type, and cost of care (Hibbert et al. 2013). They can be linked to clinical and financial measures of achievement and outcome. The range of KPIs is widening as a result of the introduction of new technologies, and of expanding the expectations of those who use and pay for services.

The place of KPIs in assessing the managerial machinery, clinical processes, and financial performance of healthcare systems is deeply entrenched. They have tended to relate to everything except the details of what the health system achieves in improving health and quality of life of patients. These provoke profound clinical scepticism, especially when accompanied with demands for copious data collection by busy clinicians. There is a tendency to measure what can easily be measured. Important

qualitative aspects are pushed aside by narrow quantitative measures. This often reduces the focus to concerns with process and activity rather than with access or outcome.

Despite these limitations, the need for KPIs is clear. Comprehensive primary healthcare using integrated, team-based services for those with multiple comorbidities is short on evidence of its effectiveness. Evaluation using performance indicators, reflecting both equity, outcome, and accessibility of care, will be crucial to advancing the effectiveness of primary healthcare (McDonald 2016).

Several serious questions should be asked of policymakers who increasingly make use of the collected data for measuring performance:

What is the purpose of the service whose performance is being assessed?
Do the KPIs relate clearly and unambiguously to that purpose?
How are performance measures being used in practice?
What are the consequences of poor performance? Education? Penalties? Dismissals?
What changes in funding, system and outcomes have occurred as a result of spending time and treasure in obtaining and analysing performance data?
What could improve the use of performance measures and the data they generate?

Although outcome data are difficult to procure and contain no magical answers, appropriate publication of such data drives improvements in the quality of care (Fung et al. 2008).

The New Zealand Health Quality and Safety Commission recently launched a dashboard of healthcare system quality indicators (Health Quality and Safety Commission New Zealand 2018). This provides data on 60 measures for each local district. Quality measures include timely access, safety, effectiveness, patient-centredness (including satisfaction) and equity. These measures and scales, which compare each district with the national average, support continuous improvement.

Conclusion

Regardless of a country's wealth, political commitment and public investment make the difference between sustainable and unsustainable access to healthcare. While less affluent countries are more likely to struggle with the sustainability of such developments, several are improving impressively.

A commitment by leadership—within countries, governments, and health organizations—to improving healthcare access is crucial if progress is to occur.

While overcoming geographical and financial barriers is important, so is a focus on workforce initiatives. Many of these seek to improve access to primary healthcare through intermediaries such as community health workers and health practitioners, who have the time to build trusted relationships with communities and patients.

Healthcare systems can support access by articulating a narrative which explains and inspires, designating expectations through strategy, policy, and funding formulas, to ensure that communities are engaged in the development, implementation, and monitoring of improvements to healthcare access. When informed by the analysis and use of local data to plan and monitor services, models of care can more closely fit the demand and the need for good access to care.

Access to healthcare is not only about entry to the healthcare system, but also about the journey through it. Committed and responsive staff is vital. For all these reasons, organizing optimal access to primary, secondary and tertiary services, in combination, warrant further research and development to achieve optimal management of communicable and NCDs, and to promote health through frontline preventive services.

ACKNOWLEDGEMENT

This chapter draws in part on the contribution to the previous edition of this textbook by Martin Guilford. We extend our thanks to him for his generosity. We are also grateful to Dr Peter Arnold OAM, of Sydney, for extensive editorial support in preparing this chapter.

REFERENCES

Abel, S. and Tipene-Leach, D. (2013). SUDI prevention: a review of Maori safe sleep innovations for future. *The New Zealand Medical Journal*, 126(1379). [Online] Available at: https://www.nzma.org.nz/journal/read-the-journal/all-issues/20I0-2019/2013/vol-126-no-1379/viewable

Aday, L.A. and Anderson, R.M. (1981). Equity of access to medical care: a conceptual and empirical overview. *Medical Care*, 19(Suppl.), 4–27.

AIHW (2018). *Deaths in Australia*. [Online] Available at: https://www.aihw.gov.au/reports/life-expectancy-death/deaths-in-australia/contents/life-expectancy

Araujo, L., La Rocca, A., and Hoholm, T., (2018). Reconfiguring the relation between primary and secondary healthcare through policy instruments. In: Hoholm, T., La Rocca, A., and Aanestad, M. (eds.) *Controversies in Healthcare Innovation*. London: Palgrave Macmillan.

Atun, R., et al. (2015). Health-system reform and universal health coverage in Latin America. *The Lancet*, **385**, 1230–47.

Babic, M., Fichtner, J., and Heemskerk, E.M. (2017). States versus corporations: rethinking the power of business in international politics. *The International Spectator*, 52(4), 20–43.

Bhutta, Z.A., Ladher, N., and Abbasi, K. (2018). Alma-Ata and primary healthcare: back to the future. *British Medical Journal*, **363**, k4433.

Bollyky, T.J. (2018). Health without wealth: the worrying paradox of modern medical miracles. *Foreign Affairs*. [Online] Available at: https://www.foreignaffairs.com/articles/2018-09-26/health-without-wealth?cid=nlc-fa_twofa-20180928

Bradford, N.K., Caffery, L.J., and Smith, A.C. (2016). Telehealth services in rural and remote Australia: a systematic review of models of care and factors influencing success and sustainability. *Rural Remote Health*, 16(4), 3808.

Braveman, P., Starfield, B., Geiger, H.J., and Murray, C.J.L. (2001). World Health Report 2000: how it removes equity from the agenda for public health monitoring and policy. Commentary: comprehensive approaches are needed for full understanding. *British Medical Journal*, **323**, 678–81.

Cochrane, A.L. (1972). *Effectiveness and Efficiency. Random Reflections on Health Services*. London: Nuffield Provincial Hospitals Trust.

Cruess, R.L. and Cruess, S.R. (2008). Expectations and obligations: professionalism and medicine's social contract with society. *Perspectives in Biology and Medicine*, 51(4), 579–98.

Daniels, N. (2008). *Just Health*. Cambridge: Cambridge University Press.

Dixon-Woods, M., et al. (2006). Conducting a critical interpretive synthesis of the literature on access to healthcare by vulnerable groups. *BMC Medical Research Methodology*, **6**, 35.

Editorial (2018). The collapse of the Venezuelan health system. *The Lancet*, **391**, 1331.

Esping-Andersen, G. (1990). *The Three Worlds of Welfare Capitalism*. Princeton NJ: Princeton University Press.

Fairall, L., Cornick, R., and Bateman, E. (2018). Empowering frontline providers to deliver universal primary healthcare using the Practical Approach to Care Kit. *British Medical Journal*, **363**, k4451.

Fairall, L., et al. (2012). Task shifting of antiretroviral treatment from doctors to primary-care nurses in South Africa (STRETCH): a pragmatic, parallel, cluster-randomised trial. *The Lancet*, **380**, 889–98.

Fraser, B. and Willer, H., (2016). Venezuela: aid needed to ease health crisis. *The Lancet*, **388**, 947–9.

Fung, C.H., Lim, Y.W., Mattke, S., Damberg, C., and Shekelle PG. (2008). Systematic review: the evidence that publishing patient care performance data improves quality of care. *Annals of Internal Medicine*, 148(2), 111–23.

Gaffney, A. (2018). *To Heal Humankind: The Right to Health in History*. London: Routledge.

Gillon, R. (2005). *Philosophical Medical Ethics*. Chichester: John Wiley.

Gulliford, M., et al. (2002). What does ‘access to health care’ mean? *Journal of Health Services Research & Policy*, **7**, 186–8.
Gunning, I. R. (1991). Modernizing customary international law: the challenge of human rights. *Virginia Journal of International Law*, 31(2), 211–48.
Gyuse, A.N., Ayuk, A.E., and Okeke, M.C. (2018). Facilitators and barriers to effective primary health care in Nigeria. *African Journal of Primary Health Care & Family Medicine*, **10** (1), 1–3.
Haines, A., et al. (2007). Achieving child survival goals: potential contribution of community health workers. *The Lancet*, **369**, 2121–31.
Hastings, G. (2012). Why corporate power is a public health priority. *British Medical Journal*, **345**, e5124.
Health Quality and Safety Commission New Zealand (2017). *Sudden Unexpected Death in Infancy (SUDI) Special Report.* [Online] Available at: https://www.hqsc.govt.nz/assets/CYMRC/Publications/CYMRCSUDIReport.pdf
Health Quality and Safety Commission New Zealand (2018). *Dashboard of Health System Quality*. [Online] Available at: https://www.hqsc.govt.nz/our-programmes/health-qualityevaluation/projects/quality-dashboards/
Health Workforce New Zealand (2018). *Workforce in Action—Developing the Health and Disability Kaiāwhina Workforce*. [Online] Available at: https://www.workforceinaction.org.nz/
Henckaerts, J-M. (2005). Study on customary international humanitarian law: a contribution to the understanding and respect for the rule of law in armed conflict, *International Review of the Red Cross*, 87(857), 175–212.
Hibbert, P., Hannaford, N., Long, J., Plumb, J., and Braithwaite, J. (2013). *Final Report: Performance Indicators Used Internationally to Report Publicly on Healthcare Organisations and Local Health Systems*. New South Wales: Australian Institute of Health Innovation, University of New South Wales.
Iacobucci, G. (2019). NHS 10 year plan pledges improvements in stroke, heart attacks, and mental health. *British Medical Journal*, **364**, l80.
Illich, I. (1976). *Limits to Medicine. Medical Nemesis: The Expropriation of Health*. Harmondsworth: Penguin Books.
Institute of Healthcare Improvement (2009). *The Triple Aim: Optimising Health, Care and Cost*. [Online] Available at: http://www.ihi.org/Engage/Initiatives/TripleAim/Documents/BeasleyTripleAimACHEJan09.pdf
International Partnership for Innovative Healthcare Delivery (2014). *Le-Nest.* [Online] Available at: https://www.innovationsinhealthcare.org/Leo/o20Nest%20Profileo/o202014.pdf
Kalantry, S., Getgen, J.E., and Koh, S.A. (2010). Enhancing enforcement of economic, social, and cultural rights using indicators: a focus on the right to education in the ICESCR. *Human Rights Quarterly*, 32(2), 253–310, 466–7.
Lahariya, C. (2017). Mohall and clinics of Delhi, India: could these become platform to strengthen primary healthcare? *Journal of Family Medicine and Primary Care*, 6(1), 1–10.
Last, J.M. (2007). *A Dictionary of Public Health*. Oxford: Oxford University Press.
Mackenbach, J.P. and Kunst, A.E. (1997). Measuring the magnitude of socio-economic inequalities in health: an overview of available measures illustrated with two examples from Europe. *Social Science & Medicine*, **44**, 757–71.
McDermott, J. (2018). An affordable necessity. *The Economist*, **427**, S3–S5.
McDonald, A. (2016). *Case Study: Harnessing Data for Local Quality Improvement in Diabetes. Te Awakairangi Health Network (TeAHN)*. New Zealand. [Online] Available at:https://www.hqsc.govt.nz/assets/Health-Quality-Evaluation/PR!Case-study-TeAwakairangi Health-Network-diabetes-Jan-2016.pdf

McKeown, T. (1979). *The Role of Medicine*. Oxford: Blackwell.
Mooney, G. (1996). And now for vertical equity? Some concerns arising from aboriginal health in Australia. *Health Economics*, **5**, 99–103.
Mossman, K., Bhattacharya, O., McGahan, A., & Mitchell, W. (2017). Expanding the reach of primary care in developing countries. Harvard Business Review online. [Online] Available at: https://hbr.org/2017/06/expanding-the-reach-of-primary-care-in-developing-countries
Murray, C.J. and Frenk, J. (2000). A framework for assessing the performance of health systems. *Bulletin of the World Health Organization*, **78**, 717–31.
North Star Alliance (2018). North Star Alliance Initiatives for mobile workers and the communities they interact with. [Online] Available at: http://www.northstar-alliance.org/initiatives/
Office for National Statistics (ONS) (1998). *Mortality Statistics: Cause, England and Wales*. Series DH2. Number 25. London: Office for National Statistics.
Office of the High Commissioner for Human Rights (2007). *International Covenant on Economic, Social and Cultural Rights*. Geneva: Office of the High Commissioner for Human Rights.
Pechansky, R. and Thomas, W. (1981). The concept of access. *Medical Care*, **19**, 127–40.
Rahman, A., Malik, A., Sikander, S., Roberts, C., and Creed, F. (2008). Cognitive behaviour therapy-based intervention by community health workers for mothers with depression and their infants in rural Pakistan: a cluster randomised control trial. *The Lancet*, **372**, 902–9.
Rawls, J. (2007). *Theory of Justice*. Cambridge, MA: Harvard University Press.
Reich, M.R. et al. (2016). Moving towards universal health coverage: lessons from 11 country studies, *The Lancet*, 387(10020), 811–16.
Sade, R.M. (1971). Medical care as a right: a refutation. *New England Journal of Medicine*, 285(23), 1288–92.
Sade, R.M. (2012). Health care reform: ethical foundations, policy, and law. *Journal of the American College of Surgeons*, 215(2), 286–96.
Sanne, I., et al.; CIPRA-SA Study Team (2010). Nurse versus doctor management of HIV-infected patients receiving antiretroviral therapy (CIPRA-SA): a randomised non-inferiority trial. *The Lancet*, **376**, 33–40.
Shengelia, B., Tandon, A., Adams, O.B., and Murray, C.J.L. (2005). Access, utilization, quality, and effective coverage: an integrated conceptual framework and measurement strategy. *Social Science & Medicine*, **61**, 97–109.
Smith, R., Beaglehole, R., Woodward, D., and Drager, N. (2003). *Global Public Goodsfor Health. Health Economic and Public Health Perspectives*. Oxford: Oxford University Press.
Starfield, B., Shi, L., and Mackinko, J. (2005). Contribution of primary care to health systems and health. *The Milbank Quarterly*, 83(3), 457–502.
Stigler, F., Macinko, J., Pettigrew, L. M., Kumar, R., and van Weel, C. (2016). No universal health coverage without primary health care. *The Lancet*, **387**, 1811.
The Economist Intelligence Unit (2017). *Global Access to Healthcare—Building Sustainable Health Systems*. [Online] Available at: http://accesstohealthcare.eiu.com/wpcontent/uploads/sites/42/2017/06/Globalaccesstohealthcare-3.pdf
United Nations (UN) (n.d.). *Millennium Development Goals*. [Online] Available at: http://www.un.org/millenniumgoals/
United Nations (UN) (2007). *Universal Declaration of Human Rights*. New York: United Nations.
United Nations (UN) (2012). Resolution 67/81. *Global Health and Foreign Policy* (A/RES/67/81). [Online] Available at: http://www.un.org/en/ga/search/view_doc.asp?symbol=A/RES/67/81

United Nations (UN) (2015). *Sustainable Development Goals—17 Goals to Transform our World.* [Online] Available at: https://www.un.org/sustainabledevelopment/sustainable-development-goals/

United Nations (UN) (n.d.). *Sustainable Development Goals.* [Online] Available at: https://www.un.org/sustainabledevelopment/health/

United Nations Human Rights, Office of the High Commissioner (n.d.). [Online] Available at: http://indicators.ohchr.org/

World Health Organization (WHO) (1978). *Declaration of Alma-Ata: International Conference on Primary Health Care, Alma-Ata, USSR, 6–12 September 1978*. [Online] Available at: http://www.who.int/social_determinants/tools/multimedia/alma_ata/en/

World Health Organization (WHO) (1988). *From Alma-Ata to the year 2000—Reflections at the Midpoint*. [Online] Available at: http://apps.who.int/iris/bitstream/handle/10665/39323/9241561246_eng.pdf?sequence=1&isAllowed=y

World Health Organization (WHO) (2018). *Top 10 Causes of Death*. [Online] Available at: http://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

SECTION 3

Public health policies, law, and ethics

3.1 Leadership in public health
Kevin A. Fenton

3.2 Ethical principles and ethical issues in public health
Nancy Kass, Amy Paul, and Andrew Siegel

3.3 The right to health supports global public health
Carmel Williams, Alison Blaiklock, and Paul Hunt

3.4 Law and the public's health
Lawrence O. Gostin

3.5 Health policy in developing countries
Thein T. Htay, Yu Mon Saw, James Levinson, S.M. Kadri, Ailbhe Helen Brady, Cecilia S. Acquin, and Aung Soe Htet

3.6 Public health policy in developed countries
John Powles † and Hebe Gouda

3.1

Leadership in public health

Kevin A. Fenton

What is leadership?

At its core, leadership is the art of motivating a group of people to act towards achieving a common goal. In the health and care setting, this can mean directing workers and colleagues with a strategy to meet either the organization's or community's needs. The leader is the inspiration and director of the action, and possesses the combination of personality and leadership skills that makes others want to follow their direction. US academic environments define leadership as 'a process of social influence in which a person can enlist the aid and support of others in the accomplishment of a common task'.[1][2] Studies of leadership have produced theories involving traits,[3] situational interaction, function, charisma, behaviour,[4] power, vision and values, [5] among others.[2]

If public health describes the organized efforts of society to improve health and tackle inequalities (UK Faculty of Public Health 2016), then leadership in public health represents the characteristics, passion, strategies, and abilities required to effect those changes. For too long public health practice and scholarship have focused on the skills required to characterize population health need, identify evidence-based approaches for addressing them, and more recently, a focus on implementation science and research to understand how evidence-based interventions are developed, implemented, and evaluated to improve population health outcomes. However, comparatively less attention had been paid in understanding the unique role, needs capacity, and importance of public health leadership and how this interfaces with policy, politics, or a range of competing interests.

In a world increasingly characterized by growing complexity and interdependencies, worsening health, social, and economic inequality,

competing social, environmental, and commercial interests, declining trust in government, and worsening social isolation and loneliness, now more than ever is the time and opportunity for public health leadership and leaders to shine. We have a unique opportunity to lead from the front: encouraging the move away from individual-focused interventions to tackling the wider (e.g. social, structural, economic) determinants of health and well-being; identifying evidence informed solutions to wicked problems (from youth violence, non-communicable disease control, to climate change); using data (both big and routine) to inform better targeting and prioritization of strategies and resources; and working with, empowering and engaging communities in new, meaningful, and more efficient ways. To achieve this public health leadership in the twenty-first century must be bold, confident, agile, innovative, inspirational, and capable of working within and across systems.

This truly is an exciting time to be a leader in public health and the ability to truly make a difference in the lives of millions at home and abroad a privilege for practitioners at home and abroad.

Leadership is not the same as management

Today, rather than viewing leadership as a set of innate qualities, we can think of the primary traits that leaders have in common which include having a positive outlook and forward thinking—being sure of themselves and their ability to influence others and impact the future.

Much literature and research conflate leadership and management, but the two are not identical: ‘Management’ derives from the Latin word ‘manus’ (meaning hand) and the subsequent Italian word ‘maneggiare’ (to control, often horses). ‘Leadership’, however, is from the Old German word ‘Lidan’ and the Old English derivation ‘lithan’ (to travel, to show the way, to guide) (Grint 2002, p. 248). Although similar, leadership and management are different: leadership goes beyond and above management. A manager can be organized and efficient in getting tasks done without a leader—someone who inspires others to achieve their best. Whereas managers manage processes; leaders lead people and organizations. Leadership is the most basic and scarcest resources in any business enterprise (Peter Drucker), of which the health and care sector is no exception.

Being an effective manager is important in public health practice. Management is no means an easy task and requires individuals to have a range of skills that may take many years to develop. Good public health managers are able to operate effectively and efficiently. They are able to take their organizations through change to accomplish goals and achieve objectives. Good managers tend to focus on the here and now as they are often expected to deliver and achieve outcomes in the short term. Good organizations rely on good managers; however, great management does not necessarily make a good organization.

Competent management is critical for successful delivery of public health programmes. It drives the ability to take care of the controlling and critical tasks required for effective programme implementation, monitoring, and evaluation. Good managers are increasingly involved in evidence informed practice, implementation science and programme research—all needed to improve management in order to strengthen the functioning of health systems (see, for example, Egger et al. 2005). This distinction between the controlling tasks associated with implementation and the visionary activities associated with setting a direction is a useful way of distinguishing management and leadership. In practice, the two are often linked either in a single individual or in a particular job description. Indeed, the vast majority senior and effective public health leaders rose through the leadership ranks based on a combination of technical, interpersonal, managerial, and leadership competences, acquired at different phases and stages of their career. As a result, separating the characteristics of good leadership from those of good management is not straightforward. However, the distinction between the leadership role and the managerial role is widely perceived as the difference between vision and implementation.

Leadership and good leadership are about having vision. The ability to look beyond the here and now to see the vast potential for one’s organization. Leadership deals with interpersonal influence, and management deals with acquiring and allocating resources to achieve work, unit or organisational goals (Campbell 2013, p. 412). Leadership is, therefore, broader than the leading of employees within organizations. Leadership takes place in all social and political contexts. Great leaders are also effective in getting things done but they accomplish their goals in different ways from managers. Leaders challenge their employees to achieve goals by creating a compelling vision for the future and then unlocking their

employees' potential to do so. Without the vision that leaders provide and without the contribution of their followers' hard work, energy, and innovation, we would never have achieve many of the excellent milestones in public health: from the global eradication of small pox and eradication of many vaccine preventable diseases, to the progress made in understanding and responding to the AIDS pandemic, and addressing more recent health threats including Ebola, and non-communicable diseases (NCDs). 'In public health, leaders are needed not only in hospitals and department divisions, but also in local, regional, and national government, in other related sectors like education and in large- and small-scale settings.

These distinctions between leaders and managers apply perfectly well in the public health sector. For example, the public health director of a health district might choose to have an operations manager who manages the finances, commissioning, and daily programme implementation, thereby allowing themselves the freedom to develop strategies and engage diverse actors and system partners in a shared vision of the district's health goals. A public health agency, to give another example, needs managers who, with limited finances, can organize and implement international, national, or local programmes; develop, implement, and evaluate public health policies; and manage programme staff and resources effectively. In both examples, health districts and public health organizations will also need leaders who can motivate staff to provide high-quality work in the context of budget restrictions and sociopolitical and economic uncertainty. Sometimes these skills and job responsibilities may exist in the same person, but the nature of management tasks differ from leadership tasks.

In this chapter, we are particularly interested in leadership rather than management. Public health requires people who can move the agenda in a world dominated by the politics of vested interest groups and where government policies may have intended and unintended consequences impacting population health and well-being. It requires individuals who can take scientific evidence and use it to help change the direction of policy. It requires leaders who can consistently and competently operate from place of authenticity, with clear vision, values and purpose, and an ability to lead across organizations and systems. These challenges call for 'big picture' skills, in addition to our core public health competences, and the ability to influence people both inside and outside the public health field.

Evolution of leadership theory

Research on leadership, like that on management, has evolved over time (Van Wart 2003). More than a century ago, thinking about leadership focused on the inherent personality of the individual leader: 'This 'Great Man Theory' approached leadership capacities as innate, fixed, and cross-contextual … the leader as solitary actor' (Turning Point 2001, p. 14).

This 'great leader' theory was deepened, in the first part of the twentieth century, by attempts to identify the key 'traits' or 'attributes' of successful leaders. Leadership attribute research was still focused on the individual, not the context. This research has continued into the present day and has generated many lists of leader traits (Zaccaro et al. 2013), including some specific to the health sector, as we show in the next section. The 'special individual' has certainly been invoked in the health sector. For example, Halfdan Mahler, a former head of the World Health Organization (WHO), and James Grant, a former head of UNICEF, were widely thought to have innate leadership qualities that largely explained the success of their agendas. When engaging colleagues who have risen to senior executive public health leadership positions within UK and internationally (as part of the background research for this chapter), a recurring characteristic is that these leaders bring a clarity of vision and strong personal values to their work, aligned with their organization's missions, are willing to take risks and to do things that defy convention, using innovation, partnerships, or disruption to achieve different outcomes. Relatively few of these leaders are autocratic or 'heroic leaders', using the power of their position to make decisions unilaterally. In contrast, they are often expert facilitators, leading from a position of authority, and utilizing a participatory decision-making style, using skilful questions to draw ideas out of others to develop shared solutions.

Research and thinking on the question of leadership moved on in the middle of the twentieth century to a 'contingency' approach (Van Wart 2003), taking into account the attributes not only of the leader, but also of the followers and the context in which he or she leads, for a more nuanced understanding of leadership success. While acknowledging that 'the essence of leadership is influence', this approach accepts that 'no single leader trait … or style explains leadership in all circumstances' and focuses instead on 'the dynamics of leadership—the coming together of leader and follower

behaviour in a social and/or organisational context' (Rumsey 2013a, pp. 1, 2, and 6; see also Rumsey 2013b).

What do leaders do?

Today the skills required to be a good leader are no longer secret, nor are they considered to be innate traits bestowed upon special individuals. Although some people seem to be born to be leaders, anyone can discover what leaders do and how to apply these skills in their everyday practice. This underscores the importance for those interested in developing their leadership skills to invest time shadowing and mentoring by experienced leaders. Types of leadership are outlined in Box 3.1.1.

Box 3.1.1 Methods of leadership

In many, or perhaps most, organizations, more than one of these conceptions may define leadership. Each implies particular ways of leading, and leaders may use a number of different methods.

Pure exercise of power. 'My way or the highway.' If you don't do what the leader demands, no matter how unreasonable, you're gone. The leader's decisions are not open to question or discussion, and no one else gets to make decisions.

Political scheming. The leader plays people off against one another, creates factions within the organization, cultivates 'allies' and isolates 'enemies', and builds up (through favours or overlooking poor performance) personal debt which can be cashed in when needed, in order to manipulate people and events as he wishes.

Using relationships. The leader develops strong positive relationships with all or most of the people in the organization and uses these relationships to steer people in particular directions. People do what they're asked because of their relationships with the leader, rather than for reasons connected to the tasks themselves.

Setting an example. The leader may or may not demand or request particular behaviour or actions, but she will demonstrate them, and expect or imply that others will follow.

Persuasion. The leader convinces people through argument, reasoning, selling techniques, or other persuasive methods that what the leader wants is, in fact, the best course, or in line with what they want to do.

Sharing power. Some leaders choose to exercise at least some leadership through the other stakeholders in the organization. In this situation they may give up some personal power in return for what they see as more ownership of decisions, goals, and the organization itself by those involved in the decision-making process.

Charisma. Some leaders are charismatic enough to simply pull others along by the power of their personalities alone. They may, in fact, advocate and accomplish wonderful things, but they do it through people's loyalty to and awe of them.

Involving followers in the goal. The leader gets others to buy into his/her vision for the organization, and to make it their own. He/she may accomplish this through charisma, through the

force of her own belief in the power and rightness of the vision, or through the nature of the vision itself.

Source: data from Rabinowitz P. 'Leadership and management' available from https://ctb.ku.edu/en/table-of-contents/leadership/leadership-ideas/leadership-styles/main

Inspire action

Most workers want to feel pride in their organization and when given the chance would give over and beyond for a cause that they believe in. This is especially so for individuals working in public health who are often drawn to the field in order to serve society, a commitment to social justice, and the desire to make a difference in their community, country, or globally. This means that there is much innate passion, creativity, and energy to draw upon in organizations. Effective public health leaders are able to use this knowledge to inspire employees to take action and achieve great things. This may be achieved by removing barriers and roadblocks to creativity, giving positive feedback and a sense of pride to employees, and creating a compelling vision for employees to strive for. Leaders help employees to tap into energy and initiative that employees did not know they had.

Communicate

Leaders make a commitment to communicate with their employees and to keep them informed about the organization. Great leaders earn the commitment of their workers by building communication links throughout the organization by providing regular feedback on performance, sharing positive news and results, publicly acknowledging the successes of individual workers and teams, keeping internal and external stakeholders well briefed of the contributions and priorities of the team, and providing opportunities for two-way communication and feedback. Duncan Selbie, Public Health England (PHE) 's Chief Executive circulates a Friday Message (e-newsletter) which informs tens of thousands of staff and agency partners nationally and globally about PHE's activities and accomplishments of the week. The agency's Health Matters resource is an effective way of engaging and communicating with broader system partners nationally on public health policies and initiatives.

Support and facilitate

Great leaders create environments in which employees are safe to speak up, tell the truth, innovate, and take risks. This is especially crucial in public health programmes where the failure of teams and the workforce to do so can result in errors with resultant harm to individual and communities. The ability to innovate and take risks is especially important for public health teams that are often faced with delivering complex population health programmes under challenging political circumstances and with financial and resource constraints and pressures. Good leaders support their employees and facilitate their ability to reach their goals. Great leaders are able to support employees, especially when times get rough. Great leaders are always there in the background—ready to assist and support workers were necessary.

Leadership traits

We have all worked with individuals who we or other consider to be great or poor leaders and seldom find it difficult to describe the traits or behaviours which contribute to this perception. The way that we manage ourselves is a central part of being an effective leader. Personal qualities like self-awareness, self-confidence, self-control, self-knowledge, personal reflection, resilience, and determination are the foundation of how we behave. Being aware of your strengths and limitations in these areas will have a direct effect on how you behave and interact with others, and they with you.

Optimism

Great leaders are more likely to envision a brighter more positive future, with great promise and optimism. This is often despite finding adversity and hard work on the way to achieving their goals. This ability to fixate on the future combined with optimism creates an aura that radiates from all leaders and touches those with whom they are in contact. Ultimately, people want to feel good about themselves and their futures and they want to work with organizations that are making a positive difference for their staff as well as the people they serve. In day-to-day practice in a busy public health department, this is not about being naïve about or ignoring the challenges

and problems faced by the team, but rather placing these within the context of other accomplishments, opportunities, and potential early wins. This reframing can be especially motivating for teams who often are unable to see the bigger or more strategic opportunities that the leader will have access to.

Confidence and decisiveness

Confident leaders make confident followers. An organization's employees mirror the behaviours of their leaders. When leaders are tentative and unsure of themselves so are the workers. When leaders display self-confidence workers follow. The nature of public health practice means that we are constantly making decisions: what evidence to use; which priorities to address; which programmes to commission; how best to evaluated them; and how to drive and improve health impact. Being decisive avoids waiting for the course of events to make decisions for the organization or individual. Often making a decision is better than making no decision at all. Great leaders are decisive. They take the time necessary to gather information on a problem, the people and resources needed to make an informed decision within a reasonable time frame. If the data are not available then the leader carefully weights the available data and information against the relative need of the decision and acts accordingly.

Integrity

Integrity is often highlighted as a most desirable trait that employees want from their leaders, and in public health, the finding is the same. Ethical behaviour, values, and a sense of fair play are key characteristics of effective leaders. Honest people want to follow honest leaders. When an organization's leaders conduct themselves with integrity, the organization can make a real difference in the lives of its employees, customers, and others who come into contact with it. Maintaining high standards of ethical and scientific integrity helps to maintain public trust in both the work of government, national and local, as well as research enterprises.

The UK General Medical Council, in articulating the duties of all doctors registered to practice, are clear that patients must be able to trust doctors with their lives and health. To justify this trust doctors must show respect for human life and make sure our practice meets the standards expected.

Maintaining trust is a one of four domains of ethical practice for UK physicians. This has implications for both public health practice and research, and the UK's Faculty of Public Health Good Public Health Practice Framework (2016) provides guidance on good professional practice in public health. It applies to all members of the core public health workforce, including public health practitioners and specialists and those training to become practitioners and specialists. It requires all involved, and especially leaders to behave in appropriate ways with a commitment to intellectual honesty and personal responsibility for one's actions as a researcher and practitioner. See Box 3.1.2.

Box 3.1.2 United Kingdoms' Faculty of Public Health. Good Public Health Practice Framework Short Guide (2016)

Maintaining the Public's Trust: Probity and Protecting the public

1. You must ensure at all times that your standards of practice justify the public's trust in your work, and in the work of the wider public health profession. Probity is the cornerstone of good public health practice. The components of probity are:
    a. Integrity: Put your service obligations above your personal interests;
    b. Honesty: Be truthful and open;
    c. Objectivity: Base advice and decisions on rigorous analysis of evidence;
    d. Impartiality: Act solely according to the merits of a case without prejudice.
2. You must disclose professionally facts which may have a bearing on your professional performance.
3. You must be vigilant about issues concerning data confidentiality and data protection.
4. You must respond to concerns and complaints.
5. You must be honest and objective when assessing the performance of those you have supervised, trained, or worked with.
6. You must be honest and trustworthy when writing reports, and when completing or signing references, forms, reports, and other documents.
7. You must co-operate with legitimate requests for information from official organizations monitoring public health.
8. You must respond, when and where you are able, in emergencies.
9. Your advertising of your services must be truthful, honest, legal, and decent.

Emotional intelligence

Emotional intelligence (sometimes referred to as EQ or EI) is the ability to comprehend, control, and develop your own feelings, while also being able to understand and manage others' feelings (Box 3.1.3). (Ref) Emotionally intelligent leadership emphasizes how one's emotions affect others and how you can use that knowledge to create positive outcomes—both personally and with the people you manage. The best leaders are self-aware of not only their emotions, but also their weaknesses and limitations, as well as their strengths. People with high emotional intelligence have a good understanding of their own emotional states, which allows them to better gauge the emotions of others. Emotional intelligence (EI) accounts for 85–90% of the difference between outstanding leaders and their more average peers (Ref). These benefits all have something in common: the enhanced EI leaders have developed produces positive results in their employees, helping to build better teams, and help prevent employee turnover.

EI includes four domains: self-awareness, self-management, social awareness, and relationship management. The first two domains determine how well we understand and manage ourselves and our emotions. The latter two dictate how well we recognize and manage the emotions of others, build relationships, and work in complex social systems. The four qualities contain 18 leadership competences all of which support the development of resonance. Together when developed these competencies have benefits to both individuals (self-regulation, less stress, more empathy, collaborative communication) as well as the organization (team management, corporate culture, better outcomes).

Box 3.1.3 Emotional intelligence domains and competences

Personal competence: These capabilities determine how we manage ourselves

Self-awareness

- Emotional self-awareness: Reading one's own emotions and recognizing their input; using gut sense to guide decisions
- Accurate self-assessment: Knowing one's own strengths and limits
- Self-confidence: Having a sound sense of one's self worth and capabilities

Self-management

- Emotional self-control: Keeping disruptive emptions and impulses under control
- Transparency: Displaying honest, integrity, and trustworthiness
- Adaptability: Demonstrating flexibility in adapting to changing situations or overcoming obstacles
- Achievement: Having the drive to improve performance to meet inner standards of excellence
- Initiative: Being ready to act and to seize opportunities
- Optimism: Seeing the 'upside' in events

Social competence: These capabilities determine how we manage relationships.

Social awareness

- Empathy: Sensing other's emotions, understanding their perspectives, and taking active interest in their concerns
- Organizational awareness: Reading the currents, decision networks, and politics at the organizational level
- Service: Recognizing and meeting follower, client, or customer needs

Relationship management

- Inspirational leadership: Guiding and motivating with a compelling vision
- Influence: Using a rage of tactics for persuasion
- Developing others: Bolstering other's abilities through feedback and guidance
- Conflict management: Resolving disagreements
- Building bonds: Cultivating and maintaining a web of relationships
- Teamwork and collaboration: Fostering cooperation and team building

Leadership styles

Today's context for health leadership is complex, rapidly evolving, and calls for new approaches to the development of leaders for today and the future. According to John Gardner, in *On Leadership*: 'Leadership is the process of persuasion or example by which an individual (or leadership team) induces a group to pursue objectives held by the leader or shared by the leader and his or her followers.' If we accept that definition, then leadership style is the way in which that process is carried out.

Leaders' styles encompass how they relate to others within and outside the organization, how they view themselves and their position, and—to a very large extent—whether or not they are successful as leaders. If a task needs to be accomplished, how does a particular leader set out to get it done? If an emergency arises, how does a leader handle it? If the organization

needs the support of the community, how does a leader go about mobilizing it? All of these depend on leadership style (Box 3.1.4).

Box 3.1.4 Summary of leadership styles

Visionary. This style is most appropriate when an organization needs a new direction. Its goal is to move people towards a new set of shared dreams. 'Visionary leaders articulate where a group is going, but not how it will get there—setting people free to innovate, experiment, take calculated risks,' write Mr Goleman and his co-authors.

Coaching. This one-on-one style focuses on developing individuals, showing them how to improve their performance, and helping to connect their goals to the goals of the organization. Coaching works best, Mr Goleman writes, 'with employees who show initiative and want more professional development'. But it can backfire if it's perceived as 'micromanaging' an employee, and undermines his or her self-confidence.

Affiliative. This style emphasizes the importance of teamwork, and creates harmony in a group by connecting people to each other. Mr Goleman argues this approach is particularly valuable 'when trying to heighten team harmony, increase morale, improve communication, or repair broken trust in an organization'. But he warns against using it alone, since its emphasis on group praise can allow poor performance to go uncorrected. 'Employees may perceive,' he writes, 'that mediocrity is tolerated.'

Democratic. This style draws on people's knowledge and skills, and creates a group commitment to the resulting goals. It works best when the direction the organization should take is unclear, and the leader needs to tap the collective wisdom of the group. Mr Goleman warns that this consensus-building approach can be disastrous in times of crisis, when urgent events demand quick decisions.

Pacesetting. In this style, the leader sets high standards for performance. He or she is 'obsessive about doing things better and faster, and asks the same of everyone'. But Mr Goleman warns this style should be used sparingly, because it can undercut morale and make people feel as if they are failing. 'Our data shows that, more often than not, pacesetting poisons the climate,' he writes.

Commanding. This is classic model of 'military' style leadership—probably the most often used, but the least often effective. Because it rarely involves praise and frequently employs criticism, it undercuts morale and job satisfaction. Mr Goleman argues it is only effective in a crisis, when an urgent turnaround is needed. Even the modern military has come to recognize its limited usefulness.

All leadership styles can become part of the leader's repertoire. Leadership styles should be adapted to the demands of the situation, the requirements of the people involved and the challenges facing the organization. In the book *Primal Leadership*, Daniel Goleman, who popularized the notion of 'Emotional Intelligence', describes six different styles of leadership. The most effective leaders can move among these styles, adopting the one that meets the needs of the moment. They can all become part of the leader's repertoire.

The style of an organization's leadership is reflected in both the nature of that organization and its relationships with the community. In many ways,

the style of its leader defines an organization. If the organization is to be faithful to its philosophy and mission, its leader's style must be consistent with them. For that reason, being conscious of both your own style as a leader and those of others you hire as leaders can be crucial in keeping your organization on the right track.

Collaborative leadership (Leadership FD P 54)

Collaborative leadership has always been a hallmark of leading in public health organizations. In part driven because of the need for leaders to move beyond working with a few individuals on the management team at the top of their organizations and understanding that to get the most of your employees and to maximize public health outcomes and outputs, we need to continue to get the most out of every employee. A collaborative leader tries to involve everyone in the organization in leadership. She is truly first among equals, in that she may initiate discussion, pinpoint problems or issues that need to be addressed, and keep track of the organization as a whole, rather than of one particular job. But decisions are made through a collaborative process of discussion, and some form of either majority or consensus agreement. Towards that end, a collaborative leader tries to foster trust and teamwork among the staff as a whole.

Collaborative leadership comes as close as possible to ensuring that members of the organization buy into its vision and decisions, since they are directly involved in creating them. See Box 3.1.5. It comes closest to the goal of servant leadership, and it also comes closest to reflecting the concepts of equality and empowerment included in the philosophy and mission of so many grass roots and community-based organizations. It thus removes much of the distrust that often exists between line staff and administrators.

Box 3.1.5 8 Principles of collaborative leadership used in Orpheus Process

- Put power in the hands of people doing the work
- Encourage individual responsibility for product and quality
- Create clarity of roles
- Foster horizontal teamwork
- Share and rotate leadership
- Discover how to listen, discover how to talk
- Seek consensus (and build creative systems that favour consensus)
- Dedicate passionately to your mission

Source: data from Seifter H (2001) 'The conductorless orchestra' *Leader to Leader Journal* No. 21.

Democratic leadership

Democratic leadership invites the participation of staff members and others, not only in decision-making, but in shaping the organization's vision. It allows everyone to express opinions about how things should be done, and where the organization should go. By bringing in everyone's ideas, it enriches the organization's possibilities. But it still leaves the final decisions about what to do with those ideas in the hands of a single person. A democratic leader understands that there is no organization without its people. Although he sees the organization as a cooperative venture, he knows that he ultimately has to face the consequences of his decisions alone.

Some models of democratic leadership might put the responsibility in the hands of a small group—a management team or executive committee—rather than an individual. What a democratic leadership doesn't necessarily do—although it can—is establish staff ownership of the organization and its goals. Although everyone may be asked for ideas or opinions, not all of those are used or incorporated in the workings of the organization. If there is no real discussion of ideas, with a resulting general agreement, a sense of ownership is unlikely. Thus, democratic leadership may have some of the drawbacks of autocratic leadership—a lack of buy-in—without the advantages of quick and clear decision-making that comes with the elimination of consultation.

Managerial leadership

The leader who sees themself as a manager is concerned primarily with the running of the organization. Where it's going is not at issue, as long as it gets there in good shape. Depending upon the nature and stability of the

organization, his/her main focus may be on funding, on strengthening the organization's systems and infrastructure (policies, positions, equipment, etc.), or on making sure day-to-day operations go well (including making sure that everyone is doing what he's supposed to).

An efficient managerial leader will generally be on top of what's happening in the organization. Depending on the size of the organization and her management level, she'll have control of the budget, know the policies and procedures manual inside out, be aware of who's doing his job efficiently and who's not, and deal with issues quickly and firmly as they come up. What she won't do is steer the organization. On the other hand, good management without a clear vision creates an organization with no sense of purpose. The organization may simply act to support the status quo, doing what it has always done in order to keep things running smoothly.

Transformational leadership

The idea of transformational leadership was first developed by James McGregor Burns in 1978 and later extended by Bernard Bass as well as others. Neither Burns nor Bass studied schools but rather based their work on political leaders, Army officers, or business executives. Transformational leadership looks at leadership differently. It sees a true leader as one who can distil the values and hopes and needs of followers into a vision, and then encourage and empower followers to pursue that vision. The transformational leader conceives of leadership as helping people to create a common vision and then to pursue that vision until it's realized. She elicits that vision from the needs and aspirations of others, gives it form, and sets it up as a goal to strive for. The vision is not hers: it is a shared vision that each person sees as his own.

The conception behind transformational leadership is thus providing and working towards a vision, but also has elements of empowerment, of taking care of people, and even of task orientation. The job of the transformational leader is not simply to provide inspiration and then disappear. It is to be there, day after day, convincing people that the vision is reachable, renewing their commitment, priming their enthusiasm. Transformational leaders work harder than anyone else, and, in the words of a spiritual, 'keep their eyes on the prize'.

The methods that transformational leaders might use to reach their goals can vary. They'll virtually always include involving followers in the goal, as well as charisma, which comes, if not from personal characteristics, from the ability to put a mutual vision into words, and to move a group towards the realization of that vision. Transformational leaders may also use sharing power, setting an example, and/or persuasion to help move a group towards its goal. Leithwood finds that transformational leaders pursue three fundamental goals:

Autocratic leadership

Autocratic leaders insist on doing it all themselves. They have all the power, make all the decisions, and don't often tell anyone else about what they're doing. If you work for an autocratic leader, your job is usually to do what you're told. An autocratic leader often maintains his/her authority by force, intimidation, threats, reward and punishment, or position. Although he/she may or may not have a clear vision, and may or may not be steering the organization in the right direction, he's not concerned with whether anyone else agrees with what he's doing or not.

Autocratic leadership allows quick decision-making, and eliminates arguments over how and why things get done. At the same time, however, it may reduce the likelihood of getting a range of different ideas from different people, and can treat people badly, or as if they don't matter. If, as is often true, the leader is concerned with his own power and status, he'll be looking over his shoulder, and moving to squelch any opposition to him or his ideas and decisions. Innovation or the use of others' ideas is only permissible if it's part of the leader's plan.

Autocratic leaders often leave fear and mistrust in their wake. Others in the organization tend to copy their protection of their position, and their distrust of others' ideas and motives. At its best (and there are decent autocratic leaders), autocratic leadership provides a stable and secure work environment and decisive, effective leadership. All too often, however, it can sacrifice initiative, new ideas, and the individual and group development of staff members for the predictability of a highly structured, hierarchical environment where everyone knows exactly what he's supposed to do and follows orders without question.

Systems leadership

Transforming a complex system—such as the transportation, housing, or food system, all wider determinants of health—is a monumental task requiring coordinated action by people with very different viewpoints. Systems change initiatives often engage hundreds of organizations—governments, companies, civil society organizations, worker associations, research institutions and others—combining their capacities to achieve a shared goal.

Systems leadership is about how you lead across boundaries—departmental, organizational, or sectoral. It's how you lead when you're not in charge, and you need to influence others rather than pull a management lever (Ref). See Boxes 3.1.6 and 3.1.7.

It describes the way the leader needs to work when faced with large, complex, difficult, and seemingly intractable problems; where you need to juggle multiple uncertainties; where no one person or organization can find or organize the solution on their own; where everyone is grappling with how to make resources meet demand which is outstripping them; and where the way forward therefore lies in involving as many people's energies, ideas, talents, and expertise as possible.

It recognizes that leadership isn't vested in people simply because of their title or position; that it is possible, and often necessary, for leadership to be shared and ceded. To achieve success, partners must come together on the basis of a shared ambition, and accept partial or clumsy solutions on the way to getting there. Working with uncertainty, and ambiguity, is a given, and it's expected that people will experiment with different methods and processes as a result.

Box 3.1.6 Key steps in effective system leadership

We distilled five key elements of the systems change process into the 'CLEAR' framework for leading systems change. These five elements are not necessarily sequential—they may overlap or repeat in cycles throughout the course of an initiative.

2. Convene and commit

Key stakeholders engage in moderated dialogue to address a complex issue of mutual concern. They define shared interests and goals, and commit to working together in new ways to create systemic

change. For example, the We Mean Business Coalition engaged nearly 1,000 leading companies to advocate for ambitious, science-based climate policy, and has made over 1,500 action commitments.

3. Look and learn

Through system mapping, stakeholders jointly build a shared understanding of the components, actors, dynamics, and influences that create the system and its current outcomes, generating new insights and ideas.

For example, The Global Alliance for Improved Nutrition targets specific gaps in the nutrition system, working to catalyse and scale market-based solutions, and targeting vulnerable populations who are most in need.

4. Engage and energize

Strong stakeholder engagement is built through continuous communication to build trust, commitment, innovation, and collaboration. Inspiration, incentives, and milestones help drive progress and maintain momentum. For example, the New Vision for Agriculture initiative engaged over 650 organizations and 1,500 individual leaders around the world, catalysing action in 21 countries including over 90 value-chain projects.

5. Act with accountability

Shared goals and principles set the direction of the initiative, while measurement frameworks help track progress. Coordination and governance structures can be developed as initiatives mature. For example, the Every Woman Every Child movement mobilized hundreds of action commitments towards its global strategy, monitoring progress through a unified accountability framework, with oversight from a high-level steering group and coordination by a global secretariat.

6. Review and revise

Stakeholders review progress regularly and adapt their strategy accordingly. Adopting an agile, flexible, innovative, and learning-centred approach allows for evolution and experimentation. For example, the 2030 Water Resources Group evolved its organizational structure through several stages, commissioning external evaluations to both review its progress and recommend opportunities to increase its impact.

While the CLEAR Framework appears quite structured, the reality of the systems change process is often messy and ambiguous. Many stakeholders describe the experience of systems leadership as a journey of discovery that evolves over time, leading to moments of discovery or insight—what we describe as ‘Aha! moments’—that crystallize each step of the journey.

Reprinted from Dreier L. et al. (2019) ‘Systems leadership can change the world—but what exactly is it?’ *World Economic Forum*, available at https://www.weforum.org/agenda/2019/09/systems-leadership-can-change-the-world-but-what-does-it-mean/

Box 3.1.7 Systems leadership

Systems leadership: a tool for our times

Systems leadership is a set of skills and capacities that any individual or organization can use to catalyse, enable, and support the process of systems-level change. It combines collaborative leadership, coalition-building, and systems insight to mobilize innovation and action across a large, decentralized network.

Two very different examples illustrate what it can look like in practice. The 2015 Paris Agreement, signed by 195 countries with support from thousands of organizations, was spearheaded by Christiana Figueres, a global diplomat who emphasized practicality, flexibility, and collaboration to bring stakeholders onboard, securing an historic agreement. At a more local level in Richmond, California, a community organizer named Najari Smith founded a new venture called Rich City Rides, galvanizing community members, local businesses, and city government to address the interconnected challenges of employment, health, and environmental sustainability among low-income communities of colour in the city.

These two leaders operated in very different spheres, but they used some similar tactics: combining a deep understanding of the systemic issues they wanted to address; an ability to engage and align diverse stakeholders around shared goals; and an emphasis on empowering action and collaboration by a broad network of organizations.

Systems leaders apply an unusual combination of skills and attributes to mobilize large-scale action for systems change. Like many leaders, they tend to be smart, ambitious visionaries with strong skills in management and execution. Unlike traditional leaders, they are often humble, good listeners, and skilled facilitators who can successfully engage stakeholders with highly divergent priorities and perspectives. Systems leaders see their role as catalysing, enabling, and supporting widespread action—rather than occupying the spotlight themselves.

Resonant leadership

This reflects the type of modern leadership required to deal with the unprecedented changes affecting our physical and social environments. Political, economic, technological, and social change are driving profound transformation in our organizational models making predictability and stability elusive if not impossible. The sheer complexity of our organizations has increased resulting in confusing organizational structures the global connectedness of our practice means that even the simplest organizations are required to reach out and be connected to faraway places. These uncertainty calls for resonant leaders—leaders capable of charting through unexpected territory and inspiring people in their organizations, institutions, and communities. They create opportunities where there are challenges. They move people—powerfully, passionately, and purposefully.

Resonant leaders are in tune with those around them resulting in people working in sync with each other’s thoughts and emotions. They either intuitively understand and use their EI (self-awareness, self-management, social awareness, and relationship management). They manage their own and other’s emotions as well as build strong trusting relationships. Resonant leaders are exciting and get results. EI is a must and great leaders build

resonant relationships around them. Great leaders are emotionally intelligent, and they are mindful—they seek to live in full consciousness of self, others, nature, and society. The face the work of uncertainty with hope; they inspire through clarity of vision, optimism, and a profound believe in there and tier people's ability to turn dreams into reality. They face sacrifice, difficulties, and challenges with empathy and compassion for the people they lead and those they serve.

See Box 3.1.8.

Box 3.1.8 Questions to ask about a resonance in leadership

Is the leader inspirational?

Does the leader create an overall positive emotional tone that is characterized by hope?

Is the leader in touch with others? Does the leader know what is on other's hearts and minds? Does the leader experience and demonstrate compassion?

Is the leader mindful—authentic and in tune with self, others, and the environment?

Leadership development

Leadership development expands the capacity of individuals to perform in leadership roles within organizations. Leadership roles are those that facilitate execution of a company's strategy through building alignment, winning mindshare, and growing the capabilities of others. Leadership roles may be formal, with the corresponding authority to make decisions and take responsibility, or they may be informal roles with little official authority (e.g. a member of a team who influences team engagement, purpose, and direction; a lateral peer who must listen and negotiate through influence).[1]

Leadership development is thought to be key to business success. A study by the Center for Creative Leadership holds that 65% of companies with mature leadership development programmes drove improved business results as compared to 6% of companies without such a programme. Similarly, 86% of companies with leadership development programmes responded rapidly to changing market conditions whereas only 52% of companies with immature programmes were able to do so.[2]

For many public health practitioners, leadership development is both through on-the-job experience combined with formal training opportunities.

The latter often involves training multisectoral teams in an applied, team-based, and collaborative leadership development model. For public health leaders, a focus on leadership core principles, working across sectors, quality improvement principles, and the use of data for planning, assessment, monitoring, and evaluation, and developing the population health perspective are increasingly seen as foundational.

Leadership development places such importance on self and peer evaluation, mentoring and coaching to ensure leaders develop greater self-insight exploring the known leadership skills and behaviours as well as those which are yet unknown, but perceived by others. Without this self-awareness, it will be much more difficult to behave in the way research has shown that good leaders do. This, in turn, will have a direct impact on your colleagues, any team you work in, and the overall culture and climate within the team as well as within the organization, and ultimately the care experience or patients and service users.

See Case Study 3.1.1.

CASE STUDY 3.1.1 Training local teams to be health leaders in poor communities: a case from the Philippines

The Philippines provides an innovative approach to leadership for the health needs of the most deprived and vulnerable populations. Begun in 2002, the Leaders for Health Program (LHP) was predicated on the understanding that health professionals must work in concert with political and community actors if basic health problems are to be addressed in low-income communities. This 'Tri-Leader' approach brought together the community doctor, the mayor, and a local community leader and trained them, both individually and as a team.

To address the need that medical leaders understand and are able to use evidence to inform policy, doctors received postgraduate training resulting in a Master in Community Health Management. Recognizing that the local political and community leaders often did not have an advanced level of education, the programme created a certificate course in community health management tailored to them, but paralleling the master's training provided to doctors. Joint workshops and training sessions were also held for the tri-leaders together. In addition, all three benefited from coaching and mentoring from a more experienced leader of their own category from another part of the country. Once this team began to identify priorities in its locality, the LHP provided assistance in identifying donors for specific interventions. For example, some communities invested in sanitation and clean water systems. One community established a local pharmacy; another invested in a boat for emergency transport of patients.

By taking this innovative approach to leadership—one that recognizes the importance of joint leadership in the medical, political, and local communities in order to address public health needs—the LHP has been able to demonstrate greater success than other initiatives that focus only on placing trained doctors in low-income areas.

Its success would seem to be due to the programme's understanding of the complexity of health systems. For example, addressing water-borne diseases properly requires sanitation systems, and sanitation systems cannot be built without political support.

Community leaders helps doctors to see beyond the narrow biomedical domain to the broader political and sociological setting in which public health problems must be tackled. One mayor reported that earlier attempts to post doctors to his village were unsuccessful because the doctor was perceived to be making demands arrogantly of a resource-poor community, rather than working with the community to reach a common understanding.

A cost analysis of the programme found that the cost per person for the close to 1 million beneficiaries in the 36 towns that participated in the LHP over its initial 4-year period was Philippine pesos 60 or USD 1.50 per year. Furthermore, for every peso invested in the programme as grant, the local government and other national government partners were able to raise a counterpart share of Philippine pesos 1.27. LHP therefore stimulated the mobilization of local resources for health system development. Most of the graduate s of the LHP have continued to work in their posts. Some joined various units of the Department of Health, the US Agency for International Development and Médecins Sans Frontières, showing a remarkable record for retaining health professionals in places of need. As a result of the LHP's success, its expansion to other local government units is now being considered as part of the national effort to strengthen local health systems.

Source: data from Baquiran et al., Leaders for Health Program 2005, Copyright 2006 and R. Baquiran, personal communication, 2013.

The UK Public Health Skills and Knowledge Framework (PHSKF)

The Public Health Skills and Knowledge Framework was first published as the Public Health Skills and Careers Framework in 2008 through the collaborative efforts of lead agencies in public health across the United Kingdom. The Public Health Skills and Knowledge Framework (PHSKF) has been revised following extensive consultation with the public health workforce and collaboration between lead agencies across the United Kingdom. The revised framework aims to reflect the current public health landscape, ensuring the public health workforce continues to develop the skills and competencies needed both now and in future. The framework is accompanied by a user guide, explaining how it can be used by individuals, employers and educational providers working in public health (Case Study 3.1.2 and Fig. 3.1.1).

Case Study 3.1.2 The Public Health Skills Framework

The PHSKF provides:

- The functional areas in which individuals, teams, and organizations operate to deliver public health outcomes
- Statements that describe what functions an individual might carry out in the course of their work. The combination of functions will vary from individual to individual and from role to role
- A benchmark or single point of reference for the UK workforce and their employers to help individuals plan their own personal development and to help employers to plan and develop their workforce
- A tool to facilitate the generation of job descriptions for new roles; templates for standard roles; and profiles for individual roles
- A common reference for the review and development of standards of practice and curricula for training and education across levels of the qualifications framework
- A description of the public health functions and subfunctions in a way that could be presented through and accessible and easy to navigate digital platform (aka: 'skills passport')

UK Faculty of Public Health Good Practice Framework

Professional bodies including the Faculty of Public Health have published extensively on guidelines for public health practice including: Good Public Health Practice—General Professional Expectations of Public Health Professions (2008) and its application. It has been reviewed and updated as necessary in line with developments in professional practice and other relevant guidance (see Table 3.1.1 and Box 3.1.9). The guidance provides the basis for good professional practice in public health. It applies to all members of the core public health workforce, including public health practitioners and specialists, and those training to become practitioners and specialists.

The standards in this guidance are designed to be complementary to those contained in other professional codes of practice. Doctors regulated by the General Medical Council must meet the standards set out in Good Medical Practice. All other professionals must similarly meet their relevant codes of professional practice.

American leadership courses are outlined in Case Study 3.1.3.

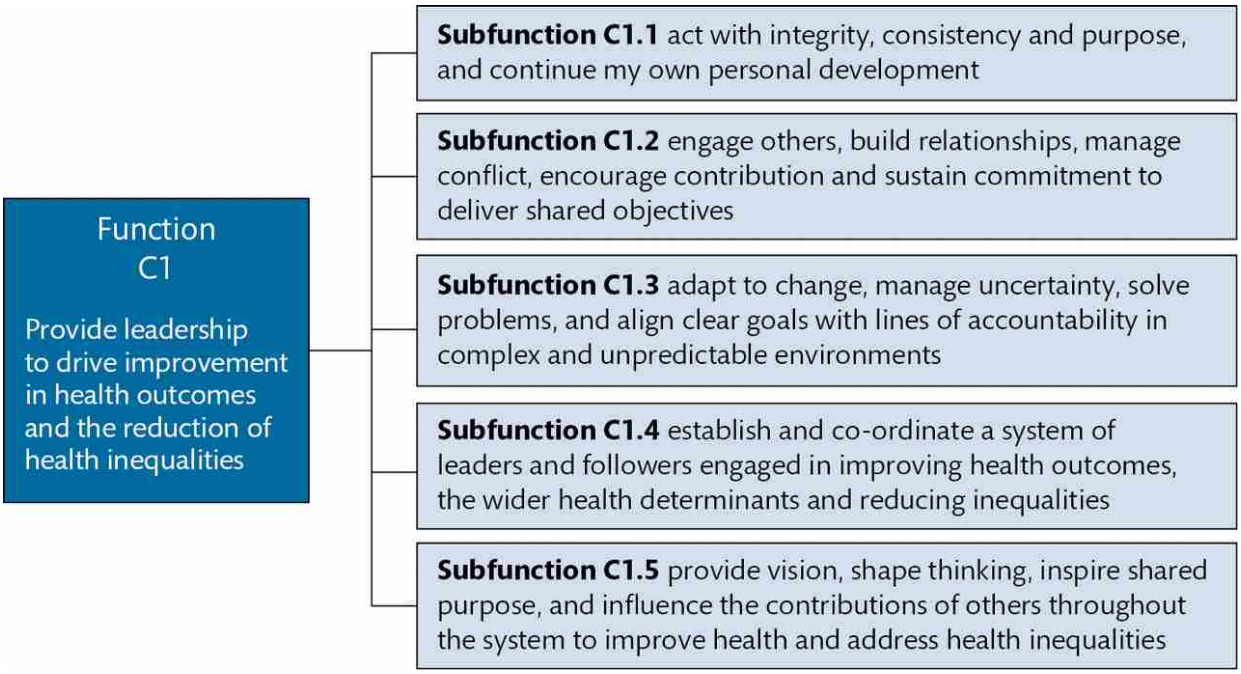


Fig. 3.1.1 Function C1, related to leadership.
Reprinted from Public Health Skills and Knowledge Framework, PHE publications gateway number 2016440, under the Open Government License 3.0.

Table 3.1.1 Competencies needed by leaders in public health

Creating change	Influencing politics	Working transorganizationally	Building teams
Visionary leadership Be able to articulate an evidence-based vision of the future and incorporate it into strategy	*Political processes* Determine appropriate actions on policy and political issues, organize key actors to cooperate for regulatory and legislative changes, and translate policies into programmes and services	*Understanding of organizational dynamics* Assess and develop systems structures based on an understanding of culture and organizational behaviour	*Develop team-oriented structures and systems* Change system infrastructure to innovative, learning teams
Sense of mission Facilitate the development of a mission, communicate it, and 'model' it through personal behaviour	*Negotiation* Mediate potential crises, bargain, and co-ordinate with key stakeholders	*Interorganizational collaborating mechanisms* Involve key actors and networks across a broad range of organizations and groups in collaborative coalitions	*Facilitate development of teams* Empower, motivate, and reward teams
Effective change agent Think creatively and analytically about systems and change strategies, and facilitate dialogue and empowerment of others to take action	*Ethics and power* Identify and use alliances based on ethical principles	*Social forecasting* Analyse trends and communicate predictions to consortium partners	*Serve in facilitation and mediation roles* Negotiate and intervene to help teams' function
	Marketing and education Use social marketing and health education to influence key audiences		*Serve as an effective team member* Through own behaviour, 'model' the key characteristics of listening, encouraging, and motivating while displaying integrity, commitment, and honesty

Source: data from Wright, K. et al., Competency development in public health leadership. *American Journal of Public Health*, Volume 90, Issue 8, pp. 1202–7, Copyright © 2000 American Public Health Association.

Box 3.1.9 Good Professional Practice Framework

A. Knowledge, skills, and performance: professionalism in public health practice

- You must be competent in all aspects of your work as a public health professional and retain your competence throughout your working life
- You must recognize and work within the limits of your competence, seeking further advice and consulting with colleagues where this is required.
- You must practice within ethical, legal, and regulatory limits that apply to your area of public health practice.
- You have a duty to monitor, protect, and improve the health of populations.
- You must be satisfied that you have valid authority, including consent w here informed consent is required f or interventions.
- You are expected to improve and develop your practice continuously.
- You must be able to receive and act upon feedback on your practice and performance, both positive and negative.
- You must take care to delegate responsibility to others safely.
- You must keep records of important decisions, actions, and consequences.

B. Safety and quality

- You have a duty to safeguard and protect the health and well-being of your population
- Your duty to vulnerable groups extends to their relatives, carers, and partners.
- You must plan and act in accordance with available evidence and use resources effectively and efficiently
- You must be professional and ethical in your public health practice at all times.
- You must make sure that you and other professionals are able to be professional and ethical in your and their public health practice.
- You must maintain financial rectitude in all your public health practice dealings.
- You must declare any relevant commercial or financial conflicts of interests you may have.
- You must not allow your own health and well-being to affect adversely your professional practice.
- You must have in place arrangements to compensate anybody who may suffer as a result of deficiencies in your work or that of your team.
- You must conduct research (including designing, organizing, and carrying out research) with honesty and integrity.

C. Communication, partnership, and teamwork. Collaborative working: organizational community and individual

- You must work in, and manage, partnerships and teams effectively and with due consideration of others.
- You must make sure that relevant information is shared.
- You must make sure information to be shared is meaningful.
- You must respect the right of people to decline to take part in interventions, initiatives, teaching, or research even if it may be of benefit to them.
- You must maintain your focus on the effectiveness of interventions.
- You must treat colleagues, individuals, and communities fairly and with respect.
- You have a professional responsibility to support trainees and colleagues in your work environment, through formal supervision, formal and informal mentoring, and modelling professional public health practice. This includes providing honest and objective assessment of performance and supporting trainees or staff you manage with appropriate levels of supervision.

D. Maintaining public trust: probity and protecting the public

- You must ensure at all times that your standards of practice justify the public's trust in your work, and in the work of the wider public health profession.
- You must disclose professionally facts which may have a bearing on your professional performance.
- You must be vigilant about issues concerning data confidentiality and data protection.
- You must respond to concerns and complaints.
- You must be honest and objective when assessing the performance of those you have supervised, trained, or worked with.
- You must be honest and trustworthy when writing reports, and when completing or signing references, forms, reports, and other documents.
- You must co-operate with legitimate requests f or information from official organizations monitoring public health.
- You must respond, when and where you are able, in emergencies.
- Your advertising of your services must be truthful, honest, legal, and decent.

Source: data from Faculty of Public Health (2016) 'Good Public Health Practice' available from https://www.fph.org.uk/professional-development/good-public-health-practice/

Case Study 3.1.3 CDC National Leadership Academy for the Public's Health, the CDC Laboratory Leadership Service, and ASTHO Leadership Institute

CDC strives to empower public health professionals with tools to improve public health practice within their jurisdictions. To further this mission, CDC's Center for State, Tribal, Local, and Territorial Support has provided support and partnership to the Public Health Institute's Center for Health Leadership and Practice External to launch and maintain the National Leadership Academy for the Public's Health (NLAPH)External.

NLAPH offers an innovative style of leadership development by bringing together teams of leaders from multiple sectors who actively engage their communities to improve population health

and achieve health equity. This 1-year programme uses blended learning modalities and an action learning process that includes a webinar series, leadership retreat, coaching support, peer networking, an applied health leadership project, and web-based resources. Competencies are embedded within the programme curriculum to encourage the advancement of leadership knowledge, attitudes, and practices.

Programme goals

The goals of NLAPH are to:

- Educate stakeholders about evidence-based policy options
- Drive the adoption of evidence-based practices in communities
- Better align medicine and public health
- Improve health outcomes in our nation through sustainable systems change

The objectives of NLAPH are to

- Provide ongoing leadership training and support to leaders that work across sectors to address community health issues
- Improve leadership teams' capacities to apply continuous quality improvement methodologies, measure progress of the projects, and document the impact of leadership development
- Build and support the development of communities with multisector leadership learners using online technology and social media tools that will contribute to community health improvement goals and inform policy change

Curriculum

Each community-based team receives leadership training and coaching from national experts and through collaboration with other teams via retreats, webinars, networking, an applied health leadership project, and web—based resources.

Competencies: Advance the leadership knowledge, attitude, and practice of individuals and teams to lead systems change efforts including the following components.

- Individual leadership mastery
- Ability to work effectively across sectors and with diverse stakeholders
- Community organizing and advocacy
- Leveraging resources for collective impact
- Policy analysis and development
- Commitment to health in all policies, social determinants of health, and advancing equity.

Four work phases of the NLAPH year:

- Inspiration
- Ideation
- Implementation
- Growing, sustaining, and transition

CDC Laboratory Leadership Service LLS

The CDC Laboratory Leadership Service (LLS) was launched in 2015 and develops laboratory scientists who integrate safety and quality as principal elements in their work, which is vital to protecting the public's health, safety, and security. CDC's LLS fellowship programme prepares high calibre PhD scientists to become public health laboratory leaders. Through on-the-job service at CDC and other public health laboratories, fellows acquire and apply skills in laboratory quality management, the science of biosafety, and leadership.

In this two-year competency-based experiential learning programme, LLS trains fellows in laboratory quality management, the science of biosafety, leadership, and other critical skills—building a strong foundation for future public health laboratory leaders.

LLS is closely aligned with CDC's Epidemic Intelligence Service (EIS) to promote interdisciplinary training, applied learning, and networking between laboratory scientists and epidemiologists. LLS provides opportunities for fellows to serve public health agencies and learn through field experiences and domestic or international emergency responses.

CDC's guiding principles for laboratory work are to ensure the safety of all staff and the community and be as transparent as possible about our work as we conduct high-quality scientific research to protect people in this country and around the world.

LLS was established as one of the multiple improvements to lab safety practices and procedures across the agency, with the goal of increasing CDC's culture of safety. As such, the core competencies of the LLS fellowship programme are designed to instil excellence through training and education in laboratory quality management, the science of biosafety, and leadership. The LLS programme's Core Activities of Learning are based on competencies published by CDC and Association of Public Health Laboratories (APHL).

LLS fellows learn through on-the-job experience, classroom instruction, case studies, exercises, and e-learning. Throughout their fellowship, LLS fellows:

- Learn skills in and apply public health laboratory competencies—particularly those focused on biosafety, quality management systems, and management and leadership—as they approach real world public health scenarios
- Begin training in July with an intensive, 1-month orientation in Atlanta, Georgia
- Attend a one-week in-person training each fall of the fellowship
- Participate in brief monthly trainings
- Have access to many other optional training opportunities that they can take based their needs
- Learn about and have access to CDC-specific tools for literature reviews and data analysis (including tools for bioinformatics and informatics)
- Obtain education through hands-on activities under mentorship and guidance from host site laboratory supervisors
- Present findings from investigations and studies during regular seminars and annual conferences
- Attend trainings, conduct investigations, and network with EIS officers and other CDC fellows

LLS fellows' learning is based on competencies published by CDC and APHL and achieved through the programme's Core Activities of Learning:

- Conduct applied laboratory research to address a public health or safety-related issue.
- Conduct a safety risk assessment to evaluate the probability and potential consequences of exposure to a given hazard.
- Evaluate a quality management system.
- Incorporate bioinformatics principles into applied public health laboratory science.
- Give a 10–20 minute oral presentation to a scientific audience.
- Give an in depth public health talk on the fellow's original LLS work or field of study.
- Write and submit, as first author, a scientific manuscript for a peer-reviewed journal.
- Participate in laboratory operations management.
- Communicate complex scientific concepts to an external lay audience.
- Provide service to public health laboratories.

ASTHO Leadership Institute

The ASTHO Leadership Institute (ALI) is a year-long, cohort-based leadership development programme designed to enhance the success of new state and territorial health officials (S/THOs).

In partnership with experts in executive leadership development, nationally recognized public health leaders, and experienced S/THOs, ALI offers a customized learning programme which includes opportunities for self-assessment; development of knowledge and skills through intensive retreats; access to a learning management system with on-demand tools and information; executive coaching; and formal mentorship by experienced S/THOs.

The leadership competencies in the framework provide a good example of the activities associated with leadership in relation to different groups, situations, settings, and intentions. All leadership stems from the ability to drive one's own actions and conduct. This area then describes action to lead and manage others; change systems; and finally around setting strategic vision and establishing collective buy-in and ownership.

The NHS Leadership Framework

The Healthcare Leadership Model has been developed by the NHS Leadership Academy, working with the Hay Group and colleagues from the Open University. It is an evidence-based research model that reflects: the values of the NHS; what we know about effective leadership; what we have learnt from the Leadership Framework (2011); and what our patients and communities are now asking from us as leaders (Box 3.1.10).

The NHS Leadership Framework (LF) provides a consistent approach to leadership development for staff in health and care irrespective of discipline, role, or function, and represents the foundation of leadership behaviour. It sets out a single model of leadership for all health and care staff to aspire to. The LF is based on the concept that leadership is not restricted to people who hold designated leadership roles and where there is a shared

responsibility for the success of the organization, services, or care being delivered.

Box 3.1.10 NHS leadership qualities framework

Cognitive capacities: general intelligence, judgement/decision-making skills, organizing skills
Social capacities: social intelligence/sensitivity, communication skills
Personality: extraversion, conscientiousness, openness, agreeableness, neuroticism/emotional stability
Motives: dominance, drive/purpose, energy, tenacity/persistence
Core beliefs: self-evaluation: self-confidence, self-efficacy
Knowledge: knowledge of the business/sector

The Healthcare Leadership Model is made up of nine ‘leadership dimensions’. Research has shown that all nine dimensions of the model are important in an individual’s leadership role. However, the type of job they have, the needs of the people they work with, and the context of their role within their organization will all affect which dimensions are most important for them to use and develop. It is also important to realize that areas identified for development within the model may be as much about how you manage yourself as about how you manage your behaviour and relate to other people.

The intention therefore is not that this model is ‘set in stone’ and will still be appropriate for healthcare staff in 2023. Instead, the intention is to make ongoing updates to the model, to make sure it remains as relevant to staff in 2 or 5 years’ time, as it is to them today. This more flexible and innovative approach will result in future versions being available over the next few years.

Leadership across the lifecourse

There are many paths to leadership in public health. Many leaders first entered the profession with a strong desire to serve. Through the course of their careers, they were exposed to circumstances that deepened their commitment. Frequently, they sought out challenging circumstances in which to gain much needed experience. They travelled, engaged in development work, often involving poor countries and communities. In the

course of this exposure, they took up more advanced public health studies, perhaps did some research, wrote articles in scientific journals, or managed a healthcare facility. Some had long periods of service in their own country before being exposed to international work. For others, international exposure came early, usually through some clinical or public health training abroad.

Some public health leaders enter the field after years of working in the clinics, practising a specialty like paediatrics, obstetrics, or surgery. Others come into positions of leadership through the field services, rising through a public health organization, or a non-governmental organization. Others were involved in academia, teaching, and research, consulting, or working with a development agency. Public health leaders could also have been involved in politics. Or they could once have been cause-oriented activists, involved in issues such as HIV prevention and primary healthcare.

Central to the growth and development of public health leaders are the years when he/she developed the personal qualities of self-awareness, self-belief, and self-management. For many leaders, their involvement in public health was imbued with higher purpose and even that they were instruments of a higher power to do good work. A desire to continue learning new things was often mentioned as a source of great satisfaction. Here are a few things you can do to choose and develop your own authentic and effective leadership style.

Plan for life-long investment in leadership development

One pursues leadership by steadily inculcating in oneself the personal qualities, competencies, and principles that make up a true leader. One does not acquire these necessary attributes and skills simply by ascending to a leadership position, often via political patronage, especially if one lacks adequate preparation. Study, training, reflection, exposure, and experience are all elements of this preparation. Thus, as one’s leadership traits flourish, the time will come when one is ready to assume a leadership position. Leadership should be sought in this way from day one of any career and hence, any strong public health organization should encourage leadership development at all levels.

Envision your future self

Public health remains an fascinating career choice with a variety of options for leadership throughout one's career. Whether deciding to work in academia, service public health, or a combined approach; working at different levels within the health and care system (e.g. global, national, regional or local); or subspecializing within particular public health disciplines (e.g. global health, epidemiology, health improvement, health protection, health inequalities). Understanding your strengths, passions, commitment to ongoing learning, maximizing opportunities for progression, and ensuring you have a strong network of professional support should all be taken into consideration when developing your 5 or 10 year career goals plans.

Similarly, one should take time to explore what kind of leader you would like to be, in which setting, and for what purpose: What kind of leader do you want to be? How would you be most effective as a leader? Who are the current system leaders you admire and why? What kind of leadership style would be of the most benefit to your organization, and would allow you to be the best leader you could be?

Committing to single leadership style isn't the only way to look at leadership. Most real leaders use a combination of styles, depending on need, situation, context, and desired outcomes. You can adapt most styles to most situations, but always remember to align these the real needs of the organization. You may need to practice a different style at the beginning from the one that you want to assume over the long term, in order to solve problems in the organization, or to get people on board. The most important practice is that you are continuously reviewing, learning, practising, and refining your craft.

Know your values and guiding principles

Individual context, culture, and background dictate how each leader will interpret their own organization's principles and will complement them with their personal values derived from their personal philosophy of life. Together, this collection of values and principles will serve to give focus and meaning to the health leader's work.

For example, PHE has defined its core mission as 'working 24/7 to protect and improve the nation's health and reduce health inequalities'. The agency in turn guided by a number of specific aims including keeping people safe,

working to prevent poor health, working to narrow the health gap and support a strong economy. The way, PHE achieves this by working as ‘One PHE, making the best possible use of the expertise and commitment of our people to deliver the biggest impact and best value for the taxpayer and our partners. We perform five key roles within the public health system, which are underpinned by a commitment to incorporate the reduction of health inequalities into all areas of our work’. Similarly, the International Association of National Public Health Institutes links and strengthens the government agencies responsible for public health. Their mission is to ‘improve the world’s health by leveraging the experience and expertise of its member institutes to build robust public health systems’.

These and other similar organizational values and principles must be understood and embedded firmly in the leader’s mind. And most importantly leaders need to understand the own personal values and principles and how they might intersect positively or negatively with their organization. Leaders need to think and act on their feet. Unless one is sure of one’s values and principles, in a time of crisis, panic and uncertainty might take over. There will not be time to think about remembering your principles when the need to make a decision is immediate. For this reason, taking the time to understand what guides oneself, whether through leadership development, coaching, or mentoring, is invaluable.

Seek mentors

Both the literature on leadership and the practical experience of today’s health leaders underline the crucial importance of mentorship in the development of the complete public health leader. Being prepared to learn from superiors, fellow colleagues, friends, and public figures is essential, however, there is no fixed prescription for mentorship. Ideally, a mentoring relationship should be mutually beneficial, with both parties learning from one another, enriching their understanding of their work and purpose.

By its very nature, mentoring is a personal arrangement between two parties. The mentoring relationship can be formal and structured, with regular meetings and logbook entries; or it could be more informal and sporadic, with the occasional chat over coffee or e-mail correspondence. The opportunity should be taken for the mentor to impart advice, the mentee to report on recent events, and for both to share thoughts on work and even

wider aspects of life, particularly if relevant to a career in health and leadership.

Pursue leadership, not the leadership position

The trajectory of an individual's path to leadership can rise steeply or gently. It is one's willingness to step forward, given one's self-awareness and knowledge of the prospective job, however incomplete, which might mark a person's readiness to lead. A potential leader will almost always be selected by having already demonstrated leadership and by showing readiness to lead. In deciding if one is ready to assume the responsibility of a public health leader, it of course helps to discuss the opportunity with mentors, co-workers, and family members.

Many public health leaders can often point to their 'stepping up to leadership roles' for as long as they can remember, while some others have taken what they have described as a circuitous, organic, and emergent, though no less determined path. A resolute focus on the outcome, being an effective leader, while being open to you route and opportunities for getting there, with regular check-ins with self and mentors, provide the most practical, implementable, and manageable choice for many.

Keep learning and developing your craft

Once in a leadership position, the leader must keep learning, whether updating technical competencies, leadership and management techniques, or investing in one's personal and emotional development. The reality is that the work environment, context, challenges, and opportunities change rapidly and it is imperative that leaders invest in their own continuing professional and personal development to improve their effectiveness, acquire new sills, and strengthen their own resilience.

Throughout the course of one leadership career, mistakes will be made, disappointments and conflicts will occur. The strong face of leadership is easily and well observed, and when times are challenging leaders must dig deep into themselves to find internal resources of strength. The capacity to affirm self-belief and overcome self-doubt is an important ingredient in surviving the difficult times. Strategies include simply pausing to reflect, listening inwardly, and reflecting on thoughts, actions, and lived experience

in the midst of complexity, adversity, and uncertainty. By learning to be truly honest with oneself and seeking objective, constructive, and critical feedback from others, a leader can begin to find an answer to this question and sustain self-improvement. In addition, a leader can and should turn outward to others for strength, listening closely to what the boss, advisors, mentors, spouse, colleagues, friends, constituencies, or even opinion surveys say.

Managing the challenges of leadership

Leadership constantly presents challenges to a leader and their abilities. These challenges are an encouragement to rise to another level, to test oneself and improve in the process, and show that you can accomplish something that may seem difficult or even impossible. Being a leader is in itself a challenge. The challenges of leadership may be viewed as being external, coming from people and situations; internal, stemming from within the leader himself/herself; and those arising from the nature of the leadership role.

It’s almost impossible to imagine a situation where a leader doesn’t have to cope with external challenges. Many public health organizations grapple with lack of funding and other resources, working with political colleagues, dealing with advocates, communicating with, and engaging the public effectively, often while dealing interpersonal problems within the organization. Social, economic, and political forces in the larger world can affect the organization as well. To some extent, the measure of any leader is how well he/she can deal with the constant succession of crises and minor annoyances that threaten the mission of his/her group. If he/she is able to solve problems, take advantage of opportunities, and resolve conflict with an air of calm and a minimum of fuss, most of the external issues are hardly noticeable to anyone else. If the leader doesn’t handle external challenges well, it is likely the organization won’t either.

While leadership presents to each of us the opportunity to demonstrate the best of what we are, it also exposes our limitations. In many cases, good leaders have to overcome those limitations in order to transmit and follow their vision. Fear, lack of confidence, imposter syndrome, insecurity, impatience, and intolerance all act as barriers to effective leadership. At the same time, acknowledging and overcoming them can turn a mediocre leader

into a great one. It's often very difficult for people, especially those who see themselves as leaders, to admit that they might have personality traits or personal characteristics that interfere with their ability to reach their goals. Part of good leadership is learning to accept the reality of those traits, and working to change them so they don't get in the way.

Real leadership makes great demands on people. As a leader, you are responsible for your group's vision and mission, for upholding a standard, often for being the group's representative to the rest of the world and its protector as well. These responsibilities might be shared, but in most organizations, one person takes the largest part of the burden. A common challenge for executive leaders is burnout. The process of giving ourselves, continuously and often in excess, leads us to become ineffective. Over time and if unchecked, the physical and emotional toll on leadership limits our ability to sustain high performance. Many people cope with stress by developing defensive routines—bad habits that keep us in denial about what is really going on inside us and around us. While others have developed strategies to cope and be more resilient, including being physically active, learning, and paying attention to their emotional and spiritual well-being. Finally, organizations, through their investment in workplace learning, workplace well-being, mentoring and peer support programmes, and early referral to support services if required, can create more positive environments to support the well-being and effectiveness of their leaders and managers.

Developing your authentic leadership journey

So thus far, this chapter has explored the definitions and concepts of leadership, explored different leadership styles, introduced frameworks for leadership training and capacity building, and discussed strategies for developing one's leadership skills across one's career and indeed the life course. This final section discusses the concept of authentic leadership—something that leaders are often asked about: How did you discover your passion for leadership? How did you develop your unique leadership voice and presence? How do you stay true to yourself while leading in diverse settings and organizations?

Authentic leadership is an approach to leadership that emphasizes building the leader's legitimacy through honest relationships with followers

which value their input and are built on an ethical foundation. It involves actually holding the beliefs you support; doing and saying what you sincerely believe. To be authentic, one must truly be himself or herself. It's not a switch that gets turned on during the commute to the office; you must be yourself and take responsibility for your behaviour. Generally, authentic leaders are positive people with truthful self-concepts who promote openness. They have the self-confidence and ability to expose their own frailties and weaknesses as well as their strengths and assets, having a genuine passion for what they are doing, caring for the organization's mission and what it means, and the desire to provide value to others, with true and unyielding commitment, absolute integrity.

Start with yourself

Use what you know about your own personality, and about how you've exercised leadership in the past—it is critical to be important to be honest with yourself about who you are. First, be clear with yourself about what your natural tendencies and talents are. Not everyone can be charismatic, but almost everyone can learn to distil and communicate a vision that reflects the hopes and needs of a group. Knowing who you are is the first step towards both choosing a style and understanding what you'll have to do to adopt it. Being truly honest with yourself is difficult—it takes an honest self-assessment, which can mean stripping away defences and facing insecurities. In addition, it may take some time with a coach, mentor or a trusted friend, or hearing feedback from colleagues and co-workers.

Becoming an authentic leader is not necessarily about learning how to lead authentically. It's about taking what is genuinely there, inside you, and letting it be visible for all to see and experience. Turning into someone you are not while at the office is all too transparent and doesn't help you build the trust and enthusiasm you expect from your staff.

Think about the needs of the organization or initiative

You can adapt most styles to most situations, but don't neglect the real needs of the organization in your calculations. You may need to practice a different style at the beginning from the one that you want to assume over the long

term, in order to solve problems in the organization, or to get people on board.

Authentic leaders are mission driven and focused on results. They are able to put the mission and the goals of the organization ahead of their own self-interest. They do the job in pursuit of results, not for their own power, money, or ego. They also focus on the long term. Leaders realize that to nurture individuals and to nurture a company requires hard work and patience, but the approach pays large dividends over time.

Observe and learn from others

Authentic leaders lead with their heart, not just their minds. They are not afraid to show their emotions, their vulnerability, and to connect with their employees. This does not mean authentic leaders are ‘soft’. In fact communicating in a direct manner is critical to successful outcomes, but it’s done with empathy.

There is much to be learnt from observing leadership behaviours and practices of others—examining what were their styles, and whether they were effective. Find a mentor. If there’s a leader whom you particularly admire, and that person is accessible, talk with him/her about leadership issues—about how they perceive what they’re doing, how they’d handle particular situations and why, etc. How did they handle different kinds of situations? How did what they did make you and others feel? Try to watch others in action and talk to them about how they see what they do. What do you like about how they operate? What don’t you like? What can you incorporate into your own style? Most people, especially if they’re good leaders and conscious of what they do and why, welcome the opportunity to help others develop their own leadership skills.

Use research on leadership

There are lots of resources available on leaders and on both the theory and practice of leadership. They’ll give you a lot more ideas about leadership styles, and help you refine your own thinking about what leadership is and what kind of leader you’d like to be. Finally, remember that the research, evaluation, and policy fields on leadership continue to evolve rapidly and there will always be more to learn. Remember that no matter how well

you're doing, none of us are perfect leaders—it never is, and never will be. Be prepared to find for yourself or hear from others the negative as well as the positive, to consider it carefully and objectively, and to make corrections if necessary. That way, you can not only become a good leader, but continue to be one.

Through ALI, participants have access to national thought leaders and the opportunity to build a trusted network of peers. All programme activities are based on best practices for adult learning and research-based core competencies for success as a chief health strategist. ALI is funded with support from the Robert Wood Johnson Foundation.

Box 3.1.11 Steps for engaging in authentic leadership

Authentic Leadership works on the principle that a leader can prove their legitimacy by nurturing sincere relationships with their subordinates and giving importance to their input.

Authentic leadership involves actually holding the beliefs you support; doing and saying what you sincerely believe. To be authentic, one must truly be himself or herself. It's not a switch that gets turned on during the commute to the office; you must be yourself and take responsibility for your behaviour.

Authentic leaders are usually the first ones to embrace the change they espouse, setting the example for others. They are visionaries who go against the grain to do what's best for their organization because it's truly in their heart.

Authentic leaders have integrity. Many people can lead, but only authentic leaders can do so with the values, morals, and integrity of a strong character. People know and feel this strong character and are drawn to it. People trust, respect, and listen to authentic leaders, making it easier to lead.

Authentic leaders create communities at work; communities that work. They foster friendships among and between co-workers to link humans to one another so they have something else and someone else to work for other than themselves.

Authentic leaders learn about themselves to foster excellence in themselves and the organization. No one is perfect and to hold perfection on a pedestal, whether it be personal, a department, or the whole organization, is setting up for failure. Instead, authentic leaders strive for excellence and use failures and mistakes as a way to gain wisdom and to move forward.

Source: data from Kruse K. (2013) 'What is Authentic Leadership?': Forbes.

Getting the public health leaders the world needs

We need public health leaders everywhere and at every level of the health system. However, good leaders are most needed in places where people are poor and suffering simply because there is such a scarcity of them there.

Committed and effective leaders often provide the vision and impetus to mobilize people and resources in needy areas. How can the international health community help developing countries develop the needed leaders?

A good starting point is to aim at strengthening educational institutions, which can provide relevant, team-based training. Schools of public health have an important role to play (Colgrove et al. 2010). Helping governments and their partners to improve public health working conditions, including financial and non-financial incentives, will also help to attract and retain talented people. From an international perspective we can work towards empowering nations to do this self-sufficiently, strengthening their own frameworks for leadership.

Health leaders have used the term ‘capacity release’ in reference to this kind of endeavour (Chan 2007). It must be stressed that training programmes are only one component of a comprehensive approach that should include strategies to retain promising public health workers and sustain the idealism of leaders who are prepared to venture forth to places of hardship. We must encourage and support them to stay where they are most needed. Examples of work being done on this include the WHO framework on training leaders and managers (Egger et al. 2005) and related work to strengthen leadership at all levels of the health system.

The challenge now is particularly great. We have not yet established systematic ways to develop public health leaders, while the context is becoming ever more complex. To address the social determinants of health, confront the ongoing communicable disease burden as well as the rising burden of NCDs, and truly ensure ‘health in all policies’ will require devoted and determined public health leaders at all levels of society and in all sectors.

Conclusions

This chapter has focused on strategies for developing, promoting public health leadership. As we reach the end of this chapter, there are some key points for the aspiring leader: Leadership is manifested in a complex and dynamic interaction between the leader, the followers, and the context. Parables about different types of leadership illustrate the diversity of contexts and behaviours that may be associated with successful leadership.

Leadership style is the way in which a leader accomplishes their purposes. It can have profound effects on an organization and its staff members and can determine whether the organization is effective or not. Leadership style depends on the leader's and organization's conception of what leadership is, and on the leader's choice of leadership methods. Depending how those fit together, a leader might adopt one of a variety styles, each reflected in the way the organization operates and the way its staff members relate to one another.

In the face of ever-increasing complexity, public health leaders are needed to prevent, treat, and eliminate disease and promote the health and well-being of all people. Such leadership can be developed and provided by local, national, and international actors. Public health leaders demonstrate personal qualities and skills, underpinned by fundamental principles, which enable them to influence actions and results. The constant manifestation and practice of these attributes and beliefs is a continuing challenge in a leader's job and career.

In closing, everyone's 'path to leadership' is unique, personal, and never-ending. However, being proactive; actively seeking out opportunities to lead; investing in life-long training, coaching, and mentoring; and understanding your personal values, mission, and life purpose are all essential parts of the personal development process. Stepping up to answer to the call to lead, and pursuing activities to develop your leadership skills, although sometimes challenging is nevertheless a privilege that should engage the mind, the heart, and spirit. It is a stimulating, rewarding, and never-ending quest. We bid the reader all success.

ACKNOWLEDGEMENTS

The authors would like to thank Amimo Agola, Anand Kurup, Chadia Wannous, Christine Lamoureux, Claudia Vivas, Delanyo Dovlo, Dick Chamla, Francis Omaswa, Guangyuan Liu, Helen Robinson, Lealou Reballos, Maria Eufemia Yap, and Miriam Were, among others, for their insights into the issues of leadership in public health. The authors would also like to acknowledge the contributions of Carmen Dolea, Daniel Shaw, and Joanne McManus to an earlier version of this chapter.

REFERENCES

Alban-Metcalfe, R.J. and Alimo-Metcalfe, B. (2000). An analysis of the convergent and discriminant validity of the transformational leadership questionnaire. *International Journal of Selection and Assessment*, 8(3), 158–75.

Avolio, B.J. and Bass, B.M. (1999). Re-examining the components of transformational and transactional leadership using the multifactor leadership questionnaire. *Journal of Occupational and Organizational Psychology*, 72(4), 441–62.

Baquiran, R., Yap, M.E., and Bengzon, A.R.A. (2006). Terminal Grant Report: Leaders for Health Program 2002–2006. Manila: Pfizer Philippines Foundation Inc.

BBC News Asia (2012). Australia Smokers Given Plain Packs. 1 December 2012. [Online] Available at: http://www.bbc.co.uk/news/world-asia-20559585.

Binney, G., Wilke, G., and Williams, C. (2005). *Living Leadership: A Practical Guide for Ordinary Heroes*. London: Pearson Education Limited.

Campbell, J.P. (2013). Leadership, the old, the new, and the timeless: a commentary. In: Rumsey, M.G. (ed.) *The Oxford Handbook of Leadership*, pp. 401–19. New York: Oxford University Press.

Cardenas, V.M., Roces, M.C., Wattanasri, S., et al. (2002). Improving global public health leadership through training in epidemiology and public health: the experience of TEPHINET. Training Programs in Epidemiology and Public Health Interventions Network. *American Journal of Public Health*, 92(2), 196–7.

Chan, M. (2007). *Address to WHO Staff*. 4 January 2007. Geneva: World Health Organization.

Colgrove, J., Fried, L.P., Northridge, M.E., and Rosner, D. (2010). Schools of public health: essential infrastructure of a responsible society and a 21st-century health system. *Public Health Reports*, 125(1), 8.

Day, D.V. (2013). Training and developing leaders: theory and research. In: Rumsey, M.G. (ed.) *The Oxford Handbook of Leadership*. New York: Oxford University Press, pp. 76–93.

De Meyer, A. (2011). Collaborative leadership: new perspectives in leadership development. In: Canals, J. (ed.) *The Future of Leadership Development*. Basingstoke: Palgrave Macmillan, pp. 44–63.

Dennis, R.S. and Bocarnea, M. (2005). Development of the servant leadership assessment instrument. *Leadership and Organization Development Journal*, 26(8), 600–15.

Egger, D., Travis, P., Dovlo, D., et al. (2005). *Strengthening Management in Low-income Countries*. Making Health Systems Work: Working Paper Series Number 1. Geneva: World Health Organization.

Faculty of Public Health (2016). *Good Public Health Practice*. [Online] Available at: https://www.fph.org.uk/professional-development/good-public-health-practice/

Firth-Cozens, J. and Mowbray, D. (2001). Leadership and the quality of care. *Quality in Health Care*, **10** (Suppl. 2), ii3–7.

Flahault, D. and Roemer, M.I. (1986). *Leadership for Primary Health Care*. Geneva: World Health Organization.

Gawad Kalinga (2012). Skoll award to Gawad Kalinga: freedom to serve. [Online] Available at: http://gk1world.com/skoll-world-forum-2012

Grint, K. (2002). Management or leadership? *Journal of Health Services Research and Policy*, 7(4), 248–51.

Habaradas, R.B. (2010). The economic and artistic flows of Gawad Kalinga. Paper presented at the Arts Congress of De La Salle University, Manila, Philippines. [Online] Available at: http://www.academia.edu/735064/The_economic_and_artistic_flows_of_Gawad_Kalinga

Hollander, E.P. (2013). Inclusive leadership and idiosyncrasy credit in leader– follower relations. In: Rumsey, M.G. (ed.) *The Oxford Handbook of Leadership*. New York: Oxford University Press, pp. 122–43.

Klimoski, R. (2013). Commentary: when it comes to leadership, context matters. In: Rumsey, M.G. (ed.) *The Oxford Handbook of Leadership*. New York: Oxford University Press, pp. 267–87.

Koh, H.K and Jacobson, M. (2009). Fostering public health leadership. *Journal of Public Health*, 31(2), 199–201.

Kotter, J.P. (2008). *Force for Change: How Leadership Differs from Management*. [eBook] New York: Free Press.

Kumar, R. and Chhokar, J.S. (2013). Cross-cultural leadership. In: Rumsey, M.G. (ed.) *The Oxford Handbook of Leadership*. New York: Oxford University Press, pp. 225–40.

Leaders for Health Program (2005). *The LHP Experience: 5 Stories of Hope and Transformation in Municipal Health Practices (Program Document)*. Makati, Philippines: Ateneo Professional Schools. [Online] Available at: http://www.leadersforhealth.ph

Marion, R. (2013). Organizational leadership and complexity mechanisms. In: Rumsey, M.G. (ed.) *The Oxford Handbook of Leadership*, pp. 184–202. New York: Oxford University Press.

NHS Leadership Centre (2006). NHS Leadership Qualities Framework. London: National Health Service Institute for Innovation and Improvement.

Plsek, P.E. and Wilson, T. (2001). Complexity, leadership, and management in healthcare organisations. *British Medical Journal*, 323(7315), 746–9.

Rumsey, M.G. (2013a). Introduction: leadership in five parts. In: Rumsey, M.G. (ed.) *The Oxford Handbook of Leadership*. New York: Oxford University Press, pp. 1–7.

Rumsey, M.G. (2013b). Commentary: a way ahead. In: Rumsey, M.G. (ed.) *The Oxford Handbook of Leadership*. New York: Oxford University Press, pp. 94–100.

Saleh, S.S., Williams, D., and Balougan, M. (2004). Evaluating the effectiveness of public health leadership training: the NEPHLI experience. *American Journal of Public Health*, 94(7), 1245–9.

Sandique-Carlos, R. (2012). Philippines adopts contraception law. *The Wall Street Journal (Europe edition)*, 29 December 2012. [Online] Available at: http://online.wsj.com/article/SB10001424127887324669104578208891390006944.html

Sendjaya, S., Sarros, J.C., and Santora, J.C. (2008). Defining and measuring servant leadership behaviour in organizations. *Journal of Management Studies*, 45(2), 402–24.

Sternberg, R.J. (2013). The WICS model of leadership. In: Rumsey, M.G. (ed.) *The Oxford Handbook of Leadership*, pp. 47–62. New York: Oxford University Press.

Stone, A.G. and Patterson, K. (2005). *The History of Leadership Focus*. *Servant Leadership Research Roundtable—August 2005*. School of Leadership Studies, Regent University.

TEPHINET (2012). *15th Anniversary Report 1997–2012*. Decatur, GA: The Task Force for Global Health, Inc.

Turning Point Program (2001). *Collaborative Leadership and Health: A Review of the Literature*. Seattle, WA: Leadership Development National Excellence Collaborative, Turning Point Program.

Umble, K., Baker, E., Diehl, S., et al. (2011a). An evaluation of the National Public Health Leadership Institute–1991–2006: Part II. Strengthening public health leadership networks, systems, and infrastructure. *Journal of Public Health Management and Practice*, **17** (3), 214–24.

Umble, K., Baker, E., and Woltring, C. (2011b). An evaluation of the National Public Health Leadership Institute—1991–2006: Part I. Developing individual leaders. *Journal of Public Health Management and Practice*, **17** (3), 202–13.

Umble, K., Steffen, D., Porter, J., et al. (2005). The National Public Health Leadership Institute: evaluation of a team-based approach to developing collaborative public health leaders. *American Journal of Public Health*, 95(4), 641–4.

Van Wart, M. (2003). Public sector leadership theory: an assessment. *Public Administration Review*, 63(2), 214–28.

Walumbwa, F.O. and Wernsing, T. (2013). From transactional and transformational leadership to authentic leadership. In: Rumsey, M.G. (ed.) *The Oxford Handbook of Leadership*. New York: Oxford University Press, pp. 392–400.

World Health Organization (1946). WHO Constitution, adopted by the International Health Conference held in New York from 19 June to 22 July 1946. [Online] Available at: http://apps.who.int/gb/bd/PDF/bd47/EN/constitution-en.pdf

World Health Organization (2003). *WHO Framework Convention on Tobacco Control*. Geneva: World Health Organization. [Online] Available at: http://www.who.int/fctc/text_download/en/index.html

World Health Organization (n.d.). *Knowledge Network on Urban Settings*. [Poster] Kobe, Japan: WHO Commission on Social Determinants of Health, WHO Centre for Health Development.

Wright, K., Rowitz, L., and Merkle, A. (2001). A conceptual model for leadership development. *Journal of Public Health Management and Practice*, 7(4), 60–6.

Wright, K., Rowitz, L., Merkle, A., et al. (2000). Competency development in public health leadership. *American Journal of Public Health*, 90(8), 1202–7.

Xirasagar, S., Samuels, M.E., and Curtin, T.F. (2006). Management training of physician executives, their leadership style, and care management performance: an empirical study. *The American Journal of Managed Care*, 12(2), 101–8.

Xirasagar, S., Samuels, M.E., and Stoskopf, C.H. (2005). Physician leadership styles and effectiveness: an empirical study. *Medical Care Research and Review*, 62(6), 720–40.

Zaccaro, S.J., LaPort, K., and Irwin, J. (2013). The attributes of successful leaders: a performance requirements approach. In: Rumsey, M.G. (ed.) *The Oxford Handbook of Leadership*. Oxford: Oxford University Press.

Yang J, Siri JG, Remais JV, et al. The Tsinghua-Lancet Commission on Healthy Cities in China: unlocking the power of cities for a healthy China. *Lancet*, **2018**, 391(10135), 2140–84.

[You can read more about this work in the report *The Revolution Will Be Improvised II*, and in the independent evaluation, The Difference that Makes the Difference. There's also more information at http://www.systemsleadership.org]

3.2

Ethical principles and ethical issues in public health

Nancy Kass, Amy Paul, and Andrew Siegel

Medical ethics, bioethics, and the emergence of public health ethics

While some cite the Hippocratic Oath as the first articulation of moral duties of physicians, and others cite John Gregory or Thomas Percival as giving birth to medical ethics, codes, and teachings about moral duties of physicians have existed for hundreds, if not thousands, of years (Percival 1985; McCullough 1998). Medical ethics focuses primarily on physicians' interactions with individual patients and is the focus of significant scholarship. The American Medical Association has longstanding codes of medical ethics, and instruction on medical professionalism is nearly universal in US and Canadian medical schools (Kao et al. 2003). The American Association of Medical Colleges requires that principles related to medical ethics should be taught in medical core curricula (Allen 2007); the UK General Medical Council requires medical schools to teach medical ethics; and the Medical Council of India requires students to 'observe medical ethics' as part of exemplary behaviour (Ravindran 2008).

Bioethics emerged more recently, with a broader focus. 'Bioethics' as a field arose in the 1960s and 1970s (Walter and Klein 2003) arising from questions of resource allocation, moral questions raised by new reproductive and genetic technologies, a growing patients' rights movement, and a lack of oversight in human subject research. Important scholarship grew, and a national commission was convened in the 1970s to examine questions of human research ethics. The National Commission drafted the Belmont Report (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research 1979) delineating three ethics principles for human research—beneficence, respect for

persons, and justice. This early work in bioethics not only went beyond the individual doctor–patient dyad but also informed public policy.

Bioethics ignited. Centres were created, journals started, meetings convened, and professionals from various disciplines started calling themselves 'bioethicists'. The three 'Belmont principles' became the foundation for a widely used bioethics text (Beauchamp and Childress 1979). While some suggested alternative approaches, 'principlism' remains a widely used framework for healthcare and research ethics (Clouser 1995; Jonsen 1995; Pellegrino 1995), and Belmont principles were adopted for the Council for International Organizations of Medical Sciences (CIOMS) international ethics guidelines (CIOMS 2002).

Early framers of these principles suggested no principle, a priori, should have moral superiority over any other (Beauchamp and Childress 1979). Nonetheless, issues animating bioethics early on—the need to tell patients and research participants the truth and the right to refuse medical care or research participation—were ones where respect for autonomy, perhaps given too little moral attention previously, was given pre-eminent moral status (Callahan 1984; Pellegrino and Thomasma 1988; Steinbock 1996). Informed consent, an application of the autonomy principle, became a hallmark of the new bioethics. Codes of ethics for clinical practice, which previously focused on not harming patients, added requirements to respect patients' right to decide about their own care (American Medical Association 1990).

The subfield of *public health ethics*, articulated as such, did not appear significantly in the literature until the 1990s, although Marc Lappe had a chapter 'Ethics and Public Health' in the 1986 Maxcy-Rosenau text *Public Health and Preventive Medicine* (Lappe 1986).[1] This chapter outlined core public health ethics challenges: fair distribution of resources, rights of individuals versus groups, and promoting efficiency while recognizing the 'special standing of those who are disadvantaged in terms of their health status' (Lappe 1986, p. 1875). Moreover, Lappe asserted that whereas medical ethics is more concerned with individual autonomy and duties of single health professionals, public health ethics focuses more on equity and efficiency. Lappe suggested individual rights can be compromised for community interests, but only with proportionality—an intervention's benefit must outweigh its harm, and the absolute level of infringement must be minimal. Related, since many public health programmes do not grant

individuals right of refusal, there must be evidence that programmes will provide the good promised. He suggested an inevitable tension between utilitarianism and justice with an ethic of public health capturing the urgency for efficiency, while recognizing the special standing of those in greatest need.

An additional exception from this early history is Dan E. Beauchamp's work, suggesting social justice and communitarian traditions should drive public health (Beauchamp 1976b). While public health practitioners had recognized, for more than a century, the relationship between social conditions and health (see Fee 1977),[2] Beauchamp's work suggested a new idea *within bioethics* that exclusive attention to individual interests, particularly through market justice, will hinder protecting the public's health. Rather, public health should 'require thinking about and reacting to the problems of disability and premature death as collective problems of the entire society' (Beauchamp 1976b, p. 9).

Furthermore, Beauchamp challenged bioethics to see public health as furthering community interests in terms of shared commitments we have to one another. Through collective actions in health and safety, for example, we share commitments to the common life (Beauchamp 1985, p. 34). While Beauchamp did not use the words 'public health ethics', his work is foundational in putting forward the idea that public health has its own set of moral priorities.

By the early 2000s a need for a more structured way to examine ethics issues in public health surfaced, and literature examined how public health ethics issues differed from those in medicine or other areas of bioethics. In 2001 and 2002, four articles were published defining the territory of public health ethics and/or offering tools and frameworks for its analysis (Kass 2001; Callahan and Jennings 2002; Childress et al. 2002; Roberts and Reich 2002), suggesting public health ethics requires identifying programme goals, determining effectiveness, minimizing burdens, proportionality, furthering social justice, and ensuring procedural justice. Publication of the first code of ethics by the American Public Health Association, an organization that had existed since 1832, followed in 2002 (*APHA Code of Ethics*; American Public Health Association 2002).

Ethics and health promotion

In the 1970s, several scholars began addressing the degree to which governments should regulate personal behaviour to promote public health. While many of the discussions were influenced by John Stuart Mill's view that government restrictions on an individual's liberty are justified only to prevent harm to others (Mill 1869), debates ensued over whether interventions to improve the health of the person targeted could be morally acceptable. Daniel Wikler suggested that promoting certain lifestyles conveyed a set of moral values when 'It is not self-evident that this vision of a safe society, with its lack of immoderation, stress, and risk-taking, is to be favoured over others whose constitutive elements have incidental adverse effects upon health' (Wikler 1978a, p. 231).

Edmund Pellegrino, one of the 'fathers of bioethics', described an ethics of prevention (Pellegrino 1981). Preventive interventions, he suggested, ranged from more voluntary approaches like health education to increasingly less voluntary ones like opinion manipulation through mass media, tax and insurance incentives and disincentives, and legal prohibitions in the name of what some consider to be 'the good life', highlighting that we are forced to make decisions about what constraints we will accept to further the common good.

Faden and Faden, similarly, described interventions ranging from facilitation to persuasion to manipulation to coercion, arguing the acceptability of different approaches rests on how rational and resistible they are (Faden and Faden 1978). Described here, to be cited frequently throughout future public health ethics writings, was the concept of *proportionality*: the burden posed by (particularly non-voluntary) interventions should be low and benefits high. As such, incentives should be favoured over disincentives, education favoured over manipulative messages, and government intervention ought not occur without considerable evidence about effectiveness.

The Society of Public Health Educators' code of professional ethics went further, calling voluntariness the *only* acceptable approach, suggesting health educators should 'empower individuals to adopt healthy lifestyles through informed choice rather than by coercion (Society for Public Health Education n.d.).

Purely voluntary interventions do not always work, however (Glanz et al. 1997; Roter et al. 1998), and the question for ethics scholars became whether or under what conditions more controlling interventions were

acceptable. Furthermore, since governments would be implementing stricter measures, how could one ensure their agenda related to public health rather than politics? (Faden 1987). Involuntary measures also assume ‘a benign, wise, and responsive government, something history finds singularly rare’ (Pellegrino 1981, p. 375).

Dan Beauchamp suggested shifting focus from the degree to which governments can regulate individual behaviour to, instead, the legal authority of public health to regulate behaviour of those marketing and distributing harmful products (Beauchamp 1976a). Wikler similarly suggested that if outside influences encourage an individual to alter his or her preferences unknowingly, this diminishes the autonomy with which those choices are made (Wikler 1978b). Coercive measures may be needed to control the manipulation of messages that run counter to the interests of public health.

The rise of chronic disease has more recently shifted the focus of public health ethics back to the acceptability of different types of behavioural interventions or public policies designed to influence behaviour. The high burdens of tobacco-related illness, injury, obesity, and diabetes have prompted interventions that target environmental, behavioural, and lifestyle determinants. In obesity prevention, for example, public health ethics must consider the justifiability of government action to restrict access to unhealthy products such as sugar sweetened beverages in various programmes or institutions as well as the effects such proposals would have on morally relevant values such as liberty, privacy, and fairness. Ethics debate surrounds whether all consumer choices constitute protected liberties, and when, if ever, restrictive policies are justifiable absent direct threats to others.

Infectious disease and public health ethics: HIV and pandemic preparedness HIV/AIDS

Few challenges have forced examination of ethics and public health as much as HIV/AIDS. Multiple public health ethics tensions surfaced, from societal rights versus individual liberties to justice and healthcare access, and arose in core functions of public health: surveillance, disease reporting, containment, and resource allocation. As policymakers navigated tough

decisions, much bioethics literature emerged, influencing what would soon be called public health ethics. Arguably, it was HIV that introduced many bioethics scholars to the world of public health.

In 1983, the Public Health Service (PHS) recommended that gay men be discouraged from donating blood (Centers for Disease Control (CDC) 1983), and soon policy documents and individual scholars examined ethics in HIV policy (Bayer 1983; Institute of Medicine 1986; Presidential Commission on the Human Immunodeficiency Virus Epidemic 1988). Fear and uncertainty precipitated proposals from tattooing infected persons to full quarantine that were highly restrictive while having little evidence of efficacy. Despite almost uniform rejection from bioethics and legal communities (Gostin and Curran 1986; Koop 1986; Macklin 1986; Musto 1986; Porter 1986), policy proposals continued to recommend exclusion of HIV-infected persons from employment, housing, insurance, and school. Bioethicists jumped into the fray to help articulate appropriate public health response and put forward ethical requirements for new screening programmes (Bayer et al. 1986; Childress 1987; Gostin and Curran 1987; Gostin et al. 1987; Hunter 1987; Bayer 1989), cognizant of the tension between potential public health benefits and concerns of exposing at-risk individuals to stigma with no treatment in return.

HIV renewed interest in the ethics of other tools of public health with calls for mandatory HIV reporting, contact tracing, and partner notification. Proponents suggested mandatory reporting would improve understanding of the disease and target education and potential treatments. Critics argued education and treatments could be provided without reporting, which was an unjustified invasion of privacy, particularly when strong antidiscrimination laws did not exist; concerns existed that at-risk individuals would avoid testing, fearing consequences. Support for mandatory reporting increased when treatments became available. HIV stimulated ethics debate in almost every area of public health including health education, duty to treat, resource allocation, access, and financing. While the phrase ‘public health ethics’ was not yet used, the foundation of the field had been laid.

Pandemic preparedness

The re-emergence of pandemic disease in the 2000s pushed public health ethics thinking to modify frameworks for the pandemic context. While the infectious nature of pandemic flu raised familiar tensions between individual interests and population health, the urgency and potential for insufficient resources highlighted additional challenges. Avian flu and severe acute respiratory syndrome outbreaks, for example, were met with insufficient healthcare workforce, vaccine shortages, and complex global governance challenges when pandemics crossed borders, raising questions of healthcare worker obligations, fair resource allocation, and justifications for restricting liberty interests (Gostin 2004; Kotalik 2005). Resulting ethical guidance (University of Toronto 2005; Thompson et al. 2006; CDC 2007, 2011) offered principles for a response while acknowledging the importance of procedural values including reasonable and evidence-based decision-making, inclusivity, transparency, and accountability.

Pandemic influenza also renewed focus within public health ethics on resource allocation as questions arose concerning how best to allocate scarce vaccines, hospital beds, ventilators, and everyday supplies. The US National Vaccine Advisory Committee and Advisory Committee on Immunization Practices recommended giving priority to those in greatest medical need (US Department of Health and Human Services 2005), some asserted priority be given to those in whom we have invested some societal resources, yet who have not had opportunity to experience multiple life-stages, meaning young adults would be prioritized above infants (Emanuel and Wertheimer 2006), others underscored the importance of considering vulnerable populations in allocation (Uscher-Pines et al. 2007; Rosoff and DeCamp 2011; DeBruin et al. 2012), and some emphasized the importance of securing access for those with important roles in maintaining societal functioning and infrastructure (Kass 2001). Questions of resource allocation in a pandemic prompted procedural interests in public engagement as a strategy to incorporate public input into policy decisions (Bailey et al. 2011; Vawter et al. 2011).

Ethical issues raised by recent global epidemics

Ebola epidemic (2014)

Localized epidemics of an haemorrhagic fever caused by the Ebola virus have been reported in central Africa previously, but usually involved no more than several hundred cases. In February 2014, a large epidemic involving 15,261 patients and causing 11,325 deaths suddenly broke out in three countries of West Africa, including Sierra Leone, Liberia, and Guinea (Gostin 2018; Dalziel et al. 2018). Because the early phase of the infection presents symptoms common to many infectious diseases, the disease is difficult to diagnose until it is well advanced. The case fatality proportion, over 70%, is very high. Because the infection is very contagious, transmitting the virus at low doses, and because of the poor infrastructure in the affected countries, the disease spread rapidly. The extremely high level of infectiousness required health personnel to wear heavy protective gear that created a barrier between them and their patients. Thus, the epidemic presented many ethical challenges that, while they may exist to smaller degrees with other infectious outbreaks, were dramatically heightened in this context.

The high risk of infection and subsequent death raised the ethics question of the limits of moral duty of health personnel to provide treatment, given the significant risk to them and, thus, to their ability to continue to provide healthcare (Dubov et al. 2016; Venkat et al. 2015). This concern had been raised previously in the context of the HIV/AIDS outbreak's early years, and indeed through the centuries when outbreaks of influenza and the plague. In West Africa, many healthcare personnel nonetheless agreed to provide care, and other healthcare personnel from other parts of Africa, and from other parts of the globe, stepped forward. And yet this created a moral obligation for the governments and health systems within the affected countries to provide protective gear and treatment facilities. Because the three affected countries are among the poorest in the world, they were hard pressed to meet this obligation, with an untenable dilemma emerging, at least in the early weeks of the outbreak. This underscored a separate issue of what moral obligation wealthy countries have to assist the affected countries (Royo-Bordonada and Lopez 2016)—although many argued that doing so was not only a compassionate response, but also a rationally self-interested one, given how rapidly outbreaks spread globally.

Further, outbreak sparked discussion about the need for wealthy nations to not only assist in the response, but make longer-term investments in the public health systems of the affected countries (Cancedda et al. 2016; Kieny

and Dovlo 2017,). The absence of adequate surveillance, laboratory, and health information systems in Liberia, Sierra Leone, and Guinea compounded the challenges of responding to and controlling the epidemic. Beyond contributing to the response, the outbreak prompted calls for the global community to make investments that would not only hasten the end of the outbreak, but leave the country better prepared to respond to future threats (Gostin 2014).

The scarcity of medical supplies meant that not everyone could receive treatment, thus the issue of priority (Satalkar et al. 2015; Thompson 2016). Related, ethics questions were raised when ex-patriate healthcare workers who became infected were airlifted to more sophisticated medical institutions outside of the country and, in some cases, were given access to experimental treatments, while healthcare providers from the affected countries who became infected were not provided with similar access. While it was clear that employers from the US and Europe, for example, generally had contracts with their employees that guaranteed them evacuation and high quality medical care should they become sick—and local employers provided no such assurances— this highlighted the longstanding disparities between rich and poor countries as an injustice in the response.

Given the high fatality rate and the lack of effective treatments, there was a surge in development of experimental vaccines and treatments. Vigorous debates focused on whether experimental treatments should be provided to all patients who were ill, or whether experimental treatments should be tested through placebo-controlled trials, with the argument that the traditional testing of experimental drugs would more quickly measure their safety and efficacy (Salerno et al. 2016; Thompson 2016; WHO Advisory Panel 2014). Some suggested that this was a context where adaptive trials made the most sense—changing a randomization allocation scheme based on interim findings and, perhaps, even starting with provision of experimental treatment to all.

That the outbreak was occurring in many remote areas, and spreading within and among households, meant that outreach, education, and community involvement was essential both to disseminate key prevention messages and to hear about new cases, rumours, or fears. (Salerno et al. 2016). And yet, without careful attention to how such outreach was to be implemented, the outreach workers themselves, particularly if they entered

in strange protective suits, could appear frightening at a time of heightened anxiety within communities. Questions emerged about whether and when it was appropriate to remove people from homes to bring to treatment centres that might be poorly equipped and whether and when individuals should be isolated or quarantined, particularly if they would then be shunned by their communities or if there was no way to guarantee food, water, and communication for them (Venkat et al. 2015; Thompson 2016).

Zika epidemic (2016)

The 2016 epidemic of congenital malformations attributable to infection by the Zika virus presented an additional set of ethical challenges. Zika is the only mosquito-borne virus known to cause congenital malformations in foetuses. Risk of neurologic malformations is highest in foetuses infected during the first (55%) and second (52%) trimesters of pregnancy, but is still high in the third trimester (29%) Risk of anencephaly is highest in those infected in the first trimester (Brasil et al. 2016; Honein et al. 2017).

Although the risk of adverse outcome of pregnancy is high (42–85%), 80% of Zika infections result in no clinical symptoms, thus making implementation of appropriate surveillance and identification of at-risk (pregnant) women difficult, especially since the risk is highest in early pregnancy when women may not know they are pregnant. Further, globally. the majority of pregnancies are unplanned. Guidelines were issued for all females of childbearing age; these included wearing long sleeves, using mosquito repellents, limiting travel to Zika-endemic areas, and delaying pregnancy (i.e. barrier sex strategies). While these guidelines are useful, the majority of women globally at risk for Zika infection are poor, poorly educated, likely to live in endemic areas, have little access to repellents on an ongoing basis, and have little control of safe sex due to cultural and religious constraints (Editorial 2016). Thus, the majority of women at risk had little meaningful authority or practical ability to protect themselves and little access to termination of pregnancy should they have chosen that if they became infected.

Absent an effective vaccine, control of the mosquito population is the primary community-targeted intervention. Challenges that are pervasive in public health ethics emerge in terms of whether it is acceptable to provide community-wide interventions, such as spraying, that may be objectionable

to some. Who gets to decide, how great the risk must be for public health officials to make unilateral decisions, and what type of consultation can be implemented, even if not all individuals can be consulted, become key ethics questions in these situations. Vaccination against Zika infection, although not available at the time this chapter was written, is likely to become available, raising the question of who should receive the vaccine. Increased attention is directed to the importance of providing vaccines to pregnant women who have traditionally been omitted from public health vaccination campaigns due to fears of fetal harm, with less parallel attention to the need to prevent both pregnant women and their future newborns from dangerous infections. All vaccines will need to be considered in terms of their risks and their benefits, and the same of course will be true in considering administration for pregnant women.

Measles outbreak (2015)

The epidemic of measles in Los Angeles in 2015 illustrates another set of public health ethics issues. Measles outbreaks in the United States have generally been rare, due to herd immunity—that is, the vast majority of individuals are vaccinated, and thus any cases or exposure are unlikely to result in secondary spread. In general, a coverage rate of over 90% is necessary to assure herd immunity and prevention of a measles epidemic. However, due to misinformation about the safety of vaccines widely available on the internet and on social media, and likely due to little awareness of the dangers that measles presents, many parents have refused to have their children vaccinated. Thus, in a particular region of Southern California, with demographically highly educated parents, the coverage proportion fell well below 90%. And indeed this was possible because, at the time, the state of California had one of the most lenient vaccination policies in the country, allowing exemptions from childhood vaccination for ‘philosophical reasons’, considered to be an extremely low bar. When one index case of an infected child attended Disneyland, thus exposing thousands of people, a measles outbreak occurred in Southern California, which then spread as far as Europe. The California legislature subsequently passed a law mandating full immunizations for children entering school with exceptions allowed only for immunocompromised children or any others for whom it would be dangerous. However, many mothers pressure

their paediatrician to write letters of exception for their child, threatening the attainment of the necessary level of herd immunity. Thus, some schools in California have exception rates as high as 58%, rendering the proportion protected well below the 90+% required to develop herd immunity (Los Angeles Times 13 July 2018a and b). What type of vaccination policy for a state to have is a longstanding and difficult ethical issue. Vaccination is one of the most effective public health preventive interventions, preventing millions of people from having a serious illness through an intervention that poses very little risk. And yet by mandating it, the autonomy of the small number of individuals who might have objected is not being respected, due to the importance of protecting others and the community at large. Indeed, because outbreaks are infectious, and because some numbers of children medically cannot be vaccinated, requiring vaccination is a means of protecting all, and protecting the most vulnerable. It is for these reasons that public health departments across the country generally require all children to be vaccinated as a condition of entering school, camp, or daycare, and allow exemptions only for medical or, in some cases, religious, reasons.

Public health and social justice

Until recently, the more theoretical literature on justice and health focused almost exclusively on issues of access to healthcare and the allocation of scarce healthcare resources. While these remain important and contested areas of inquiry, the epidemiological evidence indicates that healthcare plays a relatively limited role in determining health status (Wilkinson and Marmot 2003). It has long been understood in public health that socioeconomic status (SES) (viz. status as defined by income, education, and occupation) is a major determinant of health, giving rise to significant health inequalities between social groups. Some health disparities are due to material conditions related to poverty, such as poor nutrition and crowded and unsanitary living conditions. But studies over the past two decades suggest that poverty alone does not explain the effects of SES on health. There is a gradient in life expectancy, mortality, and morbidity across all levels of the SES hierarchy (Marmot et al. 1991; Marmot and Shipley 1996). Health inequalities in the gradient that exist in the absence of material deprivations appear to be best explained by psychosocial factors,

such as levels of social support and integration, control in the workplace, and hostility (Marmot 2004). There are also studies that, while controversial, suggest that an individual's health is influenced by the level of income inequality in society as a whole, and not just by her own level of income (Kawachi et al. 1999). Increasing attention to the social determinants of health has given rise to a heightened focus on social justice in public health.

The central social justice issue in the context of public health concerns when health inequalities between social groups are unjust. There are several approaches in the literature to addressing this question, all of which are developed from more general theoretical accounts of distributive justice. Theories of distributive justice identify a metric specifying the good that is sought and a principle for the distribution of that good. The primary metrics include welfare, resources, and capabilities, while the competing principles of distribution appeal to equality, priority, and sufficiency.

Metrics of justice vary, with some viewing health as purely instrumental to the good justice seeks and others viewing health as intrinsic to that good. Instrumentalist accounts include Norman Daniels' view that the value of health lies in its serving to secure equality of opportunity to pursue life plans (Daniels 1985, 2008) and the welfare egalitarian and contemporary utilitarian view that health is valuable insofar as it fulfils individual preferences. Each account is subject to the objection that health has value even when failing to promote these ends. Health seems important even for those near the end of life who have completed their life plans (Brock 1989; Segall 2010); health further can have value where one has a weak preference for it, particularly where the preference is due to an adaptation to conditions under which one has been persistently deprived of health (Sen 2009).

An alternative to the instrumentalist perspective conceives of health as partly constitutive of the ends of justice. Various capabilities theorists have argued that a person's ability to achieve good health is a fundamental capability inasmuch as health is essential to human life and flourishing (Sen 2002; Nussbaum 2006; Ruger 2009; Venkatapuram 2011). Madison Powers and Ruth Faden offer a related but distinct conception, viewing health as one of several dimensions of well-being (which also include personal security, respect, attachment, self-determination, and reasoning) that justice aims to secure. But they privilege the achievement of positive health

outcomes over capabilities to achieve health, both because some individuals (children, in particular) are not well positioned to make free and informed choices about their health and because there are threats to health that may not be within an individual's ability to control (e.g. exposure to environmental harms) (Powers and Faden 2006, 2011).

In assessing whether a given health disparity is unjust, one must turn to distributive principles. On a strict egalitarian account, everyone should, as far as possible, have equal health status. This account has been widely dismissed as it would permit 'levelling down' the health status of some without improving the health of others as a means to achieve equality (Parfit 1998). Alternatively, the prioritarian account permits inequalities where they benefit those who are worst off. John Rawls' Difference Principle, which allows inequalities in income and wealth (among other goods) only where they benefit the worst off, is a prominent example of prioritarianism (Rawls 1971). Daniels has applied this view to health, suggesting that distributing social goods in accordance with the Difference Principle would substantially reduce the steepness of the health gradient (2008). Another prioritarian account gives weight to personal responsibility for health differentials, positing that those who do more to promote their health should be given priority (Segall 2010). Sufficientarians, like prioritarians, are concerned with how persons fare in absolute terms; but they hold that justice should ensure that everyone has enough of the goods essential to living a decent life, without regard to inequalities beyond the sufficiency threshold (Frankfurt 1987). Powers and Faden, for example, maintain that justice demands that everyone have a sufficient amount of health, and that the most pressing health inequalities are those that are systematically accompanied by inadequately met needs in other core dimensions of well-being (Powers and Faden 2006).

Given how many people lack adequate health and how little of the vast social and economic gains to the well-off have been accompanied by gains to the worst-off, following either the prioritarian or sufficientarian views of justice likely would lead to profound improvements in population health. While public health practitioners do not bear primary responsibility for addressing the full range of the social determinants of health, some argue that public health should more explicitly incorporate a social justice orientation. Strategies to do so are of course more challenging, and include monitoring social indicators related to health inequalities to engaging in

community organization and collaborative practices that help promote the structural changes necessary to address the material and psychosocial causes of health disparities (Plough 2005).

Public health research

Literature in ethics and public health research falls into two categories. First, there is literature exploring how public health research can be distinguished, conceptually and operationally, from public health practice (CDC 1999; Casarett et al. 2000; Bellin and Dubler 2001; Amoroso and Middaugh 2003). This literature outlines criteria for when an activity is quality assurance, surveillance, or evaluation versus when it is research requiring regulatory oversight (US Department of Health and Human Services 2009). This growing body of literature recognizes the blurry lines between research and practice in many arenas, including public health, and increasingly questions whether distinguishing research from practice is the most salient approach for ensuring ethical protection (Faden et al. 2013). As more data are available electronically, and with greater emphasis on learning healthcare systems and evidence-based practice (Institute of Medicine 2007; Kass et al. 2013), blurriness between research and practice will only grow larger.

Second, there is literature examining activities that unquestionably are research, but exploring whether public health research is different from clinical research, such that it should be exempt from certain regulatory requirements; literature also has addressed whether ethical challenges differ when entire populations are targeted for research.

Distinguishing public health research from non-research activities of public health

In 1999, the Centers for Disease Control and Prevention (CDC) issued what has become a widely cited document aimed at distinguishing public health research from non-research (CDC 1999). These guidelines were written in response to queries from outsiders whether certain data gathering activities conducted by CDC or by state health departments for public health practice

should be categorized as research (Burris et al. 2003). While the guidelines were written to guide CDC employees in their own work, they have been used more broadly. The document asserts that public health practice can be distinguished from research based on what the primary intent of the activity is. If an activity is designed, primarily, to help a public health department do its job furthering the health interests of a particular community, even if it *also* produces data of more generalizable interest, the activity should be called non-research. Research, the guidance continued, is intended to produce generalizable knowledge, to have relevance beyond the population or programme from which data were collected. Surveillance, then, would be considered non-research when conducted in response to 'lawful state disease reporting or monitoring' activities. In 2001, the National Bioethics Advisory Commission echoed CDC guidance stating that projects *intended* to produce generalizable knowledge must be subject to the oversight provided by the Common Rule (National Bioethics Advisory Commission 2001).

The scholarly literature has echoed many of these concepts, especially that intention to produce generalizable knowledge characterizes research requiring federal oversight. Casarett et al. (2000), by contrast, suggest that activities should be regulated as research if: (1) the majority of patients involved are not expected to benefit directly from the knowledge to be gained and (2) additional risks and burdens exist as a result of wanting findings to be generalizable. Participants in a CDC workshop on practice versus research recommended research oversight to be determined by level of risk, rather than distinguishing based on primary intent (MacQueen and Buehler 2004). Bellin and Dubler suggested activities can be considered non-research if there is a commitment to modify programmes based on collected data (Bellin and Dubler 2001), and a recent framework suggests that the degree to which learning and collecting data increase risks or threaten important interests or well-being should dictate the degree of ethical oversight required, regardless of whether an activity is called research or practice (Faden et al. 2013).

Burris et al. contend that collection of data by government agencies fulfils their mandated functions and should be exempt from federal research regulations specifically because they are conducted by public health departments. They suggest an alternative, internal oversight system be created to ensure that 'human beings who become involved in activities that

increase our knowledge (whether defined as research or not) should be protected' and ensure that data collection activities are conducted in the least harmful, least restrictive, and most respectful way possible (Burris et al. 2003).

Fairchild and Bayer, however, suggest *all* data collection by federal or state governments should be subject to ethical review. While not suggesting institutional review boards (IRBs) necessarily perform this review, they advocate for universal review of these activities (Fairchild and Bayer 2004). With so much non-health data widely used and shared through social media, and through the internet more broadly, questions are emerging going forward about whether views about health privacy will change as the sharing of health data likely also becomes more widespread (Voyena et al. 2012).

Ethical issues raised by public health research

Other literature explores whether ethics issues in *public health research* differ from those of clinical research. Papers from the 1970s and 1980s suggested that certain types of epidemiological studies using identifiable records should be exempt from informed consent requirements (Gordis et al. 1977; Kelsey 1981; Waters 1985). Gordis et al. provide dozens of examples, from identifying the link between cigarette smoking and lung cancer, to the link between high concentrations of oxygen for premature infants and blindness, to the link between oral contraceptives and stroke, illustrating the contributions that epidemiological record review has made to public health and suggesting such methods would be nearly impossible if consent were required (Gordis et al. 1977). More recent literature examines whether and how population-based or prevention research differs in ethically relevant ways from clinical research (Taylor and Johnson 2007). Population-based research may not allow individual participants the opportunity to refuse participation, for example, when studies investigate community-wide interventions, like the impact of health promotion media campaigns in some communities and not others. In such cases, IRBs must confirm that risk to individuals is minimal and public health benefit from the research outweighs concerns imposed by individuals' inability to refuse.

Data are generally collected at the population level without individual identifiers.

Population-based research has the potential to stigmatize identifiable groups if the research targets specific geographic, religious, or ethnic communities, or groups defined by a risk behaviour such as injection drug use or sexual orientation. The potential for stigma and social harms was a prominent concern of genetic research and the launch of the Human Genome Diversity Project in 1991, as entire races and indigenous communities were the unit of interest. Social harms animated a recent case, *Havasupai Tribe v. Arizona Board of Regents*. Stored genetic samples for research ultimately undermined traditional understandings of the tribe's ancestral origin, indicated high levels of inbreeding within the tribal community, and assessed associations with schizophrenia, a highly stigmatized condition within the tribe (*Havasupai Tribe v. Arizona Board of Regents* 2008). While research ethics has traditionally allowed analysis of samples that are not individually identifiable, this case demonstrated that population-based research carries the risk that outsiders might think differently about an entire social group as a result of research findings.

It is then not surprising that population-based research is the origin of both community advisory boards (Strauss et al. 2001; Morin et al. 2003; Quinn 2004) and community-based participatory methods (Israel et al. 1998; Macaulay et al. 1999), strategies designed to include community input into planned research methods and dissemination plans. Challenges remain regarding what is meant by 'community' and 'engagement' in the context of research (Hillery 1955; CDC 1997; MacQueen et al. 2001; Sharp and Foster 2007), as well as determining when community input must be sought, when a community has been fully represented, and how to resolve disputes within communities (Sharp and Foster 2007).

Research in resource-poor settings raises additional ethical challenges, both globally and domestically, and points to why the CIOMS guidelines include sections that other research ethics guidelines do not on incentives, responsibilities when studies are over, and justice. A major controversy erupted in 1997 after US-funded HIV trials in low-income countries tested the efficacy of drugs to prevent vertical transmission of HIV. Studies were placebo-controlled even when an effective, albeit expensive and complex, treatment protocol existed. While some scholars and human rights activists asserted that use of placebo was in clear violation of ethical guidelines,

causing millions of infants to contract HIV (see Lurie and Wolfe 1997), others argued that women in the trials were no worse off by participating in trials, and placebo-controlled designs more validly tested whether locally relevant interventions were effective, to then be implemented in the field more quickly (Varmus and Satcher 1997). The controversy resulted in multiple revisions of the Declaration of Helsinki and CIOMS guidelines on the use of placebos, yet consensus on this question has not been reached.

Finally, similar to clinical trials, population-based trials—particularly those conducted in resource-poor settings—increasingly must consider what, if anything, will be made available to study communities when the research is completed. Several national and international bodies have suggested that successful research interventions should be made reasonably available to host country communities (CIOMS 2002; National Bioethics Advisory Commission 2001)—or at least justify why such availability is impossible if that is the case (Nuffield Council of Bioethics 2002), while others argued that many donor or government commitments never could have been secured until study results were in hand. An alternative approach suggested that 'fair benefits' should be provided to communities, regardless of whether those benefits are directly related to the study intervention being tested (Conference on Ethical Aspects of Research in Developing Countries 2004). Recent HIV Prevention Trial Network (HPTN) ethics guidance acknowledges that although conclusive definitions of post-study access agreements may not be possible prior to testing the intervention, the development of a detailed, preliminary plan of post-study access is nonetheless necessary (HPTN 2009).

Important in the evolution of thinking about public health research ethics was a controversial legal case of 2001 related to a lead abatement trial conducted in the 1990s in Baltimore (*Ericka Grimes v. Kennedy Krieger Institute, Inc.* 2001). This trial targeted low-income housing units in old neighbourhoods that had significant lead paint in units and where units often were in disrepair; that is, where children would be at risk of lead poisoning due to the existing poor conditions (Farfel and Chisolm 1990, 1991, 1994; Pollak 2002). Despite evidence existing for decades documenting the danger of lead paint to children, no laws existed requiring that houses be safe or abated before they could be rented to families. In the trial, households were randomized to one of several different lead abatement strategies to see the effect on both household lead dust and

children's own blood lead levels of the different abatement strategies. Two families later brought lawsuits, charging that researchers failed to warn them in a timely manner that children might still be at risk of lead poisoning. This case was paradigmatic of certain types of public health research, in that it targeted a particular community, it dealt with prevention of a major public health problem, and, significantly, it was designed to respond to a series of baseline conditions that were harmful to public health. There remains no consensus regarding whether research is exploitive when situated in settings of neglect and when testing interventions that may improve conditions, but will not improve them as much as other, existing, but clearly unavailable interventions. Scholars have questioned whether conducting research on 'partial solutions' condones the idea that partial solutions are adequate for the disadvantaged (Farmer and Campos 2004; Spriggs 2004; Buchanan and Miller 2006; Kass 2008). Miller and Buchanan suggest that rejecting all research on partial solutions out of a 'presumption that a particular conception of justice will eventually prevail'—sacrifices the welfare of 'literally millions' of (in this case) children who could benefit from the results of the partial solutions tested through public health prevention research. Related, they assert that it is inappropriate to blame researchers for 'the lack of social consensus' on the right to better social conditions and that doing so is 'misplaced indignation' (Buchanan and Miller 2006). Others, however, suggest that even the perception of exploitation raised through this case is harmful for the research enterprise and suggests a need for 'true partnership in the research enterprise, particularly when proposed research involves vulnerable communities' (Mastroianni and Kahn 2002).

Ethical issues raised by genetic research

The rapid advancement of genetic research over the past two decades has presented new challenges for public health research, including issues of consent for storage of genetic material, return of genetic test results, and assignment of ownership of genetic material.

Because biospecimens used in genetic research are frequently collected and stored in a 'biobank' from which they are accessed for a wide range of future studies, it is not always possible to know how specimens may be

used in the future. Re-contacting participants each time their sample will be used has been viewed as impractical and overly burdensome to both researchers and participants, yet 'blanket consent', in which participants agree to any future use of their samples, has been criticized as an insufficient means of respecting participant autonomy (McGuire and Beskow 2010). While empirical work shows most participants support data-sharing initiatives that enable widespread use of their samples, participants expressed lingering concerns about the capacity of data-sharing initiatives to fully protect participant data, suggesting initiatives should be transparent to maintain trust with participant populations (Trinidad et al. 2010). At the same time, there is increasing opportunity for individuals to upload their own genetic data into private non-profit and commercial databases for use by a wide network of researchers, for example, through online initiatives such as 'Consent to Research' (http://weconsent.us/) and the 'Personal Genome Project' (www.personalgenomes.org). As biobanks have sprung up across the globe and information is increasingly available electronically, the potential for global data sharing is quickly becoming a reality (Kaye 2011). While such interconnectedness leverages the resources available to researchers and opens new opportunities for research, it amplifies the potential for uncontrolled use of specimens well beyond original and anticipated uses. Consent procedures for biobanking and the ethics of global data sharing is in its infancy, and ethicists are grappling with the challenges of developing controls and accountability mechanisms (Gottweiss and Lauss 2010; O'Doherty et al. 2011).

Disclosure of results from genetic tests performed during research also raises controversy. Some argue results should be returned as a way to respect autonomy and allow participants to find meaning in results even without clinical implications; returning results may also demonstrate reciprocity for having participated in research (Ravitsky and Wilfond 2006). Others counter that supplying genetic results that are often complex and offer limited clinical value would be misleading and potentially harmful (Beskow 2006). Even when results are readily interpretable and have clinical implications, informing participants of results requires retaining linkages to individual samples. The National Bioethics Advisory Commission and the National Cancer Institute's Best Practices for Biospecimen Resources state merely that the informed consent document contain a statement explaining whether, which, and to whom individual

results would be released (National Bioethics Advisory Commission 1999; National Cancer Institute 2011). Taking a stronger stance, the National Heart, Lung and Blood Institute convened working groups in 2004 and 2009 that drafted recommendations outlining conditions under which results ethically should be returned, circumstances in which research may but should not be expected to return results, and that emphasized the importance of community engagement when working with identifiable communities (Fabsitz et al. 2010).

Finally, genetic research has given rise to questions of identity, ownership, and property in ways not seen in other types of public health research. Lawsuits filed in the early 1990s and 2000s addressed ownership of biological material used in genetic research (*Moore v. Regents of the University of California* 1990; *Greenberg v. Miami Children's Hospital Research Institute* 2003; *Washington University v. Catalona* 2007) and plaintiffs claimed a property right to the materials developed from their genetic samples. Defendants argued that samples were 'donated' and individuals relinquished ownership rights when samples were collected. Case law to date has favoured the research enterprise in rulings that do not recognize participant ownership of biospecimens collected for research, yet questions remain, ethically, as to whether viewing biospecimens as property is appropriate. McGuire and Beskow outline numerous approaches to informed consent in genetic research, (McGuire and Beskow 2010), and the National Human Genome Research Institute maintains a comprehensive web resource for conducting informed consent in genetic research. In addition, new initiatives seek to address these challenges through participant engagement (Kaye et al. 2012). Finally, emerging participatory models propose the use of social media as a way to form relationships between participants and researchers. As a means of ongoing communication, these initiatives empower participants to make informed choices over the use of their samples and provide a mechanism to communicate findings as their interpretation changes over time.

Conclusion

Public health has been given significant attention in the bioethics literature only in the last three decades. Public health must balance furthering the

health of communities through education, surveillance, interventions, and regulations with needing to restrict the freedoms or privacy of individuals affected to the minimum degree possible while also addressing core commitments to social justice. Public health is granted statutory authority to protect the public's health and, indeed extraordinary public health improvement has been achieved through sanitary measures, restaurant inspections, immunizations, and health education. At the same time, it is expressly because of this authority that ethics has such a critical role to play. Clear frameworks of ethics that outline a need for data, minimizing of harms, fair procedures, and clear principles to follow, including transparency, reciprocity, and equity, can help public health fulfil its duty to improve health on behalf of all of us, while allowing individuals to feel confident that any restrictions are appropriate and fairly defined. Public health research requires some new considerations for research ethics including the need to be more cognizant of the implications of research for populations as a whole, for finding creative ways to seek community-wide input, and to begin to consider how to ensure that communities, as well as individuals, realize research benefit.

REFERENCES

Allen, R. (2007). *Fostering Professionalism During Medical School and Residency Training. CME Report 3-A-01. Report of the Council on Medical Education*. Chicago, IL: American Medical Association.

American Medical Association. Council on Ethical and Judicial Affairs (1990). Fundamental elements of the physician–patient relationship. *Journal of the American Medical Association*, **264**, 3133.

American Public Health Association (2002). *APHA Code of Ethics*. Washington, DC: American Public Health Association.

Amoroso, P.J. and Middaugh, J.P. (2003). Research vs. public health practice: when does a study require IRB review? *Preventive Medicine*, **36**, 250–53.

Bailey, T., et al. (2011). Public engagement on ethical principles in allocating scare resources during an influenza pandemic. *Vaccine*, **29**, 3111–17.

Bayer, R. (1983). Gays and the stigma of bad blood. *Hastings Center Report*, **13**, 5–7.

Bayer, R. (1989). Ethical and social policy issues raised by HIV screening: the epidemic evolves and so do the challenges. *AIDS*, **3**, 119–24.

Bayer, R., Levine, C., and Wolf, S.M. (1986). HIV antibody screening: an ethical framework for evaluating proposed programs. *Journal of the American Medical Association*, **256**, 1768–74.

Beauchamp, D.E. (1976a). Alcoholism as blaming the alcoholic. *International Journal of Addiction*, **11**, 41–52.

Beauchamp, D.E. (1985). Community: the neglected tradition of public health. *The Hastings Center Report*, **15**, 28–36.

Beauchamp, D.E. (1976b). Public health as social justice. *Inquiry*, **13**, 1–14.
Beauchamp, T.L. and Childress, J.L. (1979). *Principles of Biomedical Ethics*. Oxford: Oxford University Press.
Bellin, E. and Dubler, N.N. (2001). The quality improvement–research divide and the need for external oversight. *American Journal of Public Health*, **91**, 1512–17.
Beskow, L. (2006). Considering the nature of individual research results. *American Journal of Bioethics*, **6**, 38–40.
Brasil, P., et al. (2016). Zika virus infection in pregnant women in Rio de Janeiro. *New England Journal of Medicine*, 375(24), 2321–34.
Brock, D. (1989). Justice, health care, and the elderly. *Philosophy and Public Affairs*, **18**, 297–312.
Buchanan, D.R. and Miller, F.G. (2006). Justice and fairness in the Kennedy Krieger Institute lead paint study: the ethics of public health research on less expensive, less effective interventions. *American Journal of Public Health*, **96**, 781–7.
Burris, S, Buehler, J., and Lazzarini, Z. (2003). Applying the Common Rule to public health agencies: questions and tentative answers about a separate regulatory regime. *Journal of Law and Medical Ethics*, **31**, 638–53.
Callahan, D. (1984). Autonomy: a moral good, not a moral obsession. *The Hastings Center Report*, **14**, 40–2.
Callahan, D. and Jennings, B. (2002). Ethics and public health: forging a strong relationship. *American Journal of Public Health*, **92**, 169–76.
Cancedda, C., et al. (2016). Strengthening health systems while responding to a health crisis: lessons learned by a nongovernmental organization during the Ebola virus disease epidemic in Sierra Leone. *Journal of Infectious Disease*, **214** (Suppl 3), S153–63.
Casarett, D., Karlawish, J.H.T., and Sugarman, J. (2000). Determining when quality improvement initiatives should be considered research: proposed criteria and potential implications. *Journal of the American Medical Association*, **283**, 2275–80.
Centers for Disease Control (1983). Current trends prevention of acquired immune deficiency syndrome: report of inter-agency recommendations. *Morbidity and Mortality Weekly Report*, **32**, 101–3.
Centers for Disease Control (1997). *Community Engagement: Definitions and Organizing Concepts from the Literature, Part 1. Principles of Community Engagement: CDC/ATSDR Committee on Community Engagement*. Atlanta, GA: CDC.
Centers for Disease Control (1999). *Guidelines for Defining Public Health Research and Public Health Non-Research*. Atlanta, GA: CDC.
Centers for Disease Control (2007). *Ethical Guidelines in Pandemic Influenza.* Atlanta, GA: CDC.
Centers for Disease Control (2011). *Ethical Consideration for Decision Making Regarding Allocation of Mechanical Ventilators during a Severe Influenza Pandemic or Other Public Health Emergency.* Atlanta, GA: CDC.
Childress, J.F. (1987). An ethical framework for assessing policies to screen for antibodies for HIV. *AIDS Public Policy Journal*, **2**, 28–31.
Childress, J.F., et al. (2002). Public health ethics: mapping the terrain. *Journal of Law and Medical Ethics*, **30**, 170–8.
Clouser, K.D. (1995). Common morality as an alternative to principlism. *Kennedy Institute of Ethics Journal*, **5**, 219–36.
Conference on Ethical Aspects of Research in Developing Countries (2004). Moral standards for research in developing countries: from ‘reasonable availability’ to ‘fair benefits’. *The Hastings Center Report*, **34**, 17–27.

Council for International Organizations of Medical Sciences (CIOMS) (2002). *International Ethical Guidelines for Biomedical Research Involving Human Subjects*. Geneva: World Health Organization.

Dalziel, B.D., et al. (2018). Unreported cases in the 2014–2016 Ebola epidemic: Spatiotemporal variation, and implications for estimating transmission. *PLOS Neglected Tropical Diseases*, 12(1), e0006161.

Daniels, N. (1985). *Just Health Care*. New York: Cambridge University Press.

Daniels, N. (2008). *Just Health: Meeting Health Needs Fairly.* New York: Cambridge University Press.

DeBruin, D., Laischenko, J., and Marshall, M. (2012). Social justice in pandemic preparedness. *American Journal of Public Health*, **102**, 586–91.

Dubov, A., Appleton, J.H., and Campbell, S. (2016). Ebola virus preparedness: emerging viruses and ethics in laboratory medicine. *Archives of Pathology & Laboratory Medicine*, **140**, 171–80.

Emanuel, E. and Wertheimer, A. (2006). Public health: who should get influenza vaccine when not all can? *Science*, **312**, 854–55.

Ericka Grimes, v. Kennedy Krieger Institute, Inc. (2001). *Wes's Atlantic Reporter*, **782**, 807–62.

Fabsitz, R.R., McGuire, A., Sharp, R.R., et al. (2010). Ethical and practical guidelines for reporting genetic research results to study participants: updated guidelines from an NHLBI working group. *Circulation: Cardiovascular Genetics*, **3**, 574–80.

Faden, R.R. (1987). Ethical issues in government sponsored public health campaigns. *Health Education Quarterly*, **14**, 27–37.

Faden, R.R. and Faden, A.I. (1978). The ethics of health education as public policy. *Health Education Monographs*, **6**, 180–97.

Faden, R., et al. (2013). An ethics framework for learning healthcare systems. *The Hastings Center Report*, **43**, S16–27.

Fairchild, A.L. and Bayer, R. (2004). Public health: ethics and the conduct of public health surveillance. *Science*, **303**, 631–2.

Farfel, M.R. and Chisolm, J.J. (1990). Health and environmental outcomes of traditional and modified practices for abatement of residential lead-based paint. *American Journal of Public Health*, **80**, 1240–5.

Farfel, M.R. and Chisolm, J.J. (1991). An evaluation of experimental practices for abatement of residential lead-based paint: report on a pilot project. *Environmental Research*, **55**, 199–212.

Farfel, M.R. and Chisolm, J.J. (1994). The longer-term effectiveness of residential lead paint abatement. *Environmental Research*, **66**, 217–21.

Farmer, P. and Campos, N.G. (2004). New malaise: bioethics and human rights in the global era. *Journal of Law and Medical Ethics*, **32**, 243–51.

Fee, E. (1977). History and development of public health. In: Scutchfield, F.D. and Keck, C.W. (eds.) *Principles of Public Health Practice*. Boston, MA: Delmar Publishers, pp. 10–30.

Frankfurt, H. (1987). Equality as a moral ideal. *Ethics*, **98**, 21–43.

Glanz, K., Lewis, F.M., and Rimer, B.K. (1997). *Health Behavior and Health Education: Theory, Research and Practice* (2nd ed.). San Francisco, CA: Jossey-Bass.

Gordis, L., Gold, E., and Seltser, R. (1977). Privacy protection in epidemiologic and medical research: a challenge and a responsibility. *American Journal of Epidemiology*, **105**, 163–68.

Gostin, L.O. (2018). New Ebola outbreak in Africa is a major test for the WHO. *Journal of the American Medical Association*, 320(2), 125–6.

Gostin, L. (2004). Pandemic influenza: public health preparedness for the next global emergency. *Journal of Law, Medicine, and Ethics*, **32**, 565–73.

Gostin, L. (2014). Towards an international health systems fund. *Lancet*, 384(9951), e49–e51.

Gostin, L. and Curran, W.J. (1986). The limit of compulsion on controlling AIDS. *Hastings Center Report*, **16**, 24–9.

Gostin, L. and Curran, W.J. (1987). Legal control measures for AIDS: reporting requirements, surveillance, quarantine, and regulation of public meeting places. *American Journal of Public Health*, **77**, 214–18.

Gostin, L., Curran, W.J., and Clark, M.E. (1987). The case against compulsory case finding in controlling AIDS—testing, screening and reporting. *American Journal of Law and Medicine*, **12**, 7–53.

Gottweiss, H. and Lauss, G. (2010). Biobank governance in the post-genomic age. *Personalized Medicine*, **7**, 187–95.

Greenberg v. Miami Children's Hospital Research Institute (2003). 264 F.Supp.2d 1064 (S.D. Fla. 2003).

Havasupai Tribe v. Arizona State University Board of Regents (2008). 204 P.3d 1063 (9th Cir. 2008).

Hillery, G. (1955). Definitions of community: areas of agreement. *Rural Sociology*, **20**, 111–24.

HIV Prevention Trials Network Ethics Working Group (HPTN) (2009). *HIV Prevention Trials Network Ethics Guidance for Research*. New York: HPTN.

Honein M.A., Dawson, A.L., Petersen,E.E., et al.; US Zika Pregnancy Registry Collaboration. (2017). Birth defects among fetuses and infants of US women with evidence of possible Zika virus infection during pregnancy. *Journal of the American Medical Association*, 317(1):59–68.

Hunter, N.D. (1987). AIDS prevention and civil liberties: the false security of mandatory testing. *AIDS Public Policy Journal*, **2**, 1–10.

Institute of Medicine (1986). *Confronting AIDS: Directions for Public Health, Health Care, and Research*. Washington, DC: National Academies Press.

Institute of Medicine (2007). *IOM Roundtable on Evidence-Based Medicine, The Learning Healthcare System: Workshop Summary*. Olsen, L., Aisner, D., and McGinnis, J.M. (eds.). Washington, DC: National Academies Press.

Israel, B.A., et al. (1998). Review of community-based research: assessing partnership approaches to improve public health. *Annual Review of Public Health*, **19**, 173–202.

Jonsen, A.R. (1995). Casuistry: an alternative or complement to principles? *Kennedy Institute of Ethics Journal*, **5**, 237–51.

Kao, A., et al. (2003). Teaching and evaluating students' professionalism in US medical schools, 2002–2003. *Journal of the American Medical Association*, **290**, 1151–2.

Kass, N. (2001). An ethics framework for public health. *American Journal of Public Health*, **91**, 1776–82.

Kass, N. (2008). Just research in an unjust world: can harm reduction be an acceptable tool for public health prevention research? In: Green, R.M., Donovan, A., and Jauss, S.A. (eds.) *Global Bioethics*, pp. 89–116. Oxford: Oxford University Press.

Kass, N., et al. (2013). The research–treatment distinction: a problematic approach for determining which activities should have ethical oversight. *The Hastings Center Report*, **43**, S4–15.

Kawachi, I., Kennedy, B.P., and Wilkinson, R.G. (1999). *The Society and Population Health Reader, Vol. I: Income Inequality and Health.* New York: The New Press.

Kaye, J. (2011). From single biobanks to international networks: developing e-governance. *Human Genetics*, **130**, 377–82.

Kaye, J., et al. (2012). From patients to partners: participant-centric initiatives in biomedical research. *Nature Reviews Genetics*, **13**, 371–76.

Kelsey, J.L. (1981). Privacy and confidentiality in epidemiological research involving patients. *IRB*, **3**, 1–4.

Kieny, M. and Dolvo, D. (2017) Beyond Ebola: A new agenda for resilient health systems. *Lancet*, 385(9963), 91–92.

Koop, C.E. (1986). Surgeon General's report on acquired immune deficiency syndrome. *Journal of the American Medical Association*, **256**, 278–89.
Kotalik, J. (2005). Preparing for an influenza pandemic: ethical issues. *Bioethics*, **19**, 422–31.
Lappe, M. (1986). Ethics and public health. In: Last, J.M. (ed.) *Maxcy-Rosenau Public Health and Preventive Medicine* (12th ed.). Norwalk, CT: Appleton-Century-Crofts, pp. 1867–77.
Los Angeles Times (13 July 2018a). Parents turn to doctors to avoid vaccines.
Los Angeles Times (13 July 2018b). A workaround for vaccine law.
Lurie, P. and Wolfe, S. (1997). Unethical trials of interventions to reduce perinatal transmission of the human immunodeficiency virus in developing countries. *New England Journal of Medicine*, **337**, 853–6.
Macaulay, A.C., et al. (1999). Participatory research maximizes community and lay involvement. *British Medical Journal*, **319**, 774–8.
Macklin, R. (1986). Predicting dangerousness and public health response to AIDS. *The Hastings Center Report*, **16**, 16–23.
MacQueen, K.M. and Buehler, J.W. (2004). Ethics, practice, and research in public health. *American Journal of Public Health*, **94**, 928–31.
MacQueen, K.M., et al. (2001). What is community? An evidence-based definition for participatory public health. *American Journal of Public Health*, **91**, 1929–38.
Marmot, M. (2004). Social causes of social inequalities in health. In: Sudhir, A. and Peter, F. (eds.) *Public Health, Ethics, and Equity*, pp. 37–62. Oxford: Oxford University Press.
Marmot, M., et al. (1991). Health inequalities among British social servants: the Whitehall II Study. *The Lancet*, **337**, 1387–93.
Marmot, M. and Shipley, M. (1996). Do socio-economic differences in mortality persist after retirement? 25 year follow up of civil servants from the first Whitehall Study. *British Medical Journal*, **33**, 1177–80.
Mastroianni, A.C. and Kahn, J.P. (2002). Risk and responsibility: ethics, Grimes v Kennedy Krieger, and public health research involving children. *American Journal of Public Health*, **92**, 1073–6.
Maynard A.D., Bowman, D.M., and Hodge, J.G. Jr. (2016). Mitigating risks to pregnant teens from Zika virus. *Journal of Law, Medicine and Ethics*, 44(4), 657–9.
McCullough, L.B. (ed.) (1998). *John Gregory's Writings on Medical Ethics and Philosophy of Medicine*. Dordrecht: Kluwer Academic Publishers.
McGuire, A. and Beskow, L. (2010). Informed consent in genomics and genetic research. *Annual Review Genomics and Human Genetics*, **11**, 361–81.
Mill, J.S. (1869). On liberty. In: Gray, J. (ed.) (1998) *On Liberty and Other Essays* (2nd ed.),. New York: Oxford University Press, pp. 1–129.
Moore v. Regents of the University of California (1990). 793 P.2d 479 (Cal. 1990).
Morin, S.F., et al. (2003). Community consultation in HIV prevention research: a study of community advisory boards at 6 research sites. *Journal of Acquired Immune Deficiency Syndromes*, **33**, 513–20.
Musto, D.F. (1986). Quarantine and the problem of AIDS. *Milbank Memorial Fund Quarterly*, **64**, 113.
National Bioethics Advisory Commission (NBAC) (1999). *Research Involving Human Biological Materials: Ethical Issues and Policy Guidance.* Vol. 1. Rockville, MD: NBAC.
National Bioethics Advisory Commission (NBAC) (2001). *Ethical and Policy Issues in International Research Clinical Trials in Developing Countries*. Vol. 1. Rockville, MD: NBAC.
National Cancer Institute (2011). Best Practices for Biospecimen Resources). [Online] Available at: http://biospecimens.cancer.gov/bestpractices/2011-NCIBestPractices.pdf
National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research (1979). *The Belmont Report: Ethical Principles and Guidelines for the Protection of*

Human Subjects of Research. Washington, DC: Government Printing Office.
Nuffield Council of Bioethics (2002). *The Ethics of Research Related to Healthcare in Developing Countries*. [Online] Available at: http://www.nuffieldbioethics.org/research-developing-countries
Nussbaum, M. (2006). *Frontiers of Justice: Disability, Nationality, Species Membership*. Cambridge, MA: The Belknap Press of Harvard University Press.
O'Doherty, K., et al. (2011). From consent to institutions: designing adaptive governance for genomic biobanks. *Social Science and Medicine*, **73**, 367–74.
Parfit, D. (1998). Equality and priority. In: Mason, A. (ed.) *Ideals of Equality*. Oxford: Blackwell Publishers, pp. 21–36.
Pellegrino, E.D. (1981). Health promotion as public policy: the need for moral groundings. *Preventive Medicine*, **10**, 371–8.
Pellegrino, E.D. (1995). Toward a virtue-based normative ethics for health professions. *Kennedy Institute of Ethics Journal*, **5**, 253–77.
Pellegrino, E. and Thomasma, D.C. (1988). *For the Patient's Good: The Restoration of Beneficence in Health Care*. New York: Oxford University Press.
Percival, T. (1985). *Medical Ethics; or a Code of Institutes and Precepts, adapted to the Professional Conduct of Physicians and Surgeons*. With an Introduction by Edmund D. Pellegrino. Birmingham, AL: Classics of Medicine Library.
Plough, A. (2005). Promoting social justice through public health policies, programs, and services. In: Levy, B.S. and Sidel, V.W. (eds.) *Social Justice and Public Health*. Oxford: Oxford University Press, pp. 418–31.
PlumX Metrics. (2016). Editorial. The right(s) approach to Zika. *The Lancet Global Health*, **4**, e427.
Pollak, J. (2002). The lead-based paint abatement repair and maintenance study in Baltimore: historic framework and study design. *Journal of Health Care Law and Policy*, **6**, 89–108.
Porter, R. (1986). History says no to the policeman's response to AIDS. *British Medical Journal*, **293**, 1589–90.
Powers, M. and Faden, R. (2006). *Social Justice: The Moral Foundations of Public Health and Health Policy*. New York: Oxford University Press.
Powers, M. and Faden, R. (2011). Health capabilities, outcomes, and the political ends of justice. *Journal of Human Development and Capabilities*, **12**, 565–70.
Presidential Commission on the Human Immunodeficiency Virus Epidemic (1988). *Final Report*. Washington, DC: Government Printing Office.
Quinn, S.C. (2004). Ethics in public health research: protecting human subjects: the role of Community Advisory Boards. *American Journal of Public Health*, **94**, 918–22.
Ravindran, G.D. (2008). Medical ethics education in India. *Indian Journal of Medical Ethics*, **5**, 18–19.
Ravitsky, V. and Wilfond, B. (2006). Disclosing individual genetic results to participants. *American Journal of Bioethics*, **6**, 8–17.
Rawls, J. (1971). *A Theory of Justice*. Cambridge, MA: The Belknap Press of Harvard University Press.
Roberts, M.J. and Reich, M.R. (2002). Ethical analysis in public health. *The Lancet*, **359**, 1055–9.
Rosoff, P. and DeCamp, M. (2011). Preparing for pandemic influenza: are some people more equal than others? *Journal of Health Care for the Poor and Underserved*, 3S, 19–35.
Roter, D.L., et al. (1998). Effectiveness of interventions to improve patient compliance: a meta-analysis. *Medical Care*, **36**, 1138–61.
Royo-Bordonada, M.A. and Lopez, F.J.G. (2016). Ethical considerations surrounding the response to Ebola: the Spanish experience. *BMC Medical Ethics*, **17**, 49.
Ruger, J. (2009). *Health and Social Justice.* Oxford: Clarendon Press of Oxford University Press.

Salerno, J., et al. (2016). Emergency response in a global health crisis: epidemiology, ethics, and Ebola application. *Annals of Epidemiology*, 26(4), 234–7.
Satalkar, P., Elger, B.E., and Shaw, D.M. (2015). Prioritising healthcare workers for Ebola treatment: treating those at greatest risk to confer greatest benefit. *Developing World Bioethics*, 15(2), 59–67.
Segall, S. (2010). *Health, Luck, and Justice*. Princeton, NJ: Princeton University Press.
Sen, A. (2002). Why health equity? *Health Economics*, **11**, 659–66.
Sen, A. (2009). *The Idea of Justice.* Cambridge, MA: The Belknap Press of Harvard University Press.
Sharp, R. and Foster, M. (2007). Grappling with groups: protecting collective interests in biomedical research. *Journal of Medicine and Philosophy*, **32**, 321–37.
Society for Public Health Education (n.d.). *Code of Ethics for the Health Education Profession*. Section 4. [Online] Available at: http://www.sophe.org/Ethics.cfm
Spriggs, M. (2004). Canaries in the mines: children, risk, non-therapeutic research, and justice. *Journal of Medical Ethics*, **30**, 176–81.
Steinbock, B. (1996). Liberty, responsibility, and the common good. *The Hastings Center Report*, **26**, 45–7.
Strauss, R.P., et al. (2001). The role of community advisory boards: involving communities in the informed consent process. *American Journal of Public Health*, **91**, 1938–43.
Taylor, H.A. and Johnson, S. (2007). Ethics of population-based research. *Journal of Law, Medicine and Ethics*, **35**, 295–9.
Thompson, A., et al. (2006). Pandemic influenza preparedness: an ethical framework to guide decision-making. *BMC Medical Ethics*, **7**, 12.
Thompson, A.K. (2016). Bioethics meets Ebola: exploring the moral landscape. *British Medical Bulletin*, 117(1), 5–13.
Trinidad S., et al. (2010). Genomic research and wide data-sharing: views of prospective participants. *Genetics and Medicine*, **12**, 486–95.
University of Toronto Joint Centre on Bioethics (2005). *Stand on Guard for Thee: Ethical Consideration in Preparedness Planning for Pandemic Influenza.* Toronto: University of Toronto Joint Centre on Bioethics.
Uscher-Pines, L. et al. (2007). Social justice and disadvantaged groups. *The Hastings Center Report*, **37**, 32–9.
US Department of Health and Human Services (HHS) (2005). *HHS Pandemic Influenza Plan. Appendix D*. Washington, DC: HHS. [Online] Available at: http://www.flu.gov/planning-preparedness/federal/hhspandemicinfluenzaplan.pdf
US Department of Health and Human Services (HHS) (2009). *Protection of Human Subjects, 45 CFR 46*. Washington, DC: HHS.
Varmus, H., and Satcher, D. (1997). Ethical complexities of conducting research in developing countries. *New England Journal of Medicine*, **337**, 1003–5.
Vawter, D.E., et al. (2011). Attending to social vulnerability when rationing pandemic resources. *Journal of Clinical Ethics*, **22**, 42–53.
Venkat, A., et al. (2015). Ethical issues in the response to Ebola virus disease in United States emergency departments: a position paper of the American College of Emergency Physicians, the Emergency Nurses Association, and the Society for Academic Emergency Medicine. *Academic Emergency Medicine*, 22(5), 605–15.
Venkatapuram, S. (2011). *Health Justice.* Cambridge: Polity Press.
Voyena, E., Mastroianni, A., and Kahn, J. (2012). Ethical issues in health research with novel online sources. *American Journal of Public Health*, **102**, 2225–30.
Walter, J. and Klein, E. (2003). *The Story of Bioethics: from Seminal Works to Contemporary Explanations.* Washington, DC: Georgetown University Press.

Washington University v. Catalona (2007). 490 F.3d 667, 675 (8th Cir. 2007).
Waters, W.E. (1985). Ethics and epidemiological research. *International Journal of Epidemiology*, **14**, 48–51.
World Health Organization Statement from Advisory Panel. (2014). Ethical considerations for use of unregistered interventions for Ebola virus disease (EVD). [Online] Available at: http://www.who.int/mediacentre/news/statements/2014/ebola-ethical-review-summary/en/
World Health Organization (2016). WHO Director-General briefs Executive Board on Zika situation. [Online] Available at: http://www.who.int/dg/speeches/2016/zika-situation/en/
Wikler, D.I. (1978a). Coercive measures in health promotion: can they be justified? *Health Education Monographs*, **6**, 223–41.
Wikler, D.I. (1978b). Persuasion and coercion for health. *Milbank Memorial Fund Quarterly/Health and Society*, **56**, 303–33.
Wilkinson, R.G. and Marmot, M. (2003). *Social Determinants of Health: The Solid Facts* (2nd ed.). Geneva: WHO.

[1] Note that the 11th edition of the same textbook (1980) had a chapter entitled ‘Legal and Ethical Issues in Public Health’ by Sidney Shindell. The bulk of this chapter is devoted to legal issues in public health, however, rather than ethics issues, and thus is not discussed here.

[2] Note that this article describes the work of Edward Chadwick in England in the early 1800s demonstrating that differences in social conditions led to a more than twofold difference in life expectancy between upper and lower classes. Also in the 1800s, governments began conducting investigations of housing conditions and garbage heaps and mapping them in relation to outbreaks of disease, and by the end of the nineteenth century, state and local boards of health were being created to enforce sanitary regulations.

3.3

The right to health supports global public health

Carmel Williams, Alison Blaiklock, and Paul Hunt

Illustrations of human rights promoting and protecting health

We commence this chapter with a selection of very short illustrations of various ways in which human rights have been deployed to improve public health. We want to encourage readers to think of human rights as practical, useful, and effective in public health. Following these introductory illustrations we then explain human rights, and in particular the right to the enjoyment of the highest attainable standard of physical and mental health, and how they support public health.

Brazil Family Health Programme

Brazil enshrined the right to health in its Constitution in 1988 and subsequently framed their family health programme with the values and features associated with health rights. The legislative and policy reforms led to reduced health inequities, and improved health outcomes, shown in the reduction of maternal mortality (53.1% by year eight of the programme) and infant and child mortality (34% decline by year eight). Although the human rights-based approach was not comprehensive, has subsequently faced setbacks and remains a work in progress, the roll out of the new programme advanced the health of poorer women and children first (Bhalotra et al. 2016). This is consistent with a human rights-based approach to health (Bustreo et al. 2013).

Treatment Action Campaign in South Africa

The decision of the Constitutional Court of South Africa in *Minister for Health v. Treatment Action Campaign* held that the Constitution required the government 'to devise and implement a comprehensive and coordinated programme to realize progressively the right of pregnant women and their newborn children to have access to health services to combat mother-to-child transmission of HIV' (CCSA 2002, para. 135 (2) (a)). This decision led to the establishment of one of the biggest programmes in Africa to reduce mother-to-child transmission of HIV (Burton et al. 2015).

Empowering people to claim their health entitlements

My Health, My Voice is a human rights-based project in Uttar Pradesh, India, which used technology to monitor and display online data regarding illegal informal payments for maternal healthcare. A toll-free hotline was established through which women could report health providers' demands for these payments. Over a 15-month period, the hotline recorded 873 reports of informal payment demands. The project enhanced women's knowledge of their entitlements, as well as their confidence to claim their rights. Anecdotal evidence suggests that health providers' demands for informal payments were reduced in response to the project (Dasgupta et al. 2015).

The State Hospital, Scotland

In 2000, The State Hospital in Carstairs, Lanarkshire, Scotland, after a critical report by the Mental Welfare Commission in 2000, decided to use the Human Rights Act as a vehicle for change (Alliance 2013). Accordingly, the human rights of everyone—staff, patients, carers, and family members—were put at the heart of the hospital's services. A working group, led by senior management and involving clinical and non-clinical members of staff, was established. Consultative methodologies were devised for operationalising a human rights-based approach at the hospital. In 2009, an evaluation of the initiative found that the human rights-based approach had supported 'a cultural change … towards an organization with a more positive and constructive atmosphere with mutual respect between staff and patients' (Scottish Human Rights Commission 2009). This led to 'increased staff and patient engagement, increased work-

related satisfaction among staff and increased satisfaction among patients over their care and treatment' (p. 6). The new approach 'coincided with staff reporting a reduction in stress and anxiety' and patients 'noted significant and sustained improvements in their care and treatment and in the overall culture' at the hospital (Scottish Human Rights Commission 2009).

Ending child marriage

Current estimates are that although child marriage is declining, 12 million girls are married in childhood each year (Cappa 2018). Eliminating child marriage is a sustainable development goal (SDG) target and important to achieving other SDGs. Child marriage is a major violation of children's rights to health and education, to protection from abuse and exploitation, as well as other human rights. The numerous health impacts for children and their babies include complications of pregnancy and childbirth; physical, sexual, and psychological violence; and HIV/AIDS. Social movements to end child marriage and support married girls are growing. More than 900 civil society organizations from 95 countries, most of whom work in communities and many identify as youth-led, have formed the Girls Not Brides global partnership. The partnership's theory of change recognizes a range of community and collective approaches at all levels are necessary. These include empowering girls to know their rights; community leaders taking action to realize girls' rights; and national laws reflecting international and regional human rights standards (Girls Not Brides 2015, 2018).

Shaping Italy's health system

Italy's Constitution of 1948 protects the right to health: 'The Republic safeguards health as a fundamental right of the individual and as a collective interest'. The courts have decided a handful of important right-to-health cases, for example, on improving access to contraception. Crucially, this human right has explicitly shaped Italy's national health service, national health plans, and community health schemes, such as the initiative that delivers women's health information and services. The right has also explicitly shaped health interventions, including cancer screening

programmes. The guidelines for cancer screening remind health professionals that ‘access to screening is an application of a right’. A recent World Health Organization study on Italy concludes that applying human rights to women’s and children’s health policies, programmes, and other interventions has not only helped the Italian government comply with its binding national and international obligations but has also contributed to improving the health of women and children (Bustreo et al. 2013).

Introduction to human rights

These aforementioned examples illustrate different ways in which human rights have been used to improve public health: they guided policy reform so that people’s right to health entitlements were fulfilled; they used judicial action to redress power imbalances and rights violations that cost lives; they protected people who were powerless and marginalized by enabling them to challenge these power imbalances; and they underpinned social movements. In sum, human rights can be used to address human suffering, and to protect people from future harm arising from rights failings and abuses.

Our overarching aim in this chapter is to demonstrate to those working in global public health how the right to health, and other human rights, can make a meaningful contribution in their field of expertise. We build on the work of many others, including the authors of the corresponding chapter in preceding editions to this volume, hoping to extend the understanding that global public health, the right to health, and human rights are natural allies with common concerns and causes.

A recent survey of the literature on health and human rights, including the right to health, finds that ‘since the turn of the century, there has been a steady stream of academic books, articles, reports, and other publications on human rights and health, including the right to health’ (Hunt 2016). But it also demonstrates that the lens of the right to health has been directed more at healthcare than either public health or global public health. We endeavour to make an original interdisciplinary contribution towards redressing this imbalance.

In the following subsections, we introduce human rights and explain their modern history and status in law. We look at different ways in which human rights are used, and identify who has entitlements to rights, who carries

human rights obligations, and what a human rights-based approach to health means. We then, in ‘What is the right to health’, examine the meaning of the right to health, adapt and refine the right to health analytical framework, as set out in several publications (Hunt 2015) and in ‘Key features of the right to health’, identify key features of the right to health. This leads to ‘The right to health in practice’ where we look at how the right to health can be used in global public health with three case studies: climate change and children; Big Data and health; and international assistance and cooperation. In our conclusion, we return to the short illustrations at the start of the chapter, examined in the light of a human rights-based approach to public health.

What are human rights?

Human rights are grounded in respect for the dignity, well-being, and equality of every human being. They are a commitment to our shared humanity, are inspired by moral values, and the recognition that all of us are inherently entitled to rights simply because we are human. But human rights go further than moral claims; they are underpinned by international law, binding agreements that all of us—all people everywhere—are legally entitled to be treated with respect, dignity, and equality. This is a key human rights principle; others include participation and accountability.

Human rights guarantee the right of everyone to the enjoyment of the highest attainable standard of physical and mental health (hereafter abbreviated to the ‘right to health’ or ‘health rights’). They also require governments and others to take collective and organized actions that promote health and equity for everyone. Inequity is avoidable because it is the result of injustice and discrimination. Thus, human rights are key to public health achieving its aims of reducing mortality and morbidity and promoting health through the collective and organized actions of communities and societies (Winslow 1920; Acheson 1988). Similarly, human rights are essential for global public health achieving its aims of health and health equity for everyone worldwide through collective and organized actions that address transnational health issues and determinants.

In 1948 the Universal Declaration of Human Rights (UDHR) was adopted by the United Nations in response to the authoritarianism and its causes, and associated atrocities, of the Second World War. It begins: ‘All

human beings are born free and equal in dignity and rights' (UN 1948). The UDHR and subsequent treaties (often called covenants or conventions), such as the International Covenant on Economic, Social and Cultural Rights (ICESCR) (UN 1966b), and the Convention on the Rights of the Child (CRC) (UN 1989), turned these values into international laws which confer entitlements on people and impose obligations on governments, transnational corporations, organizations, and people whose roles and responsibilities affect others.

Although the underpinning of human rights by international law renders them powerful, their use is not limited to the courts, as we have shown in the given illustrations. Throughout the world, human rights play a crucial role in empowering people to claim their health and other entitlements, to challenge power imbalances, and to seek social justice. Accordingly, human rights can underpin transformative public health actions that address the health consequences of inequality and discrimination. Moreover, there is emerging evidence that human rights contribute to health gains for individuals and populations (Hunt et al. 2015).

Public health and human rights are natural allies

The disciplines of public health and human rights can complement, extend, and at times challenge each other. Just as public health programmes are essential to the realization of the right to health, so too can human rights help to reinforce existing and useful health programmes and policies and identify and shape new, equitable ones.

Public health workers and human rights advocates are motivated by similar concerns for the well-being of individuals and populations. They also want effective, integrated, responsive health systems that are accessible to all. Other similarities include the recognition of:

the importance of empowerment, laws, policies, evidence and analysis, advocacy, and collaboration to bring about change;
not only access to healthcare, but also to the social, environmental, and other determinants of health including water, sanitation, nutrition, health information, and education, and much more;
the need for collective efforts to improve well-being; health is not the sole responsibility of the health ministries and health systems, but belongs to a wide range of public and private actors and requires governments, the private sector, civil society, and communities to work together;
the underlying and structural causes of problems in society;
the important struggle against discrimination and disadvantage;
cultural respect;
social justice and challenging existing unfair distributions of money, power, and resources.

Human rights-based approaches to health shift public health well beyond biomedical paradigms and expand the focus on population morbidity and mortality to the dignity and well-being of everyone. Human rights challenges utilitarian, paternalistic, and neoliberal approaches to public health. Instead of prioritizing the ‘greatest good for the greatest number’, priority is given to those who are most disadvantaged and discriminated against (Rumbold et al. 2016). As a human right, health is seen as an entitlement, not something able to be bought, traded, or bestowed by privilege or altruism.

When viewed through a human rights lens, health issues are understood in terms of rights rather than problems requiring technocratic solutions. Public health workers become defenders of human rights who promote and protect human rights, and who enable people to claim and enjoy their rights. Their practice can include human rights frameworks and tools as we have illustrated, as well as public health methodologies, and their professional codes of ethics are underpinned by universally agreed moral values and international, regional, and national human rights laws. Public health services may have various forms of political and financial accountability; they also have binding obligations under international and national law which give rise to legal accountability for human rights.

Public health workers—and indeed everyone working to improve health—can use the right to health to help devise more equitable policies and

programmes; place important health issues higher up national and international agenda; secure better coordination across health-related sectors, as well as between services within the health sector; raise more funds from the treasury; build partnerships with civil society; analyse the activities of partner agencies, businesses, and international financial institutions; leverage more funds from high-income countries to low- and middle-income countries; use internationally accepted norms and standards (e.g. definitions of population groups); improve the terms and conditions of those working in the health and other sectors; and so on. The right to health is not just a rhetorical device, but also a tool that can save lives and reduce suffering, especially among those who are most disadvantaged. These goals all remain consistent with the most recent global development agenda, the SDGs, that aims to leave no one behind.

International human rights norms and laws provide a way to challenge growing threats to public health from global power and financial asymmetries associated with multinational corporations, trade and investment treaties, and environmental degradation. The UN human rights expert on a democratic and equitable international order, has called for assessments of the impact of trade and investment treaties on human rights, health, and the environment (de Zayas 2015).

The legality of human rights

Formal expressions and legal guarantees of human rights come from treaties, customary international law, national laws, and other sources of law that impose legal obligations on governments and others for the protection, respect, and fulfilment of human rights. These recognize a wide range of civil, political, economic, social, and cultural rights, including for example, the rights to life, participation in decision-making, to an adequate standard of living, education, a healthy environment (UN Human Rights Special Procedures 2018), and many others, some of which are referred to in this chapter. Human rights laws also recognize specific rights, including protection from discrimination, for some groups of people including women, children, those with disabilities, those experiencing racism, migrant workers, refugees and asylum seekers, and indigenous peoples.

The international community has agreed that all human rights are universal, inalienable, indivisible, interdependent, and interrelated (UN 1993). Being *universal* means that everyone is entitled to equally enjoy

their human rights without discrimination on any ground such as ‘race, colour, sex, language, religion, political or other opinion, national, ethnic or social origin, property, disability, birth or other status’ (CRC 1989). Actions to tackle inequality and discrimination are at the heart of human rights just as they are at the heart of public health. Recognition that all people should be treated with dignity underpins the compassion that is integral to the work of healing and the solidarity that is integral to the work of social justice.

Being *inalienable* means that no person can give or negotiate away their human rights or have their human rights taken away other than in very specific situations. There may be occasions where it is necessary to limit someone’s human rights for the sake of other people—such as when quarantine restrictions are needed to control an epidemic; we discuss next when this can be done legitimately.

Being *indivisible*, *interrelated*, and *interdependent*, means that all human rights must be respected, not just a selection. Respect for all human rights ensures respect for the dignity and worth of every person. In practical terms, the violation of one right can affect others—for example, when children are unable to enjoy their rights to food, water, and shelter, their rights to health, education, and protection may be seriously harmed.

Notably, very few human rights are absolute: sometimes *balances* have to be struck between rights, such as the right to privacy and freedom of information. Also, most human rights may be subjected to limitations, such as those ‘determined by law … and solely for the purpose of promoting the general welfare in a democratic society’ (United Nations, Article 5, 1966b). For example, an individual may claim she has the right to receive non-fluoridated drinking water in her home, but the community in which she lives may wish to receive fluoridated water on public health grounds. In this case a balance has to be struck between the individual’s human right (to private and family life) and a possible limitation (the general welfare). The Siracusa Principles are designed to ensure fair balances are struck by way of fair processes (see ‘Siracusa Principles’, to follow).

International human rights treaties, such as ICESCR and CRC, are legally binding on the states that ratify them. Ratification means states agree to be bound by international law to the treaty and they accept concrete legal duties to comply with the treaty’s requirements and prohibitions. States are then accountable for what they do and do not do to implement the treaty’s principles and provisions. International human rights declarations,

such as the UDHR, the Declaration on the Rights of Indigenous Peoples, and the Declaration on the Right to Development are non-binding, although many of them include norms and principles that reflect obligations that are binding under international customary and treaty law. Additionally, all states have national laws that protect some human rights.

Human rights-based approaches to health

A human rights-based approach to health 'requires adoption of an approach explicitly shaped by human rights principles' (Gruskin et al. 2010). It also means that health 'policies and programs will contribute to the fulfilment of human rights' (Gruskin et al. 2010) and indeed do not undermine people's enjoyment of their rights. The underpinning of human rights in international and national law brings real strength to human rights-based approaches to public health. Additionally, human rights principles, when operationalized in policy and programmes, empower people, and provide public health with the opportunity to promote equity, and improve health outcomes. We examine three different approaches to human rights in practice (policy, judicial, and empowerment); but we stress that in reality these all overlap and none is effective without the others.

Policy approach

Human rights principles should influence policymaking processes so that health policies and programmes promote and protect human rights including rights to equality and non-discrimination, participation, and to a well-functioning health system. This we call the 'policy' approach to implementing the right to health. A right-to-health perspective should be part of all local, national, and international policymaking processes that are relevant to people's health.

As with most public health endeavours, success in bringing about changes in policy requires close cooperation across a range of disciplines, and with the communities affected. Because of its role and expertise, public health has an indispensable contribution to make to the 'policy' approach.

Judicial approach

When human rights are not respected, protected, or fulfilled, another approach can involve the courts, tribunals, and other judicial and quasi-judicial processes (the 'judicial' approach). This approach has a vital role to play and throughout the world many courts have demonstrated that they have a crucial contribution to make. Court rulings have enabled people to gain access to medicines or other health rights entitlements in thousands of cases (UNHRC 2007; Gauri and Brinks 2008; Yamin and Gloppen 2011; Ferraz 2018).

Empowerment approach

The empowerment of people to know and claim their human rights entitlements is a third approach to the right to health, one which is also essential to both the policy and judicial approaches. When people know they have a right not to be discriminated against, to equal access to health services, to be treated with dignity, and that provision of essential services is a right, not an arbitrary act of charity, they can act, individually or as part of groups and movements. Such action can change social norms, practices, and the distribution of power, within and between communities and countries.

Human rights entitlements and obligations

Human rights law confers entitlements on all people and imposes obligations on governments (states), and others, who are called 'duty bearers' in human rights discourse. States that have ratified international and regional human rights treaties are accountable for compliance with their principles and provisions and mandated to report periodically to UN monitoring bodies. In 2013, the Optional Protocol to the ICESCR came into force. This made it possible for individuals or groups in states which had ratified ICESCR and the Optional Protocol, to submit a complaint to the Committee of the ICESCR in regard to state violations of their economic, social, and cultural rights, including the right to health.

Since 2008, a Human Rights Council accountability mechanism called the Universal Periodic Review has required all UN Member States to report on the actions they have taken to improve human rights in their countries, regardless of which treaties they have ratified.

The Special Rapporteur on the Right of Everyone to the Enjoyment of the Highest Attainable Standard of Physical and Mental Health is another accountability mechanism mandated by the Human Rights Council. The Special Rapporteur presents two types of reports to the Human Rights Council and the UN General Assembly (UNGA). The first are general thematic reports (for example, corruption and the right to health (UNGA 2017) or the SDGs and the right to health (UNGA 2016)); secondly, the Special Rapporteur undertakes visits to countries, following which a country report is written on the right to health, or specific elements of the right, with a view to holding the country to account for its right to health obligations. When rapporteurs encounter specific issues on which the existing jurisprudence gives little guidance, they offer their interpretations of the international right to health (Hunt 2016).

Civil society is encouraged to engage in all of these reporting processes, and provide shadow reports, as a means of keeping the processes transparent and meaningful.

Human rights obligations, including duties relating to the right to health, extend beyond states to members of society (UN 1948, preamble; UNCESCR 2000, para. 42). This includes health workers, families, communities, inter- and non-governmental organizations, civil society groups, and businesses including multinational and transnational companies, for example, pharmaceutical companies. One of the human rights responsibilities of states is to provide an environment in which the actions of all these individuals and groups is in accord with their human rights duties, and where people are able to claim their human rights entitlements.

When states enter into bilateral, multinational, or global agreements, including trade agreements, they are duty bound to ensure such agreements respect and protect human rights. The SDGs, adopted by all UN Member States in 2015, explicitly refer to human rights in their preamble, stating, ‘The 17 Sustainable Development Goals … seek to realize the human rights of all and to achieve gender equality and the empowerment of all women and girls’ (UNGA 2015). Furthermore, this global initiative uses as its catchphrase the mission to ‘leave no one behind’. This echoes the human rights imperative of equality and non-discrimination, and the public health commitment to reduce inequities.

Human rights are not without controversy

The right to health—like other economic, social, and cultural rights—does not escape controversy and ideological objections. Some states are reluctant to see it as a right of similar weight to others, such as the right to a fair trial. However, under international law, the right to health is an integral part of the international code of human rights, a right that is interdependent and indivisible from other rights and must be given equal treatment and attention (UN 1993, para. 5).

Despite the promise of human rights, they cannot by themselves resolve many of the complex problems pertaining to their realization. For example, despite human rights accountability mechanisms, there is little that most states can do to force other states to meet their human rights duties. Furthermore, in our globally interconnected world, where transnational companies control and own much of our health and general information, nation states remaining the primary human rights duty bearers has severe limitations. Nor do human rights present easy solutions to decisions to guide the prioritization of scarce health resources. However, the human rights principles of equality, non-discrimination, participation, and accountability are useful to guide responses to these challenges.

What is the right to health?

The right to health does not mean a right to be healthy (UNCECSR 2000, para 8). It is however, as recognized in international law, the right to facilities, goods, services, and conditions that are conducive to the realization of the highest attainable standard of physical and mental health. In some quarters, there has been a presumption that the right to health relates to personal health services. Such a narrow definition is in fact inconsistent with international human rights law, which encompasses clinical care, public health, and the determinants of health, such as access to water, sanitation, nutrition, housing, and education, as confirmed by Article 12 of ICESCR and Article 24 of CRC.

Under the right to health people have freedoms, such as the freedom from discrimination or non-consensual medical treatment and experimentation. They also have entitlements including to maternal, child, and reproductive health services; healthy workplaces and environments; the prevention,

treatment, and control of diseases, including access to essential medicines; and access to safe and potable water and sanitation.

The right to health is closely related to and dependent upon the realization of other fundamental human rights contained within international law. The rights to education, an adequate standard of living, life, privacy, and access to information, as well as other rights and freedoms, relate to and address integral components of the right to health.

International human rights law

The origins of the right to health in international law can be traced back to the 1946 World Health Organization's (WHO) Constitution which says: 'The enjoyment of the highest attainable standard of health is one of the fundamental rights of every human being without distinction of race, religion, political belief, economic or social condition' (WHO 1946). Two years later, article 25(1) of the Universal Declaration of Human Rights laid the foundations for the international legal framework for the right to health. Since then, the right to health has been codified in legally binding international and regional human rights treaties, and enshrined in many national laws, some of which are signalled in the following paragraphs.

The cornerstone protection of the right to health in international law is found in Article 12 of ICESCR. By 2018, 164 states had ratified this treaty, legally binding themselves to its implementation (OHCHR 2018).

Additional right to health protections are contained in international treaties that address issues specific to marginalized groups, or groups whose health is at risk from the effects of marginalization.[1] All countries have ratified at least one treaty with right to health provisions and with the exception of the United States, all countries have ratified the CRC (OHCHR 2018).

Authoritative and interpretive guiding principles of several of these treaty provisions on the right to health shed further light on the parameters and content of the right generally, and in relation to the application of the right to specific groups. Of special relevance are General Comment 14 on the right to the highest attainable standard of health adopted in 2000 by the UN treaty-body responsible for monitoring ICESCR (UNCESCR 2000); General Comment 15 on the right of the child to the enjoyment of the highest attainable standard of health adopted in 2013 by the UN treaty-body

responsible for monitoring the CRC (UN CRC 2013); international guidelines on HIV/AIDS and human rights (OHCHR and UNAIDS 2006); the Technical Guidance on the Application of a Human Rights-Based Approach to the Implementation of Policies and Programmes to Reduce Preventable Maternal Morbidity and Mortality (UN OHCHR 2012b); the Technical Guidance on the Application of a Human Rights-Based Approach to the Implementation of Policies and Programmes to Reduce and Eliminate Preventable Mortality and Morbidity of Children under 5 years of age, (UN HRC 2014); and *Ensuring Human Rights in the Provision of Contraceptive Information and Services: Guidance and Recommendations* (WHO 2014).

Far-reaching commitments relating to the right to health have been made in the outcome documents of numerous UN and WHO world conferences, such as the Alma-Ata Declaration (WHO 1978), the International Conference on Population and Development (UN 1994), the Fourth World Conference on Women (UN 1995), the Bangkok Charter for Health Promotion in a Globalized World (WHO 2005), the Rio Political Declaration on Social Determinants of Health (WHO 2011), the SDGs (UNGA 2015), and the Action Agenda from the Second Global Conference on Climate and Health (WHO 2016). These conferences and global commitments have helped to place international problems—including issues such as maternal mortality, mental health, non-communicable diseases (NCDs), health inequities, and poverty—on global agendas and their outcome documents influence international and national policymaking. Several refer to the right to health and/or health-related rights.

The UN Declaration on the Rights of Indigenous Peoples (DRIP) has particular importance. For indigenous peoples, colonization has often led to dispossession from their lands and territories; cultural subjugation; social, economic, and political marginalization; and consequential serious harm to their health. DRIP recognizes that for indigenous peoples, well-being is shared, holistic, and includes the relationship between people and environment, and is aligned with indigenous models for public health (see Chapter 10.6) (United Nations General Assembly 2007).

Siracusa Principles

Tragically, there have been situations such as the United States Public Health Service ‘Tuskegee Study of Untreated Syphilis in the Negro Male’

that began in 1932 and continued for 40 years (Centers for Disease Control and Prevention 2013) in which public health objectives have been used to justify what have been human rights violations, such as the withholding of information and treatment to develop an understanding of a disease. There are some occasions in which the public health measures necessary to respond to a serious threat to the health of a population or individual do require the limitation of other human rights. International human rights law sets out narrowly defined standards, called the Siracusa Principles (UN Commission on Human Rights 1984), as to when such limitations may be implemented legitimately. Such measures should only be taken when they are: necessary for the promotion of the general welfare; provided for in law and carried out according to the law; there is no other less intrusive and restrictive alternative; reasonable, based on evidence, and not discriminatory in any other way; subject to review; and of limited duration (UN ECOSOC 1985, Annex; UNCESCR 2000, paras. 28–9).

Regional human rights law

The right to health is recognized in human rights treaties adopted and monitored by regional human rights systems. These have effect only in their respective regions and include: the African Charter on Human and Peoples' Rights (Article 16); the African Charter on the Rights and Welfare of the Child (Article 14); the Additional Protocol to the American Convention on Human Rights in the Area of Economic, Social and Cultural Rights, known as the 'Protocol of San Salvador' (Article 10); and the European Social Charter (Article 11). Other regional instruments, such as the American Convention on Human Rights (OAS 1969) and the European Convention for the Protection of Human Rights and Fundamental Freedoms (Council of Europe 1950), also provide, through health-related rights and other provisions, protection of the right to health.

At the regional level there are also judicial and other mechanisms that adjudicate cases involving the right to health. A notable case in 2002 was the finding by the African Commission on Human and Peoples' Rights (ACHPR) of a violation of the right to health by the Federal Republic of Nigeria, on account of violations against the Ogoni people in relation to the activities of oil companies in the Niger Delta (ACHPR 2001). In 2009, the same tribunal held that Sudan had violated the right to health as a result of

exposing members of indigenous tribes in the Darfur region to serious health risks by destroying homes, livestock, farms, and poisoning wells. In so doing, the Commission recognized that the right to health extends to underlying determinants of health, including access to safe water, adequate and safe food, and housing (ACHPR, *Sudan Human Rights Organisation, et al. v. Sudan* Comm. No. 279/03-296/05 (2009).

Another important development has come from the Inter-American Commission on Human Rights (IACHR) which expressed its willingness to 'take into account' provisions of the regional treaty (the Protocol of San Salvador) related to the right to health when analysing the merits of a case, even though it lacked competence to determine violations under these provisions (IACHR 2000). In the case of *Gonzales Lluy et al. v. Ecuador*, **2015**, the American Court of Human Rights (IACtHR) recognized that the state has an obligation to supervise and monitor not only public hospitals but also private entities in the healthcare sector. The Ecuadorian government was found negligent in its supervisory duties for failing to properly test blood banks leading to a child becoming infected with HIV after a blood transfusion. It was found to have violated both the right to life and the right to humane treatment. Finally, in a recent case, *Poblete Vilches et al. vs. Chile*, the IACtHR ruled that 'the right to health is an autonomous right protected by Article 26 of the American Convention' and that the state was responsible for violations in relation to rights within article 26 (IACtHR 2018).

National laws

A study has shown that 55% of the constitutions of all nations have provisions regarding health rights (Heymann et al. 2013). In addition, 27 states have provisions regarding sexual and reproductive health and rights, and nine of these explicitly link the right to sexual and reproductive health to the rights to education, science, and/or to make autonomous decisions about sexuality and reproduction (Pizzarossa and Perehudoff 2017).

In some jurisdictions, these constitutional provisions have generated significant jurisprudence. The Constitutional Court of South Africa in *Minister for Health v. Treatment Action Campaign*, illustrated at the start of this chapter, is such an example. In many South American countries, with the right to health guaranteed in their constitutions, there has been 'an

‘explosion’ of litigation … (thousands of cases in Costa Rica, and hundreds of thousands in Colombia and Brazil) and a heightened disposition of judges to enforce the right to health through strong remedies in comparison to non-Latin American countries’ (Ferraz 2018).

These and numerous other laws and decisions at the international, regional, and national levels confirm that the courts have an important role to play in the protection, and fulfilment of the right to the highest attainable standard of health.

Key features of the right to health

In recent years, the Committee on Economic, Social and Cultural Rights, WHO, Special Rapporteurs on the right of everyone to the enjoyment of the highest attainable standard of physical and mental health, human rights experts, civil society organizations, academics, and many others, have developed ways of ‘unpacking’ or analysing the right to health with a view to making it easier to understand and apply to health-related policies, programmes, and projects in practice.

Accordingly, the following paragraphs set out 10 key features of the right to health. Some of the features overlap, for example, accessibility (in key feature nine) overlaps with non-discrimination (in key feature two). Also, some of the features have technical names, such as ‘obligations of immediate effect’, which reflect their legal provenance. Nonetheless, the features also reflect an emerging discourse for understanding some of the key characteristics of the right to health. ‘The right to health in practice’ applies these features, by way of illustration, in a more practical, operational way to the three case studies that follow on.

Laws, norms, and standards

As explained in greater detail, international, regional, and national laws, norms, and standards relevant to the particular issue, programme, or policy must be identified. These will include both general standards relating to the right to health, in addition to instruments that relate to specific groups and contexts (UNCHR 2003a, Annex 1).

Dignity, equality, and non-discrimination

Non-discrimination and equality are fundamental to all human rights, and to achieving the core human right entitlement of being treated with dignity. These entitlements are conferred from the first two articles of the Universal Declaration of Human Rights: ‘All persons are born free and equal in dignity and rights’ and ‘Everyone is entitled to all the rights and freedoms set forth in this Declaration, without distinction of any kind … ’ (UDHR, Arts 1, 2).

Participation

Participation is grounded in internationally recognized human rights, such as the rights to participate in the formulation and implementation of government policy (CEDAW, Article 7(b)), to take part in the conduct of public affairs (ICCPR, Article 25(a); UN 1966a), and to freedom of expression and association (ICCPR, Articles 19 and 22). All individuals and communities are entitled to active and informed participation on issues relating to their health, including policy making and accountability. Health professionals, and their associations, are also entitled to participate in decisions that will affect their employment and the work they do.

Accountability

Accountability lies at the heart of human rights where it has two main components: addressing past grievances and correcting systemic failure to prevent future harm (Tech Guidance HRC14/39, par 33.) Without accountability, human rights obligations can become empty promises. Accountability mechanisms provide individuals and groups (people with rights entitlements) with an opportunity to understand how states and others with human rights obligations (duty bearers) have discharged their obligations, and it also provides duty bearers with an opportunity to explain their actions. In this way, accountability mechanisms are transparent, and they help to identify policy or programme adjustments that are necessary to realize human rights entitlements. Accountability also tends to encourage the most effective use of limited resources, as well as a shared responsibility among all parties. Sometimes accountability is conflated with

monitoring but monitoring is only one step towards accountability. Accountability can be understood as monitoring, review (including independent review) in relation to human rights and other standards, and remedial action (Williams and Hunt 2017). Transparent, effective, and accessible accountability mechanisms are among the most crucial characteristics of human rights and they should extend to clinical care, public health, and the determinants of health.

International assistance and cooperation

In line with obligations envisaged in the UN Charter and some human rights treaties, such as the ICESCR (Article 2) and the CRC (Article 4), countries that need international support to meet their right to health obligations have a responsibility to seek cooperation and international assistance, and states that can provide such assistance have responsibilities to do so (Hunt et al. 2010; Maastricht Principles on Extraterritorial Obligations of States in the Area of Economic, Social and Cultural Rights 2011). Global inequality cannot be reduced without such support, as explained in the Alma-Ata Declaration, and this is also acknowledged in the SDGs (Agenda 2030, par 17.9). In times of emergency, people are entitled to disaster relief and humanitarian assistance, including public health services. States have a responsibility to cooperate in providing this support, as well as assistance to refugees and internally displaced persons (UNCESCR 2000, para. 40).

Respect, protect, fulfil all human rights

States have obligations to respect, protect, and fulfil all human rights, including the right to health, as explained and used by CESCR, the Committee on the Elimination of Discrimination Against Women (CEDAW), the Technical Guidance on the reduction of preventable maternal morbidity and mortality, among others (UN Human Rights Council 2012). These concepts are also enshrined in the Constitution of South Africa.

With regard to the right to health, the obligation to *respect* places a duty on states to refrain from interfering directly or indirectly with the enjoyment of the right to health. The obligation to *protect* means that states must

prevent third parties from interfering with the enjoyment of the right to health. The obligation to *fulfil* requires states to adopt the necessary measures, including legislative, administrative, and budgetary measures, to ensure the full realization of human rights, including the right to health.

This feature also draws attention to the indivisibility of human rights, and highlights the need to acknowledge the broad cross section of civil, political, social, economic, and cultural rights that underpin the wider determinants of health. Efforts to fulfil one right do not justify a failure to respect or protect other rights.

As we have explained, the right to health encompasses medical care and public health (see 'What is the right to health?' earlier) including the determinants of health. However, there is a tendency to give more attention to medical care and to neglect determinants. Because of that, in the application of this feature (see 'The right to health in practice', to follow) we give particular attention to determinants to ensure they are not neglected, marginalized, or overlooked.

Resource constraints and progressive realization

International human rights law recognizes that the realization of the right to health is subject to resource availability. Thus, what is required of a better resourced state today is of a higher standard than what is required of a poorer-resourced state. However, a state is obliged—whatever its resource constraints and level of economic development—to realize progressively the right to the highest attainable standard of health (UN 1966b, Art. 2(1)). In essence, this means that a state is required to be doing better in 2 years' time than it is doing today. In order to measure progress (or the lack of it) over time, indicators and benchmarks must be identified, and public health has a lot of experience in developing such tools. Progressive realization often involves making difficult decisions to ensure priority goes to those who are most disadvantaged. In human rights, it is not permissible for states to take steps that reverse human rights achievements. For example the present levels of service provision in healthcare may not be reduced, the current level of access to health services may not be lowered, and so on. If this situation, called retrogression, occurs, states must explain to UN monitoring committees why it has happened, as it is impermissible unless the state had no alternative actions available to it.

Obligations of immediate effect

Despite resource constraints and progressive realization, the right to health also gives rise to some obligations of immediate effect, such as the duty to avoid discrimination (UNCESCR 2000, para. 43). These are obligations without which the right would be deprived of its *raison d'être* and as such they are not subject to progressive realization, even in the presence of resource constraints (UNCESCR 1990, para. 10). Guidelines adopted by the UN monitoring committees are helping to clarify what these might be and have highlighted: agreement of a national plan of action relating to health, including sexual and reproductive health; provision of essential drugs as identified by WHO; provision of reproductive, maternal, and child healthcare; immunization against major infectious diseases; prevention, treatment and control of epidemic and endemic diseases; and training of health personnel. Clarification of minimum core obligations remains an ongoing and important debate.

Available, accessible, acceptable, and good quality

All health services, goods, and facilities should comply with each of these four requirements. For example, all essential immunizations should be *available* within the country. Additionally, they should be *accessible*. Accessibility has four dimensions: accessible without discrimination, physically accessible, economically accessible (i.e. affordable), and accessible health-related information. As well as being available and accessible, immunization campaigns should be provided in a culturally *acceptable* manner. This requires, for example, planning and coordination with community leaders. The *good quality* element will require, for example, ensuring the integrity of vaccine cold chains.

Note the similarity between these requirements and the four 'As' of public healthcare envisaged since the 1978 Alma-Ata Declaration: geographical accessibility, financial accessibility, cultural accessibility, and functional accessibility (WHO 1978b).

Strengthening health systems

At the heart of the right to the highest attainable standard of health lies an effective and integrated health system, encompassing healthcare and the underlying determinants of health, responsive to national and local priorities, and accessible to all (UN HRC 2008). The right to health cannot be fulfilled without a well-functioning health system (Backman et al. 2008). In essence, this means that the six components of a health system as identified by WHO—service delivery, health workforce, information, medical products, vaccines, and technologies, financing, and leadership and governance—are available, accessible, acceptable, and of quality to all, without discrimination. However, a health system is more than simply these six components. It also includes the people for whom it exists and a human rights perspective of a health system keeps these people at its centre. As Alicia Ely Yamin has written, 'If health systems are understood as core social institutions, it follows that the choices to be made cannot be seen just as technical choices but rather as reflecting and embedding values and norms of the larger society, including in relation to equality, dignity, and human rights' (Yamin 2016, p. 104).

The right to health in practice

The key features of the right to health underscore what the right to health contributes to public health. The preoccupation with dignity, equality, and non-discrimination requires a state to take effective measures to address the health inequities that characterize some populations. The focus on active and informed participation requires a state to adopt, so far as possible, a 'bottom–up' participatory approach in health-related sectors. The requirement of accountability can help to ensure that health policies, programmes, and practices are observant of human rights, and offer remedy when necessary.

Crucially, the key features of a right-to-health framework are not merely to be followed because they accord with sound management, ethics, social justice, or humanitarianism. States are required to conform to the key features *as a matter of binding law*. Moreover, they are to be held to account for the discharge of their right-to-health responsibilities arising from these legal obligations.

Having looked at the legality of human rights, the meaning of the right to health, and having identified its key features, we now turn our attention to what this means in practice. We apply this approach to critically important global public health issues—the impact of climate change on children's right to health, the digitization of health through Big Data, and overseas development assistance.

Case Study 3.3.1 The Impact of Climate Change on Children's Right to Health

Climate change is an enormous threat to global health (see Chapter 2.9). The greatest burden of climate change is borne by low resource countries and especially by poor children (Smith 2014). Children are especially vulnerable because of their growth and development, unique physiology and metabolism, reliance on adults and families, and the conditions in which they live, learn, work, and play. Consequently their health is at greater risk from climate-related events, infections, and pollutants, with the youngest and most disadvantaged children being most vulnerable (Ahdoot et al. 2015).

Climate change impacts on children's right to health and indeed all their human rights–their rights to life, survival, development, health, food, water, sanitation, a healthy environment, family, education, housing, safety, protection, participation, culture, heritage, and so forth (Knox 2016; United Nations General Assembly 2017). Without effective mitigation and adaptation strategies, climate change has life-long consequences for today's children and brings increasing harm to future generations of children. However, well-designed climate-related actions bring opportunities to improve the health and well-being of children and future generations (Hosking et al. 2011).

Laws, norms, and standards

Human rights laws, norms, and standards are used to reduce the harm that climate change causes on the well-being of children. Under the United Nations CRC, children are entitled to special protection measures and assistance to ensure their survival and development and that they are treated as human beings equal in dignity and worth with adults. Their evolving capacities as they grow and develop require different priorities and approaches at different stages of development throughout their childhood to ensure their right to health. Children's best interests should be central in considering all decisions that may affect their health and development (United Nations Committee on the Rights of the Child 2013). There are obligations in international law for actions that protect the rights and best interests of children, including their right to health, from the impact of climate change (United Nations General Assembly 2017).

The Paris Agreement on climate change recognizes the rights of children and people in vulnerable situations, intergenerational equity, and the right to health in its preamble (Conference of the Parties 2015). In its General Comment on the rights of the child to health, the UN Committee for the Rights of the Child included their right to a healthy environment and said climate change 'is one of the biggest threats to children's health and exacerbates health disparities. States should, therefore, put children's health concerns at the centre of their climate change adaptation and mitigation strategies (para. 50)' (United Nations Committee on the Rights of the Child 2013). The

UN Committee regularly makes recommendations about climate action in its in reviews of states' progress.

Dignity, equality, and non-discrimination

The burden of climate change is spread inequitably within and between countries. Those who have done the least to cause climate change are the worst affected and yet have the least economic and infrastructure resources for adaptation. Mothers and children, especially girls, are particularly affected (Metcalfe 2015). Climate change aggravates existing inequities, even leading to violent conflict over reduced resources and forced migration (Watts et al. 2015) and compounding other discrimination and injustices experienced by children.

If all children are to equally enjoy their right to health free of discrimination, climate-related actions must prioritize those children who are most disadvantaged, so that they are among the first to benefit. In practice this may require changing current priorities and ensuring more resources for these children, especially those experiencing multiple discrimination and living in vulnerable situations.

Participation

It is only recently that international climate negotiations have considered the rights of children and children have had some—albeit limited—opportunities to be heard. Children, especially those most affected, have the human right to participate and contribute to decision-making about matters which affect them and their health in accordance with their age and maturity. This requires having safe, child-friendly, and age-appropriate processes to support children to have their views heard and taken seriously. It may include them participating in legal and policy processes about health and the environment; priority-setting; the development and delivery of mitigation and adaptation policies and programmes; and accountability mechanisms. Participation has practical benefits: if children are involved in the development, delivery, and evaluations of programmes, the programmes are more likely to be effective (Children in a Changing Climate 2015).

Public health services should support children and their families' participation, including in the design, development, and evaluation of programmes. For example, although in times of extreme weather events and other disasters children carry out family and community duties, it is uncommon for them to be involved in disaster risk reduction planning. Yet their practical and creative ideas and unique knowledge of local environments mean they can make important contributions (Mudavanhu et al. 2015).

Many children are active environmental advocates. In several countries children are taking legal action, arguing that the state's obligations to hold some resources in common for all people, includes an obligation to protect the earth's atmosphere for children now and in the future. For example, Rabab Ali, a 7-year-old girl in Kawachi, petitioned Pakistan's Supreme Court about plans to develop a coal field, saying that atmospheric pollution harms and threatens the rights of Pakistani children and future generations, including their right to mental and physical health (Sabin Center for Climate Change Law 2016).

Accountability

There are some opportunities to hold states accountable at an international level for the impact of their climate-related actions on children through mechanisms under the Paris Agreement and the SDGs. These are weak mechanisms, especially for the magnitude of the task. The human rights mechanisms of the UN Human Rights Council, human rights treaty monitoring bodies and special procedures are also weak but can supplement the others. As countries develop and strengthen

national mechanisms for accountability there will be increasing opportunities to hold people, businesses, and other organizations accountable for the impact of actions on children.

Accountability mechanisms at international, regional, national, and local levels should include opportunities for the participation of children and child-friendly processes, and for consideration of the right to health of future generations of children. Unless data are disaggregated by age, sex, and other characteristics, it is not possible to hold states accountable in relation to their health rights obligations for *all* children.

Accountability requires consideration of the differentiated responsibilities for climate change and remedial action. High-income countries have benefited economically from their high greenhouse gas emissions. It is plainly unfair that the health of poor children throughout the world suffers the consequences. Countries with the highest emissions for their population must rapidly reduce their own emissions as part of remedying past, present, and future injustices against children. They also have urgent human rights duties to provide remedial action (including finance) for losses and damage suffered already by climate vulnerable countries and support their adaptation efforts (Hunt and Khosla 2009; Office of the UN High Commissioner for Human Rights 2015).

International assistance and cooperation

Unprecedented international cooperation is needed to reach the Paris Agreement's aspirational goal of limiting global warming below 1.5° and avoid catastrophic damage to global health. The human rights responsibilities of states include actively supporting—and not obstructing—the realization of the right to health for people in other countries. This requires working for an equitable international order and healthy environmental conditions for all including measures that will stop and reverse climate change (Hunt and Khosla 2009).

Human rights require states to act to the 'maximum extent of available resources'. High-income countries have a human rights obligation to share resources with less resourced countries so they have the resources they need to address climate change. Considerable increases in international assistance are needed to build climate-resilient health services and programmes. Health adaptation has however been largely overlooked in international assistance: between 2003 and 2017 less than 1% of the development funds available for adaptation through global climate change financing mechanisms, was for health sector adaptation (Watts et al. 2018). International assistance for health adaptation should pay special attention to poor children in low-income countries because they bear the greatest health impact of climate change.

Respect, protect, and fulfil all human rights

Tackling the impact of climate change on children's right to health requires action on determinants to respect, protect, and fulfil all of children's human rights. For example, for indigenous peoples, including children, climate change may mean displacement from their traditional lands and resources; disruption of their heritage, family life, cultural identity, and knowledge, and spiritual relationship with the natural environment; increasing poverty and marginalization; and direct and indirect harm to their health (Jones et al. 2014). Climate-related actions have potential to support or harm the well-being of indigenous peoples: for example, forced displacement of indigenous peoples from their traditional lands to build hydroelectric dams has caused serious human rights violations (Tauli-Corpuz 2017). When a state respects the right of indigenous peoples to guardianship of their forests; ensures the right is protected from interference by others (for example, multinational logging companies), and takes the legal, administrative, and budgetary measures necessary to fulfil the right, then this also respects traditional health practices and other health-related rights—as well as benefiting global climate mitigation.

Mitigation and adaptation plans made by states, regional, and intergovernmental organizations should address the human rights impacts of climate change, with a strong focus on respecting,

protecting, and fulfilling the rights of children and other vulnerable groups. Plans should include progressive improvements for child health services and programmes and the underlying determinants of health (Williams 2015). Climate-related actions should not cause retrogression of human rights: for example, the intention of increasing biofuel crop production may be to reduce use of fossil fuels but if done by displacing poor farmers from their land, may cause serious human rights violations, including of children's rights to health, food, education, a family environment, an adequate standard of living, safety, and so forth.

Available, accessible, acceptable, and good quality

There is considerable potential for public health co-benefits from climate-related actions (Watts et al. 2015). Climate-related public health services and programmes should be *available* to children, especially children living in climate vulnerable countries and situations. They should be *accessible*. This means prioritizing those children experiencing the greatest disadvantage; ensuring that access is not affected by physical barriers from location, distance and so forth, or economic cost; and enabling children to easily access information about health and climate change (for example, in the school curriculum). They should be provided in ways that are appropriate for the age and culture of the children served, i.e. *acceptable*. They should be informed by evidence, of sound *quality*, and provided in ways that support the best interests of children. In doing this, services should recognize the diverse needs of children—for example, protocols should include availability, accessibility, acceptability, quality criteria (AAAQ) for children who live with disabilities or whose caregivers do, and child-headed families.

Strengthening health systems

Health services can contribute to mitigation through reducing their own greenhouse gas emissions. This can include ways that come with co-benefits for the children: for example, when clinics are held in communities, travel-related emissions are reduced and physical accessibility and affordability can increase; and when hospitals stop using coal as a source of power, children are less exposed to harm from air pollution.

Health systems face considerable challenges in adaptation to climate change. Climate resilience has to be operationalized into each of WHO's six building blocks of health systems (World Health Organization 2015). The basic public health actions that are the most effective immediate general adaptation measures—such as vaccination, essential child health services, clean water, and sanitation, increasing disaster risk management capacity, and reducing poverty (Smith 2014) are all key to improving child health.

Case Study 3.3.2 An Impact Assessment of Big Data Using the Right to Health

Big Data is beginning to radically transform healthcare in multiple ways—through its use in electronic health records, data sharing, diagnostic tools, and treatment. These new technologies are changing the health and public health landscapes. Examples of the roles of digitization in healthcare include personalized medicine (a patient's treatment can be specifically tailored to their genetic, genomic, social and environmental profile) (Parry and Greenhough 2018), DNA sequencing (where the resulting vast amount of data is stored in bio banks), forensic, genetic or medical databases, including all the data from public health studies and clinical trials—and all of which can be re-

purposed for various technical inventions. These large data sets are driving the burgeoning artificial intelligence (AI) industry in healthcare. Where available and accessible these technologies offer improved public health, clinical care, and patient management; where they are not, inequities in health will likely result.

Blockchain technology also plays a role in health, with the technology, previously more associated with crypto currencies, being used to keep large health data sets more secure when data is shared. A focus on data security, although crucially important, must not become a distraction from other right to health threats posed by Big Data. In particular these threats may arise from data ownership and how equitably data will be used in the future.

Here we examine Big Data through a right to health lens and identify opportunities and threats to health equity and to the right to health in a new data-driven economy. We suggest that just as governments have been expected to carry out human rights impact assessments prior to adopting and implementing policies (MacNaughton and Hunt 2009; MacNaughton and Hunt 2011), so they should before embarking on the use of Big Data in health and other domains.

Laws, norms, and standards

While some health data sets are collected and controlled nationally, particularly people's medical records, much of the data feeding into global databases is derived from individuals contributing their own data to organizations beyond the borders of their resident state. This includes fitness and lifestyle data derived from wearable devices, or personal information uploaded into Facebook and other social media, as well as far more clinical and medically detailed data such as DNA sequencing, or clinical trials information. The rules, regulations, and laws governing how data can be used, and how individuals providing the data are protected, are not universal, although there are some regional regulations, for example, the European Union's General Data Protection Regulation (GDPR) https://www.eugdpr.org/. Collection and use of data for Big Data sets may be overseen by professional ethics, developed to guide fair use of data, but such codes of ethics are not legally binding.

National laws need to ensure that personal data is collected, stored, and used in ways that respect, protect, and fulfil human rights entitlements; but how can this be achieved beyond national borders? International human rights law provides such a facility. All use of Big Data, however the data sets were developed, even if ethically or with individual consent, must not subsequently be used in ways that violate human rights, including the right to health. For example, this means that, as a matter of law, Big Data must not be used, either when gathered or at a later time, to discriminate against people in terms of their right to health, including their participation in the health system, or to increase health inequities.

Dignity, equality, and non-discrimination

There are vast global inequalities in relation to data: from the personal (being literate, owning mobile phones, having internet access), to national information systems (functional national statistical offices, electronic information systems, electronic health information systems, a Big Data-driven sector that uses and develops electronic applications). Those countries, usually low-income countries, without sophisticated and extensive electronic infrastructure will be left further behind from the benefits of data-driven health gains, resulting in increased global health inequality, unless there is specific international assistance and technical support provided from high-income countries to address information systems deficits. There are great risks that, for example, wealthy people in high-income countries will have access to personalized medicine, tailored to their own genomes, while millions in poor countries continue to die prematurely from neglected tropical diseases, and other conditions of poverty. This is an example of gross inequity in public health, and a human rights failing at an international level, which requires urgent remedy.

Big Data enthusiasts cite its usefulness in many health domains, aside from clinical purposes. These uses include social media's role in monitoring and tracking epidemics and monitoring SDGs (Harvard Humanitarian Initiative, MIT Media Lab, and Overseas Development Institute 2016). Because of the human rights imperatives that inequality must not increase and there is no discrimination in the delivery of healthcare, it is vitally important to disaggregate data collected by not just looking at the data to determine the characteristics of those who contributed to the data sets, but also to understand who has not. When considering the use of Big Data to track epidemics, for example, there are still large pockets in the world where communities live without internet access (Williams and Hunt 2017). The percentage of the global population with access to the internet at the end of 2017, was 54.4%, ranging from 95% in North America to 35.2% in Africa (Internet World Stats 2018). Even these statistics require far greater disaggregation: for example, several African nations have less than 10% internet coverage, with Eritrea the lowest on 1.4%. If there is a global shift to monitoring epidemics through social media, widely considered more efficient and accurate than traditional health surveillance, those communities or countries living beyond internet penetration are at risk of invisibility during health or humanitarian crises.

As discussed earlier in 'Key features of the right to health', there are national and international human rights obligations to address these inequalities and discriminations, including through international assistance and cooperation (see 'Dignity, equality, and non-discrimination' and 'International assistance and cooperation')

Participation

The human right to participation means people are entitled to be part of decision-making processes that affect their lives and their health, and also to decision-making processes about health policy. At the least, when it comes to the use of their own data, people have the right to decide whether or not their data can be used for any purposes beyond the stated reason for its collection. Ownership and terms of use of the data must be clear and obvious, so that contributors to data sets are not later surprised to find their provision of personal information has been used to discriminate against them. This could, for example, occur if details of a person's dietary or other lifestyle factors, gleaned from social networks or sporting achievements applications, are shared with agencies that could withhold services on the basis of this information, or charge varying rates for the service on the basis of this information which was not provided directly to the service provider. Health insurance premiums could be an example of this.

Accountability

Over the past few years, use of Big Data has been plagued with scandals: for example, in the United Kingdom in 2016 over one million confidential patient files were given freely, and without hiding personal identities, to a Google-owned company to develop privately owned apps for clinical use; throughout 2016–2018 details were released about the extent of data mining through Facebook networks to influence voters and US and Brexit election results. As mentioned already, rules have now been introduced to control aspects of data collection, to protect privacy. However, such rules focus more on rights to privacy than they do on the longer-term consequences, for example, on who will own this data in its re-purposed presentations in the future. It is difficult to predict how data collected now will be used in the future, but our social and economic transactions, and healthcare, are now completely dependent on data. In order to protect social rights, such as the right to health, use of Big Data must be transparent, and there must be accountability for the human rights implications of how this data is subsequently used. Data processors need to be held to account for the human rights implications of the products they develop: for example, will data be gathered in discriminatory ways (perhaps from poorer communities and countries who use public health services), but will the resulting products be used only by individuals or their insurers who can afford

to pay for them? As discussed earlier, human rights require that duty bearers are held accountable by way of a process which includes monitoring, review against human rights and other standards, and remedial action (see ‘Key features of the right to health’).

Respect, protect, and fulfil all human rights

Big Data arises from multiple data sets in social, economic, and political domains. It is imperative that people’s rights in each of these domains are respected and protected during the collection and use of data, and that the integrated data are not used to discriminate against anyone in each of the domains. For example, domains of data about ethnicity, poverty, and crime might be combined in order to locate emergency health services where they are most needed which would be advancing equitable access to healthcare. However, the data could also result in biased algorithms that lead to disproportionate policing in areas where a minority population resides, leading in turn to increased arrests in that area. This could be discriminatory, and a violation of people’s rights to privacy and to justice.

Available, accessible, acceptable, and good quality

The benefits of Big Data should be available, accessible, and acceptable and based on good quality information and methodologies. When a state has determined, in a participatory way, the health services and facilities that can be provided from its maximum available resources, it is then obligated to ensure these are of good quality and available, accessible (including financially accessible), and acceptable to all. This covers such issues as ensuring that the datasets from which applications develop are built from quality, robust data, with privacy protected. Is the use of the data agreed to and acceptable to the people from whom the data was collected? Health services that arise from the use of Big Data, for example health apps to assist with diagnosis of diseases, or personalized medicine based on a person’s genome, must meet the ‘available and accessible’ criteria and avoid increasing inequities. For example, if patient information and notifications switch to digital systems, does this discriminate against people without smart phones, and result in them losing access to these services?

Strengthening health systems

The health system is the core institution through which the right to health can be fulfilled. Data is an essential component of health systems, used for improving patient management, providing health information, and importantly, to monitor the health system’s performance. As data sets increase in size, and as multiple data sets are integrated, it is essential to monitor and review that this use of Big Data continues to strengthen the health system. Right to health impact assessments should be undertaken before installing or implementing data-driven new technologies. Assessments are needed to question whether such technologies are protecting and respecting people’s rights; they can guide their use so technology strengthens, not weakens, health systems and thereby can fulfil people’s health rights entitlements. Questions about Big Data-driven technologies and their impacts on health systems would cover both technical infrastructure aspects, such as the robustness of the new systems and their vulnerability to hackers or power outages, through to ownership of the systems and the data contained within them.

Case Study 3.3.3 International Assistance and Cooperation and the Right to Health

One element of the right to health pertains to the obligation on states to provide international assistance and cooperation. It is incumbent upon states that can afford to do so, to offer assistance and cooperation to states who request it in order to meet their human rights obligations. States have an obligation to seek such assistance when they do not have available resources to meet their right to health duties. There are also global aid agreements arising from the Alma-Ata Declaration, the SDGs, and a commitment made by OECD countries in 1970 to spend 0.7% of gross national product (GNP) on aid (OECD, n.d.). Most OECD countries have not yet achieved that level of expenditure on aid, although global health spending has increased dramatically since 2000. WHO estimated $19 billion was spent on development assistance for health (DAH) in 2015. After a decade of rapid growth from 2000 to 2010 (up 11.4% annually), DAH grew at only 1.8% annually between 2010 and 2016. In low-income countries, DAH made up 34.6% of total health spending in 2016 (IHME 2016; WHO 2017). This is partially in response to the attention brought to global health and other inequalities from the Millennium Development Goals (MDGs), and SDGs, but also because of funding from global health initiatives such as The Global Fund to Fight AIDS, TB and Malaria, Global Alliance for Vaccines and Immunization (GAVI) and the Gates Foundation. Non-government organization (NGO) funding made up 30.1% of total DAH disbursed in 2016, whereas in 1990 it contributed only 7.6% of total DAH funding. NGOs collectively disbursed more funds than any other group by a wide margin (IHME 2016, pp. 32–3). Usually NGO funding is directed to the health initiatives pertinent to their own missions. The capacity of such organizations, and the non-state sector as a whole, to have an impact of people's right to health entitlements is therefore considerable.

In addition to the international human rights obligation to seek and provide international assistance and cooperation, when such assistance and cooperation is provided its delivery must conform to the requirements of human rights law. In this case study, we use the right to health framework to explore issues that global partners must consider when shaping aid interventions, so they comply with duties to respect, protect, and fulfil health rights.

Laws, norms, and standards

Development assistance must be compliant with international human rights law, as well as national laws and regulations. In the former, this pertains to such considerations as delivering assistance without discrimination (see next section), and at the national level, it will mean ensuring that, inter alia, health personnel will not be trained to deliver services beyond their regulatory scope. An example might be not training nurses to provide minor surgery if they are not permitted to perform such procedures.

A human rights-based approach to health recognizes that health is a right, and people have human rights entitlements to health services. International providers of assistance for health services are therefore assisting the fulfilment of people's rights, and empowering people to recognize and claim these rights entitlements. This is a different conceptualization of aid initiatives from ones of charity or philanthropy where beneficiaries may be viewed as fortunate to receive services but are not considered to have equal rights to them.

Dignity, equality, and non-discrimination

Assistance must not be provided in a discriminatory way, even when national law results in such discrimination. Examples of discrimination in health can be found when only married women can

receive some health services, or husbands' permission is required for services, or services are prioritized for people working in the formal sector, and so on.

International actors must therefore examine issues of inequality and discrimination in the countries they work to identify where interventions are most needed to reduce health inequities in that country. Treating all people with dignity is a fundamental value on which all human rights are based, and in aid programmes, that extends to the way in which people are presented in fundraising activities in donor countries.

Participation

When providing assistance and cooperation, international agencies need to ensure there is participation at many levels: that the state welcomes and invites their support, that local partners are leaders in the design and implementation of support, and the people for whom services will be provided have contributed meaningfully to all aspects of the initiative and its delivery.

Accountability

Human rights require effective accountability arrangements to be put in place for the human rights obligations arising from international assistance and cooperation. Accountability in aid initiatives is frequently limited to financial accountability, but human rights accountability is much more than this. People have human rights entitlements to know how decisions are made about the health system and provision of services. This includes for international assistance, at the least, transparency around: approval processes for international assistance support; the nature of services, products, or funds being delivered; location of services; how the quality of health services will be monitored; and whether services are provided in a non-discriminatory and equitable way. Human rights accountability necessitates the use of transparent processes to check that aid initiatives respect, protect, and fulfil human rights, remembering that rights are indivisible—so it is not acceptable to promote one health service at the expense of other health or other services, unless this is an agreed approach to the national plan. (For example, a country as a whole may decide to forego some tertiary level care to ensure primary healthcare reaches everyone). This is explained in 'Strengthening health systems'.

Respect, protect, and fulfil all human rights

Health aid cannot be sustainable without considering the broad determinants of health and their related rights. For example, assistance to improve a healthcare service such as child health is of limited use if children live in communities that do not have potable water, sanitation systems, or schools. Similarly, health aid programmes to educate midwives are unlikely to succeed in the long term if midwives are not paid adequately or if they are subject to violence at home or at work. It is therefore important that determinants and related human rights are assessed as part of international assistance and cooperation, and ways of addressing these rights are identified.

Available, accessible, acceptable, and good quality

International assistance and cooperation for health must be delivered or directed so that their support for health services meet the AAAQ criteria.

- Under availability, questions should be asked about what is required to make a health service fully available. This is not just about where the service is to be provided, but, for example, what medicines and other supplies are needed, and how will they be made available in a sustainable way. Similarly, how will staff be trained and employed so they are available for these services?
- Under accessibility, aid initiatives might consider who can access the service in terms of gender, disability, place of dwelling, or knowledge of such services. Care must be taken that groups, for example, indigenous people, are made aware of services and are helped to access them.
- Acceptability calls for culturally appropriate care to be made available. For example, are new services planned to be delivered in a way that is acceptable to women in minority or indigenous communities? Will health workers speak the languages spoken by disadvantaged groups?
- Quality healthcare is critically important and lower standards are not acceptable simply because recipient countries have fewer economic resources. This means that health workers must be well trained and supported, medicines must be within use-by dates, and countries that receive assistance should not be treated as dumping grounds for obsolete second hand equipment that will not be serviced, or as a means of claiming tax rebates on the 'value' of the goods donated.

Donors of health aid should include a right to health impact assessment of their initiatives before implementation, to consider the possible impact of a new or expanded service on the AAAQ of other health services. For example, well-funded stand-alone interventions for some health services may draw health professionals and other resources to them, weakening other core services—thereby resulting in a reduction in availability and quality (Williams and Brian 2013). At all times, donors must consider the sustainability of any interventions, in order to avoid commencement of services that will not endure, thereby later contributing to a regression in health rights fulfilment.

Strengthening health systems

Because the health system is the key institution through which the right to health can be realized, international assistance must take care to work in ways that promotes the long-term sustainability of the system. Initiatives must work in keeping with the national health plan. They should also strengthen the health system's capacity to provide quality health services for all, addressing inequities so that the most disempowered and worst off have their rights fulfilled first. Health impact assessments should be undertaken before implementing aid initiatives to systematically check the health system will be strengthened, not weakened, by international assistance.

Conclusion

We began this chapter with illustrations of successful right to health initiatives from around the world in high-, middle- and low-income countries, and concluded with detailed case studies demonstrating the right to health in practice. Together, these are examples of human rights supporting global public health.

The right to health entitles all people to a package of health services, facilities, and goods, extending to healthcare, and public health, including the underlying determinants of health, such as access to safe water, adequate sanitation, and health-related information. This package must be available, accessible, and of good quality. Also, it must be sensitive to different cultures, thus, acceptable. Health systems are the key institution through which healthcare and public health are delivered. They must be sites of inclusion, where people can participate in policy making and planning. In this way, health systems will have many features that are common to all countries, but there will also be differences between countries that will reflect their different disease burdens, determinants of health, cultural contexts, resource availability, and people's priorities.

Whichever approach and combination of approaches is employed, the objective is to respect, protect, and fulfil people's right to health.

Key points

The right to the highest attainable standard of health is enshrined in several international treaties and numerous national constitutions.

International law recognizes that all people have human rights entitlements and states have binding obligations to uphold these rights.

Public health and the right to the highest attainable standard of health are natural allies both seeking effective, integrated, responsive health systems that are accessible to all.

The key features of the right to health assist in understanding and responding effectively to public health issues and reducing global and national inequities.

The right to the highest attainable standard of health can be understood as a right of everyone to public health and health services and the underlying determinants of health.

ACKNOWLEDGEMENTS

In a number of ways, this chapter builds on the corresponding chapter, 'The right to the highest attainable standard of health', in the 5th and 6th editions of the *Oxford Textbook of Global Public Health*. We gratefully acknowledge our indebtedness to the co-authors—Gunilla Backman,

Judith Bueno de Mesquita, Louise Finer, Rajat Khosla, Dragana Korljan, and Lisa Oldring—who worked with the current co-author, Paul Hunt, on the chapter in the two preceding editions.

Thanks are also due to Amanda Gray Meral, University of Essex, for her research assistance in preparing this chapter for the 7th edition.

REFERENCES

Acheson, D. (1988). *Public Health in England: The Report of the Committee of Inquiry into the Future Development of the Public Health Function*. London: HMSO.

ACHPR (African Commission on Human and Peoples' Rights) (2001). *Communication No. 155/96: The Social and Economic Rights Action Center for Economic and Social Rights v. Nigeria*. Fifteenth Annual Activity Report of ACHPR, 2001–2002, Annex V.

ACHPR (African Commission on Human and Peoples' Rights) (2009) *Sudan Human Rights Organisation, et al. v. Sudan*, Comm. No. 279/03-296/05.

African Charter on Human and Peoples' Rights (OAU Doc. CAB/LEG/67/3 (1982). [Online] Available at http://www.achpr.org/files/instruments/achpr/banjul_charter.pdf

African Charter on the Rights and Welfare of the Child [Online]. Available at: http://www.achpr.org/instruments/child/

Ahdoot, S., Pacheco, S. E., and Council on Environmental Health. (2015). Global climate change and children's health. *Pediatrics*, 136(5), e1468–84.

Alliance Scotland (2013). *Being Human: A Human Rights-based Approach to Health and Social Care in Scotland*, Alliance, p. 10. [Online] Available at: http://lx.iriss.org.uk/sites/default/files/resources/human_rights_20pp.pdf

Backman, G., et al. (2008). Health systems and the right to health: an assessment of 194 countries. *Lancet*, **372** (9655), 2047–85.

Bhalotra, S., Rocha, R., and Soares, R. (2016). *Does Universalization of Health Work? Evidence from Health Systems Restructuring and Maternal and Child Health in Brazil. Institute for Social and Economic Research*, Working paper No 2016-16, University of Essex. [Online] Available at: https://www.iser.essex.ac.uk/research/publications/working-papers/iser/2016-16.pdf

Burton, R., Giddy, J., and Stinson, K. (2015). Prevention of mother-to-child transmission in South Africa: an ever-changing landscape. *Obstetric Medicine*, 8(1), 5–12.

Bustreo, F., et al. (2013). *Women's and Children's Health: Evidence of Impact of Human Rights*. Geneva: WHO.

Cappa, C. (2018). *New Global Estimates of Child Marriage, 15 March 2018 Webinar*. [Online] Available at: https://www.girlsnotbrides.org/resource-centre/unicef-data-on-child-marriage-2018/

CCSA (Constitutional Court of South Africa) (2002). *Minister of Health* et al *v. Treatment Action Campaign* et al. Case CCT 8/02, decided on 5 July 2002.

Centers for Disease Control and Prevention (2013). US Public Health Service Syphilis Study at Tuskegee. Washington DC: CDC. [Online] Available at: https://www.cdc.gov/tuskegee/timeline.htm

Children in a Changing Climate (2015). *Child-centred Adaptation: Realising Children's Rights in a Changing Climate*. [Online] Available at: http://www.childreninachangingclimate.org/

Conference of the Parties (2015). *Adoption of the Paris Agreement, 21st Conference of the Parties, 12 December 2015, UN Doc. FCCC/CP/2015/L.9/Rev.1*. [Online] Available at: http://unfccc.int/documentation/documents/advanced_search/items/6911.php?priref=600008831

Constitutional Court of South Africa (2002). *Minister of Health v Treatment Action Campaign (TAC)* 5 SA 721 (CC).

Council of Europe (1950). Convention for the Protection of Human Rights and Fundamental Freedoms. ETS No. 5. Strasbourg: Council of Europe.

Council of Europe (1997). European Court of Human Rights. *D. v. United Kingdom*, Doc. 146/1996/767/964.

Dasgupta, J., Sandhya, Y.K., Lobis, S., Pravesh, V., and Schaaf, M. (2015). Using technology to claim rights to free maternal health care: lessons about impact from the My Health, My Voice pilot project in India. *Health and Human Rights*, 17(2), 135–47.

Data-Pop Alliance (2016). *Opportunities and Requirements for Leveraging Big Data for Official Statistics and the Sustainable Development Goals in Latin America*. Cambridge, MA: Harvard Humanitarian Initiative, MIT Media Lab, and Overseas Development Institute.

de Zayas, A.-M. (2015). Thematic report to the Human Rights Council on the adverse impacts of free trade and investment agreements on a democratic and equitable international order, Human Rights Council, 14 July 2015, UN Doc. A/HRC/30/44. [Online] Available at: http://www.ohchr.org/EN/HRBodies/HRC/RegularSessions/Session30/Documents/A_HRC_30_44_ENG.docx

Ferraz, O. (2018) Editorial: Health in the Courts of Latin America. *Health and Human Rights*, 20(1), 67–77.

Gauri, V. and Brinks, D. (2008). *Courting Social Justice: Judicial Enforcement of Social and Economic Rights*. Cambridge: Cambridge University Press.

Girls Not Brides (2015). *Theory of Change*. [Online] Available at: https://www.girlsnotbrides.org/child-marriage-theory-of-change/

Girls Not Brides (2018). *About Girls Not Brides/Our Members*. [Online] Available at: https://www.girlsnotbrides.org/about-girls-not-brides/our-members/—who

Gruskin, S., Bogecho, D., and Ferguson, L. (2010). 'Rights based approaches' to health policies and programs: articulations, ambiguities, and assessment. *Journal of Public Health Policy*, 31(2), 129–45.

Heymann, J., Cassola, A., Raub, A., and Mishra, L. (2013). Constitutional rights to health, public health and medical care: the status of health protections in 191 countries. *Global Public Health*, 8(6), 639–53.

Hosking, J., Jones, R., Percival, T., Turner, N., and Ameratunga, S. (2011). Climate change: the implications for child health in Australasia. *Journal of Paediatric Child Health*, 47(8), 493–6.

Hunt, P. (2016). Interpreting the international right to health in a human rights-based approach to health. *Health and Human Rights*, 18(2), 109–30.

Hunt, P. and Khosla, R. (2009). Climate change and the right to the highest attainable standard of health. In: Humphreys, S. (ed.) *Human Rights and Climate Change*. Cambridge: Cambridge University Press, pp. 238–56.

Hunt, P. and Leader, S. (2010). Developing and applying the right to the highest attainable standard of health: the role of the UN Special Rapporteur. In: Harrington, J. and Stuttaford, M. (eds.) *Global Health and Human Rights: Philosophical and Legal Perspectives*. New York: Routledge, pp. 28–61.

Hunt, P. and MacNaughton, G. (2007). A human rights-based approach to health indicators. In: Baderin, M. and Mccorquodale, R. (eds.) *Economic, Social and Cultural Rights in Action*. Oxford: Oxford University Press, pp. 303–30.

Hunt, P., Mesquita, J., and Khosla, R. (2010). The human rights responsibilities of international assistance and cooperation in health. In: Gibney, M. and Skogly, S. (eds.) *Universal Rights and Extraterritorial Obligations*. Pennsylvania Studies in Human Rights. Pennsylvania, PA: University of Pennsylvania Press, pp. 104–29.

Hunt, P., et al. (2015). The right to the highest attainable standard of health. In: Detels, R., et al. (eds.) *Oxford Textbook of Global Public Health* (6th ed.), Chap. 4.1. Oxford: Oxford University Press.

Hunt, P., Yamin, A.E., and Bustreo, F. (2015). Making the case; what is the evidence of impact of applying human rights-based approaches to health? *Health and Human Rights*, 17(2), 1–9.

IACHR (Inter-American Commission on Human Rights) (1997). *Victor Rosario Congo v. Ecuador.* Case 11.427, Report No. 12/97, OEA/Ser.L/V/II.95 Doc. 7 rev at 257.

IACHR (Inter-American Commission on Human Rights) (2000). *Jorge Odir Miranda Cortez et al v. El Salvador*. Case 12.249, Report No. 29/01, OEA/Ser.L/V/II.111 Doc. 20 rev. at 282.

IACHR (Inter-American Commission on Human Rights) (2015) *Gonzales Lluy et al. v. Ecuador.*

IACtHR (Inter-American Court of Human Rights) (2018). *Poblete Viches et al vs Chile*. [Online] Available at: http://www.corteidh.or.cr/docs/casos/articulos/seriec_349_esp.pdf)

Institute for Health Metrics and Evaluation (IHME) (2017). *Financing Global Health 2016: Development Assistance, Public and Private Health Spending for the Pursuit of Universal Health Coverage*. Seattle, WA: IHME.

Internet World Statistics (2018). [Online] Available at: https://www.internetworldstats.com/stats.htm

Jones, R., Bennett, H., Keating, G., and Blaiklock, A. (2014). Climate change and the right to health for Maori in Aotearoa/New Zealand. *Health and Human Rights Journal*, 16(1), 54–68.

Knox, J. H. (2016). Report of the Special Rapporteur on the issue of human rights obligations relating to the enjoyment of a safe, clean, healthy and sustainable environment. Human Rights Council, 1 February 2016, UN Doc. A/HRC/31/52. Geneva. [Online] Available at: https://undocs.org/A/HRC/31/52

Maastricht Principles on Extraterritorial Obligations of States in the Area of Economic, Social and Cultural Rights (2011). Maastricht: International Commission of Jurists, Maastricht University.

MacNaughton, G. and Hunt, P. (2009). Health impact assessment: the contribution of the right to the highest attainable standard of health. *Public Health*, 123(4), 302–5.

MacNaughton, G. and Hunt, P. (2011). A human-rights based approach to social impact assessment. In: Vanclay, F. and Esteves, A.M. (eds.) *New Directions in Social Impact Assessment: Conceptual and Methodological Advances*. Cheltenham: Edward Elgar, pp. 355–68.

Metcalfe, S. (2015). Fast, fair climate action crucial for health and equity. *New Zealand Medical Journal*, 128(1425), 13–17.

Mudavanhu, C., et al. (2015). Taking children's voices in disaster risk reduction a step forward. *International Journal of Disaster Risk Science*, **6**(3): 267–81.

OAS (Organization of American States) (1969). *American Convention on Human Rights*. OAS Treaty Series No. 36, 1144 U.N.T.S. 123. Adopted at the Inter-American Specialized Conference on Human Rights, San Jose, Costa Rica, 22 November 1969. Washington DC: OAS.

OECD, Development Assistance statistics. [Online] Available at: http://www.oecd.org/dac/stats/the07odagnitarget-ahistory.htm

Office of the United Nations High Commissioner for Human Rights (OHCHR) and UNAIDS (2006). *International Guidelines on HIV/AIDS and Human Rights: 2006 Consolidated Version*. Geneva: OHCHR and UNAIDS.

Office of the United Nations High Commissioner for Human Rights (2015). *Key Messages on Human Rights and Climate Change*. [Online] Available at:

http://www.ohchr.org/Documents/Issues/ClimateChange/KeyMessages_on_HR_CC.pdf
Office of the High Commissioner for Human Rights (OHCHR). Status of Ratification Interactive Dashboard, as of May 2018. [Online] Available at: http://indicators.ohchr.org/
Office of the High Commissioner for Human Rights (2014). *Technical Guidance on the Application of a Human Rights Based Approach to the Implementation of Policies and Programmes to Reduce and Eliminate Preventable Mortality and Morbidity of Children Under 5 Years of Age*. UN Doc. A/HRC/27/31.
Parry, B. and Greenhough, B. (2018). *Bioinformation*. Cambridge, UK: Polity Press.
Pizzarossa, L.B. and Perehudoff, K. (2017) Global Survey of National Constitutions: mapping constitutional commitments to sexual and reproductive health and rights. *Health and Human Rights*, 19(2), 279–93
Rumbold, B., et al. (2017). Universal health coverage, priority setting, and the human right to health. *The Lancet*, 390(10095), 712–14.
Sabin Center for Climate Change Law (2016). *Ali v. Federation of Pakistan*. [Online] Available at: http://climatecasechart.com/non-us-case/ali-v-federation-of-pakistan-2/
Scottish Human Rights Commission (2009). *Human Rights in a Health Care Setting: Making it Work: An Evaluation of a Human Rights-based Approach at The State Hospital*, p. 5.
Smith, K.R., et al. (2014). Human health, impacts, adaptation, and co-benefits. In: Field, C.B., et al. (eds.) *Climate Change 2014: Impacts, Adaptation, and Vulnerability Part A: Global and Sectoral Aspects Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge, UK: Cambridge University Press, pp. 709–54.
Tauli-Corpuz, V. (2017). *Report of the Special Rapporteur on the Rights of Indigenous Peoples—Impacts of Climate Change and Climate Finance on Indigenous Peoples' Rights.* Human Rights Council, 1 November 2017. UN Doc. A/HRC/36/46. [Online] Available at: http://www.ohchr.org/EN/Issues/IPeoples/SRIndigenousPeoples/Pages/AnnualReports.aspx
UN (United Nations) (1948). *Universal Declaration of Human Rights*. GA Resolution 217A (III), UN GAOR, Resolution 71, UN Document A/810. New York: UN.
UN (United Nations) (1966a). *International Covenant on Civil and Political Rights* (ICCPR). UN GA Resolution 2200A (XXI), 16 December 1966. New York: UN.
UN ICCPR (2007). *LMR v. Argentina* (2007) UN Doc. CCPR/C/101/D/1608/2007.
UN (United Nations) (1966b). *International Covenant on Economic, Social and Cultural Rights* (ICESCR). UN GA Resolution 2200A (XXI), 16 December 1966. New York: UN.
UN (United Nations) (1979). *Convention on the Elimination of All Forms of Discrimination Against Women* (CEDAW). GA Resolution 34/180, UN GAOR, 34th Session, Supplement No. 46 at 193, UN Document A/34/46. New York: UN.
UN (United Nations) (1989). *Convention on the Rights of the Child* (CRC). UN GA Resolution 44/25, 20 November 1989. New York: UN.
United Nations Committee on the Rights of the Child (2013). *General Comment No. 15 (2013) on the right of the child to the enjoyment of the highest attainable standard of health (art. 24), 17 April 2013, UN Doc. CRC /C/GC/15*. Geneva. [Online] Available at: http://tbinternet.ohchr.org/_layouts/treatybodyexternal/TBSearch.aspx?TreatyI
United Nations General Assembly (2007). United Nations Declaration on the Rights of Indigenous Peoples. Resolution adopted by the General Assembly on 13 September 2007, UN Doc. A/RES/61/295. [Online] Available at: https://documents-dds-ny.un.org/doc/UNDOC/GEN/N06/512/07/PDF/N0651207.pdf?OpenElement
United Nations General Assembly (2016). *Report of the Special Rapporteur on the Right of Everyone to the Enjoyment of the Highest Attainable Standard of Physical and Mental Health*; Doc no A/71/304 [SDGs]. [Online] Available at: https://undocs.org/A/71/304

United Nations General Assembly (2017). Analytical study on the relationship between climate change and the full and effective enjoyment of the rights of the child. Report of the Office of the United Nations High Commissioner for Human Rights. 4 May 2017. UN Doc. A/HRC/35/13. Geneva: United Nations Office of the High Commissioner for Human Rights.

United Nations General Assembly (2017). Report of the *Special Rapporteur on the Right of Everyone to the Enjoyment of the Highest Attainable Standard of Physical and Mental Health*; Doc no A/72/137 [corruption]. [Online] Available at: https://undocs.org/A/72/137

UN (United Nations) (1993). United Nations General Assembly. *Vienna Declaration and Programme of Action. World Conference on Human Rights*, Vienna 14–25 June 1993, UN Document A/CONF.157/23. New York: UN.

United Nations General Assembly. (2007). *United Nations Declaration on the Rights of Indigenous Peoples*. Resolution adopted by the General Assembly on 13 September 2007, UN Doc. A/RES/61/295. New York. [Online] Available at: https://documents-dds-ny.un.org/doc/UNDOC/GEN/N06/512/07/PDF/N0651207.pdf?OpenElement

United Nations General Assembly. (2017). Analytical study on the relationship between climate change and the full and effective enjoyment of the rights of the child. Report of the Office of the United Nations High Commissioner for Human Rights. 4 May 2017. UN Doc. A/HRC/35/13. [Online] Available at: http://ap.ohchr.org/documents/dpage_e.aspx?si=A/HRC/35/13

UN (United Nations) (1994). *International Conference on Population and Development.* 5–13 September 1994. Cairo, Egypt.

UN (United Nations) (1995). *Report of the Fourth World Conference on Women*. Beijing, China 4–15 September 1995. UN Document A/CONF.177.20.

UNCESCR (United Nations Committee on Economic, Social and Cultural Rights) (1990). *General Comment No. 3 (Fifth Session). The Nature of States Parties Obligations (Art.2, par.1)*. UN Document E/1991/23. Geneva: UN.

UNCESCR (United Nations Committee on Economic, Social and Cultural Rights) (2000). *General Comment No. 14 (Twenty Second Session). The Right to the Highest Attainable Standard of Health*. UN Document E/C.12/2000/4. Geneva: UN.

UNCHR (United Nations Commission on Human Rights) (2003). *The Right of Everyone to the Enjoyment of the Highest Attainable Standard of Physical and Mental Health*, Report of the Special Rapporteur, 13 February 2003, UN Document E/CN.4/2003/58). Geneva: UN.

UN ECOSOC (United Nations Economic and Social Council) (1985). *Siracusa Principles on the Limitation and Derogation Provisions in the International Covenant on Civil and Political Rights*. UN Document E/CN.4/1985/4, Annex.

UNGA (2015). *Transforming Our World: The 2030 Agenda for Sustainable Development*, resolution adopted by the GA on 25 September 2015, UN Doc. A/Res/70/1 (2015).

UNHRC (United Nations Human Rights Council) (2007). *Report of the Special Rapporteur on the Right of Everyone to the Enjoyment of the Highest Attainable Standard of Physical and Mental Health*, Paul Hunt. 17 January 2007, UN Document A/HRC/4/28. New York: UN.

UNHRC (United Nations Human Rights Council) (2008). Report of the Special Rapporteur on the Right of Everyone to the Enjoyment of the Highest Attainable Standard of Physical and Mental Health, Paul Hunt. 8 January 2008, UN Document A/HRC/7/11. New York: UN. [Online] Available at: http://www.who.int/medicines/areas/human_rights/A_HRC_7_11.pdf

UNOHCHR (United Nations Office of the High Commissioner for Human Rights) (2012b). *Technical Guidance on the Application of a Human Rights-Based Approach to the Implementation of Policies and Programmes to Reduce Preventable Maternal Morbidity and Mortality*. UN Document A/HRC/21/22. Geneva: UN.

UN Human Rights Special Procedures (2018). *Framework Principles on Human Rights and the Environment*. [Online] Available at:

https://www.ohchr.org/Documents/Issues/Environment/SREnvironment/FrameworkPrinciplesUserFriendlyVersion.pdf
Watts, N., et al. (2015). Health and climate change: policy responses to protect public health. *Lancet*, 386(10006), 1861–914.
Watts, N., et al. (2018). The Lancet Countdown on health and climate change: from 25 years of inaction to a global transformation for public health. *Lancet*, 391(10120), 581–630.
Williams, C. (2015). Climate change and the right to health. In: Luber, G. and Lemery, J. (eds.) *Global Climate Change and Human Health: From Science to Practice*. San Francisco, CA: Jossey-Bass, pp. 601–15.
Williams, C and Brian G. (2013). Using a rights-based approach to avoid harming health systems: a case study from Papua New Guinea. *Journal of Human Rights Practice*, 5(1), 177–94.
Williams, C. and Hunt, P. (2017). Neglecting human rights: accountability, data and sustainable development goal 3. *The International Journal of Human Rights*, 21(8), 1114–43.
Winslow, C.E. (1920). The untilled fields of public health. *Science*, 51(1306), 23–33.
World Health Organization (1946). *Constitution of the World Health Organization*, adopted by the International Health Conference, New York, 19 June–22 July 1946, and signed on 22 July 1946. Geneva: WHO.
World Health Organization (1978a). Declaration of Alma-Ata: International Conference on Primary Health Care. 6–12 September, USSR.
World Health Organization (1978b). *A Joint Report by the Director-General of the WHO and the Executive Director of UNICEF Presented at the International Conference on Primary Health Care*, **1978**, Alma-Ata. Geneva: WHO.
World Health Organization (2017). *New Perspectives on Global Health Spending for Universal Health Coverage*. Geneva: WHO.
World Health Organization (WHO) (2011). *The Rio Political Declaration on Social Determinants of Health*. World Conference on Social Determinants of Health, Rio de Janeiro, 19–21 October 2011. [Online] Available at: http://www.who.int/sdhconference/declaration/en/
World Health Organization (WHO) (2016). *The Second Global Conference on Health and Climate*. Conference Conclusions and Action Agenda. Paris, 7–8 July 2016. [Online] Available at: http://www.who.int/globalchange/conferences/second-global/en/
World Health Organization (2014). *Ensuring Human Rights in the Provision of Contraceptive Information and Services: Guidance and Recommendations*. Geneva: WHO.
World Health Organization and the Thai Ministry of Health. Bangkok Charter for Health Promotion in a Globalized World. Sixth Global Conference on Health Promotion. Policy and partnership for action: addressing the determinants of health. Bangkok, Thailand, 7–11 August 2005. [Online] Available at: http://www.who.int/healthpromotion/conferences/6gchp/en/
World Health Organization, & United Nations Framework Convention on Climate Change Secretariat. (2015). *Climate and Health Country Profiles 2015: A Global Overview.* [Online]. Available at: http://www.who.int/globalchange/resources/countries/en/
World Health Organization (2015). Operational framework for building climate resilient health systems. Geneva: World Health Organization.
Yamin, A.E. (2016). *Power, Suffering and the Struggle for Dignity*. Philadelphia: PENN Press, p. 104
Yamin, A.E. and Gloppen, S. (eds.) (2011). *Litigating Health Rights: Can Courts Bring More Justice to Health?* Cambridge, MA: Human Rights Program, Harvard Law School.

[1] International Convention for the Elimination of All Forms of Racial Discrimination (CERD)., Article 5 (e) (iv); the International Convention on the Elimination of All Forms of Discrimination Against Women (CEDAW, Articles 11(1)f, 12 and 14(2) (b); the Convention on the Rights of Persons with Disabilities (CRPD, Articles 25 and 26, International Convention on the Protection of the Rights of All Migrant Workers and Members of Their Families. The preamble of WHO's first treaty, The Framework Convention on Tobacco Control, reiterates international law on the right to health in its preamble (2003).

3.4

Law and the public's health

Lawrence O. Gostin

Introduction to law and the public's health

Public health law can be defined as the legal powers and duties of government, in collaboration with multiple partners (e.g. healthcare, business, the community, the media, and academe), to assure the conditions for people to be healthy and safe (to identify, prevent, and ameliorate risks to health in the population). In many countries, there is a constitutional right to health, while in others government has the power to safeguard the public's health. The prime objective of public health law is to pursue the highest possible level of physical and mental health in the population, consistent with the values of social justice (Gostin and Wiley 2016).

Although government needs adequate power to safeguard the public's health, it should not interfere with the civil rights and liberties of individuals without adequate justification. Government therefore has limitations—both legal and ethical—in exercising power that constrains the autonomy, privacy, liberty, proprietary, or other legally protected interests of individuals. Some of the most challenging, and enduring, debates in public health law are the conflicts between individual liberties and the common good.

Public health laws have deep roots in social history, ranging from the ancient use of quarantines at land and at sea, to sanitary laws in the wake of the Industrial Revolution, through to injury prevention, and non-communicable disease (NCD) control in the late twentieth century. Domestic public health laws operate at the local (e.g. municipal codes), state (provincial), and national level. Often, federalist states such as Brazil, Canada, India, and the United States have complex constitutional arrangements between local, regional, and national public health laws.

Domestic public health laws and regulations also intersect with international health law (e.g. the International Health Regulations and Framework Convention on Tobacco Control), 'soft' law (e.g. World Health Organization (WHO) codes of practice and global strategies), and wider international regimes that impact on health (e.g. trade, migration, human rights, and the environment). At the 2019 World Health Assembly, the WHO adopted a Global Action Plan on Migrant Health. This is an area of considerable public health concern as many migrants are excluded from national health systems (Gostin et al. 2019). The complex intersections between local, national, and international law to protect the public's health cannot be overemphasized.

This chapter first examines the powers and duties of government to protect the public's health and safety, together with the limits on its authority to protect a sphere of individual rights and liberties. Second, it explores the central value of public health law—social justice, with particular attention to unconscionable inequalities in health based on socioeconomic status. Third, the chapter examines the statutes and regulations that create public health agencies—defining their mission, powers, funding, and scope. Fourth, the chapter categorizes seven legal 'tools' at the state's disposal to promote the health, safety, and well-being of the population: taxing and spending, the informational environment, the built environment, the socioeconomic environment, direct regulation, indirect regulation, and deregulation. Finally, the chapter turns to how legal interventions can be used to prevent or ameliorate three major health hazards: infectious diseases, NCDs, and injuries (both unintentional and intentional).

Public health: powers, duties, and limits

Governments have the power and duty to safeguard the public's health and safety. The international human right to health, which virtually all countries accept as part of treaty obligations, imposes on governments the primary responsibility for the public's health. In many newer democracies (for example, in Africa and Latin America), national constitutions also guarantee the right to health, as well as other socioeconomic rights such as the right to life and to a safe environment (Friedman and Gostin 2012;

Gostin et al. 2019). Emerging economic powers such as Brazil, India, and South Africa have had robust 'right to health' litigation in their judicial systems. Whether through international law or national constitutions, the state's highest obligation is the defence of the social and economic well-being of the population, with health and security being among the most vital. In most theories of democracy health and security are seen as quintessential aspects of the social contract between individuals and their governments (Walzer 1983).

Beyond socioeconomic rights, virtually all states around the world—older constitutions such as in the United States, newer ones such as in South Africa, and even those without written constitutions such as the United Kingdom—protect civil and political rights—for example, bodily integrity, privacy, liberty, and non-discrimination. At the same time, national constitutions, as well as public health laws, also usually afford individuals civil and political rights, which can constrain the power of ministries to use compulsion. Provided that agencies act for an important public health purpose, and use means that are proportionate to the ends, they can require conformance with publicly established standards of conduct.

Protecting and preserving the public's health and safety is often not possible without constraining a range of private activities that pose unacceptable risks. Private actors can profit by engaging in practices that damage the rest of society: individuals can spread infectious disease; industry has a profit motive, which can result in unsafe workplaces, environmental pollution, or unhealthy foods. Some industries are inherently dangerous (e.g. tobacco), while others can have decidedly harmful effects (e.g. alcoholic beverages). In each instance, individuals or organizations act in their own interests, but their actions may adversely affect communal health and safety. In the absence of governmental authority and willingness to coerce, such threats to the public's health could not easily be reduced.

Although regulation in the name of public health is intended to safeguard the health and safety of populations, it often benefits those most at risk of injury and disease. Everyone gains value from public health regulations, such as food and water standards, but some regulations protect the most vulnerable. For instance, eliminating a toxic waste site, enforcing a building code in a crowded tenement, or closing an unhygienic restaurant holds particular significance for those at immediate risk. Frequently, those at

increased risk are particularly vulnerable due to their race, gender, or socioeconomic status.

Individuals and businesses often oppose government regulation (Reeve and Gostin 2019). Resistance is sometimes based on philosophical grounds of autonomy, choice, or freedom from government interference. Citizens, and the groups that represent them, claim that regulating self-regarding behaviour, such as the use of seat belts or motorcycle helmets, is not the business of government. Or they harbour suspicions about government's intent, such as fluoridation of drinking water or compelled vaccination. Public health agencies sometimes have to defend charges of paternalism, or the 'Nanny State', such as with Mayor Bloomberg's decision to limit portion size of sugary drinks in New York City (Gostin and Gostin 2009).

Public health has historically constrained the rights of individuals and businesses to protect community interests in health. Reporting and surveillance diminish privacy; mandatory testing, vaccination, and treatment invade bodily integrity; environmental standards reduce property values; industrial regulation impedes economic freedom; and isolation and quarantine deprive individuals of liberty. In each of these cases, and more, public health has not shied from regulating individuals and businesses for the common good.

Public health powers can legitimately be used to restrict human freedoms and rights, but they must be exercised consistent with constitutional and statutory constraints on state action. Governments cannot over-reach, such as constraining liberty without sufficient evidence of public benefit. When agencies coerce, they must do so proportionately, such as by using the least restrictive measure needed to avert the health threat. And they must always act fairly, without discrimination on grounds such as race, gender, or disability. Achieving a just balance between the powers and duties of the state to advance the public's health and constitutionally protected rights poses an enduring challenge for public health law.

Social justice

Social justice is viewed as so central to the mission of public health that it has been described as the field's core value: 'The historic dream of public health … is a dream of social justice' (Beauchamp 1999, p. 105). Among

the most basic and commonly understood meanings of justice is fair, equitable, and appropriate treatment in light of what is due or owed to individuals and groups (Rawls 1971). Yet, while many global health indicators have gone up, the 'equity gap' between higher and lower income groups has remained stagnant within and among nations. Brazil, for example, has considerable poverty but yet has enormous income held by the most wealthy.

Social justice captures the twin moral impulses that animate public health: to advance human well-being by improving health and to do so particularly by focusing on the needs of the most disadvantaged (Gostin and Powers 2006). This account of justice has the aim of bringing about the human good of health for all members of the population. An integral part of that aim is the task of identifying and ameliorating patterns of systematic disadvantage that profoundly and pervasively undermine prospects for well-being of oppressed and subordinated groups—people whose prospects for good health are so limited that their life choices are not even remotely like those of others (Powers and Faden 2006). These two aspects of justice—health improvement for the population and fair treatment of the disadvantaged—create a richer understanding of public health law and ethics. Seen through the lens of social justice, the central mission of the public health system is to engage in systematic action to assure the conditions for improved health for all members of the population, and to redress persistent patterns of systematic disadvantage.

Imagine a government policy designed to reduce tobacco consumption in the population. If the government were to lower the smoking prevalence from, say, 30% to 15%, would that be acceptable if marginalized groups such as institutionalized populations (e.g. prisoners and persons with mental illness) continue to have high smoking prevalence? What if the government explicitly targets smoking prevention messages to the well-off but not in poor neighbourhoods? Is that acceptable as a matter of social justice?

These are the quintessential values of public health law—government power and duty, coercion and limits on state power, and the value of social justice. To achieve the goals of population health under the rule of law requires sound legal foundations. Statutes and regulations—as interpreted by the judicial system—establish the infrastructure for public health agencies ranging from their mission, functions, and powers, to their organization and funding.

Public health statutes: legal foundations of public health agencies

The field of public health is typically regarded as a scientific endeavour, and undoubtedly our understanding of the aetiology and response to disease is heavily influenced by technical inquiry. Less well understood is the role of law in public health. Law defines the jurisdiction of public health officials and specifies the manner in which they may exercise their authority. Public health statutes create public health agencies, designate their mission and core functions, appropriate their funds, grant their power, and limit their actions to protect a sphere of freedom. They authorize the collection of health information and enable monitoring and regulation of dangerous activities. The most important public health debates take place in legal forums—legislatures, courts, and administrative agencies—and in the law's language of rights, duties, and justice (Grad 2004).

The quality and modernity of public health laws around the world are highly variable (Institute of Medicine 2011; Gostin et al. 2019). Some states have relatively recent and comprehensive public health laws; others have older statutes, while still others have virtually no public health laws. Overall, public health statutes often have a number of deficiencies. They may be quite old, reflecting an era when infectious and sexually transmitted infections were the major health hazards. They can be highly fragmented, built up in layers over time, for example, beginning with tuberculosis and sexually transmitted infections (e.g. syphilis and gonorrhoea), through to polio and then AIDS, and more recently traffic safety and diet and physical activity. These laws may not afford the state sufficient authority to protect the public's health and safety and they may fail to adequately protect individual rights and freedoms. For example, many older statutes do not ensure natural justice or procedural due process before the exercise of compulsory powers.

Given the importance of sound, modern public health laws, many (if not most) states should examine the corpus of their public health laws and regulations. Low- and middle-income countries may need the capacity and knowledge to reform their laws. The influential Institute of Medicine (now the National Academy of Medicine) in the United States issued a major report on public health law reform in 2011, urging states to adopt the

‘Turning Point’ Model Public Health Act (Hodge et al. 2006). That model act could provide a template for other countries. Additionally, the WHO and partners have developed a report on using law to advance the right to health (WHO 2017). The WHO report offers law reform guidance particularly in low- and middle-income countries.

The WHO Constitution requires member states to notify the organization of its major public health statutes and regulations, even though few states actually comply. The WHO historically published a *Health Law Digest*, but the organization has drastically cut back resources for the *Digest*. The Director-General’s reorganization several years ago removed ‘law’ from the title of the department charged with publishing the *Digest*. Today, the *Digest* is solely in electronic form and has lost its utility to provide a systematic account of national public health laws.

The lawmaking process: building constituencies and forming partnerships

The methods and goals of public health are often misunderstood and undervalued within government and society. Public health needs opportunities to draw attention to its resource requirements and its achievements so that it can develop constituencies for programmes. The lawmaking process provides such an opportunity. A bill is the first step towards a coalition. It is an occasion for contact with interest groups and affected communities, some of whom may be motivated to act in support of the bill. Contacts and collaborative efforts also help to establish long-term ties and identify important sources of support for other programmes. Moreover, the process of negotiating for support can be a useful and concrete way for health agencies to incorporate the views of persons who receive public health services or are subject to regulation.

Legal reform also has the potential to enhance health agencies’ relationships with the legislature. Positive lawmaking offers a different sort of contact with legislators than tends to occur in the appropriations process. Public health law reform may offer an occasion to deal with a far greater range of legislators outside the context of contentious budget discussions. The drafting, negotiating, and hearing processes provide a variety of forums for educating lawmakers and their staffs about public health needs and

methods and also provide health planners with better information about legislative views and priorities.

Law reform, of course, cannot guarantee better public health. However, by crafting a consistent and uniform approach, carefully delineating the mission and functions of public health agencies, designating a range of flexible powers, and specifying the criteria and procedures for using those powers, the law can become a catalyst, rather than an impediment, to reinvigorating the public health system.

Models of legal interventions: law as a tool for the public's health

Law can be empowering, providing innovative solutions to the most implacable health problems. Many of the great public health achievements in the twentieth century were realized, at least in part, through law reform or litigation, including vaccinations, safer workplaces, safer and healthier foods, motor vehicle safety, control of infectious diseases, tobacco control, and fluoridation of drinking water (CDC 2000).

The study of public health law, therefore, requires a detailed understanding of the various legal tools available to prevent injury and disease and to promote the health of the populace. There are a number of models for using law as a tool for the public's health and safety, including: taxation and spending, the informational environment, the built environment, the socioeconomic environment, direct regulation, indirect regulation through the tort system, and deregulation. Although in each case the law can be a powerful agent for change, the interventions can also raise critical social, ethical, or constitutional concerns (Gostin and Gostin 2009).

Model 1: the power to tax and spend

The power to tax and spend—the quintessential governmental function found everywhere—provides government with an important regulatory technique. Most governments invest in a broad array of services to promote public health ranging from surveillance to education and research. Although funding is far too limited, many governments spend to establish and maintain a public health infrastructure consisting of: a trained workforce;

electronic information and communications systems; rapid disease surveillance; laboratory capacity; and response capability. In some states, governments can attach ‘conditions’ to public spending that are intended to be health promoting. For example, in the United States, the federal government requires states to have a drinking age of 21 years as a condition of the receipt of highway funds. Yet, in 2012, the US Supreme Court on the Affordable Care Act (ACA) ruled that the federal government cannot constitutionally require states to expand Medicaid as a condition of funding. In the ACA, Congress had authorized the Department of Health and Human Services to withhold all Medicaid funding if states did not expand Medicaid. The Court ruled that was too ‘coercive’ (*National Federation of Independent Business v. Sebelius* 2012).

The power to tax is primarily designed to raise revenue, but it also can provide significant inducements to engage in beneficial behaviour and disincentives to engage in risk activities. Tax relief can be offered for health-producing activities such as healthcare services, childcare, and charitable contributions. At the same time, tax burdens can be placed on the sale of hazardous products such as cigarettes, alcoholic beverages, and firearms. The WHO Framework Convention on Tobacco Control, for example, supports high taxes on tobacco products. A tax on sugar sweetened beverages has also been shown to reduce consumption of calorie-laden beverages (Gostin 2018; Roache et al. 2016). Of course, taxation can create perverse incentives such as tax relief for the purchase of unsafe and fuel-inefficient motor vehicles. A carbon tax offers a good example of an economic vehicle that would be beneficial to the public’s health and the environment.

Model 2: the power to alter the informational environment

The public is constantly exposed to images and information that influence their life choices, and this undoubtedly affects health and behaviour. The reliability and usefulness of information is variable. Some messages are educational, such as child nutrition campaigns. Other information can be biased—either inaccurate or misleading. For example, the corporate profit motive can result in advertisement of harmful products, such as foods with excess sugar and salt. Messages such as ‘low fat’, ‘low calorie’, or ‘light’ can be confusing or deceptive.

Regulating the flow in information can be challenging. In practical terms, governments do not have the available resources to compete with large advertising budgets that market unhealthy foods, alcoholic beverages, or tobacco. In constitutional terms, there can be questions about whether advertising restrictions violate the corporation's right to free expression.

Governments have several tools at their disposal to alter the informational environment thereby encouraging people to make more healthful choices about diet, physical activity, tobacco, and alcohol. First, government, as a health educator, can use communication campaigns that warn consumers of high risks (e.g. smoking or driving while intoxicated) or encourage safer activities (e.g. wearing seat belts or motorcycle helmets). Second, government can require businesses to label their products to inform consumers: instructions on safe product use, disclosure of contents or ingredients, or warnings about hazards. Third, government can regulate advertising and promotions for potentially harmful products such as cigarettes and firearms. They can also limit marketing certain products to vulnerable populations, such as children and adolescents.

Regulating tobacco packaging and advertising offers a good illustration. The WHO's Framework Convention on Tobacco Control (WHO 2005a) requires states to adopt supply and demand reduction strategies, including advertising restrictions. Yet, the tobacco industry has challenged advertising restrictions and packet labels in countries as diverse as Australia, Uruguay, and the United States. Australia was the first country to introduce plain packaging legislation with graphic images of the harms of tobacco. Australia's highest court upheld the legislation, and the government won both in an international Investment Treaty dispute and the World Trade Organization challenge (*British American Tobacco Australia Limited v. The Commonwealth* 2012). Uruguay's tobacco control legislation was also being challenged in a separate Investment Treaty case (*FTR Holding v. Oriental Republic of Uruguay* 2010). The tribunal also ruled in favour of Uruguay. 'Big Tobacco' is claiming that the state is unfairly 'taking' their property or diminishing the value of their private investments. Although this may seem to be an outrageous claim, it is expensive for states to defend their public health laws (Roach et al. 2016).

In the United States, the Food and Drug Administration's requirement of graphic warning labels was put on hold following a court ruling. Tobacco companies are claiming that the graphic image requirement violates their

rights to free expression (*R.J. Reynolds Tobacco Company v. Food and Drug Administration* 2012). The United States is one of the world's outliers in that its highest court ardently defends the right of corporations to free expression, thus thwarting public health restraints on advertising hazardous products.

Model 3: the power to alter the built environment

The design of the 'built' or 'physical' environment can hold great potential for creating healthier and safer communities. Public health has a long history of altering the built environment to reduce injuries (e.g. workplace safety, traffic calming, consumer protection, and fire codes), infectious diseases (e.g. sanitation, nuisance abatement, and housing codes), and environmentally associated harms (e.g. lead paint and toxic emissions). Sanitary laws go back hundreds of years, designed to ensure more hygienic conditions in growing cities; occupational health and safety became prominent following the Industrial Revolution, which exposed workers to toxins and unsafe working conditions; and more modern laws provide patients and consumers with protection against unsafe products.

The epidemiological transition from infectious to chronic diseases raises new challenges in the design of neighbourhoods. For example, urban design can encourage more active lifestyles (walking lanes, bicycle paths, parks, green fields, and playgrounds); improve nutrition (zoning to limit fast food); discourage the use of hazardous products ('tobacco-free' indoor spaces); reduce violence (well-lit streets, safe storage of firearms); and increase social interactions (helping neighbours and building social capital) (Perdue et al. 2003).

Many zoning laws in the mid and late twentieth century were unfavourable to healthy lifestyles, such as those giving preference to motor vehicles over pedestrians and bicycle riders, or preferring commercial development over playgrounds. Municipalities often facilitated the expansion of unhealthy fast foods, forming 'food deserts' in poor neighbourhoods. These are political choices, and debates still rage over whether government should use health as a major factor in designing living spaces, rather than, say, economic development. What seems certain is that it is within the power of government to structure the environment in ways that are more conducive to a healthy diet and physical activity.

Model 4: the power to alter the socioeconomic environment

A strong and consistent finding of epidemiological research is that socioeconomic status (SES) is correlated with morbidity, mortality, and functioning (Marmot 2005). SES is a complex phenomenon based on income, education, and occupation. The relationship between SES and health often is referred to as a 'gradient' because of the graded and continuous nature of the association; health differences are observed well into the middle ranges of SES. These empirical findings have persisted across time and cultures and remain viable today.

Some researchers go further, concluding the overall level of economic inequality in a society correlates with (and adversely affects) population health. That is, societies with wide disparities between rich and poor tend to have worse health status than societies with smaller disparities, after controlling for per capita income. These researchers hypothesize that societies with higher degrees of inequality provide less social support and cohesion, making life more stressful and pathogenic. Drawing upon this line of argument, some ethicists contend, 'social justice is good for our health' (Daniels et al. 2000).

The WHO issued a report on social determinants of health, urging more attention to the underlying factors in influencing health outcomes (WHO 2008b). However, the WHO never put significant resources into social determinants and did not follow the Commission Report with international law or softer norms. From government's perspective, ensuring a healthy population that is fair and equitable requires an 'all of government', indeed 'all of society' approach that transcends the health sector.

Although social determinants powerfully affect health, they can require redistribution of wealth and hence are often hotly contested. Opponents of redistributive policies argue such policies punish personal accomplishment, thereby discouraging economic growth. Still, it is impossible to make significant inroads into health disparities without careful attention to the deeper social and economic determinants of health.

Model 5: direct regulation of persons, professionals, and businesses

Government has the power to directly regulate individuals, professionals, and businesses. In a well-regulated society, public health authorities set clear, enforceable rules to protect the health and safety of workers, consumers, and the population at large.

Regulation has long been a staple of injury control. In most Western countries, for example, the rate of motor vehicle-related injuries and deaths has plummeted due to law reform. Governments required individuals to engage in safer behaviour, such as by using seat belts and motorcycle helmets. The law also compelled manufacturers to design and sell safer vehicles, such as cars with airbags and with the ability to better withstand crashes. The law also set standards for governmental and commercial development such as safer roads and traffic calming. Yet, with the explosion of motorized vehicles in newly emerging states, the injury and death rates from motor vehicles are alarming. The streets of Bangkok, Delhi, and Hanoi (among many others) are teeming with unsafe vehicles, chaotic roads, and motorcycle drivers and passengers without helmets.

Perhaps the most ancient use of direct public health regulation is in the control of infectious diseases. Governments often mandate vaccination, screening, partner notification, directly observed therapy, and isolation or quarantine. Originating from the ancient scourges such as leprosy, malaria, and tuberculosis, through to syphilis and herpes, and to the modern AIDS pandemic, states—for good or bad—have enacted civil and criminal laws to deter or punish perceived risk behaviours. Early in the AIDS pandemic, for example, many states enacted laws to require testing (e.g. prisoners), travel and immigration restrictions, and even quarantine (e.g. Cuba). Criminal laws were altered to punish behaviours conducive to transmitting HIV, and to targeting risk populations such as injecting drug users and commercial sex workers.

Direct regulation also plays a major role in ensuring that professionals and businesses operate safely and with the requisite level of skill and knowledge. Licences and permits enable government to monitor and control the standards and practices of healthcare professionals, hospitals, and nursing homes. By licensing and accrediting health professionals, public health agencies can help ensure safe, competent, and well-qualified practitioners. Finally, inspection and regulation of businesses helps to assure humane conditions at work, reduced toxic emissions, and improved consumer product safety.

The control of chronic diseases is becoming the new battle-ground over the wisdom of direct regulation. Some jurisdictions have banned harmful ingredients such as trans fatty acids; others have reduced harmful additives, such as excess sugar or sodium; while still others have regulated portion size, such as the size of a sugary soft drink. Recent research suggests, 'sugar is toxic', which may require public health agencies to set or recommend safer levels of sugar intake (Basu et al. 2013).

When regulating what people eat or their physical activity, governments often face the stinging criticism that they are acting paternalistically, with the charge of running a 'Nanny State'. To public health advocates, these are all necessary measures to make health the 'easier' or 'default' choice. But to many in society, it restricts their freedoms to make their own decisions about their health and well-being (Gostin and Gostin 2009).

Model 6: indirect regulation through the tort system

Ministries of justice and health, as well as private citizens, possess a powerful means of indirect regulation through the tort system. Civil litigation can redress many different kinds of public health harms: environmental damage (e.g. air pollution or groundwater contamination), exposure to toxic substances (e.g. pesticides, radiation, or chemicals), hazardous products (e.g. tobacco or firearms), and defective consumer products (e.g. children's toys, recreational equipment, or household goods).

The most prominent illustration of using judicial remedies for a public health purpose is tobacco litigation. Tort litigation against the tobacco industry began in the United States. This litigation had multiple benefits. It resulted in a 'Master Agreement' between states and tobacco companies, with large sums given to states. It uncovered highly damaging evidence about the deceptive behaviour of the industry, such as raising the nicotine level of cigarettes, hiding information about the hazards of smoking, and targeting young girls, racial minorities, and minors with cigarette advertisements and promotions.

More modern litigation has been waged under the constitutional right to health in places such as India and Latin America. Public health advocates in these and other countries have sued tobacco companies for knowingly misleading the public. They have also sued their own governments to force

them to adopt more rigorous tobacco control measures, such as clean air acts (Cabrera and Gostin 2011).

Perhaps it would be predictable for Big Tobacco to use the law as a sword against government regulation. Increasingly, the industry is using the constitutional right to property or commercial speech, bilateral investment treaties to safeguard their investments, or World Trade Organization treaties to stall or block tobacco control (Mitchell and Studdert 2012).

Using tobacco litigation as a model, public health advocates have sought to use tort law to attack the deceptive and harmful practices of other powerful industries. This litigation ranges from lawsuits against the food industry (e.g. marketing unhealthy products to minors), the firearms industry (e.g. trafficking in dangerous weapons), and the alcohol industry (e.g. glamorizing underage or excessive drinking). Thus far, most of these lawsuits have not been successful. One difference is that food, firearms, and alcoholic beverages when used responsibly are not harmful, while tobacco use is always hazardous.

To what extent should the harmful activities of these and other industries be immune from tort litigation on the theory that autonomous consumers are free to make their own decisions—whether healthful or harmful? Given that individuals are embedded in communities and societies that profoundly affect their personal choices, laws should enable consumers to litigate if the private sector has exposed them to unnecessary and serious risks.

Model 7: deregulation: law as a barrier to health

Sometimes laws are harmful to the public's health and stand as an obstacle to effective action. In such cases, the best remedy is deregulation. Politicians may urge superficially popular policies that have unintended health consequences. Consider laws that prohibit or penalize needle exchange programmes and pharmacy sales of sterile syringes. Even though access to sterile injection equipment significantly reduces the risk of needle-borne infections (e.g. HIV and hepatitis B or C) and does not encourage illicit drug use, many governments discourage or ban the distribution of these products, including by public health agencies.

A further illustration is the decision to close bathhouses, making it more difficult to reach gay men with condoms and safe sex literature. Rather than

banning these activities, government could use them as venues for education and harm reduction.

Finally, but perhaps even more importantly, governments throughout the world criminalize high-risk activities. They punish sex workers, drug users, and even those who engage in same-sex sexual activities. Criminal laws that punish activities that risk transmission of HIV, for example, do not appear to reduce unsafe behaviour, but rather drive the epidemic underground.

Criminal law is usually a blunt, ineffective, and unfair tool when used for an allegedly public health purpose. Deregulation, however, can be controversial because it often involves a direct conflict between public health and other values that society may favour, such as crime prevention or morality. Drug laws, the closure of bathhouses, and HIV-specific criminal penalties represent society's disapproval of disfavoured behaviours. Deregulation becomes a symbol of weakness that is often politically unpopular. Public health advocates may believe passionately in harm-reduction strategies, but the political community may want to use the law to demonstrate social disapproval of certain activities such as illicit drug use, commercial sex, or unprotected sex.

The government, then, has many legal 'levers' designed to prevent injury and disease and promote the public's health—taxing and spending, the informational environment, the built environment, the socioeconomic environment, direct regulation, indirect regulation, and deregulation. Legal interventions can be highly effective and need to be part of the public health officer's arsenal. However, legal interventions can also be controversial, raising vital ethical, social, constitutional, and political concerns.

Enforcing public health laws

Law on the books that promotes the public's health is important, but it has little utility if it is not well implemented and enforced. Enforcement is a problem in all countries, but particularly many low- and middle-income countries (LMICs). Enforcement in LMICs can be impeded by many forces, such as lack of political will, corruption, and lack of legal capacities. All governments should make implementation of public health laws a high priority because enforcement is needed to secure the common good. This will require tackling corruption and funding capacity building. The South

African Constitutional Court required an anticorruption mechanism that was badly needed for holding public officials accountable. Legal capacities include well trained legislators, judges, public health officials, inspectors, and others who can ensure that laws are implemented rigorously but fairly.

Case studies of four major health hazards

The previous sections often used major health hazards as illustrations of the use of public health law. This chapter concludes with four brief case studies: infectious diseases, NCDs, unintentional injuries, and violence to self or others. This final section shows how the intersection between public health law at the local, national, and international levels can influence health outcomes.

Infectious disease law

As mentioned earlier, infectious disease law has ancient origins (Markel 2005). Epidemic disease has historically had profound implications, not only for animal and human health, but also for security, trade, tourism, and social relationships. Infectious diseases often bring out the worst in people and societies, provoking stigma and discrimination against disfavoured groups such as immigrants and ethnic or religious minorities. Rapidly emerging diseases (such as novel influenzas) often provoke unjustified trade or travel restrictions. At the same time, governments are likely to spend disproportionate public health resources on highly transmissible or novel infections—disproportionate at least insofar as infectious pathogens receive so much more attention than equally destructive health hazards, such as NCDs, injuries, and mental illness.

This dynamic can be seen in the emphasis in domestic and international laws on infectious diseases. The oldest international health law—dating from mid-nineteenth-century sanitary conferences in Europe—is the International Health Regulations, which are devoted almost exclusively to health security. The first time the International Health Regulations (WHO 2005b) was used was in relation to the influenza A (H1N1) pandemic. In total, WHO has declared the following emergencies under the IHR: swine flu (2009), polio (2014), Ebola (2014), and Zika (2016). As of June 2019,

WHO has not declared an emergency in response to the Ebola epidemic in the Democratic Republic of Congo.

The emphasis on epidemic diseases can also be seen in domestic law, where most public health statutes are devoted to infectious disease powers, such as testing, screening, partner notification, directly observed therapy, and isolation or quarantine. During the SARS outbreaks, states—particularly in North America and Asia—exercised public health powers extensively including travel restrictions, trade barriers, border security, and quarantine (Fidler 2004). When the Republic of Korea experienced a major Middle Eastern Respiratory Disease outbreak in 2015, it also exercised extensive powers.

Even in the realm of international relations, infectious diseases have dominated funding and political attention. The AIDS pandemic fundamentally changed the political and funding calculus for global health. The United States launched the greatest effort by a single country against a single disease—PEPFAR. The United Nations (UN) formed a new agency to coordinate work on AIDS—UNAIDS. And powerful new public/private partnerships emerged to provide unprecedented funding for work on AIDS, tuberculosis, and malaria, such as the Global Fund and UNITAID, while the GAVI Alliance supports another infectious disease initiative—immunizations. Most impressively, the international community has vastly ramped up research, funding, and access to antiretroviral treatment for HIV/AIDS. AIDS law at the national level can be highly impactful. Laws, from Africa and Asia to Europe and the Americas can ensure access to treatment, while combatting non-discrimination.

Following the West Africa Ebola epidemic in 2013–16, the WHO fundamentally reformed it health emergency capacities, including a new programme and an emergency contingency fund (Gostin et al. 2016).

Non-communicable disease law

Four NCDs—cardiovascular disease, cancer, respiratory disease, and diabetes—account for 63% of global deaths annually. The conventional wisdom, which conceives of NCDs as a ‘First World’ problem, is starkly belied by data: 80% of the 35 million people who die annually from NCDs live in LMICs. The death toll is projected to rise by 17% over the next decade, unless meaningful steps are taken urgently. With the exception of

sub-Saharan Africa, NCD mortality now exceeds that of infectious, maternal, perinatal, and nutritional conditions combined (WHO 2011).

The suffering and early deaths from NCDs, and the disproportionate burden in lower-income countries, will only grow. And they will interact with existing health challenges in poorer countries, creating the ‘double burden’ of hunger alongside obesity. Although longer lifespans partly explain the increased burden of NCDs, human behaviour plays a central role: unhealthy diets, sedentary lifestyles, smoking, and excessive alcohol consumption account for more than two-thirds of all new NCD cases.

The moral tragedy lies in the fact that much of this suffering and early death is preventable, and many of the solutions are to be found in law—at both the domestic and international levels. Local and national governments have all the public health law tools at their disposal in fighting the NCD epidemic. They can tax unhealthy products such as tobacco, alcoholic beverages, sugary drinks, and sodium-laden processed foods. Equally, rich states could cease subsidizing unhealthy foods, such as high-fructose corn syrup, cane sugar, and even red meat and dairy, while subsidizing healthy fruits and vegetables.

Governments can also alter the informational environment, counteracting incessant commercial messages—for example by package labelling, health warnings on cigarettes and alcohol, and comprehensible nutritional information on food packages and restaurant menus. The United Kingdom Food Standards Agency developed a voluntary system ‘traffic light’ system that is visible and simple to follow. By putting a series of bright red lights on food packages with excessive saturated or trans fats, sodium, sugar, and calories, it incentivizes food companies to be more responsible, while clearly informing consumers. The state can restrict advertising unhealthy products, particularly when aimed at minors (Hawkes 2007).

Individual choices are not made in isolation but reflect the environment in which people live. Government could use its powers to plan cities and zones to facilitate healthy lifestyles, including supermarkets and farmers’ markets. It can build bike and walking paths, designate pedestrian areas, subsidize mass transit, and discourage motor vehicle use—such as the ‘congestion tax’ instituted by the city of London in the United Kingdom. Government could require ‘health impact assessments’ as a condition of new government or commercial development.

Although incentives and voluntary approaches are often more politically palatable, there is also a role for direct regulation. Some products, such as trans fatty acids, are so hazardous to health that they should be removed from the market, as Denmark has done (see Table 3.4.1).

Table 3.4.1 Domestic strategies to address NCD risk factors

Domain	Goal	Intervention	Example
Healthy lifestyles	Optimal nutrition	Agricultural production Food manufacturing, processing, and distribution	Incentives to produce healthier foods Regulation of food industry and retailers
		Disincentives for buying and selling unhealthy food	Fat tax
		Marketing unhealthy foods	Restricting advertisements targeting children
		Nutritional information disclosure and education	Providing government nutritional guidelines; improving package labelling; menu labelling
		Direct regulation	Banning unhealthy ingredients (such as trans fats)
		Public–private partnership	Voluntary targets for nutritional values
	Physical activity	Incentives for exercise	Subsidies for physical exercise and organized activities
		Flexible spending accounts	Corporate allowances for fitness activities
		Monitor community physical activity	Public health surveillance of monitoring prevalence of NCDs
Healthy places	Places to buy and eat Food	Access to affordable, healthy foods	Farmers' markets; zoning of unhealthy fast foods
		Healthy foods in schools, workplaces, etc.	Vending machines, snacks, meals
		School/childcare curricula and programmes	Body mass index (BMI) surveillance; nutrition education
	Places to walk, recreate, and play	Workplace/education settings and programmes Urban land use and planning	Organized activities; fitness memberships Attractive, accessible, safe public places and paths
		Tax incentives for building design	Businesses incorporating building designs that promote physical activity
		Transportation	Mass transit and safe routes

Domain	Goal	Intervention	Example
		Health impact assessments	Consultation with affected communities; public health evaluations
Healthy societies	Social justice	Antidiscrimination laws	Proscribe medical condition-related discrimination
	Access to services	Services to support life functions of ill individuals	At-home support services
		Encourage prevention and treatment	Metabolic screening; counselling
		Monitoring disease in the community	Surveillance through BMI reporting

It is equally important to marshal a transnational response. The impact of globalization—globalized food and tobacco markets and the homogenization of cultures and behaviours—means that no country can effectively tackle NCDs without the support and cooperation of the international community. Global governance for health requires a multifaceted approach (Gostin 2014). The international community could set norms for healthy activities, such as the Framework Convention on Tobacco Control (WHO 2005a). The WHO has also developed 'global strategies' covering: NCD prevention and control (2000); diet, physical activity, and health (2008); and the harmful use of alcohol (2010). In 2008, the Assembly adopted the 2008–2013 Action Plan for its Global Strategy on NCDs, and in 2010, it adopted recommendations on the marketing of foods and non-alcoholic beverages to children (WHO 2008a, 2010).

The decade of WHO governance on NCDs failed to raise NCDs to the same high profile and impact as HIV/AIDS. But the UN General Assembly (2011) hosted a high-level meeting on NCDs in September 2011, adopting a *Political Declaration on the Prevention and Control of Non-Communicable Diseases*—a 65-point document, cataloguing many of the global challenges, and calling for a whole-of-government/whole-of-society response. Broadly speaking, the Declaration recommends preventative measures, health system strengthening, international cooperation, research and development, and monitoring and evaluation of progress. Sadly, however, the UN failed

to set clear targets, a tangible road-map to meet those targets, and the resources needed to tackle the global NCD pandemic.

Just as domestic governments need to adopt an 'all-of-government/all-of-society' approach, so too must global institutions look beyond WHO to a more expansive view of governance. In short, there is a need for 'total global engagement', encompassing trade, investment, development, intellectual property, and human rights—all international law regimes in their own right. The private sector is a vital part of the solution, but the major companies operate all over the world, so corporate governance must also operate at the global, as well as the national, level.

Injury control law

Injuries are often portrayed as 'accidents' or 'random events' that are unpredictable and therefore unpreventable, as opposed to a serious public health concern, which explains the field's historical neglect. However, injuries are a leading cause of death and disability worldwide, resulting in nearly 6 million deaths annually and representing 17% of the global burden of disease; by 2020 the global burden of injuries is projected to increase to 20% (WHO 2011). The injury epidemic has been compounded by rapid unplanned urbanization, increased motorization, and dramatic changes in the built environment in LMICs (Lozano et al. 2012; Murray et al. 2012; Salomon et al. 2012).

There are marked inequalities in the global distribution of injuries, with >90% of all deaths and disability-adjusted life years occurring in LMICs, mostly among young people. The economic burdens for those countries are staggering: traffic incidents alone cost $65 billion or 1–1.5% of GDP annually (Norton et al. 2006).

The WHO differentiates between unintentional injuries (no predetermined intent such as road traffic incidents, poisonings, falls, fires, and drowning) and intentional injuries (self-directed such as suicide or self-mutilation, interpersonal such as domestic violence or murder, or collective violence such as war-related injuries). Law reform can help prevent both of these.

The UN Convention on the Rights of the Child (UN 1989) guarantees all children the right to a safe environment and to protection from injury and violence. Yet, injuries are the single greatest cause of death and disability

among children in LMICs. Drowning is a leading cause of morbidity and premature mortality in children under 5 years of age. With increasing age, mortality resulting from road traffic crashes surpasses that from drowning, with traffic deaths being the greatest overall killer of children aged 15–19.

Men often suffer injuries in the course of paid labour or occupational pursuits, while women suffer burns as a result of their customary domestic role—in the kitchen, from fires or scalding water, and as a result of cooking on open fires. The home use of kerosene or paraffin lamps, which are easily knocked over and ignited, is also a large contributor. Above all, underlying gross poverty heavily influences many of these injuries, irrespective of age and gender.

National public health laws are perhaps the single most potent interventions to prevent injuries. Developed countries have used a multitude of regulations to vastly curtail traffic injuries and deaths—e.g. driving licences, speed limits, vehicle safety, motorcycle helmets, and road design. For example, when Viet Nam passed and enforced a law requiring helmet use for motorcycles, the number of head injuries declined dramatically (WHO et al. 2017). Yet, in many LMICs, the roads are chaotic and very dangerous, with vehicles unfit for travel and roads poorly planned. The mix of pedestrians, bicycles, and motorized vehicles on unsafe roads that do not separate the types of traffic is lethal. To further compound the problem, roadways are usually designed with the convenience of the drivers in mind and little thought for the safety of pedestrians. In Nairobi, residents of Kibera, one of the largest urban slums in the world, suffer inordinate road traffic injuries and deaths, as they have to run across dangerous roads to get to work, school, or shops to buy food and clothing.

Work-related injuries are ubiquitous in low-resourced countries. The rise in industrial development and the lack of attention to safety will ensure that occupational injuries double by 2025. Again, this situation is very different in the developed world, which has steadily decreased occupational injury rates over the years. Most industrial work in low-resourced countries such as agriculture, manufacturing, and mining is already associated with high rates of injury from electrical, mechanical, and physical hazards. Now, the steady expansion of chemical and biotechnology industries has introduced new injury risks.

Yet, occupational health and safety laws cover only 10% of the population in low-resourced countries. Without adequate regulations,

businesses do not invest in quality equipment or train workers to ensure safety. Companies often expect workers to take inordinate risks because labour is cheap and plentiful. If developing countries do not soon implement effective safety standards, the health costs to workers in high-risk jobs will outweigh the financial benefits of industrial development.

Intentional injuries are also a vast and growing problem, particularly in LMICs. Rates of robberies, assaults, rapes, and homicides can be devastating in major cities such as Johannesburg. India has experienced an epidemic of gender-related violence, which also occurs in many other parts of the world. In some African countries, gay men are subjected to brutal assaults and murders. Yet, it is possible to prevent most injuries through technological innovation, safety regulation, and law enforcement. However, the same ingenuity and resources that make the home, workplace, and the streets safer in richer countries have not been applied in low-resourced countries. The WHO devotes <1 per cent of its budget to injuries and violence.

Developed countries, of course, are not immune from violence, which could be dramatically lowered with sound firearm regulation and enforcement. In the United States, for example, firearm-related deaths are roughly equal to road traffic deaths, with few checks on gun buyers or the lethality of weapons and bullets.

If the public health community wishes to close the ‘injury gap’ between rich and poor, it will have to devise evidence-based solutions, demonstrate their effectiveness, engage the private sector, gain the cooperation of police and the courts, devote the resources, and advocate for strong occupational health and safety laws throughout the developed world.

Conclusion: the future of public health law reform

This chapter has sought to demonstrate the powerful potential of law to prevent injuries and disease. Law can build the infrastructure of public health agencies, ensuring a broad mission and adequate powers. It can protect individuals from excessive or unjustified coercion or discrimination. It can promote social justice by narrowing health disparities. At the same time, government has multiple legal tools available to improve the public’s health and safety, ranging from its economic power to tax and spend, to its

authority to alter the environments in which people live (e.g. informational, built, and socioeconomic), through to direct and indirect regulation, including, where necessary, deregulation.

With the potential for rapid dissemination of pathogens, products, services, and marketing throughout the world, the need for global governance cannot be overstated. No country acting alone can stem the tide of injuries and disease. Global health leadership could establish health-promoting norms, set benchmarks, monitor progress, and encourage compliance by state and non-state actors. Global leadership, moreover, must act beyond the health sector, engaging legal regimes such as trade, intellectual property, food, migration, and the environment.

When law is used in a systematic and coordinated way at the local, national, and international levels, it can go a long way towards preventing infectious diseases, NCDs, and injuries (both unintentional and violence-related). For public health practitioners, the lesson is to effectively use the battery of legal tools at their disposal. And when those tools are inadequate, public health agencies must work with civil society and communities to raise the level of political awareness. In a complex and globalized world, public health is a deeply legal and political pursuit—well beyond purely science and technology.

REFERENCES

Basu, S., Yoffe, P., Hills, N., and Lustig, R.H. (2013). The relationship of sugar to population-level diabetes prevalence: an econometric analysis of repeated cross-sectional data. *PLoS One*, 8(2), e57873.

Beauchamp, D.E. (1999). Public health as social justice. In: Beauchamp, D.E. and Steinbock, B. (eds.) *New Ethics for the Public's Health*. New York: Oxford University Press, pp. 105–14.

British American Tobacco Australia Limited v. The Commonwealth (2012). HCA 43.

Cabrera, O.A. and Gostin, L.O. (2011). Human rights and the Framework Convention on Tobacco Control: mutually reinforcing systems. *International Journal of Law in Context*, 7(3), 285–303.

Centers for Disease Control and Prevention (CDC) (2000). *Ten Great Public Health Achievements in the 20th Century.* [Online] Available at: http://www.cdc.gov/about/history/tengpha.htm

Daniels, N., Kennedy, B., and Kawachi, I. (2000). Justice is good for our health. *Boston Review*, **25**, 6–15.

Fidler, D.P. (2004). *SARS, Governance and the Globalization of Disease*. Houndmills: Palgrave Macmillan.

FTR Holding v. Oriental Republic of Uruguay (2010). ICSID, Request for Arbitration, 3–5, 77.

Friedman, E.A. and Gostin, L.O. (2012). Pillars for progress on the right to health: harnessing the potential of human rights through a framework convention on global health. *Health and Human Rights: An International Journal*, 14(1), 1–16.

Gostin, L.O., et al. (2019). The Lancet Commission on global health and the law: the legal determinants of health: harnessing the power of law for global health and sustainable development. *The Lancet*, 393(10183), 1857–910.
Gostin, L.O., Abubakar, I., Guerra, R., Rashid, S.F., Friedman, E.A., and Jakab, Z. (2019). WHO takes action to promote the health of refugees and migrants. *The Lancet*, 393(10183), 2016–18.
Gostin, L.O., et al. (2016). Toward a common secure future: four global commissions in the wake of Ebola. *PLoS Medicine*, 13(5), 1–15.
Gostin L.O. (2018). Tackling obesity and disease: the culprit is sugar; the response is legal regulation. *Hastings Centre Report*, 48(1), 5–7.
Gostin, L.O. and Wiley LF (2016). *Public Health Law: Power, Duty, Restraint* (3rd ed.). Berkeley, CA: University of California Press.
Gostin, L.O. (2014). *Global Health Law.* Cambridge, MA: Harvard University Press.
Gostin, L.O. and Gostin, G.G. (2009). A broader liberty: JS Mill, paternalism, and the public's health. *Public Health*, **123**, 214–22.
Gostin, L.O. and Powers, M. (2006). What does justice require for the public's health? Public health ethics and policy imperatives of social justice. *Health Affairs*, **25**, 1053–60.
Grad, F.P. (2004). *The Public Health Law Manual* (3rd ed.). Washington, DC: American Public Health Association.
Hawkes, C. (2007). Regulating food marketing to young people worldwide: trends and policy drivers. *American Journal of Public Health*, **97**, 1962–73.
Hodge, Jr. J.G., Gostin, L.O., Gebbie, K., and Erickson, D.L. (2006). Transforming public health law: the Turning Point Model State Public Health Act. *Journal of Law, Medicine and Ethics*, 33(4).
Institute of Medicine (2011). *For the Public's Health: Revitalizing Law and Policy to Meet New Challenges.* Washington, DC: National Academy Press.
Lozano, R., et al. (2012). Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet*, **380**, 2095–128.
Markel, H. (2005). *When Germs Travel: Six Major Epidemics that Have Invaded America and the Fears They Have Unleashed.* New York: Vintage.
Marmot, M. (2005). Social determinants of health inequalities. *Lancet*, **365**, 1099–104.
Mitchell, A.D. and Studdert, D.M. (2012). Plain packaging of tobacco products in Australia: a novel regulation faces legal challenge. *JAMA*, **307**, 261–2.
Murray, C., et al. (2012). Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet*, **380**, 2197–223.
National Federation of Independent Business v. Sebelius, 567 U.S. 519 (2012).
Norton, R., Hyder, A.A., Bishai, D., and Peden, M. (2006). Unintentional injuries. In: Jamison, D.T., et al. (eds.) *Disease Control Priorities in Developing Countries*. New York: Oxford University Press, pp. 737–54.
Powers, M. and Faden, R. (2006). *Social Justice: The Moral Foundations of Public Health and Health Policy*. New York: Oxford University Press.
Perdue, W.C., Stone, L.A., and Gostin, L.O. (2003). The built environment and its relationship to the public's health: the legal framework. *American Journal of Public Health*, **93**, 1390–4.
Rawls, J. (1971). *A Theory of Justice*. Cambridge, MA: Harvard University Press.
Reeve, B. and Gostin, L.O. (2019). 'Big' food, tobacco, and alcohol: reducing industry influence on noncommunicable disease prevention laws and policies. *International Journal of Health Policy Management*, 8(7), 450–4.
R.J. Reynolds Tobacco Company v. Food and Drug Administration (2012). US Court of Appeals for the District of Columbia. Available at:

http://www.cadc.uscourts.gov/internet/opinions.nsf/4C0311C78EB11C5785257A64004EBFB5/$file/11-5332-1391191.pdf
Roach, S.A., Gostin, L.O., and Bianco Fonsalia, E. (2016). Trade, investment, and tobacco: Philip Morris v Uruguay. *The Journal of the American Medical Association*, 316(20), 2085–6.
Salomon, J.A., et al. (2012). Healthy life expectancy for 187 countries, 1990—2010: a systematic analysis for the Global Burden Disease Study 2010. *Lancet*, **380**, 2144–62.
United Nations (UN) (1989). *Convention on the Rights of the Child*. UN GA resolution 44/25, UN Doc. A/RES/44/25, 20 November 1989. New York: United Nations.
United Nations (UN) (2011). *Political Declaration of the High-Level Meeting of the General Assembly on the Prevention and Control of Non-Communicable Diseases*. New York: United Nations.
Walzer, M. (1983). *Spheres of Justice: A Defense of Pluralism and Equality*. New York: Basic Books.
World Health Organization (WHO) (2005a). *Framework Convention on Tobacco Control.* Geneva: World Health Organization.
World Health Organization (WHO) (2005b). *International Health Regulations*. Geneva: World Health Organization.
World Health Organization (WHO) (2008a). *2008–2013 Action Plan for the Global Strategy for the Prevention and Control of Noncommunicable Diseases.* Geneva: World Health Organization.
World Health Organization (WHO) (2008b). *Commission on Social Determinants of Health. Closing the Gap in a Generation: Health Equity Through Action on the Social Determinants of Health.* Geneva: World Health Organization.
World Health Organization (WHO) (2010). *Set of Recommendations on the Marketing of Foods and Non-Alcoholic Beverages to Children*. Geneva: World Health Organization.
World Health Organization (WHO) (2011). *Global Status Report on Noncommunicable Diseases 2010.* Geneva: World Health Organization.
World Health Organization (WHO), International Development Law Organization, O'Neill Institute for National and Global health Law, University of Sydney (2017). Advancing the Right to Health: The Vital Role of Law. Geneva: World Health Organization.

3.5

Health policy in developing countries

Thein T. Htay, Yu Mon Saw, James Levinson, S.M. Kadri, Ailbhe Helen Brady, Cecilia S. Acquin, and Aung Soe Htet

Introduction

A strong health policy is an essential platform for the formation of a comprehensive health system in developing countries. Three essential purposes of such a policy are (a) assuring adequate and accessible health services; (b) encouraging responsiveness by the country's population; and (c) protecting the population against health-related catastrophes (WHO 2000). Such a health policy, in turn, is dependent on responsible governance and adequate financial resources.

International health policies focused primarily on communicable diseases (HIV, Tuberculosis and Malaria), non-communicable diseases (NCDs), and maternal and child health problems. In addition, attention is given to malnutrition, problems of drug resistance, mental illness, and chronic NCDs.

Invariably, economic, social inequalities, and often cultural differences existing in developing countries result in poor health status among the disadvantaged, but also, increasingly, chronic NCDs including heart disease, hypertension, stroke, and diabetes along with a growing epidemic of obesity, these problems not limited to industrialized countries.

An increasing number of countries have undertaken health sector reform involving an integrated stewardship process based on a decentralization of healthcare services and the inclusion of social insurance or protection. Accordingly, such stewardship and high standards of governance are essential to effective health policies.

Such policies and the programmes emanating from them require rigorous monitoring and evaluation. China has been successful in decentralizing responsibility and services to local government entities, yet these local governments have generally failed in their monitoring and evaluation

responsibilities. By contrast, Mexico, which also has implemented health sector reform, has been successful in integrating monitoring and evaluation mechanisms into these systems.

The stages of health transition

With decreasing infant and child mortality and improved disease control in many countries, the proportion of elderly persons in these countries has increased (UNFPA 2017). Changes also have been observed in the patterns and causes of death in many countries, with the burden of NCD increasing (WHO 2017). These have been accompanied by the 'nutrition transition' now taking place both in industrialized and developing countries, with countries as diverse as India, China, Brazil, and Russia facing both under- and overnutrition problems (Kapoor and Anand 2002; de Neuezer et al 2001). As a result, health policies and programmes are evolving to tackle these emerging health needs (Martens 2002).

The health transition includes what have been called demographic and epidemiologic transitions (Omran 1998). (The ecological and economic transitions which often accompany health transitions are not included in this article.)

The demographic transition describes changes over time in fertility and mortality rates (Notestein 1945; Demeny 1968; Borgerhoff 1998).

The epidemiologic transition describes the changes in mortality and morbidity patterns (moving increasingly from infectious to chronic NCDs) relative to socio-demographic and economic changes (Omran 1971). Epidemiologic transition theory includes three models: (a) the 'western or classical model' (e.g. Western Europe and Northern America), (b) the 'accelerated model' (e.g. Japan and Eastern Europe), and (c) the 'delayed or contemporary model' (e.g. most developing countries) (Omran 1998).

Overall, the health transition, internationally, has evolved through the following three stages:

Stage I, 'The Age of Pestilence and Famine' was characterized by high and unstable mortality that impedes population growth. During this stage, average life expectancy at birth was low and inconstant, ranging from 20 to 40 years (Omran 1971). The stage was characterized by a high prevalence of infectious diseases and high rates of undernutrition. Women of childbearing age were prone to pregnancy-related illnesses and more vulnerable to

tuberculosis. Epidemics such as plague, cholera, and smallpox were often managed by isolation and quarantine measures as experienced in the late seventeenth century in several European cities (Goudsblom 1986), and these measures also became common in addressing contagious diseases, with rules adopted for trade quarantine and isolation of the sick in US port cities in the eighteenth century. A law for isolation of smallpox patients and for ship quarantine was enforced beginning in 1701 in Massachusetts (Hanlon and Pickett 1984). The recording and reporting of deaths by the plague and other diseases were established in the late seventeenth century (Goudsblom 1986).

With increased urbanization in the nineteenth century, infectious diseases like smallpox and tuberculosis spread easily. Under-five mortality rates were particularly high. During this stage of health transition, public health measures against infectious diseases were developed, and hospitals for physical and mental illnesses were established in the US in the eighteenth century (Turner 1977). Information on health issues facing developing countries during this stage was limited.

In Stage II, 'The Age of Receding Pandemics', we see a significant reduction in mortality. Infectious disease epidemics were less common and sometimes eliminated. Life expectancy at birth increased to 30–50 years. As a result, population growth rates increased significantly. In the early phase of this stage, endemic, parasitic and nutrition deficiency-related diseases, and maternal and child health problems were the leading causes of death (Caldwell 2001). During the nineteenth century, sanitation and personal hygiene improved markedly as increasing numbers of people better understood the causality of infectious diseases and their modes of transmission (Winslow 1923). Although epidemics remained a public health problem, infectious disease control measures evolved from quarantine and isolation to prevention, e.g. personal hygiene and environmental sanitation. Environmental sanitation including waste and sewage disposal and sanitation promotion was adopted in the UK by the Public Health Act of 1848 (Chave 1984).

At the same time in the US, comprehensive public health systems evolved from initiatives such as Shattuck's Report of the Massachusetts Sanitary Commission (Winslow 1923; Rosenkrantz 1972), and boards of health and state health departments began to be established shortly thereafter (Hanlon and Pickett 1984; Fee 1987).

Industrialization, however, became an increasingly prevalent mode of transmission for infectious diseases (Wohl 1983). Although sanitation improved, death from urbanization-related diseases, particularly tuberculosis, increased, as seen in the state of Massachusetts (Hanlon and Pickett 1984). By late in the nineteenth century, however, bacteriology, the germ theory, and laboratories to investigate the causes of infection were being developed in the United States (Winslow 1923). And mortality due to such infectious diseases as yellow fever and typhoid were significantly reduced by improved water sources and sanitation and through an understanding of the concept of vector-borne diseases (Winslow 1923; Blakes 1956).

Stage III 'The Age of Degenerative and Man-Made Diseases' is characterized by declining mortality together with low fertility, resulting in a population pyramid change and a higher proportion of elderly persons. In Stage III, life expectancy at birth increases, on average, to more than 70 years (Omran 1998). In developing countries increased attention is given to tuberculosis. In developed countries, increased attention is given to chronic NCDs and cancer. In the United States, the Food and Drug Act was passed in 1906, the National Institute of Health (NIH) was established in 1930, the National Cancer Institute in 1937, and the Social Security Act was passed in 1935.

During the twentieth century, most developed countries moved into Stage III of the health transition. In developing countries, the health transition has varied significantly among countries, with some still experiencing low life expectancy, an insufficient infrastructure for healthcare delivery and for water, hygiene, and sanitation. Meanwhile, the double burden of communicable and NCD is observed in many of these countries along with concurrent problems of undernutrition and overweight/obesity.

For a summary of the three stages, please see Table 3.5.1 and Figure 3.5.1.

Table 3.5.1 Stages of the health transition and associated activities

	Stage I The age of pestilence and famine	Stage II The age of receding pandemics	Stage III The age of chronic diseases
Health profile	• High, but fluctuating mortality rates due to epidemic, famines, and wars • High fertility rates • High rates of child mortality—with infectious disease the major cause of death • Among women of childbearing age, high risk of maternal mortality resulting from complications in pregnancy and labour • Life expectancy between 20–40 years	• Decreased prevalence of infectious diseases and malnutrition resulting in dramatically reduced mortality rates • Increased population growth rates due to high fertility rates with decreasing mortality • Life expectancy at birth increased to 35–50 years	• Low mortality and fertility rates resulting in reduced population growth • Reduced prevalence of infectious disease resulting in reduced mortality rates • An increasingly ageing society facing chronic degenerative diseases with relatively high morbidity • Significant changes in lifestyles • Life expectancy at birth increases to over 70 years
Health activities	• Absence of planned healthcare systems	• Healthcare and health technologies were developed, e.g. immunization programmes and the development of antibiotics for control of infectious diseases including acute bronchitis, influenza, and syphilis	• Social and national health insurance systems developed • Increased healthcare addressing needs of the elderly

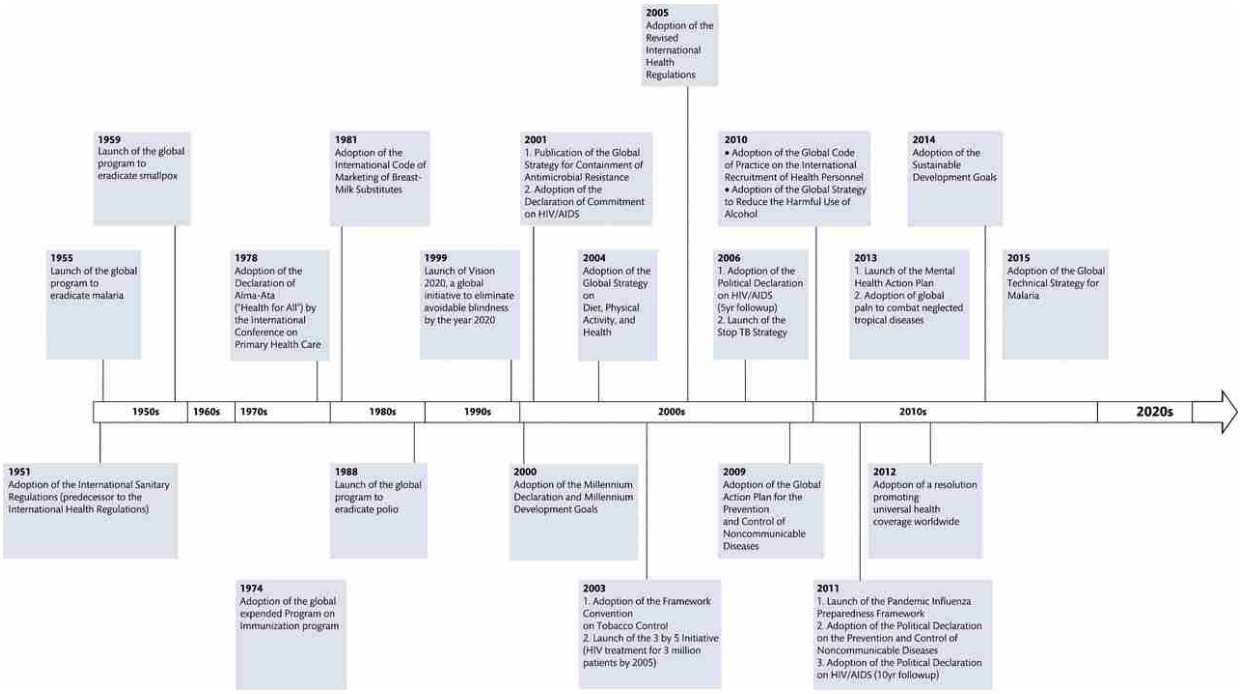


Fig. 3.5.1 Milestones of health policy.

Globalization and health policy

Globalization has been found to have direct and indirect impacts on public health (Welander et al. 2014). The development of global health policies also has been significantly affected by globalization and its impacts.

Globalization and the resulting surges in trade and migration have increased public health threats (infectious diseases, long-term illness, and environmental stress). Higher flows of goods and travel contribute to spread of diseases such as pandemic influenza, drug resistant tuberculosis and severe acute respiratory syndrome (SARS). The most recognizable example was the SARS epidemic in 2003. This epidemic spread most quickly during the three months when international travel peaked (Kovačić 2014). And, with more lower- and middle-income countries adopting the lifestyles common in industrialized countries, the prevalence of NCD has been steadily increasing.

On the plus side, production processes associated with globalization have sometimes triggered actions with positive effects on child health outcomes (Owen et al. 2007). During the nineteenth century, these included measures designed to reduce the health consequences of industrialization. Globalization also has brought with it an enhanced flow of information and

technology improving both health literacy and disease surveillance in developing countries, although the usage of such scientific findings is sometimes hindered by intellectual property rights, and despite the fact that the distribution of health services in most developing countries is highly skewed away from those in greatest need.

Global warming, likely increased by globalization, is one of the strong social determinants of health (Watts et al. 2015). The agricultural losses attributable to climate change have already had indirect health impacts in poorer countries which rely so heavily on their agricultural products. And a weakened agricultural sector inevitably has negative effects on the economy, reducing employment, increasing migration, and increasing food insecurity (Kovačić 2014; Watts et al. 2015).

In the 1980s and 90s, economic globalization had positive effects on the economies of some developing countries, particularly Costa Rica and Vietnam, with resulting increases in healthcare service delivery (Cornia Giovanni 2001; Díaz-bonilla 2002; Prabhat et al. 2002). China also has been a major beneficiary. By contrast, in many other developing countries, particularly in Africa, globalization has led to increases in income disparities coupled with and a sluggish, insecure and uneven pattern of growth and stagnation in health indicators has been observed with a general increase in mental illnesses, in alcohol and drug related health problems, and in stress-related deaths attributable to cardiovascular diseases and suicides (Cornia Giovanni 2001). Another frequent consequence of globalization has been higher numbers of insecure workplaces with poor safety regulation. And while global tariff reductions have had positive effects on food consumption, they also have had the effect of increasing tobacco consumption in some countries.

Current health challenges in developing countries

Developing countries are generally characterized by wide economic disparities which, in turn, are reflected in health status. These disparities, along with cultural and political factors have resulted in a far greater vulnerability to diseases among the poor in these countries (Orach 2009). Low-income households and individuals are rarely able to afford and access the quality healthcare received by the economically advantaged, and are far more likely to die from infectious disease and undernutrition. Such negative

health effects, in turn, have deleterious effects on productivity and national economic growth (Worldwash Institute 2018).

The spread of infectious diseases in these countries is, in part, an outcome of fragile health systems characterized by poor surveillance, low immunization rates, and problems of drug resistance. In some countries, new infectious diseases such as Middle East Respiratory Syndrome (MERS) and Bourbon virus have become prevalent; in others more common diseases have spread to other geographic areas or develop new clinical patterns resistant to existing treatment. As indicated, increased migration and changing social practices can contribute to the emergence or re-emergence of infectious diseases. Among the most serious, at present, are HIV/AIDS, SARS, H1N1 influenza and Chikungunya disease all of which take their toll both on individuals and on national economies (Heymann 2009; Dikid et al. 2013).

WHO estimates that more than 95% of HIV infections occur in developing countries with roughly 60% of them in sub-Saharan Africa. Over 7 million persons in Asia and the Pacific are infected with the virus. WHO data indicate that the primary mode of HIV transmission differs according to region (WHO 2017). Homosexual transmission is more common in Southeast Asia and in Central and South America than in Africa. Transmission from intravenous drug usage and transfusion of contaminated blood are also common in sub-Saharan Africa and South Asia. In developing countries as a whole, sex work, poor public health facilities, and poverty have been key factors in the spread of HIV infection (Morens and Fauci 2013).

HIV/AIDS is now the leading cause of death among women of reproductive age. UNAIDS has estimated that, in 2016, over 6,500 teenagers and young women between the ages of 15 and 24 were infected with HIV every week, the majority in sub-Saharan Africa (UNAIDS 2018). In some regions, women who have been subjected to violence are one and a half times more likely to become infected with HIV than women who have not. A UNAIDS project in Zimbabwe found that gender inequality, itself, is a major contributor to women's HIV vulnerability.

In addressing the HIV/AIDS problem in developing countries, governments and assistance organizations have, increasingly, been focusing both on prevention and control strategies with special attention to disease-specific intervention (particularly condom promotion and antiretroviral

therapy) and women's empowerment (focused both on education and employment). These efforts will be enhanced by overall economic growth and improved income distribution (Ruger 2004).

Another disease epidemic of recent decades has been SARS, a respiratory disease caused by the SARS coronavirus (SARS-CoV). In 2002–03, an epidemic originating in Guangdong Province of China spread globally, with over 8,000 known cases reported in 8 months from 33 countries on five continents (of which 21% were healthcare workers). The Asian Development Bank estimated that SARS resulted in the loss of US$20 billion in Gross Domestic Products among Asian countries in 2003 (Heymann 2009). The last known community case occurred in the USA in July 2003, but another localized SARS-related crossover from animals occurred in 2004. At the time of press in 2020, the SARS-CoV-2 pandemic is developing rapidly, with over 200,000 dead at the time of writing.

Another animal-related influenza was the re-emergence of the H1N1 pandemic in 2009 (the first outbreak took place in 1918). This pandemic, lasting from early 2009 to late 2010, was of a new strain resulting from a combination of a Eurasian pig flu virus with a previous reassortment of avian, swine and human flu viruses (Morens and Fauci 2013).

Some infectious diseases such as Dengue Fever, earlier considered to be lower level concerns, are now re-emerging with increased incidence, or spreading to new areas. This has resulted from combinations of drug resistance, the genetic evolution of agents of the diseases, inadequate public health measures, and the globalization issues discussed earlier. Drug resistance problems often result from over or under utilization of antibiotics. Some bacteria undergo genetic mutation from drug resistance, while others acquire the genes in question without a mutation. The same species of bacteria can sometimes receive the transfer of a resistance gene by bacterial genetic exchange mechanisms (NIH 2007). The contributing shortcomings in public health measures usually relate to low immunization coverage or to the absence of adequate international health regulations. These problems are exacerbated by population movement, conflict, and poor living standards (Heymann 2009). In the case of Dengue Fever, the spread of the infection in India has been attributed primarily to rapid population growth, poor sanitation, and migration. In Malaysia, nearly 60,000 cases of Dengue Fever and 165 deaths were reported in 2015 (Singh et al. 2017).

Among diseases of concern in Southeast Asia, is chloroquine-resistant *Plasmodium falciparum* malaria which began spreading in 1973. Although combination therapy with two antimalarial drugs is being used currently, ongoing surveillance for the evolution of antimalarial drug resistance of the protozoa is required. Primary prevention strategies include removal of risk factors through vector eradication, immunizations, and provision of uncontaminated drinking water, while, secondary prevention strategies include surveillance, and early therapeutic interventions. Underlying both sets of strategies is the need for infrastructure development to assure proper distribution of such care to all segments of communities (Singh et al. 2017).

Overall, the control of emerging and re-emerging infectious diseases requires fully functioning immunization efforts (accompanied by efforts to address cultural and trust-related resistance to such immunization), and strong surveillance systems supported by continuing region-specific research efforts and ongoing advocacy efforts aimed at policymakers in these countries. Control of these diseases also requires increased coordination among public health officials, clinicians and research scientists (Sarma 2017).

As indicated earlier, developing countries presently are in a stage of epidemiologic transition from common communicable disease-based mortality to one of increasing incidence of NCDs. WHO estimates that 71% of all deaths result from NCDs, with 15 million persons between the ages of 30 and 60 dying from these diseases. Over 85% of these ‘premature’ deaths occur in low and middle-income countries. While urbanization and industrialization have been boosting life expectancy in these countries, they also increase exposure to NCD risk factors. The World Economic Forum has estimated that the cumulative cost of NCDs will reach US$47 trillion or 5% of global GDP by the year 2030 (WHO 2018).

As part of their responses, national and international health policies are seeking to reduce alcohol and tobacco use through health education and with taxes and pictorial warnings on products. Such taxes have also been placed on sugary soft drinks in some countries to reduce overweight and obesity problems associated both with cardiovascular diseases and diabetes mellitus (Yamada et al. 2013; Baldwin 2015).

Maternal and child health

In February 2015, WHO published ‘Strategies towards ending preventable maternal mortality (EPMM)’ (EPMM Strategies), a direction-setting report outlining global targets and strategies for reducing maternal mortality under the Sustainable Development Goals (SDGs). The result of extensive consultations with stakeholders around the globe, the targets and strategies in the EPMM Strategies report are grounded in a human rights approach to maternal and newborn health and a focus on eliminating significant inequities that lead to disparities in access, quality and outcomes of healthcare within and among countries (Maternal Health Task Force).

As the agenda of the SDGs emerges, healthy societies where women, adolescent girls, newborns, and children thrive, and pregnancy and childbirth are safe everywhere, should be at the heart of its ambition. Ending preventable maternal and newborn deaths and stillbirths is possible within a generation and requires focused attention on high impact interventions and strategies to improve access and quality of care. Effective care around the time of childbirth is most critical for survival and health, but comprehensive strategies and high-impact interventions should span the continuum of care including before pregnancy. Since 2014, Every Newborn: an action plan to end preventable deaths (ENAP) came into being endorsed by 194 member states, at the 67th World Health Assembly. The plan identifies eight specific milestones that guide immediate actions needed by 2020 to best impact the ambitions of the SDGs and the Global Strategy for Women’s, Children’s, and Adolescents Health by their close in 2030. ENAP is a road map of strategic actions that harness evidence-based solutions to end preventable newborn mortality and stillbirths, and contribute to reducing maternal mortality and morbidity.

At the International Confederation of Midwives Triennial Congress in Toronto, a joint WHO and UNICEF publication—Reaching the Every Newborn National Milestones by 2020: Country Progress, Plans and Moving Forward—was released. The report details the country leadership and actions taken towards meeting the milestones throughout 2016. Progress was informed by the participating countries’ use of the Every Newborn Tracking Tool, which captures the processes of country ownership and action. The tool was adopted by 33 new countries between 2015 and 2016 (a 150% increase) and, as a result, the report provides a rich and internationally informed picture of progress made in addressing maternal and newborn health outcomes. Newborn health plays an increasingly important role in the

SDG period; the day of birth and the first 28 days of life are the most vulnerable time for a child's survival and health. Despite three-quarters of all newborn deaths resulting from preventable and treatable conditions, within the bracket of under-five deaths, newborn deaths declined at the slowest rate during the MDGs period (1990–2015), and now account for 45% of under-five deaths globally (up from 40% in 1990). The detailed strategic guidance, specific targets, interventions, and milestones from the ENAP and EPMM global action plans provide guidance for countries to use in their formulation of national health plans and funding priorities (Chou et al. 2015).

In many studies across the world especially in developing countries, differential analyses were conducted in order to better understand under-five mortality levels and patterns. Several variables were considered as differentials of under-five mortality rates. The most important variable was women's parity: the higher the number of children already born, the lower the probability of survival of the child. The 2013 Human Development Index (HDI) varies from 0.337 (Niger) to 0.944 (Norway). The value for East Asia and the Pacific is 0.707 (UNDP 2014). It is unlikely that mortality in developing countries with high mortality, and in particular under-five mortality, can be significantly reduced unless socioeconomic development experiences a substantial and sustained advance.

Under-five mortality rates are a leading indicator of the level of child health and overall development in countries. Worldwide, child deaths are falling, but not quickly enough. The global under-five mortality rate has declined by a little more than half, dropping from 91 to 43 deaths per 1,000 live births between 1990 and 2015. However, the current rate of progress was well short of the Millennium Development Goal (MDG) target of a two-thirds reduction by 2015. Under-five mortality remains the focus of the post-2015 MDG agenda, particularly in the newly adopted SDGs (Chou et al. 2015). It is important to remember that infant and under-five mortality estimates frequently differ between censuses and surveys as surveys usually use full birth histories which tend to underestimate infant and under-five mortality, while censuses use indirect methods which tend to slightly overestimate infant and under-five mortality.

Substantial progress was made in achieving the MDGs. Maternal mortality fell by almost 44% worldwide, from 385 deaths to 216 deaths per 100,000 live births over the period 1990 to 2015 (WHO 2015). This translates into an average annual rate of reduction of 2.6 percent. Although

this was a substantial decline, it was less than half of the 5.5 percent rate needed to achieve the three-quarters reduction in maternal mortality targeted for 2015 in MDG 5. In general, the ignominy of poverty has not ended at all. In 2016, members of the United Nations agreed on 17 SDGs to be achieved by 2030. This new set of goals aims to complete the agenda of the MDGs so that no country is left behind. Health is the third goal of the SDGs; one of its targets is to reduce the maternal mortality ratio to less than 70 per 100,000 live births by 2030 (WHO 2016).

Nutrition

An essential ingredient in the improvement of maternal child health is the improved nutritional status of these individuals.

Attention to maternal and child undernutrition—stunting, underweight and wasting, and deficiencies of essential micronutrients—increasingly focuses on the crucial period from conception to the child's second birthday, the 1,000 days during which the nutritional status of mothers and children have lasting effects on the life of the child.

The Lancet Maternal and Child Nutrition Series, 2008 and 2013 (Bhutta et al. 2008; Black et al. 2008; Ruel 2013; Black et al. 2013; Bhutta 2013; Victora et al. 2013) underlines these lifetime effects of nutritional conditions during this critical period, plus the adolescent period for girls, and reviews the evidence supporting interventions seeking to address the undernutrition problem.

The series also focuses on the so-called 'double burden of malnutrition'—these undernutrition problems coupled with problems of overweight and obesity, occurring in the same geographic areas, and with the latter sometimes a consequence of the former.

Increasingly, it is understood that, undernutrition problems, rather than being encompassed within the confines of public health, are multisectoral in their causality and, in turn, require multisectoral interventions. Where combinations of multisectoral interventions are put in place in vulnerable areas of developing countries, the effects on nutritional status have been found to be synergistic and often highly impressive. Figure 3.5.2, often referred to as the UNICEF Conceptual Framework, summarizes these multisectoral relationships

(https://www.unicef.org/nutrition/training/2.5/4.html, accessed September 2018).

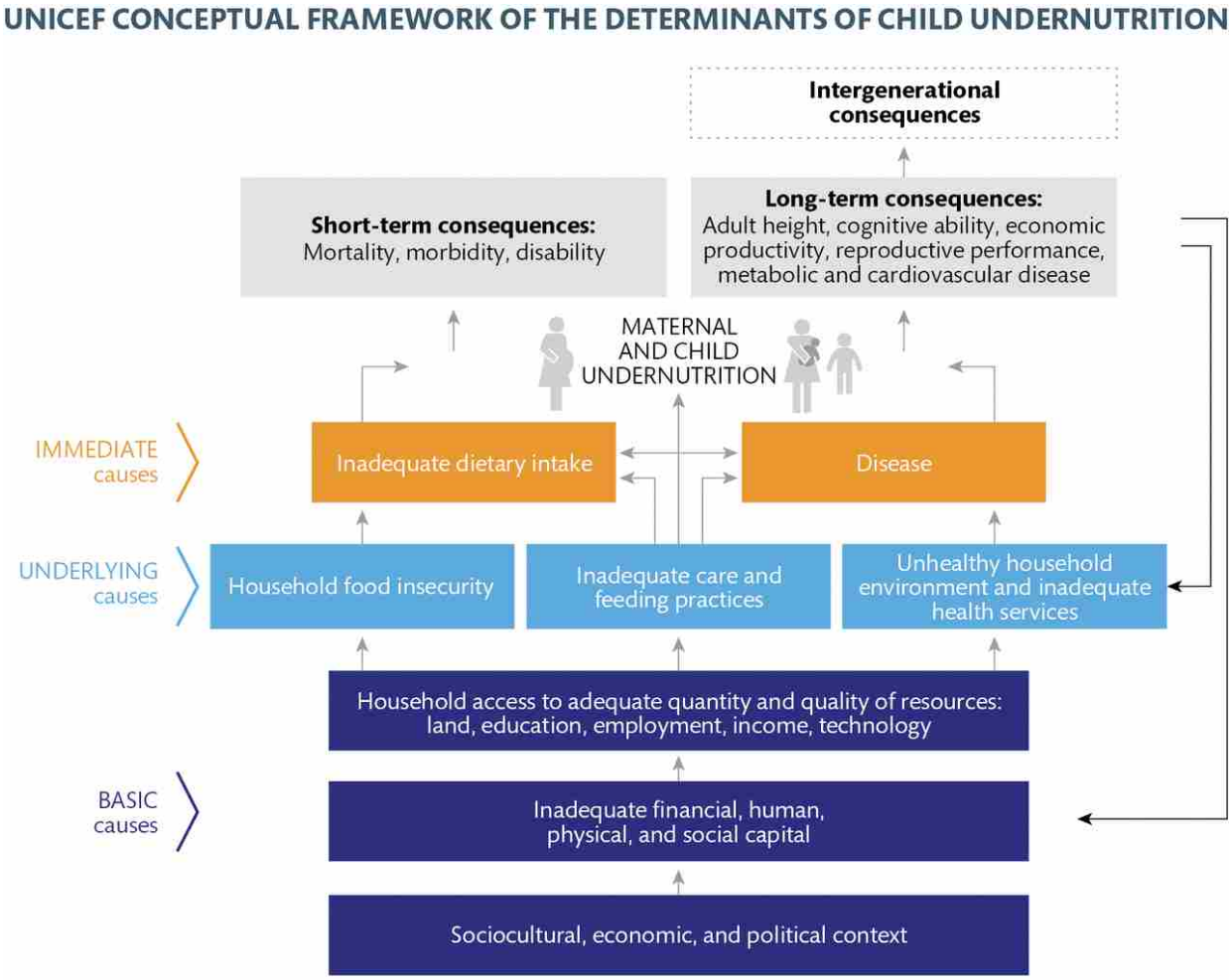


Fig. 3.5.2 Conceptual framework for the causes of malnutrition in society.
Reprinted from United Nations Children's Fund. UNICEF's approach to scaling up nutrition for mothers and their children. Discussion paper. Programme Division, UNICEF, New York, June 2015.

As indicated in the diagram, malnutrition is the result of some combination of inadequate dietary intake and infectious disease. These, in turn, are the result of a series of vitally important 'underlying causes'—which, depending on their relative importance in a given geographic area, require effective interventions. Accordingly, nutrition policies and strategies

often include project activities addressing food insecurity and poverty, behavioural change communications, health service delivery and water, sanitation, and hygiene (WASH) interventions, plus maternal and child feeding and micronutrient provision which address the ‘immediate causes’.

Interventions addressing undernutrition that fall outside of the health sector are generally referred to as *nutrition sensitive*, addressing, primarily, the determinants of undernutrition (Pinstrup-Anderson 2013). Examples of these are:

> *In the agricultural sector*—homestead gardens, small livestock, and small pond fish farming.
>
> *In the education sector*—inclusion of essential nutrition information in primary and secondary school curricula, and, more generally, generating increases in female enrolment and attendance at both levels.
>
> *In the social protection sector*—the coupling of social protection benefits with counselling that facilitates the translation of income supplements into improved nutrition and health status.
>
> *In the WASH sector*—assurance that these services are provided to those households at highest risk, and that the physical infrastructure provided is accompanied by adequate counselling on its use.
>
> In the health sector, nutrition-related interventions are often described as *nutrition specific*, addressing more directly the effects of undernutrition with a particular focus on prevention. These interventions include child feeding, behavioural change communications with special attention to optimal breastfeeding and complementary feeding practices, micronutrient supplementation and fortification, and the treatment of severe acute malnutrition.

Such nutrition-specific interventions in the health sector are particularly important given existing problems. Despite improvements in some countries, breastfeeding practices continue to be highly problematic and contribute significantly to high child mortality rates. Deficiencies of vitamin A, iron, iodine, and zinc adversely affect maternal and child health, while addressing these problems through supplementation and fortification have been found to be among the most cost-effective interventions functioning in developing countries.

Overall, over 165 million children in these countries suffer from stunted growth. The highest prevalence of stunting is found in eastern and western Africa and in South Asia. The largest number of children affected by stunting, over 70 million live in South Asia. The numbers of stunted children in Africa has actually been increasing in recent years (Black 2013). When micronutrient deficiencies are added in, undernutrition, resulting both in losses in cognition and reduced schooling has been found to reduce the

economic growth of a developing nation by at least 8% through productivity losses (Horton 2013).

There has been a significant response to these problems from governments, international and bilateral assistance agencies, and non-governmental organizations. These interventions, however, are unlikely to achieve the target of the World Health Assembly calling for a 40% reduction in stunting in children under age 5 between 2010 and 2025 (Black 2013).

Actors in health policy

Governments in developing countries play a primary role in the formulation of health policy formation and its implementation, often with the support of development partners. These partners include United Nations (UN) agencies, other international and bilateral donors, and international and domestic non-governmental organizations. International and national donors provide financial as well as technical expertise, often with an eye on international targets such as the SDGs set for the year 2030 (Koduah et al. 2015).

National governance and the orientation of key national political leaders inevitably have considerable influence on both the direction of health policies and their implementation (Shiffman and Smith 2007; Erasmus and Gilson 2008; Gilson et al. 2014; Koduah et al. 2015). At the same time, the role of officials and institutions at subnational levels, the extent of community involvement, and the capacity and commitment of front-line workers can be critical in this regard (Erasmus et al. 2017).

Multiple studies have highlighted the gaps and challenges in health policy implementation in developing countries (Belaid and Ridde 2012; Lehmann and Gilson 2013; Gilson et al. 2014). These shortcomings often relate to financial shortfalls or to the inadequate authority or commitment of implementing actors (Belaid and Ridde 2012). Limited responsibility and discretionary power at subnational levels have proven to be a particular liability (Lehmann and Gilson 2013; Gilson et al. 2014). More successful health policy implementation has often come on the heels of health sector reform including some combination of structural changes, e.g. decentralization, and changes in the model of healthcare delivery system (Atkinson et al. 2005).

Increasing attention also is being given to the role of social media in health policy. Where there are limitations on media freedom and, in turn, on

government transparency, or on the activities of civil society advocacy organizations there have usually been consequences (Ha et al. 2010; Green et al. 2011). By contrast, social media have played key roles in the success of behavioural change efforts relating to the consumption of alcohol and tobacco products and to traffic accidents (Grilli et al. 2002; Elder et al. 2004; Bala et al. 2013). Effective health policies generally include the role of social media in disseminating key elements of these policies and in providing the public with essential health information (Roland 2017).

Policy agenda setting is usually determined by combinations of health sector data (ideally disaggregated by geographic area), by the availability of funding, and by the orientation of policymakers themselves (Koduah et al. 2015). As indicated, donor agencies often have a key role in both policy formation and in its implementation (Grépin et al. 2017). The nature and scope of donor influence at different levels of policymaking have been well documented in Pakistan and Cambodia (Khan et al. 2018). Primary influences often include the utilization of financial resources and mechanisms to assure adequate monitoring and evaluation of policies and programmes (Hafner and Shiffman 2013; Khan et al. 2018).

Two other important influences on health policy development are the international agenda (e.g. SDGs), and the availability and quality of scientific evidence (Wallace et al. 2012).

Reviews in some African countries and in Myanmar have recommended an upgrading of evidence-based decision-making and policy development, and have underlined the vital importance of increased receptivity to such evidence, and of universities in the necessary capacity development (Lavis et al. 2005; Orton et al. 2011; Young et al. 2016; Ellen et al. 2018). To fill gaps in essential knowledge transfer, such training is necessary both for policymakers and research organizations, and improved communication between the two should be encouraged (Whitehead et al. 2004).

Political and economic impacts on health systems

Government health systems are significantly affected by the country's economy, and by its political structure and orientation. Where political instability exists, political systems are weak (e.g. overly centralized, bureaucratized decision-making, weak regulations), economic growth is

sporadic, or welfare systems are dysfunctional, health consequences are inevitable (Guha-Sapir and D'Aoust 2010).

In some countries with strong political commitment to reduce hunger, poverty, or inequality, the health status of citizens has improved markedly even without highlighted attention to health and nutrition. This has been seen during particular periods of recent history in China and Cuba, and during the early years of the Lula presidency in Brazil (Field and Levinson 1975). In Vietnam, health sector management until the mid-1980s was state–socialist or universalist. Health expenditure emanated both from the central and subnational governments and from other local resources. All citizens could enjoy preventive and curative healthcare free of charge as citizens' rights. As a result, child mortality rates dropped markedly. Unfortunately, in some regions of the country with more limited resources, health improvement was less dramatic and was often characterized by inequality in service delivery (London 2008).

Commercialization in the healthcare sector began in the mid-1980s. Often resulting from macroeconomic reform, such commercialization often led to a lower quality of healthcare delivery particularly in public hospitals where, despite an outward appearance of free services, the informal fee charges levied—emanating from the new profit orientation of these hospitals, often reduced hospital utilization by the poor. Similarly, financial decentralization in the social sector as a whole often resulted in inadequate budgetary allocation for actual health services at subnational levels, with most of the funding going for high level medical staff salaries, and more sophisticated equipment rarely available to those most in need (Forsberg 2011). The associated poor management and weak regulatory systems have usually led to reduced immunization rates and to the increased prevalence of particular infectious diseases.

In South Africa, racial discrimination was the primary cause of poor health services for the African population along with inequality in most services during the Apartheid period. The Bantustan regions set aside for large segments of the African population had their own healthcare systems which were invariably inadequate and resulted in high mortality rates (Maillacheruvu and McDuff 2014).

In 1994, with the election of Nelson Mandela and the end of Apartheid, a new National Health Plan was developed with comprehensive primary health services (Coovadia et al. 2009). Many health sector challenges

persisted, however. With inefficiencies in the health system, the percentage of Gross Domestic Product devoted to health was much higher than in most developed countries but without the quality of service delivery desired, and the expansion of health infrastructure resulted in an overly heavy work load for the often underpaid healthcare professionals—many of whom left the government for better jobs elsewhere. Meanwhile, continued poverty and health illiteracy among large segments of the population resulted in low immunization rates and high use of tobacco and alcohol. Over time, new economic policies emerged with a focus both on economic growth and redistribution, and results have improved accordingly (Mayosi and Benatar 2014).

With these new policies in South Africa, improved benefits for health staff reduced migration to the private sector (George 2013), and intensified community mobilization significantly increased awareness of communicable diseases, particularly HIV/AIDS and tuberculosis (TB). The country's Strategic Plan addressing HIV/AIDS 2007–2011 served as a model for many other countries (South Africa Department of Health 2007). At the same time, a new council of healthcare professionals became the regulatory body for new drug approval (Coovadia et al. 2009).

In sum, even countries with similar geographical conditions are likely to have marked differences in their health outcomes depending on the political and economic systems functioning and their quality (Chongsuvivatwong et al. 2011) In general, socio-democratic political systems with redistributive policies result in improved coverage of health systems and better health outcomes (Borrell 2007).

Challenges in implementing global and national health policies

Governance: responsibility and accountability

Public health policies in developing countries need to be compatible both with regional and global policies to increase cohesiveness, integration, and support. Where this is done, national policies will better assure that health priorities in local settings are addressed, and that international support can support health sector priorities. With the creation of MDG and now SDG goals, more attention is being given in these countries to policies and programmes which are results- and outcome-oriented. There remains, however, a tendency in many countries to be satisfied with the achievement

of input and activity targets (more easily achieved and less threatening to health system management likely to be concerned about the extent to which these activities actually will achieve outcome targets—- and the repercussions should these outcome targets not be achieved). There also is a tendency for health systems to address only disease symptoms and not their underlying causes.

Even where outcome targets are pursued and underlying determinants are identified, a country's ability to address these adequately is often limited by financial capacity.

Using the analogy of building a sound physical structure, a national health system needs to ensure a concrete foundation with (a) a clear understanding of its population's socio-cultural norms and its health problems—both symptoms and causes (disaggregated by geographic area), and of existing population disparities and inequalities, and (b) a functioning system of governance and accountability.

Accountability has to exist both in national and subnational governance, and applies both to financial disbursements and to health service delivery staffing and infrastructure. These, in turn, require skilled management.

Sound strategy development, actions plans, and programmes plus effective review processes might constitute the first layer of building blocks above the foundation. These can be problematic when they are formulated by a limited inner circle of officials and do not take into account the needs of local governance and communities, the latter, as indicated, often diverse in their socio-cultural norms and values. Even where efforts are made to involve communities in decision-making and to elicit their responsibility, those representing the communities are often only persons from upper income groups and (where relevant) from more advantaged religious groups, races, or castes.

The second layer of building blocks (above the foundation) would be the programme administrators, managers and service providers/front-line workers, with these persons fully cognizant of the building structures below them, clear about their own responsibilities (and, where relevant, their own accountability), competent to perform their tasks, and able to avail of the necessary resources to accomplish them.

The cement, if you will, holding these building blocks together and firmly attached to the foundation, consists of such values as honesty, vision, and the ability to keep results and outcomes in sight, fairness, mutual respect,

adherence to commitments, motivation, and sound judgement, the ability to instil confidence in others, and, perhaps binding *these* together, a sense of ownership of the undertaking as a whole. Exemplifying such cementing qualities might be segments of the health sector professionals in Brazil in recent years, believing so firmly in their country's healthcare pursuits that they regularly eschewed more lucrative offers from the private sector (Levinson and Balarajan 2013).

Health financing and the utilization of existing resources

The maximum utilization of existing resources through leveraging and aligning these interventions among stakeholders are necessary steps in health system modernization. These often can be accomplished by eliminating wasteful expenditures, reducing the cost of service provision, and encouraging new and more efficient means of providing care. Modernized healthcare also requires effective information technology (IT), which provides a base upon which to build managerial, staffing, and financial suprastructures. At the same time, cost-effectiveness studies comparing treatment and prevention modalities offer providers with valuable information to use in the selection of these modalities.

Health information

Health information system (HIS), which is an integral part of the health system strengthening needs a strong institutional regulatory framework. It is also an entry point into the health sector reform through provision of information and knowledge on the quantity as well as quality of health services. It is of paramount importance to have the reliable, relevant, and real-time information in order to make proper decision for improving the quality of services through proper and effective allocation of available resources.

Fragmentation of HIS is hampered further as donor funding sometimes targets specific areas and creates new information systems which are not integrated with the existing HIS (Braa Jørn et al. 2007). The role of the HIS in the health system is not just routine collection of health services data and reporting to higher levels of the healthcare system, but to be used by the health staff at various level and apply and facilitate evidence-based decision-making at all levels, most importantly at the point of collection but it is not

usually the case in real life situation. Moreover, health policies and plans should be based on such information, which also needs to be ensured to be aligned with the needs at the operational level. Many developing countries are still facing with the issues of data privacy versus data sharing as both are equally important and supportive to each other in realizing its function effectively. Proper regulation of the creation and use of healthcare information is imperative to everyone in the system.

It is crucial to strengthen health information including human resource for long-term cost-effectiveness in spite of having potential resistance to change and cost issues (Mpofu et al. 2014). Improved health worker capacity to monitor and evaluate programmes would improve data quality, management, and reporting, which further increased use of health data for disease surveillance and planning purposes. Quality data within the health sector itself is crucial for monitoring and evaluation (M&E) of health programme performance and for developing appropriate policies, plans, and strategies to ensure quality health services. It usually covers the dimension, comprised of data accuracy, reliability, completeness, timeliness, precision, and integrity (USAID 2007). Many developing countries experience challenges related to producing quality health data (Chaulagai et al. 2005). The contributing factors include weak linkages between M&E systems and data generation points and inadequate harmonization of multiple data collection and reporting systems (UNAIDS 2010). Additional challenges include poorly coordinated M&E systems characterized by duplication of data collection and underutilization of nationally-defined tools [12–15]. A frequently documented challenge related to producing quality health data in low and middle-income countries is the shortage of trained human resources (UNAIDS 2010). In addition, improved data quality better informed evidence-based planning at the all level. One approach that has been effectively used to address skilled health workforce shortages is task shifting (WHO 2008). In Botswana case, the district M&E officers took on M&E responsibilities that were previously conducted by nurses and other health workers as secondary activities to core programme and clinical responsibilities. This allowed such health workers to concentrate on delivering quality clinical services and focusing on implementation of quality, priority health programmes (Mpofu et al. 2014).

Several other studies also indicated that data quality improved by addressing human resource capacity issues, mainly through training of health

workers on the importance of public health information and regular feedback (Mphatswe et al. 2012). In Malawi and South Africa, it was found that data quality improved through provision of supportive supervision (Mate et al. 2009) and institutional research (Harries et al. 2011). Other initiatives, such as the transition from paper-based to electronic records, have also been shown to improve data quality (Ekouevi et al. 2011).

Infrastructure, technology, medical supplies, and logistics

Healthcare infrastructure must be adequate to ensure a reduced burden of diseases and an improvement in the health status of the population. A solid broad-based health infrastructure is composed of necessary physical structures accessible to the entire population served; other support facilities including laboratories, medical record systems, and transport; and a competent health workforce supported by systematic and effective resource allocation practices, supervision, and supportive monitoring. An effective healthcare infrastructure needs to be both human-centred and client-responsive. The proper location and management of health infrastructure components are essential to the health system's effectiveness in order to achieve the desired 'universal health coverage' and to meet basic health needs, primary healthcare centres should be within the reach of the entire community and fully cognizant of prevailing disease patterns. These primary, secondary, and tertiary healthcare centres, in turn, require the necessary facilities and equipment, competent and efficient staff, and the ability to deliver an essential health services package at affordable costs. These centres also should be characterized by a welcoming environment for both clients and service providers.

Proper formative research and area-specific baseline information on disease prevalence, collected through necessary surveys and research, should be in place prior to the formulation of actual health plans in order to determine the optimal location of each type of health centre, the anticipated workload, and the transportation that will be necessary. Often the establishment of fewer more efficient tertiary centres at strategic locations, with transportation available to assure access by more remote population groups, can provide greater benefit than establishing a larger number of tertiary centres with insufficient infrastructure and human resources. In the building of hospitals and other health facilities, a premium must be placed on proper design facilitating cost-effective, quality-assured service delivery,

and should include durable, built-in water and sanitation facilities which are adequately equipped.

The global trend towards high-end medical technologies presents a different type of challenge for developing countries, which, on the one hand, would wish to improve the quality of treatment by using these modern technologies—and save more lives in the process, but recognize also the exorbitant costs associated with much of this technology. Where resources are limited, and they usually are, choices often have to be made between acquiring such equipment—likely to benefit a relatively small portion of the population, and using available resources to upgrade the quality and availability of health services and preventive public health utilized by larger portions of the population. High-end medical technology also needs a cadre of sophisticated medical practitioners requiring high salaries, and well-equipped maintenance staff to keep the equipment functioning properly.

Some middle-income countries have found a workable balance in working with these choices. A prime example is Turkey, which has managed to expand high-quality diagnostic and curative services efficiently and effectively but without spending on them a disproportionate percentage of the country's health budget. This has been done, in part, by successfully enlisting the involvement of the private sector including pharmaceutical companies, as well as biomedical engineering societies and philanthropic organizations. Turkey also has put in place a reasonably effective referral system—including timely transportation arrangements—to increase the likelihood that those most in need of this more sophisticated technology will be able to benefit from it. See Case Study 3.5.1.

Case Study 3.5.1 Turkey's Health Reform

It shows an evidence of how a poor health system with health indicators below par changed into an improved system with population satisfaction for the health services they received. The fruitful outcomes within a decade were attributable to the decisive political action and effective reform, resulting from wisdom and forward thinking vision of health minister in 2002. Today, it has improved its health outcomes, achieving health-related Millennium Development Goals well before the 2015 deadline. The number of women dying from pregnancy-related causes more than halved since 1990 to about 20 per 100,000 live births in 2013. The number of infants dying before their first birthday dropped from 55.7 in 1990 to 16.5 per 1,000 live births in 2012.

Strong economic growth and political stability were also found contributed a lot to the strategic directions and its implementation in health reform. A tactful confidence building and support by doing 'easy first and difficult later' reform interventions was very thoughtful and noteworthy.

Major success reasons: Country-led strategic directions and realistic goals, embedded in a strategic management framework with effective implementation and M&E system.

Emergency Action Plan of the 58th Government, which under the heading Health for All was adopted shortly after the elections on 3 November 2002, contains 11 dimensions of reform and on the basis of the Emergency Health Action Plan, the HTP was prepared in early 2003 under eight themes. It had been transformed from working as provider of health services to steward of the health sector. Turkey's Health Transformation Programme met health-related MDGs quite ahead of the 2015 deadline. The lessons learned from Turkey centre on the importance of the sequencing of health reforms and their rapid implementation, and the benefits of primary healthcare reforms. The 'patient-centredness' of the Turkish system with a strong emphasis on health as a right was the cornerstone of reform.

Policy lessons

- Effective sequencing in health reform steps by ensuring a good start and winning people's hearts
- Rapid and timely interventions with right pace for future policy shift, for example dual practice of physicians
- Expanding public health expenditure and giving a space for non-contributory insurance scheme for the poor
- Ending fragmentation and duplication in public sector service delivery and introduction of family medicine programme
- Unification of various health insurance schemes and health financing reforms along with the changes in health service delivery

(http://www.worldbank.org/en/news/feature/2015/02/17/lessons-from-turkey-how-bahrain-is-looking-to-improve-its-healthcare)
(http://www.euro.who.int/__data/assets/pdf_file/0017/272321/Strategic-Planning-for-Health_Turkey.pdf)

An effective health workforce requires adequately detailed job descriptions, and then training to assure that these positions can be professionally met. Pre-service as well as in-service training needs to be both job-based and results-oriented. At the same time, supportive supervision and M&E systems are always necessary. As in the case of equipment support discussed previously, the private sector may, at times be in a position to support or augment a health workforce, although effective regulations and systems of accountability must be in place to assure that such a private sector role not being solely profit-oriented.

Retention of public health staff is often as important as recruitment and training. Cases exist in numerous countries of health professionals who are trained for public service and then, utilizing such training, seek better paying jobs in the private sector. Accordingly motivation through recognition and incentives is often a vital pre-requisite for sustaining an effective workforce.

These should be systematically included in protocols, guidelines, and standard operating procedures (SOPs).

The allocation of a country's health budget reveals much about a government's orientation, priorities, and level of commitment to its population as a whole. Where budgets are disproportionately weighted towards high-end urban hospitals and health services and with poor access to quality services by rural and lower income population groups, the overall health status of the country is likely to suffer. Responsible allocation should be based largely on the burden of diseases in the country, and on the relative urgency of preventing and combating each disease category. This, in turn, requires proper disease surveillance and antibiograms, effective treatment guidelines, and the use of generic drugs wherever possible. Prevention itself deserves high priority in a country's health budget, and involves health education, effective food and drug regulations, and attention to safe water, hygiene, and sanitation plus interministerial collaboration to assure, for example, adequate efforts to reduce road traffic accidents.

The existence of affordable health insurance systems can, in many country contexts, assure that adequate medical care can be provided for population groups across the income spectrum.

Service delivery—overcoming geographical factors and conflict

In most developing countries, the majority of the population still live in rural areas, even with rapid urban migration. Rural areas have higher death rates from chronic obstructive pulmonary disease, suicide, and unintentional injury compared to highly urbanized counties (Eberhardt 2001). As indicated, it is essential that this population have access to affordable quality both through health centres and some combination of health extension personnel and community health workers. Included in this health delivery should be pertinent behavioural change counselling, preventive health screening, and the opportunity to avail of low-cost health insurance. The active engagement of the population in these processes and some level of social cohesiveness also have been identified as important ingredients in health sector success stories and is a key element of the Alma Ata Declaration.

One of the primary causes of development failure and, in turn poor health status is conflict. Paul Collier notes that nearly 75% of people in the 'bottom billion' countries are engaged in a civil war or have recently been through

one. (Collier indicates that the primary factors that make a country prone to civil war are low-income, slow economic growth, and dependence upon primary commodity exports.) Increases in disease prevalence resulting from such conflict are likely to continue well after the fighting has ceased (Collier 2007). Accordingly, health plans in such areas need to include the capacity for emergency responses followed by sustained health delivery. At the same time, it is clear that no health policy is capable of adequately responding to a sustained environment of conflict such as that found in Libya (Nathan 2005). Overall, responses to armed violence and its prevention require intense efforts involving criminal justice, public health, anthropology, journalism, human rights, and education (Lemard 2004; Kaldor 2007; Arya 2008).

The direct consequences of armed violence pose multifaceted challenges for the health system (Valenti et al. 2007). The economic and human costs significantly burden resource-limited services. This correlates with a reduced capacity for quality health service provision in the post-conflict environment (Kopel 1995). The indirect effects of armed violence include the disruption of livelihoods through population displacement leading to pervasive insecurity and adverse health consequences. Moreover, such factors as psychological trauma, inadequacy of essential services, and a resurgence in vaccine-preventable disease adversely affect security, health, and development (OECD 2009; Krause et al. 2011). Globally, armed violence control is increasingly recognized as a public health asset, especially in the post-conflict environment (Gilligan 2000; Krug et al. 2002).

The health workforce

Workforce development

A well-built health workforce is essential to overcome the complex threats of rapidly spreading infectious diseases, emerging and re-emerging disease and the increasing trend of NCDs in a developing country—these NCDs spreading as a result of globalization and interdependence among countries, changing lifestyles, and the improper use of chemicals and medicinal products. The unfinished agenda of the MDGs regarding infectious diseases, malnutrition, and maternal complications persists in a diversity of settings. These are further jeopardized by the emergence of chronic vascular disease, diabetes, cancers, and lower respiratory conditions, which now often constitute the most common causes of death. Significantly, there also are

sharp increases in accidents and injuries, and a growing recognition of mental illness and its consequences (Adjuik et al. 2006; Tollman et al. 2008).

Globally, countries are facing a Human Resources for Health (HRH) crisis. However, those likely to suffer most are the underserved poor, rural, and marginalized individuals and groups in the rural areas of developing countries. The development of quality HRH takes time and requires careful needs-based planning, access to skills-based knowledge, and effective training.

Health workforces often can be divided into three basic categories: (a) community-based health workers (CHWs); (b) mid-level workers (MLWs); and (c) health workers with particular expertise on an aspect of health delivery. CHWs are involved in providing preventive medical services, monitoring the community's health, identifying patients at particular risk, and sometimes providing basic curative services. These CHWs are referred to by a variety of names: health auxiliaries, health volunteers, health promoters, family welfare educators, village health workers, and community lay health aides. These are often volunteers with limited training in commodity and information dissemination and disease prevention efforts with the associated behavioural change counselling (Tollman 2008).

In many low resource countries, MLWs are the backbone of their health service. Studies on the contribution of mid-level providers indicate that they often play a pivotal role in improving access to care for underserved populations, a role also seen in some higher income countries (Keni 2006; Staton et al. 2007). Indeed the significantly increased access to HIV treatment in most African countries was achieved through the use of MLWs —and in some instances, by staff with even less formal training (Médicins Sans Frontières 2007).

These MLWs are certificated individuals who have been trained to provide primary healthcare services at the community level, to diagnose and treat common health problems, to manage emergencies, to refer appropriately, and to transfer the seriously ill or injured for further care. These workers go by a range of titles depending on the country in which they work. In Myanmar, health assistants, lady health visitor, and midwives provide preventive services and the management of minor ailments; lady health workers in Pakistan provide maternal and child health services; técnicos de cirugía in Mozambique provide surgical care; clinical officers in East African countries provide diagnostic and treatment services at the

primary care level and in hospitals. In Nepal, health assistants staff health posts while auxiliary health workers are responsible for subhealth posts. In the United States, physician assistants provide a range of specific services (Lehmann 2008; McAuliffe et al. 2009).

The training of these MLWs is likely to vary considerably. Those who enter this work with pre-existing medical training, often nurses, are likely to require only 'top-up' training which is likely to be relatively brief in duration. Those without such a background receive training that usually ranges from 2 to 4 years including some that is field-based with supervision (Lehmann et al. 2008).

The third category, health workers with particular expertise in particular aspects of health delivery, represents an essential ingredient in the management of specific diseases, such as tuberculosis and malaria, for highly specific tasks such as the use of DOTS (Directly Observed Therapy, Short-course) in tuberculosis management, and in management of chronic diseases.

Building the human resource base for health delivery requires, in addition to physicians and nurses, these categories of health workers in order to provide the necessary services and to respond to changing health conditions in a country. Policies are therefore necessary to address the dynamics of the health labour market and the development and management of a skilled health workforce.

The challenges faced in optimizing these HRH are numerous and include carefully synchronizing a country's health workforce with its primary health needs and the inherent geographic variations in these needs (Dussault and Franceschini 2006).

With such training, health staff is often able to assume higher level responsibilities. In Rwanda, sets of tasks in HIV management have been transferred from physicians to nurses trained in HIV management. Subsequent analysis found that improvements in CD4 count, correct prescriptions, patient monitoring, and overall survival rates were similar between these nurses and the physicians they replaced in this work (AIDSMap 2010). In Malawi, clinical officers carry out nearly 95% of major obstetric operations in government hospitals with no significant differences in postoperative outcome, well-being, stillbirths, or neonatal mortality, compared with such operations carried out by medical officers. Evaluations

of the performance of MLWs in Mozambique showed no significant difference in outcome when compared to that of medical specialists.

Finally, there is often a critical shortage, particularly in developing countries of professionals trained in medical and nursing schools, of health managers and researchers trained in the schools of public health and of MLWs, lay community health workers, midwives, and other healthcare providers. (These shortages are also affected by those trained professionals who emigrate to higher income countries or seek employment in better paying private sector jobs.) Addressing these shortages often requires the attention and explicit involvement of the political leadership in these countries (Speybroeck et al. 2006).

Way forward and strategies to improve health policy and status of health in developing countries (recommendations)

This section recommends strategies to improve health policy and the overall status of health in developing countries. These recommendations are listed as follows and will be explored in turn.

Risk pooling to minimize catastrophic out of pocket expenditure
Forging and strengthening public-private partnerships
Evidence-based decision-making, strengthening of public providers in reporting and the introduction of updated technology
Decentralization and triage of workload, implementing proper referral system, empowering community healthcare and capacity building
Training, capacity building, and investment in infrastructure
Reducing disparities and inequity in minorities and other vulnerable population groups
Achieving Universal Health Coverage and the SDGs

Risk pooling, minimizing catastrophic out of programme experience (OOPE)

Risk pooling aims to achieve equitable access to health services with financial protection. It allows for redistribution of prepaid resources to individuals with the greatest health needs. The concept is that money flows from richer to poorer, and from healthier to sicker. When a poor or socially vulnerable person becomes ill or injured, the entire household can become

trapped in a downward spiral of lost income and high healthcare costs. Poor people are more vulnerable to this downward spiral as they are more prone to diseases and have more limited access to healthcare and social insurance. These catastrophic out of pocket payments arise under the conditions outlined in Figure 3.5.3.

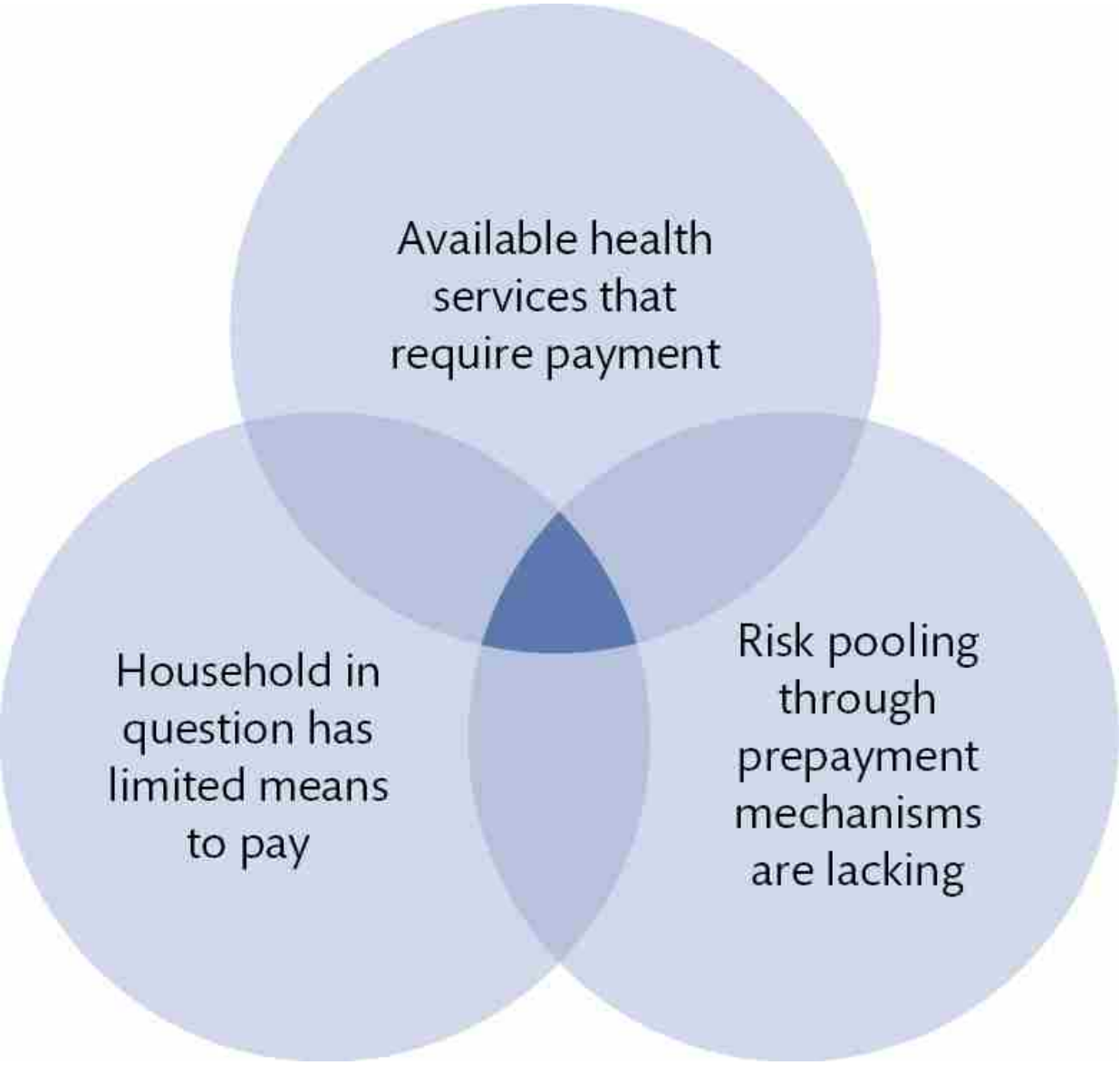


Fig. 3.5.3 Risk pooling.

As a consequence of inability to pay, households may refrain from using services. This will adversely impact their health, and this morbidity will adversely impact their earning potential. Reduced earnings have a knock on effect throughout the household. Not only the ill person is affected; their family and future generations can be affected too. High expenditure can result in reduced food and clothing purchases, and reduced investment in education, with associated sequelae.

Risk pooling addresses this point. Pooling of funds occurs through various prepayment mechanisms, which come in many forms. For instance there is tax-based financing of healthcare by governments, compulsory insurance contributions, external sources, and social health insurance. These mechanisms are common within developed nations, for instance the use of the Bismarck model in Europe. The Bismarck model uses an insurance system and is usually financed jointly by employers and employees through payroll deduction. Bismarck-type health insurance plans do not make a profit and must include all citizens (Wallace 2015). A larger and more diverse risk pool is more favourable. To create diversity compulsory participation in a health coverage scheme is required; if participation was to be voluntary there would be less likelihood of healthier individuals contributing, which ultimately would limit cross subsidy from healthy to sick and would undermine the financial stability of the system (OECD 2003).

When risk pooling, care needs to be taken to avoid fragmentation which can result in system inefficiency. Fragmentation is noted to take multiple forms, for instance local government health agencies, externally funded vertical programmes and overlapping and uncoordinated projects between different funding agencies such as Ministry of Health and health insurance funds.

Ultimately, health systems requiring lower out-of-pocket payments for healthcare offer better protection to the poor against catastrophic spending. Prepayment entails predictable expenditure for a given set of services. This allows opportunity to budget for healthcare and share risk with those more able to pay (e.g. as a ‘Sin Tax’; see Case Study 3.5.2). There is a reduced need for sudden acquisition of funds, and the payments required are reduced.

It is important to note that a prepayment model will not eliminate catastrophic out of pocket payments, as other factors; for instance health

services offered and copayments; need to be accounted for. Additionally, catastrophic expenditures do not automatically disappear with rising income.

Case Study 3.5.2 Philippines Sin Tax

In 1995, the Government of the Philippines set up the National Health Insurance Programme 'PhilHealth'. All Filipinos are automatically enrolled in and therefore entitled to benefits of the programme. There are formal and informal groups, and the formal are obliged to pay a premium, which essentially cross subsidizes other members. All senior citizens are mandatorily covered.

PhilHealth is mandated with improving access to quality care and reducing OOPE. To achieve this, it introduced a 'Sin Tax' in 2012. The basic premise is that taxes on alcohol and tobacco are earmarked for expenditure, mostly on healthcare:

- 85% of revenues are allocated for health, of which:
- 80% are for achieving UHC
- 20% for medical assistance and improvement of health facilities

Within its first year, Sin Tax raised more than 1.2billion USD, allowing enrolment of an additional 14 million families—totalling 45 million Filipinos—into PhilHealth.

What can be learned from PhilHealth?

PhilHealth is an example of a sustainable financing mechanism for healthcare, providing an alternative to out of pocket payments. Research suggests that even at low levels of public spending, countries can make significant steps towards UHC. Revenue raising needs to be coupled with the pooling of funds and the purchasing of health services; it is the combination of reforms which drive improvements in health system performance (Jowett et al. 2016).

The Programme has confirmed the need both to understand history and local environment, and to use a clear technical framework when analysing and assessing key drivers and impediments towards achieving UHC. There will be a struggle to attain UHC without a basic sense of solidarity in the community.

Criticism

While the country's UHC agenda is commendable for its pro-poor approach, other PhilHealth members must be kept in mind and feel the benefits of PhilHealth to ensure political support and stability. In this regard PhilHealth and the accredited hospitals need further strengthening to increase patients' access to and utilization of quality health services and establish transparent and fair provider payment processes.

Additionally, facilities need to be technically and managerially competent and empowered to respond to PhilHealth in its role as a powerful purchaser. Facilities should be held accountable for the delivery and quality of care that they provide.

Private–public partnership and strengthening

It is widely accepted that private–public partnerships are essential in the journey towards Universal Health Coverage (UHC). Governments are encouraged to recognize that they cannot achieve UHC on their own (OECD 2003).

UHC requires a resilient health system, one that is strong and responsive. Engagement between public and private sectors to achieve these qualities can take multiple forms, for instance capacity development, innovations, outsourcing, and consulting-all with varying degrees of commitment and financial risk sharing. Engagement of the private sector should ideally maximize strategic areas such as those shown in Figure 3.5.4.

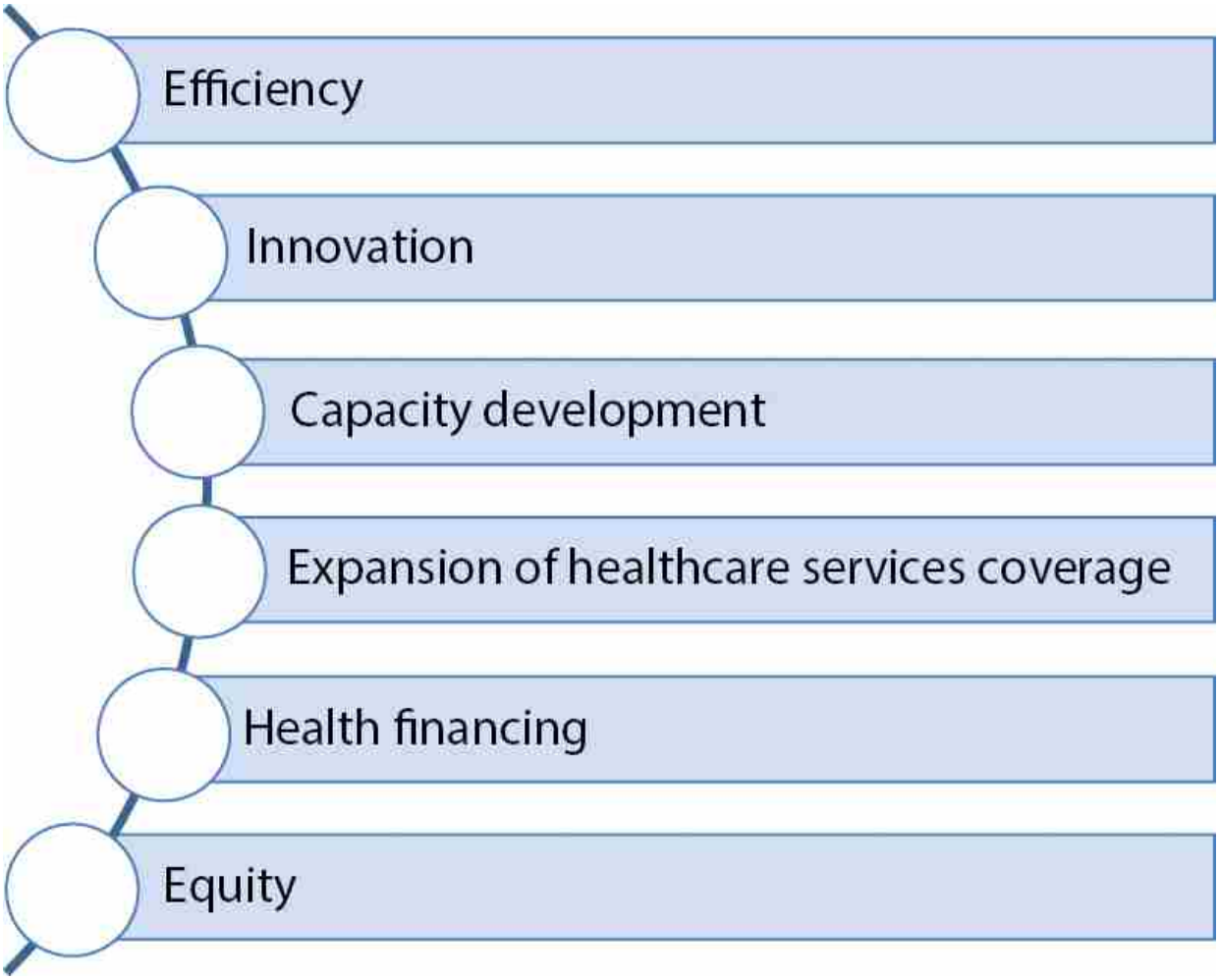


Fig. 3.5.4 Strategic areas to the engagement of the private sector.
Source: data from Ota, M.O.C., et al. (2018) 'Proceedings of the first African Health Forum: effective partnerships and intersectoral collaborations are critical for attainment of Universal Health Coverage in Africa'. *BMC Proceedings*, 12 (Suppl 7):8, 1–8.

Ultimately, governments can use the private sector to increase healthcare service provision. When embarking on these partnerships governments need to ensure that there are adequate regulatory frameworks, policies, and protocols in place to facilitate good communication across sectors. Poor communication may hinder engagement which subsequently risks poor understanding of government policies and procedures. In addition to communication, central to fostering enduring private–public partnerships is building trust between sectors. The potential benefits of collaboration are vast, and it would be a pity to lose out on such due to a failure by policymakers to ensure this trust. This could be facilitated by an organizational culture of openness and sharing.

The private sector has different needs to the public sector, which entails need for regulatory frameworks and accreditation systems to aid collaboration and foster partnerships. Literature suggests that development of protocols for communication across sectors could make coordination of projects smoother and more efficient (Suchman et al. 2018). To this effect, the private sector should be afforded regular opportunity to input into policy design, and both public and private entities should have standardized processes in place for when they collaborate. Shared goals, mutual understanding and structured ongoing communication that facilitates a strong working relationship when collaborating is crucial in making these endeavours a success (Baicker and Chandra 2017).

Considering achievement of the MDGs, a challenge was the limited focus on interventions that affect health however are not within the remit or oversight of the respective Ministry of Health, such as infrastructure. What should be learned from this in pursuit of UHC and the SDGs is the importance of partnerships and intersectoral collaboration (Garg 2018).

Of course private–public partnerships are not without challenges. Literature to date has highlighted a number of areas for improvement, notably:

Lack of information sharing
Weaknesses in management capacity
Funding insecurity
Mismatched organizational styles and differing priorities
Corruption (Suchman et al. 2018)

Evidence-based decision-making, strengthening of public providers in reporting, introduction of updated technology

Evidence-based policymaking has a fundamental role in improving the health status of countries. Evidence facilitates decision-making for the achievement of goals. Goals need to be clear, and it must be accepted that a rich body of evidence does not guarantee that these goals will be achieved. Often, the process of evidence-based policymaking will be uncertain (Baicker and Chandra 2017). As important as evidence is to sound policy choices, goals are determined based on values and priorities—a given body of evidence can be used to support very different policy positions.

The role of technology in healthcare improvement should not be underestimated. The literature highlights multiple applications:

Telemedicine
Training
Health education dissemination
Behaviour change promotion
Vaccination
Drug inventory and storage (Garg 2018)

Other applications of technology within developed nations are Electronic Patient Registries (EPR) and Clinical Decision Support Systems (CDSS). EPR benefits include greater accuracy, a higher proportion of correct information, time saved in locating information, more economical use of financial resources, greater ease, and speed of recovery of patient data. CDSS has been found to result in increased physician adherence to standardized treatment plans, cost reduction, easier standardization and regulation of requests for secondary and tertiary healthcare and examinations, thus reducing variability between services. From an administration point of view, it is possible to obtain greater adherence to public policies. Standardized programmes for earlier detection of diseases tend to have greater diagnostic value thereby contributing towards the promotion of equity, and the reduction of complications and costs related to more complex treatments (Tomasi et al. 2004).

Countries should adopt an open data sharing policy and develop a supportive legal framework to facilitate sharing and use of data in accordance with agreed standards for confidentiality and data sharing.

In general there is greater efficiency of management processes within the primary healthcare setting as a result of technology. Of course there is the issue of security and training to use systems. Ultimately, harnessing the opportunity that modern technology presents is key to improving the health status of countries.

Decentralization and triage of workload, implementing proper referral system, empowering community healthcare and capacity building

Health status and policy can be improved from the community level by empowering communities to take responsibility for their healthcare services.

According to the WHO,

> community empowerment is the process of enabling communities to increase control over their lives … It is the process by which they increase their assets and attributes and build capabilities to gain access, partners, networks, and/or a voice, in order to gain control … It assumes that people are their own assets and the role of the external agent is to catalyse, facilitate or accompany the community in acquiring power … Community empowerment implies community ownership and action that explicitly aims at social and political change … a process of re-negotiating power to gain more control … power is a central concept in community empowerment and health promotion invariably operates within the arena of a power struggle … seeks to build partnerships with other sectors in finding solutions.

The WHO concludes that empowerment can reduce health disparities (WHO 2018).

Globalization adds another dimension to the concept of community empowerment; the local and global settings are inextricably linked. Community empowerment recognizes and strategically acts upon this interlinkage and ensures that power is shared at both local and global levels.

Key to achieving community empowerment is communication. Approaches that encourage discussion and debate result in increased knowledge and awareness, and a higher degree of critical thinking. Critical thinking enables communities to understand the interplay of forces operating on their loves and helps them take their own decisions (WHO 2018).

Technology can help communities and individuals to develop opportunities, capacities and tools that benefit them and it can ensure that communities can mobilize targeted populations to obtain needed health resources by fostering awareness of a given problem (WHO 2018).

Policymakers must recognize and facilitate empowerment at a variety of levels. This can be done by involving community members and the underserved in the initiation and implementation of policy, as well as by relying on people who hold more traditional power. It is a responsibility of policymakers to achieve shared policymaking, to increase the sustainability of these advances towards improved health status and ultimately UHC (Thompson et al. 2016).

Training, capacity building, and investment in infrastructure

Central to improved health status and UHC is quality care. Quality matters as we are seeing a shift in disease burden to more complex conditions which require greater utilization of a health system and its services. These conditions are becoming an increasingly bigger influence on population

health, which makes addressing them an increasing priority. Countries require the appropriate infrastructure to meet quality standards, and governments are encouraged to establish a national quality guarantee for healthcare services. Literature states that poor quality care is now a bigger barrier to reducing mortality than insufficient access (Kruk et al. 2018).

For a health system to function and develop a quality service, there needs to be an adequate workforce. This can be achieved by increasing the number of training places for given professions, leveraging technology to provide required training, and ensuring respect for and good morale among staff. A concern however is brain drain. The workforce needs to be available, accessible, and have the capacity to deliver quality integrated care. Investment in primary healthcare is most needed and is recognized as the most cost-effective investment in improving equity in access to essential health services (Garg 2018).

Development agencies can play a significant role in increasing focus on human resource strategies—which are integrated into the health sector in wider civil service reforms—by:

Analysing the mix of personnel and skills
Reorienting training
Addressing pay and non-pay incentives
Improving job opportunities and conditions
Improving and expanding the use of information and communications technology in training and education programmes (OECD 2003)

Development agencies also have an interest in capacity building to enhance the effectiveness of their partnerships and to aid achievement of pro-poor health objectives.

Capacity building is necessary for multiple players within the healthcare arena. Taking research for example, it is crucial to align priorities and strategies of those investing in such, ensuring that programmes are coordinated and resources not wasted due to duplication. It is desirable to be able to detect and track research and innovations to facilitate progress. Health systems need to develop the capacity to measure and use data to learn. Additionally, countries should make it a priority to ensure sustainable financing mechanisms to optimize function. With investment in appropriate infrastructure and human resources, research priorities can be addressed (Ota et al. 2018).

Ultimately, capacity building should go beyond the health sector. It requires viewing pro-poor health approaches in a larger context of:

Political and economic restructuring
Fiscal policy
Administrative reform
Strengthening of participation and democratic systems (Ota et al. 2003)

Next there is investment in infrastructure. Looking at this from a health system perspective, the literature suggests various key facets of a high-quality health system. A commission by *The Lancet* concluded that high-quality health systems should be informed by four values—they are for people, equitable, resilient, and efficient (Kruk et al. 2018).

The WHO building blocks of a health system are shown in Figure 3.5.5.

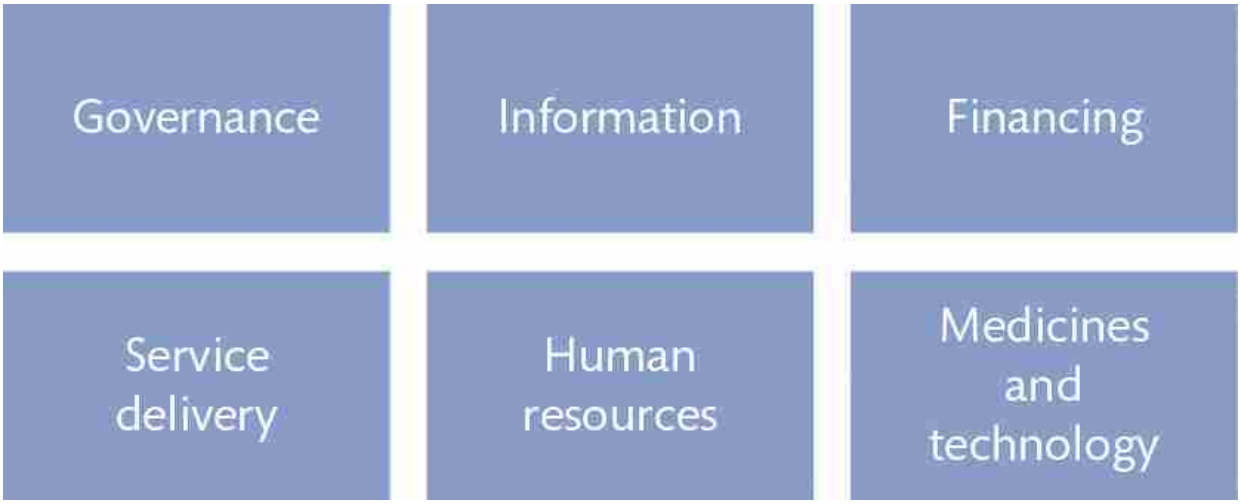


Fig. 3.5.5 System building blocks.
Source: data from World Health Organization (2010) Monitoring the building blocks of health systems: a handbook of indicators and their measurement strategies. Geneva, Switzerland: WHO.

It follows from this that strengthening these aspects will result in stronger health systems. Stronger health systems with strong infrastructure will help to provide quality care and UHC. By contrast, poor-quality health systems are posited to result in catastrophic expenditures and waste of resources.

The Lancet Global Health Commission 2018 endorses four universal actions to raise quality across health systems

Health system leaders need to govern for quality by adopting a shared vision of quality care, a clear quality strategy, strong regulation, and continuous learning. Ministries of health cannot achieve this alone and need to partner with the private sector, civil society, and sectors outside of health such as education, infrastructure, communication, and transport.
Countries should redesign service delivery to maximize health outcomes rather than geographical access to services alone.
Countries should transform the health workforce by adopting competency based clinical education, introducing training in ethics and respectful care, and respecting all workers to deliver the best care possible.
Governments and civil society should ignite demand for quality in the population to empower people to hold systems accountable and actively seek high-quality care (Kruk et al. 2018).

These improvement efforts should ideally be adapted for local context and monitored, in a bid to achieve sustainable change (Kruk et al. 2018).

Ultimately, national governments need to invest in high-quality health systems for their own people and make such systems accountable to people through legislation, education about rights, regulation, transparency and greater public participation (Kruk et al. 2018) Countries need to innovate to determine how they can efficiently utilize the available limited resources to provide access to essential healthcare services while exploring domestic sources of revenue to finance UHC. In achieving such, limited complexity and fragmentation is desirable. See also Case Study 3.5.3.

CASE STUDY 3.5.3 Ayushman Bharat programme, India

India is a vast, diverse, and a highly populated country. Its problems are complex and unique to itself. It is a challenge to reach out to vast different regions across India as local beliefs, culture, and practices influence every region's outlook.

On 1 February 2018, the Indian Government announced two major initiatives in the health sector, as part of the Ayushman Bharat programme.

The initiatives intervene to address health holistically at the primary, secondary and tertiary levels, and incorporate both disease prevention and health promotion. They are expanded on as follows:

1. Health and Wellness Centres
    a. India's National Health Policy 2017 envisaged Health and Wellness Centres as the foundation of India's health system, bringing the system and its services closer to the homes of people. The centres are intended to provide medical care to the poor, providing comprehensive healthcare ranging from diagnostic services to free essential drugs—including the realms of NCDs and maternal and child health. The national budget has set aside funds for this project, in addition to contribution of the private sector through CSR (?) and philanthropic institutions. To disseminate their services optimally, the centres must be accessible and appealing, providing a sense of comfort and belonging to the people of the regions where they are located.
2. National Health Protection Scheme
    a. This is the world's largest government funded healthcare programme Launched on 15th August 2018 and providing cover to more than 10 crore poor and vulnerable families. Coverage is up to 5 lakh rupees per family per year for secondary and tertiary care hospitalization. Adequate funds have been set aside to facilitate smooth implementation of the scheme. Benefits of the scheme are available across the country and a beneficiary covered under the scheme will be allowed to take cashless benefits from any public or enrolled private hospital.
    b. In partnership with think tank NITI Aayog, a robust, modular, scalable, and interoperable IT platform has been made operational, entailing a paperless, cashless transaction.

In order to further enhance accessibility of quality medical education and healthcare, 24 new Government Medical Colleges and Hospitals will be set up, by upgrading existing district hospitals in the country. This would ensure there is at least 1 medical college for every 3 parliamentary constituencies and at least 1 government medical college in each state of the country.

The aim of the Ayushman Bharat programme is to build a New India by 2022 and ensure enhanced productivity, well-being, and avert wage loss and impoverishment. These schemes will also generate jobs, especially for women.

Reducing disparities and inequity in minorities, vulnerable population

Health policy can be utilized strategically to target vulnerable population groups in society. Those identified as vulnerable populations are indigenous populations, minority and socially excluded groups, refugees, or displaced persons, the mentally or physically disabled and people living with HIV AIDS. These groups are among the poorest of the poor in many societies and require special attention in policy action for poverty reduction (OECD 2003).

A pro-poor health approach is one that gives priority to promoting, protecting, and improving the health of the poor. It builds on health systems, health financing and key policy areas beyond the health sector promoting policy coherence global public goods (OECD 2003).

Benefits of investing in the health of the poor include:

Higher labour productivity
Higher rates of domestic and foreign investment
Improved human capital
Higher rates of national saving
Demographic change (OECD 2003)

In addition to their beneficial macroeconomic impact, health improvements have intergenerational spillover effects (OECD 2003).

Gender inequality is a major determinant of poverty and ill health. Poor women and girls are worse off in relation to assets and entitlements; within the household and in society. Socio-cultural beliefs regarding the roles of men and women contribute to this inequality. Women are also major producers of healthcare through their roles as household managers and carers. But their health, including reproductive health suffers from inadequate nutrition, heavy workloads, and neglect of basic healthcare, factors aggravated by exposure to sexual abuse and interpersonal violence. Action on gender inequalities is therefore an essential element of a pro-poor approach to health.

It must be noted that focusing solely on the poor may not achieve desired results. Common shortcomings include that benefits packages with only limited services and high copayments. This supports focus on elderly, disabled, and those with chronic health diseases.

Ways to UHC and SDGs

UHC is a target of the 2015 SDGs. As defined by the WHO:

> UHC means that all individuals and communities receive the health services they need without suffering financial hardship. It includes the full spectrum of essential, quality health services, from health promotion to prevention, treatment, rehabilitation and palliative care. (WHO 2017)

It is widely accepted that countries who make progress towards UHC will as a consequence make progress towards other SDGs.

The rationale behind seeking UHC is avoidance of people being pushed into poverty due to unexpected illness that requires them to use up savings or sell assets. This can have catastrophic consequences for the current generation and indeed future generations. UHC exists to protect people from

adverse financial consequences of utilizing health services. Good health allows children to learn and adults to earn, helps people escape from poverty, and provides the basis for long-term economic development (WHO 2017).

Components of UHC are shown in Figure 3.5.6.

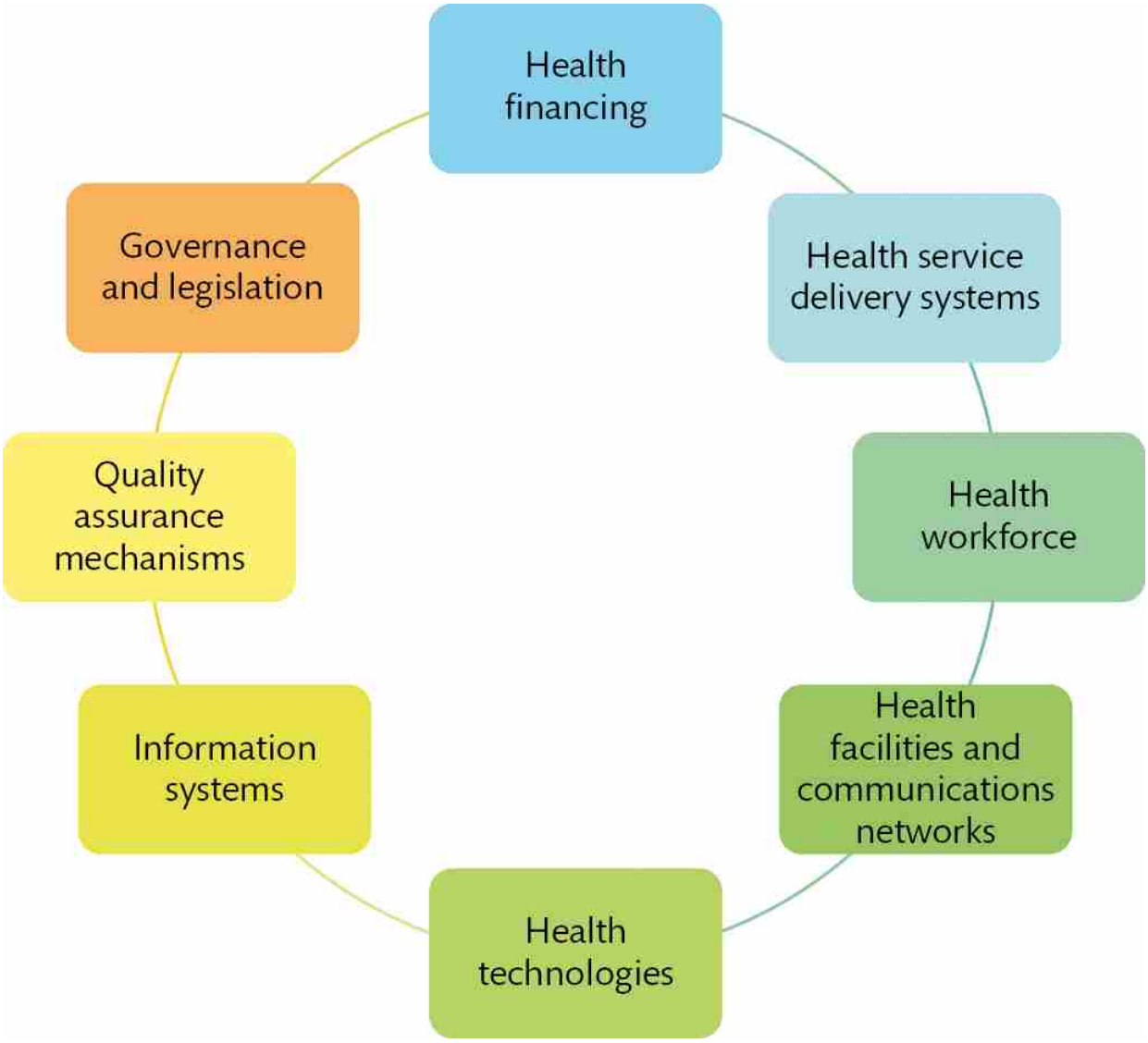


Fig. 3.5.6 Components of Universal Health Coverage.

Taking steps towards UHC means steps towards equity, development priorities, and social inclusion and cohesion. UHC requires quality care, a clear quality strategy, strong regulation, and continuous learning. Many

countries are already making progress towards UHC. All countries can take actions to accelerate the process or to maintain gains already made. Ways to UHC encompass what has already been explored in this chapter:

Strengthening health systems, ensuring robust financing and risk pooling. Limited complexity and fragmentation.
Workforce to improve health service coverage and health outcomes.
Good governance.
Sound systems of procurement.
Adequate supply of medicines—use of generics to ensure this.
Utilizing advances in health technologies.
Well-functioning HISs.
Addressing vulnerable groups.

(WHO 2017)

When developing healthcare services social and cultural norms need to be accounted for. Incorporating traditional medical services organized around the comprehensive needs and expectations of people and communities will empower communities to take a more active role in their health and health systems.

As developments in healthcare take place, it is essential that progress is measured. In the case of UHC this can be done by measuring:

Proportion of population that can access essential quality health services
Proportion of the population that spends a large amount of household income on health (WHO 2017)

Conclusion

An effective health system, especially in developing countries, needs strong health policy as a basic foundation. The chapter has highlighted the stages of health transition**,** which led to health activities and health policy development in both developed and developing countries.

Milestones of global health policy and global trends of health policy shifting were also described given its importance in changing global health trends from primary healthcare in the 1970s, global health initiatives in 1990s, the MDGs in 2000s, and then to current era of SDGs. How global

health policy and globalization impact and shape the health of people in developing countries were discussed through various health policies, health outcomes and achievements, and current health challenges and problems. The three case studies were chosen, namely Turkey's health system reform, the Philippines' Sin Tax, and the Ayushman Bharat programme, India, from which impressive information and inputs were received to be considered in formulation of health policies with a notion of country's adaptation wherever necessary.

Implementation of global policy as well as country's own health policy has challenges and pitfalls in various settings. It was explained within the context of health system building blocks and other related issues. Policy recommendations were made as way forward and strategies to improve health policy and status of health in developing countries in terms of; risk pooling, minimizing catastrophic OOPE; private–public partnership and strengthening; evidence-based decision-making, strengthening of public providers in reporting, introduction of updated technology; decentralization and triage of workload, implementing proper referral system, empowering community healthcare and capacity building; training, capacity building, and investment in infrastructure; reducing disparities and inequity in minorities, vulnerable population and ways to UHC and SDGs.

Policymaking for the effective, efficient, and equitable health systems in developing countries is more critical than ever to be realized through its optimal matrix of strategic plans and implementation modalities. New challenges are being addressed, including SDGs. We hope that this chapter will contribute the information for policymakers to be able to formulate not only knowledge-base, but also results-orient policy design.

REFERENCES

Adjuik, M., et al. (2006). Cause-specific mortality in sub-Saharan Africa and Bangladesh. *Bulletin of the World Health Organization*, **84**, 181–92.

Atkinson, S., et al. (2005). Prevention and promotion in decentralized rural health systems: a comparative study from northeast Brazil. *Health Policy Planning*, **20**, 69–79.

AIDSMap. *Task Shifting: Successes from Mozambique and Rwanda*, 5 September 2008. [Online] Available at: https://www.aidsmap.com/news/sep-2008/task-shifting-successes-mozambique-and-rwanda

Arya, N. and Santa Barbara, J. (2008). *Peace Through Health: How Health Professionals Can Work for a Less Violent World*. Sterling: Kumarian Press.

Bärnighausen, T. and Bloom, D.E. (2009). Financial incentives for return of service in underserved areas: a systematic review. *BMC Health Services Research*, **9**, 86.

Bala, M.M., Strzeszynski, L., and Cahill, K. (2013). Mass media interventions for smoking cessation in adults. *Cochrane Database of Systematic Review*, (6), CD004704.

Baicker, K. and Chandra, A. (2017). 'Evidence-Based Health Policy', *New England Journal of Medicine*, 377(25); 2413–2415.

Baldwin, W. (2015). *Population Health: Behavior and Social Science Insights: The Burden of Non-communicable Diseases in the Developing World*. Agency for Health Care Research and Quality. [Online] Available at: https://www.ahrq.gov/professionals/education/curriculumtools/population-health/baldwin.html

Belaid, L., and Ridde, V. (2012). An implementation evaluation of a policy aiming to improve financial access to maternal health care in Djibo district, Burkina Faso. *BMC Pregnancy and Childbirth*, **12**, 143.

Bhutta, Z., et al, 'Evidence-Based Interventions.' *The Lancet*, vol. 382, no. 9890 http://www.thelancet.com/journals/lancet/issue/vol382no9890/PIIS0140-6736(13)X6037-4, 2013, pp. 452–477.

Bhutta ZA, Ahmed T, Black RE, et al, for the Maternal and Child Undernutrition Study Group. What works? Interventions for maternal and child undernutrition and survival. Lancet 2008; **371**: 417–40.

Black, R., et al., 'Maternal and Child Undernutrition and Overweight in Low-Income and Middle-Income Countries,' *The Lancet*, vol. 382, no. 9890 http://www.thelancet.com/journals/lancet/issue/vol382no9890/PIIS0140-6736(13)X6037-4, 2013, pp. 427–451.

Black RE, Allen LH, Bhutta ZA, et al, for the Maternal and Child Undernutrition Study Group. Maternal and child undernutrition: global and regional exposures and health consequences. *Lancet* 2008; **371**: 243–60.

Blake, N.M. (1956). *Water for the Cities: A History of the Urban Water Supply Problem in the United States*. Syracuse, NY: Syracuse University Press.

Caldwell, J.C. (2001). Population health in transition. *Bulletin of the World Health Organization*. **79**(2), 159–60.

Campbell, O.M.R. What are maternal health policies in developing countries and who drives them? A review of last-half century. [Online] Available at: https://www.gov.uk/dfid-research-outputs/what-are-maternal-health-policies-in-developing-countries-and-who-drives-them-a-review-of-the-last-half-century

Chongsuvivatwong, V., et al. (2011). Health and health-care systems in southeast Asia: diversity and transitions. *The Lancet*, **377**, 429–37.

Chave, S.P.W. (1984). The origins and development of public health. In: Detels, R., et al. (eds.) *Oxford Textbook of Public Health, Vol. 1: History, Determinants, Scope, and Strategies*. New York: Oxford University Press.

Chaulagai, C.N., et al. (2005). Design and implementation of a health management information system in Malawi: issues, innovations and results. *Health Policy Planning*, 20(6), 375–84.

Chou, D., Daelmans, B., Jolivet, R.R., Kinney, M., Say, L. (2015). Ending preventable maternal and newborn mortality and stillbirths. *British Medical Journal*. 351: h4255.

Collier, Paul. (2007). The Bottom Billion: Why The Poorest Countries Are Failing and What Can Be Done About It. Oxford: Oxford University Press.

Coovadia et al. (2009). The health and health system of South Africa: historical roots of current public health challenges. *Lancet*, 374: 817–34.

Cornia Giovanni, A. (2001). Globalization and health: results and options. *Bulletin of the World Health Organization*, 79(9), 834–41.

de Neuezer, T.M., et al. (2001). Obesity and undernutrition in a very low income population in the city of Maceid, northern Brazil. *British Journal of Nutrition*, **86**, 277–88.

Demeny, P. (1968). Early fertility decline in Austria-Hungary: a lesson in demographic transition. *Daedalus*, **97**, 502–22.

Dikid, T., Jain, S.K., Sharma, A., Kumar, A., and Arain, J.P. (2013). Emerging and re-emerging infection in India: An overview. *Indian Journal of Medical Research*, **138**, 19–31.

Díaz-Bonilla E., et al. (2002). Globalizing Health Benefits for Developing Countries. No. 108 TMD Discussion Papers, International Food Policy Research Institute (IFPRI). [Online] Available at: https://EconPapers.repec.org/RePEc:fpr:tmddps:108

Dussault, G. and Franceschini, M.C. (2006). Not enough there, too many here: understanding geographical imbalances in the distribution of the health workforce. *Human Resources and Health*, **4**, 12.

Eberhardt, M.S. (2001). *Urban and Rural Health Chartbook*. Hyattsville, MD: National Center for Health Statistics.

Ekouevi, D.K., Karcher, S., and Coffie, P.A. (2011). Strengthening health systems through HIV monitoring in Sub-Saharan Africa. *Current Opinion in HIVAIDS*, 6(4), 245–50.

Elder, R.W., et al. (2004). Effectiveness of mass media campaigns for reducing drinking and driving and alcohol-involved crashes: a systematic review. *American Journal of Preventive Medicine*, **27**, 57–65.

Ellen M.E., Lavis, J.N., Horowitz, E., and Berglas, R. (2018). How is the use of research evidence in health policy perceived? A comparison between the reporting of researchers and policy-makers. *Health Research Policy and Systems*, **16**, 64.

Erasmus, E. and Gilson, L. (2008). How to start thinking about investigating power in the organizational settings of policy implementation. *Health Policy Planning*, **23**, 361–8.

Erasmus, E., Gilson, L., Govender, V., and Nkosi, M. (2017). Organisational culture and trust as influences over the implementation of equity-oriented policy in two South African case study hospitals. *International Journal for Equity in Health*, 16(1), 164.

Fee, E. (1987). *Disease and Discovery: A History of the Johns Hopkins School of Hygiene and Public Health 1916–1939*. Baltimore, MD: Johns Hopkins University Press.

Field JO, Levinson FJ. (1975). Nutrition and development: dynamics of political commitment. *Food Policy*, **1**(1):53–61.

Forsbereg, Le Thanh (2011). *The Political Economy of Health Care Reform in Vietnam*. Oxford-Princeton Global Leaders Program, Working Paper.

Garg, S. (2018). 'Universal health coverage in India: Newer innovations and the role of public health.', *Indian Journal of Public Health*, **62**(3); 167–170.

Green A, et al. (2011). Health policy processes in maternal health: a comparison of Vietnam, India and China. *Health Policy*, 100(2–3), 167–73.

Guha-Sapir, D. and D'Aoust, O. (2010). 'Demographic and health consequences of civil, LT 2011, 'Political Economy of Health care reform in Vietnam', paper presented in Oxford symposium on urbanization, health and human security: Challenges conflict', *World Development Report*, pp. 12–34.

Gilligan J. (2000). Violence in public health and preventive medicine. *Lancet*, **355**, 1802–4.

Gilson, L., Schneider, H., and Orgill, M. (2014). Practice and power: a review and interpretive synthesis focused on the exercise of discretionary power in policy implementation by front-line providers and managers. *Health Policy Planning*, **29**, iii51–69.

Gilson, L., Schneider, H., and Orgill, M. (2014). Practice and power: a review and interpretive synthesis focused on the exercise of discretionary power in policy implementation by front-line providers and managers. *Health Policy Planning*, **29**, iii51–69.

Grilli, R., Ramsay, C., and Minozzi S. (2002). Mass media interventions: effects on health services utilization. *Cochrane Database of Systematic Review*, (1), CD000389.

Goudsblom, J. (1986). Public health and the civilizing process. *The Milbank Quarterly*, 64(2), 161–88.

Grépin, K.A., et al. (2017). Donor funding health policy and systems research in low- and middle-income countries: how much, from where and to whom. *Journal of Health Services Research & Policy*, **15**, 68.

Hafner. T. and Shiffman, J. (2013). The emergence of global attention to health systems strengthening. *Health Policy Planning* 28: 41–50.

Ha, B.T., Green, A., Gerein, N., and Danielsen, K. (2010). Health policy processes in Vietnam: a comparison of three maternal health case studies. *Health Policy*, 98(2–3), 178–85.

Hanlon, G. and Pickett, J. (1984). *Public Health Administration and Practice*. St Louis, MO: Times Mirror/Mosby.

Harries, A.D., et al. (2011). Operational research in Malawi: making a difference with cotrimoxazole preventive therapy in patients with tuberculosis and HIV. *BMC Public Health*, **11**, 593–10.

Heymann. D.L. (2009). *Emerging and Re-Emerging Infections*. Oxford: Oxford University Press, pp. 1264–76.

Horton. et al. (2013). "The Economic Rationale for Investing in Stunting Reduction." GCC Working Paper Series, GCC 13–08.

Jørn, B., et al. (2007). Health information systems in developing countries. *MIS Quarterly*, **31**, 1–XXX.

Kapoor, S.K. and Anand, K. (2002). Nutritional transition: a public health challenge in developing countries. *Journal of Epidemiological Community Health*, **56**, 804–805.

Kaldor, M. (2007). *Human Security: Reflections on Globalization and Intervention*. Cambridge: Polity Press.

Kawonga, M., Blaauw, D., and Fonn, S. (2012). Aligning vertical interventions to health systems: a case study of the HIV monitoring and evaluation system in South Africa. *Health Research Policy Systems*, **10**, 2.

Keni, B.H. (2006). Training competent and effective primary health care workers to fill a void in the outer islands health service delivery of the Marshall Islands of Micronesia. *Human Resources and Health*, **4**, 27.

Khan, M.S., et al. (2018). How do external donors influence national health policy processes? Experiences of domestic policy actors in Cambodia and Pakistan. *Health Policy Planning*, **33**, 215–23.

Koduah, A., van Dijk, H., and Agyepong, I.A. (2015). The role of policy actors and contextual factors in policy agenda setting and formulation: maternal fee exemption policies in Ghana over four and a half decades. *Health Research Policy and Systems*,13, 27.

Kopel, D.B. (1995). Guns, germs and science: public health approaches to gun control. *Journal of the Medical Association of Georgia*, 84(6), 269–73.

Krause, K., Muggah, R., and Gilgen, E. (2011). *Global Burden of Armed Violence 2011: Lethal Encounters*. Cambridge: Cambridge University Press.

Kruk, M.E., Gage, A.D., Arsenault, C., Jordan, K., Leslie, H.H., Roder DeWan, S., et al. (2018). 'High-quality health systems in the Sustainable Development Goals era: time for a revolution', *The Lancet Global Health*, 6(11); PE1196-E1252

Krug, E., Dahlberg, L., Mercy, J., Zwi, A., and Lozano, R. (2002). *World Report on Violence and Health*. Geneva: World Health Organization.

Kovačić, N. (2014). Globalization and the impact of globalization on the health industry. *Interdisciplinary Management Research*, **10**, 684–95.

Lavis, J., et al. (2005). Towards systematic reviews that inform health care management and policy-making. *Journal of Health Services Research & Policy*, 10 Suppl 1, 35–48.

Lee, K. (2001). A dialogue of the deaf? The health impacts of globalisation. *Journal of Epidemiology and Community Health*, 55(9), 619.

Lee, K. (2000). The impact of globalization on public health: implications for the UK faculty of Public Health Medicine. *Journal of Public Health and Medicine*, 22(3), 253–62.

Lemard, G. (2004). *Violence as a Global Health Challenge: A Case Analysis of Jamaica Using an Interdisciplinary Approach*. PhD International Studies, University of Miami, FL.

Lehmann, U. (2008). *Mid-level Workers. The State of the Evidence on Programmes, Activities, Costs and Impact on Health Outcomes. A Literature Review*. Geneva: World Health Organization.

Lehmann, U., Dieleman, M., and Martineau, T. (2008). Staffing remote rural areas in middle- and low-income countries: a literature review of attraction and retention. *BMC Health Services & Research*, **8**, 19.

Levinson, F. J. and Yarlini Balarajan, *Addressing malnutrition multi-sectorally. What have we learned from recent international experience?* United Nations, 2013.

London, J.D. (2008). Reasserting the state in Vietnam health care and the logics of market-Leninism. *Policy and Society*, **27**, 115–28.

Martens, P. (2002). Health transitions in a globalizing world: towards more disease or sustained health?. *Futures*, 37,635–48.

Maternal Health Task Force (n.d.). *Ending Preventable Maternal Mortality*. [Online] Available at: https://www.mhtf.org/projects/ending-preventable-maternal-mortality/

Maillacheruvu, P. and McDuff, E. (2014). South African's return to primary care: the struggles and strides of primary health care system. *Journal of Global Health*. [Online] Available at: https://www.ghjournal.org/south-africas-return-to-primary-care-the-struggles-and-strides-of-the-primary-health-care-system/

Mate, K.S., Bennett, B., Mphatswe, W., Barker, P., and Rollins, N. (2009). Challenges for routine health system data management in a large public programme to prevent mother-to-child HIV transmission in South Africa. *PLoS One*, 4(5), e5483.

Mayosi, Bongani M and Benatar, Solomon R. Health and health care in South Africa—20 years after Mandela; NEJMsr1405012, 2014 Oct 2;371(14):1344–53.

Mphatswe, W., et al. (2012). Improving public Health information: a data quality intervention in KwaZulu-Natal, South Africa. *Bulletin of the World Health Organization*, 90(3), 176–82.

Mpofu, M., et al. (2014). Strengthening monitoring and evaluation (M&E) and building sustainable health information systems in resource limited countries: lessons learned from an M&E task-shifting initiative in Botswana. *BMC Public Health*, **14**, 1032.

Morens, D.M. and Fauci, A.S. (2013). Emerging infectious diseases: threat to human health and global stability. *PLoS Pathogens*, **9**, e1003467.

McAuliffe, E., et al. (2009). Measuring and managing the work environment of the mid-level provider—the neglected human resource. *Human Resources for Health*, **7**, 13.

Médicins Sans Frontières (2007). *Help Wanted. Confronting the Health Care Worker Crisis to Expand Access to HIV/AIDS Treatment*. Johannesburg: Médicins Sans Frontières.

Notestein, F. W. (1945). Population — The Long View. In Theodore W. Schultz, Ed., Food for the World. Chicago: University of Chicago Press.

NIH Curriculum Supplement Series (2007). National Institute of Health (US); Biological science curricular study. *Understanding Emerging and Re-emerging Infectious Diseases*. [Online] Available at: https://www.ncbi.nlm.nih.gov/books/NBK20370/

Nathan, L. (2005). *The Frightful Inadequacy of the Statistics: A Critique of Collier and Hoeffler on the Causes of Civil War, Discussion paper no. 11*. London: Crisis States Research Network.

OECD (2003). "Investing in Health to Reduce Poverty", in Development Co-operation Report 2002: *Efforts and Policies of the Members of the Development Assistance Committee*, O Publishing, Paris.

OECD (2009). *Armed Violence Reduction: Enabling Development*. Paris: Organisation for Economic Co-operation and Development. [Online] Available at: https://www.oecd.org/dac/conflict-fragility-resilience/docs/armed%20violence%20reduction.pdf

Organisation for Economic Co-operation and Development (2003). DAC Guidelines and Reference Series- Poverty and Health: OECD Publications.

Omran, A. R. (1971). The epidemiologic transition: a theory of the epidemiology of population change. Milbank Memorial Fund Quarterly, **49**, 509–38.

Omran, A.R. (1998). The epidemiological transition revisited thirty years later. *World Health Statistics Quarterly*, **51**, 99–191.
Orach, C.G. (2009). Health equity: challenges in low income countries. *African Health Sciences*, **9**, 49–51.
Orton, L., Lloyd-Williams, F., Taylor-Robinson, D., O'Flaherty, M., and Capewell, S. (2011). The use of research evidence in public health decision-making processes: systematic review. *PLoS One*, **6**, e21704.
Owen, A., Wu, S., and Owen, A.L. (2007). Is trade good for your health ? Is trade good for your health? *Review of International Economics*, **15**(4), 660–82.
Pang, T. and Guindon, G.E. (2004). Globalization and risks to health. *EMBO Report*, **5**, S11–16.
Prabhat, J., et al. (2002). *Improving Health Outcomes of the Poor: The Report of Working Group 5 of the Commission on Macroeconomics and Health*. [Online] Available at: https://apps.who.int/iris/handle/10665/42488
Rosenkrantz, B.G. (1972). Public health and the State; changing views in Massachusetts, 1842–1936. Cambridge, Harvard University Press.
Roland (2017). Social Media, Health Policy, and Knowledge Translation; J Am Coll Radiol 2018;15:149–152. Crown Copyright © 2017 Published by Elsevier on behalf of American College of Radiology.
Ruger, J.P. (2004). Combating HIV/AIDS in developing countries. *British Medical Journal*, **329**, 121.
Ruel MT, Alderman H, and the Maternal and Child Nutrition Study Group. Nutrition-sensitive interventions and programmes: how can they help to accelerate progress in improving maternal and child nutrition? *Lancet* 2013; published online June 6. http://dx.doi.org/10.1016/S0140-6736(13)60843–0.
Sarma, N. (2017). Emerging and re-emerging infectious diseases in Southeast Asia. *Indian Journal of Dermatology*, **62**, 451–5.
Singh, M., Chakraborty, A., Kumar, S., and Kumar, A. (2017). The epidemiology of Dengue viral infection in developing countries: a systematic review. *Journal of Health Research and Reviews*, **4**, 104–7.
Shiffman, J. and Smith, S. (2007). Generation of political priority for global health initiatives: a framework and case study of maternal mortality. *The Lancet*, **370**, 1370–9.
South African Department of Health (2007). *HIV & AIDS and STI Strategic Plan for South Africa, 2007–2011*. Pretoria: Department of Health.
Staton, F.S., Bhosle, M.J., Camacho, F.T., Feldman, S.R., and Balkrisnan, R. (2007). How PAs improve access to care for the underserved. *JAAPA*, **20**, 32–36.
Speybroeck, N., Kinfu, Y., Dal Poz, M.R., and Evans, D.B. (2006). *Reassessing the Relationship Between Human Resources for Health, Intervention Coverage and Health Outcomes. Evidence and Information for Policy*. Geneva: World Health Organization. [Online] Available at: https://www.who.int/hrh/documents/reassessing_relationship.pdf
Suchman, L., Hart, B.A. and Montagu, D. (2018). 'Public-private partnerships in practice: collaborating to improve health finance policy in Ghana and Kenya.', *Health Policy and Planning*, 33(7); 777–785.
Tollman, S.M., Kahn, K., Sartorius, B., Collinson, M.A., Clark, S.J., and Garenne, M.L. (2008). Implications of mortality transition for primary health care in rural South Africa: a population based surveillance study. *The Lancet*, **372**, 893–901.
Thompson, B., Molina, Y., Viswanath, K., Warnecke, R. and Prelip, M.L. (2016). 'Strategies to Empower Communities to Reduce Health Disparities', *Health Aff (Millwood)*, 35(8); 1424–1428.
Tomasi, E., Facchini, L.A. and de Fatima Santos Maia, M. (2004). 'Health information technology in primary health care in developing countries: a literature review', *Bulletin of the World Health Organisation*, 82(11); 867–874.

Turner, J.B. (ed.). (1977). *Encyclopedia of Social Work* (17th ed). Washington, DC: National Association of Social Workers.

Tulchinsky, T.H. and Varavikova, E.A. (2008). *The New Public Health* (2nd ed). San Diego, CA: Elsevier Academic Press, Chapter 15.

Lehmann, U. and Gilson, L. (2013). Actor interfaces and practices of power in a community health worker programme: a South African study of unintended policy outcomes. *Health Policy Planning*, **28**, 358–66.

USAID (2007). *Data Quality Assurance Tool for Program-Level Indicators.* [Online] Available at: https://www.measureevaluation.org/resources/publications/ms-07-19

UNAIDS (2010). *Progress Report of the National Response to the 2001 Declaration of the Commitment on HIV and AIDS: Botswana Country Report, 2010*. [Online] Available at: https://data.unaids.org/pub/report/2010/botswana_2010_country_progress_report_en.pdf

UNAIDS (2018). Fact Sheet-July 2018; 2017 global HIV statistics. [Online] Available at: http://www.unaids.org/sites/default/files/media_asset/UNAIDS_FactSheet_en.pdf

Valenti, M., Ormhaug, C.M., Mtonga, R.E., and Loretz, J. (2007). Armed violence: a health problem, a public health approach. *Journal of Public Health Policy*, **28**, 389–400.

Victora., et al. 'Maternal and Child Undernutrition and Overweight in Low-Income and Middle-Income Countries,' *The Lancet*, http://www.thelancet.com/journals/lancet/issue/vol382no9890/PIIS0140-6736(13)X6037-4. 382, no. 9890, 2013, pp. 427–451.

Watts, G. (2009). The health benefits of tackling climate change—an executive summary for *The Lancet* Series. *The Lancet*, 1–8. [Online] Available at: https://www.thelancet.com/pb/assets/raw/Lancet/stories/series/health-and-climate-change.pdf

Wallace, L.S. (2013). 'A View of Healthcare Around the World', *Annals of Family Medicine*, 11(1), 84.

Wallace, J., Nwosu, B., and Clarke, M. (2012). Barriers to the uptake of evidence from systematic reviews and meta-analyses: a systematic review of decision makers' perceptions. *BMJ Open*, **2**, e001220.

Watts N, et al. (2015). The Lancet Commissions Health and climate change: policy responses to protect public health. *The Lancet*, 386(15). doi.org/10.1016/S0140-6736(15)60854-6.

Welander, A., Lyttkens C, and Nilsson T. (2014). *Globalization and Child Health in Developing Countries: The Role of Democracy*. [Online] Available at: http://www.ifn.se/wfiles/wp/wp1016.pdf

Winslow, C.E.A. (1923). *The Evolution and Significance of the Modern Public Health Campaign.* South Burlington, VT: Journal of Public Health Policy.

Whitehead, M., Petticrew, M., Graham, H., Macintyre, S.J., Bambra, C., and Egan, M. (2004). Evidence for public health policy on inequalities: 2: assembling the evidence jigsaw. *J Epidemiol Community Health*, **58**, 817–21.

World Health Organization. (2015). Trends in maternal mortality: 1990–2010 WHO, UNICEF, UNFPA and The World Bank estimates. WHO Press, Geneva, Switzerland. [Online] Available at: https://www.afro.who.int/sites/default/files/2017-05/trends-in-maternal-mortality-1990-to-2015.pdf

World Health Organization. Regional Office for the Western Pacific. (2016). Sustainable development goals (SDGs): Goal 3. Target 3.1: By 2030, reduce the global maternal mortality ratio to less than 70 per 100,000 live births [poster]. Manila, WHO Regional Office for the Western Pacific. [Online] Available at: https://apps.who.int/iris/handle/10665/208272

Wohl, A.S. (1983). *Endangered Lives: Public Health in Victorian Britain*. Cambridge, MA: Harvard University Press.

World Health Organization (2018). Non-communicable diseases, key facts. [Online] Available at: http://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases

World Health Organization (2017). WHO HIV updates 2017—Global and regional epidemic; progress and cascade; policy updates and accountability. [Online] Available at: https://www.who.int/hiv/pub/progressreports/en/

Worldwatch Institute (2018). Study highlights four key health challenges in developing countries; China struggling with all. [Online] Available at: http://www.worldwatch.org/study-highlights-four-key-health-challenges-developing-countries-china-struggling-all

World Health Organization (2000). *World Health Report 2000: Health Systems—Improving Performance*. Geneva: World Health Organization. [Online] Available at: http://www.who.int/whr/2000/en/

World Health Organization (2014). *Health Statistics and Information Systems*. [Online] Available at: https://www.who.int/healthinfo/facility_information_systems/Facility_Community_Data_Toolkit_final.pdf?ua=1

World Health Organization (2008). *Tobacco Use and its Impact on Health* [Online] Available at: http://www.who.int/tobacco/publications/gender/en_tfi_gender_women_prevalence_tobacco_use.pdf

World Health Organization WHO 2008 (2008). *Task Shifting: Global Recommendations and Guidelines*. Geneva: World Health Organization.

Yamada, T., Chen, C.C., Chiu, I.M., and Rizvi, S.W. (2013). Non-communicable diseases in developing countries: causes and health policy/ program assessments. *Journal of Tropical Diseases, 1*, 117.

3.6

Public health policy in developed countries

John Powles † *and Hebe Gouda*

> ... the best strategy to identify the social and political processes that drive change in public health policies is by cross-national comparison. (Nathanson 2009)

The nature and scope of public health policy

The scope and purpose of public health policy may be seen as implicit in widely used definitions of public health. Winslow's definition, as adapted by the Acheson Report in England, is: '*The science and art of preventing disease, prolonging life and promoting health through organised efforts of society*' (Secretary of State for Social Services 1988).

Public health policies might thus be thought of as the policies that guide these 'organized efforts' to protect and improve health. The scope of such policies depends a good deal, however, on what is considered to be entailed by 'organized efforts' and on how centrally 'organized efforts' are understood to be related to efforts that are more decentralized, more informal, less organized, perhaps even 'spontaneous'. The relative importance and legitimacy of centralized versus decentralized uses of knowledge in protecting and enhancing health is a common underlying theme in discussions of public health policy.

We need first to define our geopolitical scope by deciding which countries are currently classifiable as 'developed'. Many answers are clearly possible, but for the purposes of this chapter the boundary will be drawn as follows: the 33 members of the Organisation for Economic Co-operation and Development (OECD) not qualified as 'emerging' (OECD n.d.) plus 5 member states of the EU that are not included in the OECD list (indicated by (EU), plus Singapore. These 39 countries are grouped as follows: 21 in Western Europe (Austria, Belgium, Cyprus (EU)), Denmark, Finland,

France, Germany, Greece, Iceland, Ireland, Italy, Luxembourg, Malta (EU), Netherlands, Norway, Portugal, Spain, Sweden, Switzerland, United Kingdom (plus Israel); 11 in Eastern Europe (Bulgaria (EU), Croatia (EU), Czech Republic, Estonia, Hungary, Latvia, Lithuania, Poland, Romania (EU), Slovakia, Slovenia); 3 in Asian Developed (Japan, South Korea and Singapore); 3 in Other Developed (Canada, Australia, New Zealand). To reduce potential heterogeneity in the last group, and because of its geopolitical importance, we treat the United States (USA) separately.

As a topic, policy gains interest from the assumptions that it guides action and that deliberate actions to protect and enhance health are major determinants of actual health trends. A brief review of health trends sets the context for our consideration of health policy.

Context: survival trends since 1950

The unit of analysis for policy is typically the nation state. Estimates are available from the Global Burden of Disease project for trends in child mortality (risk of death before the fifth birthday) and premature adult mortality (risk for a 15-year-old of dying before 60, if age specific mortality rates were to remain unchanged) from 1950 to 2017.

Child survival

See Figure 3.6.1.

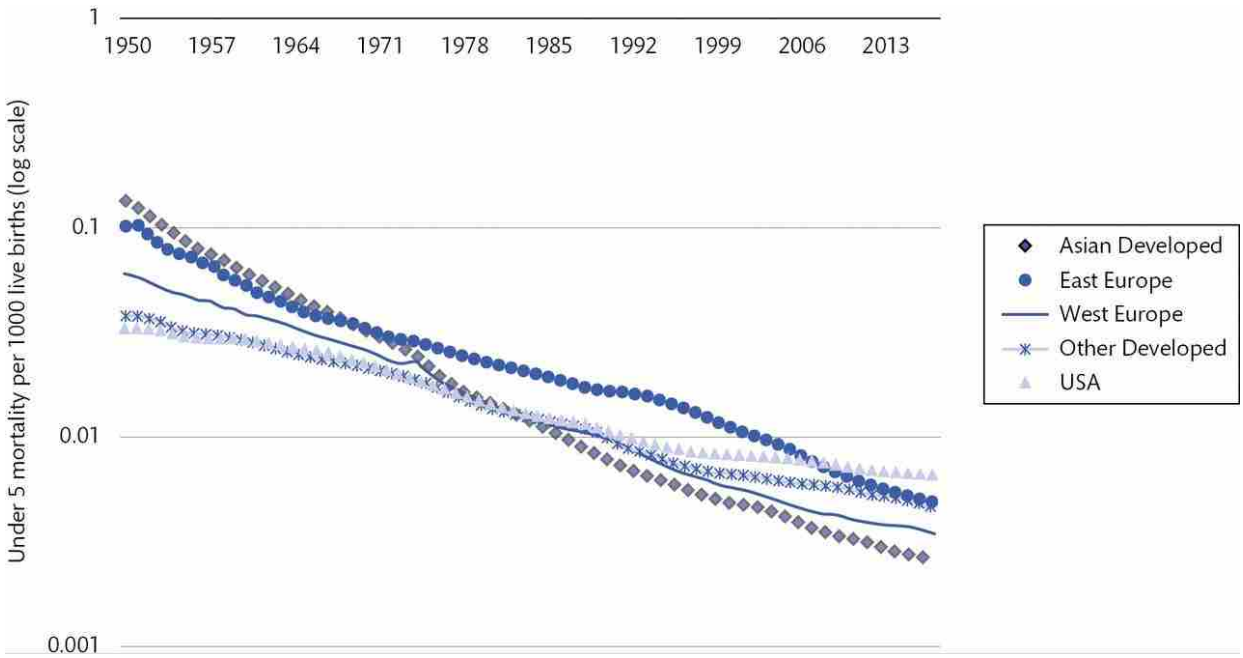


Fig. 3.6.1 Risk of death before the fifth birthday, unweighted means of four groups of countries with the USA shown separately, 1950–2017. For membership of country groups see Section 1.
Source: data from Global Burden of Disease Collaborative Network (2017). (Seattle, United States: Institute for Health Metrics and Evaluation (IHME)).

The most salient finding is the marked and widespread gains in child survival achieved since the mid-twentieth century, extending the even larger gains of the first half of the century. Risks of death before the fifth birthday have fallen, in proportional terms, by a remarkable average of around 95% since 1950—in absolute terms, from around 7.5% to around 0.04% (unweighted averages across included countries). These transformations suggest continuing pervasive improvements in the determinants of child health. Around the central trend there is also some interesting variation across country groups, brought out by the log scale used in Figure 3.6.1: most improved are the Asian Developed countries, moving from highest to lowest for group averages; least improved is the USA, moving from lower than any of the group averages in 1950 to higher than any in 2017. The East European group, which also started with high risks, improved rapidly to the mid-1960s and again from the mid-1990s, with a period of stagnation between.

Adult survival

See Figure 3.6.2.

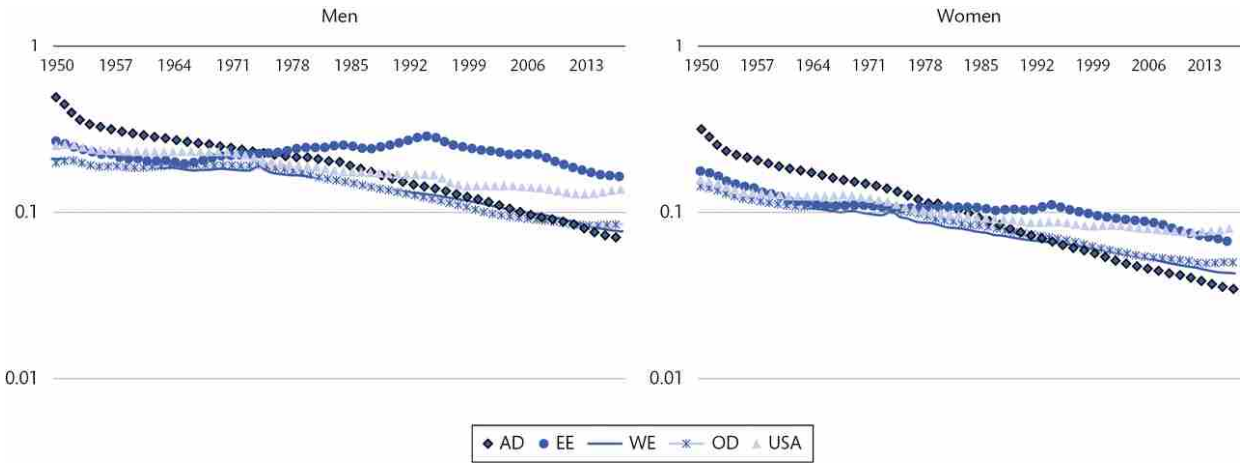


Fig. 3.6.2 Risks for 15-year-olds of dying before 60, unweighted means of four groups of countries with the USA shown separately, 1950–2017, semi-log scale; a. Men, b. Women
Source: data from Global Burden of Disease Collaborative Network (2017). (Seattle, United States: Institute for Health Metrics and Evaluation (IHME)).

Adult mortality estimates for the 39 countries in 1950, 1990, and 2017 are shown in Table 3.6.1.

Table 3.6.1 Risks for 15-year-olds of dying before 60, assuming mortality rates were to remain at levels for designated years, men, and women in 1950, 1990, and 2017. Countries ranked for each sex by estimates for 2017

Men					Women				
		1950	1990	2017			1950	1990	2017
1	Switzerland	0.200	0.122	0.054	1	Singapore	0.254	0.080	0.034
2	Singapore	0.407	0.133	0.060	2	Switzerland	0.135	0.061	0.034
3	Sweden	0.158	0.114	0.063	3	South Korea	0.424	0.098	0.035
4	Italy	0.204	0.131	0.064	4	Japan	0.254	0.053	0.036
5	Netherlands	0.145	0.117	0.065	5	Italy	0.143	0.061	0.037
6	Israel	0.106	0.103	0.067	6	Iceland	0.125	0.068	0.037
7	Norway	0.152	0.131	0.067	7	Spain	0.175	0.059	0.037
8	Japan	0.307	0.107	0.069	8	Israel	0.088	0.066	0.038
9	Ireland	0.222	0.131	0.070	9	Cyprus	0.145	0.065	0.039
10	Iceland	0.188	0.112	0.071	10	Sweden	0.124	0.065	0.040
11	Malta	0.202	0.103	0.073	11	Malta	0.162	0.062	0.041
12	Luxembourg	0.236	0.158	0.074	12	Norway	0.108	0.066	0.042
13	Spain	0.260	0.144	0.075	13	Slovenia	0.229	0.083	0.043
14	Austria	0.241	0.153	0.079	14	Portugal	0.181	0.079	0.043
15	Cyprus	0.237	0.120	0.080	15	Finland	0.165	0.070	0.043
16	Australia	0.209	0.126	0.080	16	Austria	0.159	0.074	0.044
17	South Korea	0.755	0.231	0.082	17	Ireland	0.195	0.079	0.045
18	United Kingdom	0.207	0.129	0.083	18	Australia	0.138	0.067	0.045
19	Denmark	0.159	0.153	0.084	19	France	0.149	0.066	0.046
20	Canada	0.201	0.129	0.085	20	Greece	0.140	0.054	0.047
21	Belgium	0.237	0.137	0.086	21	Croatia	0.227	0.087	0.048
22	New Zealand	0.191	0.141	0.087	22	Luxembourg	0.156	0.079	0.048

Men					Women				
		1950	1990	2017			1950	1990	2017
23	France	0.224	0.160	0.092	23	Germany	0.131	0.076	0.050
24	Germany	0.212	0.156	0.093	24	Belgium	0.152	0.075	0.050
25	Finland	0.321	0.182	0.096	25	Netherlands	0.113	0.068	0.051
26	Greece	0.197	0.114	0.097	26	Canada	0.143	0.070	0.052
27	Slovenia	0.348	0.208	0.097	27	Denmark	0.128	0.100	0.052
28	Portugal	0.271	0.174	0.101	28	Czech Republic	0.161	0.095	0.054
29	Czech Republic	0.246	0.230	0.108	29	New Zealand	0.147	0.094	0.054
30	Croatia	0.326	0.219	0.117	30	United Kingdom	0.140	0.078	0.055
31	United States	0.259	0.170	0.140	31	Poland	0.188	0.100	0.061
32	Slovakia	0.246	0.268	0.140	32	Estonia	0.145	0.108	0.062
33	Poland	0.283	0.259	0.158	33	Slovakia	0.189	0.103	0.062
34	Hungary	0.254	0.302	0.165	34	Hungary	0.184	0.131	0.075
35	Estonia	0.275	0.299	0.166	35	Romania	0.162	0.112	0.079
36	Bulgaria	0.220	0.215	0.188	36	Lithuania	0.143	0.106	0.080
37	Romania	0.216	0.235	0.193	37	Latvia	0.136	0.116	0.082
38	Latvia	0.248	0.306	0.225	38	United States	0.159	0.090	0.082
39	Lithuania	0.241	0.285	0.240	39	Bulgaria	0.175	0.096	0.086

Estimates for 1950 and 1990 for countries which became independent after 1990 have mostly been derived from vital statistics for populations living within the post-independence boundaries. In some cases they have been imputed using hierarchical Bayesian modelling with greater margins of uncertainty.

Source: data from Global Burden of Disease Collaborative Network (2017). (Seattle, United States: Institute for Health Metrics and Evaluation (IHME)).

The trends for premature mortality in Figure 3.6.2 and Table 3.6.1 reveal some striking divergences: From the mid-1960s mortality in the East European group rises substantially in men and stagnates in women before resuming declines from the mid-1990s. Also striking is the stagnation in adult mortality decline in the USA, apparent from the early 1980s. In relative

terms this is more notable for women: By 2017 only in Hungary, Latvia, Lithuania, and Bulgaria were rates of premature mortality in women higher than in the USA—which were about twice the median levels of all the other country groups apart from Eastern Europe.

Declines in risks of premature death in adults have been less profound and less uniform than the declines in child mortality. Overall risks roughly halved between 1950 and 2017 for both women and men. For men, proportional reductions were greatest in the Asian Developed group (almost 85%) compared with 36% for men in East Europe and 45% in men in the USA. In 2017 the risks for men ranged from around 8% in the Asian Developed, West European, and Other Developed groups to twice that in the USA and Eastern Europe.

Some qualitative similarities can be seen in the trends in child and adult survival: the USA and Eastern Europe each show lags relative to other groups—but with the lags occurring in different periods. In Eastern Europe, improvements were rapid until the mid-1960s after which they slowed (for children), stagnated for adult women and reversed (in some cases catastrophically) for adult men. From the mid-1990s these adverse trends in Eastern Europe changed course but with rates of improvement increasingly diverse for countries within the group.

These manifest changes in survival provide much food for thought. The part played by ‘policy’ is not self-evident and nor are the roles played by governments at different levels nor by officials, experts, and citizens considered as individuals, nor as collectivities such as state and professional organizations, corporations, social movements, and households.

Institutions, knowledge, and policies

Institutions

Because national (and subnational) boundaries separate populations with distinctive health trends, such trends can be assumed to reflect not just the characteristics of the individuals constituting such aggregates but also the differing national and subnational institutional contexts in which the individuals live. Here we understand institutions to be not just formally constituted organizations, such as those of states, but also the less formal and more fundamental rule-based structures which pattern social life (Searle

2005, 2010). Such institutional structures may have their own, typically longer term, patterns of evolution, in many cases beneath the surface of overt developments in the political realm (Williamson 2000; Thelen 2004).

Constance Nathanson has produced a monumental, systematic study comparing responses in the USA, Canada, Britain, and France, to two public health challenges in the early 20th century—infant mortality and tuberculosis—and two late in the century—cigarette smoking and HIV infection in injecting drug users. Her study brings out the diversity, complexity, and richness of societal responses to risks to health (Nathanson 2009). In spite of the widespread assumption that the agency of the state has been central to public health responses, she finds no simple relationship between the strength of state institutions (presumptively high in France, low in the USA) and the nature and strength of societal responses to risks to health. Her work helps to clarify the relationship between state and non-state 'actors', bringing out the often limited contribution of the former, at least during the earliest, creative phases when the societal responses to health hazards are just beginning to take shape.

> . . . in only one of the eight contemporary country-case combinations (AIDS drugs in Canada) did leadership in recognizing and addressing a serious public health issue come from local public health officers or departments. (p. 254)

At a national (as distinct from subnational) level she does find one striking example of rapid and successful state-initiated action in the face of a public health crisis—the British government's response to the threat of HIV/AIDS in injection drug users (p. 181). Although the contributions of state actors may be limited in the earliest phases they are typically engaged dynamically with civil organizations as soon as the policy process begins to evolve. State agencies usually lead in programme implementation.

Nathanson invokes three broad determinants of public health action: 'states, collective actions, and constructions of risk' and sees her task as understanding 'how structure, opportunity, and human agency interact to create, or inhibit, the possibility of change' (p. 203). Writing as a 'social constructivist', narrating public health 'stories', she makes little attempt to assess effects on objective health trends and pays little attention to the advance of objective knowledge about the causes of disease and injury. She concentrates instead on 'constructions of risk' which follow their own timetables, somewhat independent of advances in objective knowledge. In relation to the earlier influence of the germ theory of disease, she notes

> French public health was reenergized in the late nineteenth century, due in large part to the extraordinary impact of Pasteur on French medicine and science . . . Nevertheless, apart from the campaign against infant mortality, which had relatively little to do with Pasteur, the consequences on the ground were minimal. (p. 35)

While her approach serves well to bring out the importance of human agency and the ways in which social mobilization has taken place in response to public health challenges, her relative neglect of the escalating output and diffusion of new knowledge and her sketchy coverage of the links between actions and outcomes leaves her interpretation incomplete.

The growth of knowledge

The modern era has been characterized by an enormous increase in the production of new knowledge and this process has tended to accelerate through time. Global scientific output is estimated to be now doubling every 9 years (Bornmann and Mutz 2015). Medical research output has been highest in the USA and other English-speaking countries and in Scandinavia. This remains true when adjusted for population size. Research output in public health sciences follows a similar overall pattern but also with notably low outputs in Eastern and Southern Europe (Clarke et al. 2007).

Knowledge: ‘interventions’ plus less direct effects

The broad temporal coincidence between the increase of knowledge and the improvement of health suggests that the former is likely to be a fundamental cause of the latter. But the question of just how the advances in knowledge have actually contributed to gains in health remains unclear. One prominent idea is that the most important uses of knowledge are those embodied in ‘interventions’. The increasing use of idea of ‘interventions’ is linked to increasing reliance on randomized controlled trials to provide trustworthy evidence on ‘what works’. In each field of activity, summaries of the evidence identifying ‘effective interventions’ are marshalled for the benefit of ‘policymakers’.

The implication here is that the experimental arms of controlled trials provide a model for understanding how knowledge brings benefits in populations at large. But the ways in which knowledge about the causes of disease and injury is used in everyday life often differ from what happens in the experimental arm of a controlled trial. In a controlled preventive

experiment, agency—the capacity to decide ‘what is to be done’ and the execution of that decision—rests with the ‘intervenor’. The idea of ‘interventions’ thus carries with it the idea that agency lies with the ‘intervenors’, that is, with professionals and officials. In this framework, all that is needed is for ‘policymakers’ to take heed of what the experts are telling them about the scope for ‘effective interventions’ for the benefits of existing knowledge to be realized. This interpretation is inadequate.

Issues of agency are central to thinking about public health policy. They came the fore earlier in debates provoked by Thomas McKeown’s thesis on the causes of rapid mortality decline in England and Wales in the late nineteenth and early twentieth century. McKeown noted, in a series of publications, that there were no new effective interventions (‘medical measures’ in his terminology) that could explain the favourable shifts in mortality. The shifts must have been due to something else. His *deus ex machina* was improved nutrition. He justified his choice by invoking the logic of Sherlock Holmes: ‘*When you have eliminated the impossible, whatever remains, however improbable, must be the truth*’ (McKeown 1988, pp. 9–10, italics in original).

McKeown’s implicit, though fallible, logic was: if not ‘knowledge as embedded in interventions then not knowledge’. His choice of improved nutrition as the alternative explanation has been judged implausible by social, demographic, and economic historians. These critics have noted the lack of evidence for sudden changes in diet and emphasized instead the importance of changes in civic and state organizations facilitating sanitary reform (Szreter 1988), and in risk reducing changes in childrearing practices implemented by households. In the late nineteenth and early twentieth century households were key actors:

> The revolution was not just in the way physicians thought about disease, but in the growing awareness that households could control their destinies by their own actions and prevent disease by avoiding certain well-understood sources of infection. (Mokyr 2002, p. 206)

The nature of the relevant changes in childrearing practices during the period of rapid mortality decline has been documented further in Mokyr and Stein (1997, pp. 143–205), Cowan (1983), Tomes (1998), and Preston and Haines (1991).

The acceptance of new knowledge by households was not automatic, but depended on persuasion:

> Persuasion requires shared standards of evidence, chains of authority, networks of trust, and accepted rules of logic and evidence. Changes in the rules of discourse and communication, no less than the knowledge unearthed by science, are the background to the changes in health and longevity that are the mark of the 'modern' age. (Mokyr 2002, p. 180)

Put another way: it is not just the development of knowledge that is important but also the diffusion of knowledge and the associated evolution of institutions which increase receptivity to new knowledge and facilitate its use.

The economic historian Simon Kuznets saw the advance of knowledge as fundamental to modern economic development (Kuznets 1973); Mokyr (2002) and Deaton (2013) each see economic development and health improvement as the joint but distinct products of advancing knowledge, propelled in separate ways by advances in knowledge, and not always in synchrony with each other.

Two important conclusions, relevant to the understanding of more recent trends, thus emerge: (1) the advance of knowledge (understood broadly as 'useful knowledge', extending beyond the content of textbooks to knowledge useful in everyday life) has been fundamental to the improvement of health by both direct and indirect pathways; and (2) the persuasiveness of relevant knowledge has depended on institutional contexts evolving in ways that enhanced receptivity to new ideas.

Causes of health trends

Child health

Trends in child mortality have not exhibited the large swings and reverses seen in adult survival and are now in asymptotic approaches to very low levels. Measures to reduce risks to the health of children have generally faced little resistance and have been widely accepted and applied. A partial exception is the antivaccination movement's effects on immunization coverage with associated rises in measles incidence and deaths (Kmietowicz 2018).

Although absolute risks of death before the fifth birthday are now small, relative differences remain appreciable. Within Western Europe risks are almost twice as high in the United Kingdom as in Sweden. Most of this difference has been related to 'birth characteristics' such as prematurity and

birthweight (Zylbersztejn et al. 2018). These are more related to differences in the health of mothers than to professional services provided after birth. The 'biological standard of living' (Komlos 1993) of mothers is important and differences of this kind are likely to have origins that are embedded in the social structure.

Adult health

The large differences in long-term trends and recent levels of premature mortality in adults between different groups of countries shown in Figure 3.6.2 resist easy explanation. The Global Burden of Disease project produces estimates of the contributions of proximate causes (risk factors) but the role of more distal causes, such as policies, and the institutional supports for those policies, remains poorly understood.

We shall proceed by considering parts of this complex picture. The reduction of health losses from tobacco, traffic injuries and cardiovascular disease have accounted for a substantial part of the total health gains since the mid-twentieth century. They are relatively well documented, as are the collective measures that have sought to achieve these outcomes. Case studies of these may provide some insights into the links between advancing knowledge, institutional evolution, and health outcomes.

Tobacco

The first English language demonstrations of the quantitative relationships between smoking and lung cancer risk were reported in 1950 (Doll and Hill 1950; Wynder and Graham 1950). By 2012, all of the 39 developed countries except the USA and Switzerland had ratified an international treaty on tobacco control, committing themselves to a full repertoire of evidence based measures to reduce tobacco smoking (Conference of the Parties to the WHO FCTC 2003). Over a 60 year interval, the currently established repertoire of measures for tobacco control were invented, tested, and refined —ultimately gaining political support at the highest level.

The early 1950s to the 1970s was a period of gradual clarification at a scientific, professional, political, and public level of the nature and meaning of the health risks arising from tobacco use. The epistemology of professional medicine was still pre-quantitative. As reports from

epidemiological studies accumulated, high level policy debates continued to revolve around a largely illusory search for ‘proof of causation’ (Pollock 1999). Many physicians remained unconvinced of the hazards. In the USA a 1960 poll found that only one-third of physicians were convinced that smoking was a ‘major cause’ of lung cancer and 43% were still regular smokers (Proctor 2011). In the United Kingdom, physicians and public health professionals who appreciated the magnitude of the risks and who were in a position to do something about it struggled to find a way forward.

In the late 1950s, George Godber was Deputy Chief Medical Officer in Britain and was very keen to help reduce harm from cigarette smoking. But his superior, the Chief Medical Officer, did not want to take the matter forward with the Minister of Health who was known to be unenthusiastic about taking on the Chancellor of the Exchequer (Treasurer). The Treasury was keen to protect the 15 or so per cent of central government revenue provided by tobacco tax (Berridge 2007). Godber visited Charles Fletcher, a respiratory physician, at a London teaching hospital and invited him to his club for lunch so that they could discuss strategies. They decided on working through the Royal College of Physicians in order to by-pass the Chief Medical Officer. The College took up the matter energetically and decided to produce a report aimed at a large audience. In 1962, Smoking and Health appeared and quickly sold out (Pollock 1999), (Lock et al. 1998). The United States followed suit and in 1964 the Surgeon-General’s report on The Health Effects of Smoking appeared. The mass media covered these reports extensively and Pierce and colleagues have shown that smoking cessation rates in US middle-aged adults moved in synchrony with the news coverage (Pierce and Gilpin 2001). Coverage of the Surgeon-General’s report was not the beginning of mass media coverage of smoking and health in the USA—as early as 1946 a news magazine story ‘Cigarettes cause cancer?’, based on experimental work on the carcinogenicity of cigarette tar was ‘widely read’ (Proctor 2011). However, there is good reason to see the Surgeon-General’s report as a turning point in increasing public awareness in the USA of the hazards of smoking.

A major difficulty, in the 1950s, was that public health authorities were reluctant to tell adult men how they should behave. Reluctance of this kind had not previously inhibited anti-spitting campaigns to reduce risks of transmitting a communicable disease (tuberculosis). Nor had it inhibited strong and detailed directions to mothers, early in the twentieth century on

how they should care for their children (Lewis 1980). But advising adult men to abandon a widely accepted habit was seen as different. Crossing this frontier would ultimately change ‘the discourse of public health’ in Britain and legitimize government activity to promote health favouring lifestyles (Berridge 2007).

In Britain, early tobacco control activism was largely confined to professional and political elites who aimed to build constituencies ‘from above’ in support of action by parliament. This turned out to be a slow process. Kenneth Robinson who was health minister through the mid-1960s said that lack of public support had prevented him acting (Berridge 2007). In 1971 the Royal College of Physicians established a pressure group—Action on Smoking and Health—‘to make non-smoking the norm in society and to inform and educate the public about the death and disease caused by smoking’ (Action on Smoking and Health 2017). Most of the early office holders were activist physicians.

Major breakthroughs in mobilizing a broader base of public support came in the mid to late 1980s in Victoria, Australia. The 1987 Victorian Tobacco Act was the first to secure a tax hypothecated to health promotion—in this case to the Victorian Health Promotion Foundation (Vichealth). Vichealth then used this money (AUD 23m per year—about USD 7 per person per year) to buy out tobacco sponsorship of sporting and arts events and to fund health promotion programmes. The close passage of this Act was the outcome of careful planning, initially conducted in strict secrecy in order to catch the tobacco industry off-guard and prevent it deploying its enormous resources against the Act. Once the campaign was public, the large contributor base of the Anti-Cancer Council of Victoria was mobilized. Newspaper coverage was sympathetic—the main morning paper reported more correspondence on the topic than any other in the paper’s history (Borland et al. 2009). Over 150,000 letters of support were sent to state politicians and church leaders were engaged in the campaign. The idea of an hypothecated tax was taken to California and included in Proposition 99 which won voter approval in November 1988.

Nathanson notes the marked differences between countries in the groups who led the opposition to tobacco. In the USA ‘lay activists’ led the way. In California, an early victory was the passage in 1977 by the Berkeley City Council of an ‘indoor air ordinance’ instigated by the local branch of GASP (Group Against Smoke Pollution). (GASP had been launched nationally in

1970 by environmental activist Clara Gouin who attributed her father's death a year before to smoking.) From the 1970s local groups, many now constituted, in California, as Americans for Nonsmokers Rights, had achieved 117 local tobacco control ordinances by 1986 (Glanz and Balbach, p. 39). In the lead in to the Proposition 99 referendum, private polling commissioned by the Tobacco Institute reported as its 'bottom line' that '*people do not like smoking or smokers*' (V. Lance Tarrance and Associates (1987) cited in Glanz and Balbach, p. 42, italics in original). It was against this background that the 1988 vote was won, despite industry spending to defeat it being 13 times greater than the spending in support (Glanz and Balbach, p. 72).

The formal state-wide tobacco control programme mandated by Proposition 99 was funded by a 25c per pack levy on cigarettes, sufficient to pay for mass media campaigns and to fund activities developed and run by local groups. It did not fund professional and pharmacological assistance for smoking cessation. During the 1990s it was noted by programme evaluators that a substantial fraction of households were reporting that they had decided to make themselves 'smoke free' even though this had never been promoted by the formal programmes. This imitation by households of 'clean air ordinances' enacted for public venues provides a nice illustration of the complex decentralist—centralist dynamic at the heart of the California's response to the challenge of tobacco (Pierce et al. 2010).

From the 1990s the tobacco industry has suffered further losses of legitimacy and support in the developed world. In the USA, the 1998 Tobacco Master Settlement Agreement between 46 states and the largest cigarette manufacturers, resulted in the industry agreeing to restrict some marketing practices, to dissolve its three main front organizations and to pay the states $206 billion (sic) over 25 years to compensate for the extra state medical expenditures attributable to cigarette smoking. Interpretation of the potential effects on the Settlement on tobacco control have been mixed (Schroeder 2004) but the public release of previously secret industry documents during the discovery phase of the litigation has led to further damage to the industry. Proctor used these documents to show how the industry deliberately and secretly modified the composition of cigarettes to make the smoke more inhalable and the smoker more dependent (Proctor 2011).

In 2006 Philip Morris was convicted by a US federal court under legislation designed to counter organized crime (the Racketeer Influenced and Corrupt Organizations Act) for having: falsely denied that they market to youth, falsely denied that they manipulated cigarette design, falsely represented that light and low tar cigarettes deliver less nicotine and tar, falsely denied that ETS (second-hand smoke) causes disease and suppressed documents, information, and research (Proctor 2011). In November 2012 they were ordered to make public statements admitting that they had lied to the public (Associated Press 2012). Elsewhere, the closing off of marketing opportunities in high-income countries was taken one step further when, in December 2012, Australia enacted a law requiring that all cigarettes be sold in plain standardized packaging. As of mid-2018, seven more countries had followed suit: France, United Kingdom, Norway, Ireland, New Zealand, Hungary, and Slovenia.

The tobacco wars are not over yet. Despite declining deaths from smoking, tobacco was still ranked first for men and fifth for women among causes of lost healthy life in ‘high socio-demographic index’ countries in 2017 (GBD Compare 2018). Deaths from lung cancer provide a robust and comparable indicator of health losses from tobacco and those at younger ages reflect more recent trends in tobacco exposure. Table 3.6.2 gives years of life lost from lung cancer deaths in men and women under the age of 50. It shows that, for men, progress was almost universal between 1990 and 2016 with only Portugal and Cyprus showing increases. Reductions, both proportional and absolute, were biggest in the post-communist countries, with the Czech Republic and Hungary both showing very large reductions (from 659 to 143 and from 922 to 436 respectively). Trends for women are lagged relative to those for men with 21 out of 39 countries showing increased lung cancer mortality between 1990 and 2016. Meanwhile, the tobacco industry has moved its focus from developed countries to emerging markets in low and middle-income countries where prevalence rates are high and, in some cases, growing.

Table 3.6.2 Years of life lost to lung cancer from deaths at ages 15 to 49, men and women in 39 developed countries, 1990 and 2016: age-standardized rates per 100,000 person years with percentage changes from 1990 to 2016

Men					Women				
Countries ranked by rates for 2016		Rate 1990	Rate 2016	Change	Countries ranked by rates for 2016		Rate 1990	Rate 2016	Change
1	Sweden	149	60	–59%	1	Malta	50	54	8%
2	Finland	185	61	–67%	2	Finland	66	57	–14%
3	Norway	152	84	–44%	3	Cyprus	45	57	26%
4	New Zealand	139	96	–30%	4	Estonia	59	57	–4%
5	Australia	161	99	–39%	5	Latvia	64	59	–8%
6	Singapore	196	108	–45%	6	Lithuania	63	64	1%
7	Iceland	152	113	–26%	7	Japan	84	67	–20%
8	Japan	154	118	–23%	8	Israel	80	74	–7%
9	Canada	235	118	–50%	9	Sweden	133	82	–39%
10	Denmark	255	121	–53%	10	Singapore	105	83	–21%
11	United Kingdom	217	123	–43%	11	Australia	93	87	–6%
12	Ireland	165	124	–25%	12	Czech Republic	129	89	–31%
13	Switzerland	288	128	–56%	13	Slovakia	93	90	–3%
14	South Korea	197	129	–35%	14	Ireland	88	96	9%
15	Cyprus	114	134	17%	15	South Korea	84	97	16%
16	Czech Republic	659	143	–78%	16	Norway	99	97	–1%
17	Malta	180	143	–47%	18	Poland	128	103	–19%
18	Austria	272	145	–47%	18	United Kingdom	123	109	–11%
19	United States	312	152	–51%	19	Switzerland	121	111	–8%

Men					Women				
Countries ranked by rates for 2016		Rate 1990	Rate 2016	Change	Countries ranked by rates for 2016		Rate 1990	Rate 2016	Change
20	Estonia	412	153	–63%	20	Portugal	78	119	53%
21	Israel	173	155	–10%	21	New Zealand	129	126	–3%
22	Netherlands	286	168	–41%	22	Slovenia	110	126	15%
23	Slovakia	509	170	–67%	23	Austria	101	130	28%
24	Italy	358	177	–51%	24	Greece	100	133	32%
25	Slovenia	432	178	–59%	25	Italy	102	136	34%
26	Luxembourg	401	178	–56%	26	Spain	65	138	110%
27	Germany	352	184	–48%	27	Luxembourg	118	143	21%
28	Belgium	410	198	–52%	28	United States	205	144	–30%
29	Poland	494	202	–59%	29	Croatia	121	147	22%
30	Lithuania	391	217	–44%	30	Germany	121	149	23%
31	Latvia	428	228	–47%	31	Denmark	234	150	–36%
32	Spain	388	245	–37%	32	Canada	206	150	–27%
33	Croatia	589	288	–51%	33	Romania	100	151	51%
34	Greece	406	294	–27%	34	Bulgaria	108	153	41%
35	France	447	318	–29%	35	Belgium	127	158	24%
36	Portugal	266	322	21%	36	France	78	183	134%
37	Bulgaria	592	428	–28%	37	Iceland	177	191	8%
38	Hungary	922	436	–53%	38	Netherlands	157	218	39%
39	Romania	506	448	–11%	39	Hungary	293	324	11%

Source: Global Burden of Disease Study 2018 (GBD 2018) Results. (Seattle, United States: Institute for Health Metrics and Evaluation (IHME)).

The century long epidemics of cigarette smoking and its consequent disease burdens illustrate the time frame over which chronic disease determinants may rise and fall, making campaigns to reduce them into long

wars of attrition. Tobacco control also illustrates how the development of quantitative methods has supported appropriate policy responses. With increased acceptance of epidemiological reasoning and its transmission to political and wider publics (for example, Peto et al. 2013), quantitative assessments have become more central to policy deliberations, facilitating effective institutional responses, locally, nationally, and globally.

Traffic injuries

Deaths from traffic injury per unit of exposure (e.g. per 10,000 vehicles per year) have fallen up to 10-fold in many developed countries since the mid-twentieth century. Despite this outstanding success, traffic injuries still ranked sixth among causes of lost healthy life in males and fourteenth in females in 2017 (GBD Compare 2018). Because of the visibility of control measures and the short time lags between their implementation and their effects, traffic injury control provides a sensitive field in which to explore the relationship between policy adoption, programme implementation, and health effects.

First, it is appropriate to note that success has not flowed exclusively from centrally initiated policies, especially in the early decades of motorization. Smeed writing in the early 1970s showed that deaths from traffic crashes in relation to the number of registered vehicles had followed a general, and pronounced, downward trend as the number of motor vehicles increased in relation to population (Smeed's Law: Deaths/vehicle = 0.0003 $(\text{vehicles/population})^{-0.66}$) (Smeed 1972). However, over the past four or five decades organized policy responses have clearly made a bigger contribution. The Australian state of Victoria has been a pioneer in traffic injury control and its experience will be briefly reviewed here.

Consistent with Smeed's prediction, deaths per 10,000 vehicle years fell in Victoria as the number of vehicles increased in relation to population—from over 20 in the 1930s to about 9 in the early 1960s. During these early decades, Victoria generally had rates in excess of Smeed's prediction. The recurrent funerals resulting from avoidable deaths in young adults provoked a growing conviction that 'something needed to be done'. Early policy responses relied on exhortation—for example, encouraging motorists to sign a pledge to abstain from 'careless and discourteous behaviour' on the roads. This approach was not successful. A campaigning young police surgeon engaged journalists in the cause by taking them to the scenes of alcohol

related accidents on Friday and Saturday nights. At this time, the editor of ‘The Age’ a Melbourne daily newspaper threw the paper’s weight behind the campaign. A committee of the Royal Australasian College of Surgeons made a case before the parliamentary committee on road safety for more rigorous surveillance and for compulsion. In 1970, these submissions convinced the main political parties to support legislation, the first in the world, compelling Victorians to fit and wear seat belts. Compliance with the new law was high and road deaths fell. In 1976, the Victorian parliament was the first in the world to legislate for random breath alcohol testing, responding again to advice from professionals. Deaths declined over the following decade. In 1986 a Transport Accident Commission was established, with secure funding from the compulsory third party injury insurance levy on motorists being made payable to it. The Commission is required to pay no-fault compensation to those injured, giving it a financial incentive to reduce injuries. A specified proportion of its spending is required to be spent on programme evaluation—thus supporting the development of an independent accident research institution. In late 1989, after traffic deaths had begun to rise again, it launched one of the most intensive mass media campaigns ever run in Australia—combined with intensified police enforcement (Davison and Yelland 2004, ch 6). In 2009 police performed 1.24 million random breath tests in a total population of 5.4 million persons (Transport Accident Commission 2013). Since 1960 traffic fatality rates in Victoria have fallen to levels that are among the lowest of the developed world.

Local historians commented in 2004:

> Probably nowhere else in the world was the conduct of the individual motorist more closely monitored or more rigorously controlled. Yet, with only occasional grumblings, motorists overwhelmingly accepted these measures.
>
> (Davison and Yelland 2004, p. 166)

These strategies of education and enforcement were combined with those of engineering. In 1965, Ralph Nader denounced US cars as ‘unsafe at any speed’ (Nader 1965). In Australia, the national government negotiated against the resistance of the international automobile industry for ‘design rules’ to make cars safer—e.g. collapsible steering columns, shatter-proof windshield glass and the removal of dangerous protrusions from the passenger cabin (personal communication: Tony Ryan). For the US, Robertson has documented the continuing resistance of the motor industry to manufacturing safer cars (Robertson 2010).

Cardiovascular disease

Cardiovascular diseases, mainly ischaemic heart disease and stroke, remain the leading causes of lost life years in developed countries. The overall change from 1990 to 2016 in the level of lost life years from these diseases in our 39 countries is shown in Table 3.6.3. The median reduction was 58%. In both time periods Japan, France, Spain, and Switzerland occupy the top four ranks (i.e. have the lowest mortality). In 1990 the Eastern European countries were at the bottom of the ladder. Over the 26-year interval reductions are largest in South Korea (76%) and Israel (71%). The post-communist group shows increasing heterogeneity with Slovenia and the Czech Republic showing above average declines while, at the other end, Bulgaria and Lithuania only achieved half the average reduction. In 2016 rates of lost life years in Bulgaria were over three times those in Slovenia. The USA also performed poorly with a reduction of just 39%, falling in rank from thirteenth to twenty-eighth.

Table 3.6.3 Years of life lost to cardiovascular diseases, 39 developed countries, 1990 and 2016: age-standardized rates per 100,000 person years with percentage changes from 1990 to 2016

Men					Women				
Countries ranked by rates for 2016		Rate 1990	Rate 2016	Change	Countries ranked by rates for 2016		Rate 1990	Rate 2016	Change
1	Sweden	149	60	–59%	1	Malta	50	54	8%
2	Finland	185	61	–67%	2	Finland	66	57	–14%
3	Norway	152	84	–44%	3	Cyprus	45	57	26%
4	New Zealand	139	96	–30%	4	Estonia	59	57	–4%
5	Australia	161	99	–39%	5	Latvia	64	59	–8%
6	Singapore	196	108	–45%	6	Lithuania	63	64	1%
7	Iceland	152	113	–26%	7	Japan	84	67	–20%
8	Japan	154	118	–23%	8	Israel	80	74	–7%
9	Canada	235	118	–50%	9	Sweden	133	82	–39%
10	Denmark	255	121	–53%	10	Singapore	105	83	–21%
11	United Kingdom	217	123	–43%	11	Australia	93	87	–6%
12	Ireland	165	124	–25%	12	Czech Republic	129	89	–31%
13	Switzerland	288	128	–56%	13	Slovakia	93	90	–3%
14	South Korea	197	129	–35%	14	Ireland	88	96	9%
15	Cyprus	114	134	17%	15	South Korea	84	97	16%
16	Czech Republic	659	143	–78%	16	Norway	99	97	–1%
17	Malta	180	143	–47%	18	Poland	128	103	–19%
18	Austria	272	145	–47%	18	United Kingdom	123	109	–11%
19	United States	312	152	–51%	19	Switzerland	121	111	–8%
20	Estonia	412	153	–63%	20	Portugal	78	119	53%

Men					Women				
Countries ranked by rates for 2016		Rate 1990	Rate 2016	Change	Countries ranked by rates for 2016		Rate 1990	Rate 2016	Change
21	Israel	173	155	–10%	21	New Zealand	129	126	–3%
22	Netherlands	286	168	–41%	22	Slovenia	110	126	15%
23	Slovakia	509	170	–67%	23	Austria	101	130	28%
24	Italy	358	177	–51%	24	Greece	100	133	32%
25	Slovenia	432	178	–59%	25	Italy	102	136	34%
26	Luxembourg	401	178	–56%	26	Spain	65	138	110%
27	Germany	352	184	–48%	27	Luxembourg	118	143	21%
28	Belgium	410	198	–52%	28	United States	205	144	–30%
29	Poland	494	202	–59%	29	Croatia	121	147	22%
30	Lithuania	391	217	–44%	30	Germany	121	149	23%
31	Latvia	428	228	–47%	31	Denmark	234	150	–36%
32	Spain	388	245	–37%	32	Canada	206	150	–27%
33	Croatia	589	288	–51%	33	Romania	100	151	51%
34	Greece	406	294	–27%	34	Bulgaria	108	153	41%
35	France	447	318	–29%	35	Belgium	127	158	24%
36	Portugal	266	322	21%	36	France	78	183	134%
37	Bulgaria	592	428	–28%	37	Iceland	177	191	8%
38	Hungary	922	436	–53%	38	Netherlands	157	218	39%
39	Romania	506	448	–11%	39	Hungary	293	324	11%

Source: Global Burden of Disease Study 2016 (GBD 2016) Results. (Seattle, United States: Institute for Health Metrics and Evaluation (IHME)).

These trends prompt three policy-related questions:

What have been the main contributors to reductions in cardiovascular mortality over recent decades?
Why have the trends in the post-communist group been so diverse?
Why has the USA been less successful?

As with road traffic injuries and tobacco caused disease, endeavours to control circulatory diseases have been pursued in conjunction with determined attempts to better understand their causes. In the United States, President Harry Truman signed the National Heart Act in June, 1948. Between 1950 and 2018 the budget of what was to become the National Heart Lung and Blood Institute rose from US$10 million to US$3,380 million (NIH, n.d.). The deliberate rapid expansion of research into the causes of ischaemic heart disease contributed strongly to the modernization of the scientific base of public health, including the emergence of modern epidemiology and the concept of ‘risk factors’.

The causation of these diseases has turned out to be complex but much of the variation in cardiovascular risk, at individual and population levels, continues to be explicable by the three classic risk factors—blood pressure, blood cholesterol concentration, and cigarette smoking—whose quantitative relationship to risk is now well understood. Smoking epidemic in young adult males before 1980 in most of the developed countries. Less data is available on trends in blood pressure and blood cholesterol concentrations. Where documented, trends in both sexes in developed countries have tended to follow a broadly similar time course to those of cigarette smoking in males—that is, to have been generally favourable, especially since around 1990 (Danaei et al. 2013). The development of drugs that are effective in lowering blood pressure and blood cholesterol concentrations has enabled clinicians to also make an important contribution to the prevention of these diseases. More effective treatments for heart attacks and for precursor coronary artery disease also mean that the trends in mortality from heart attack reflect not only the extent of success in preventing disease onset but also in reducing the risk of death after onset.

In seeking to characterize societal responses to heart disease and stroke one might start with an hypothetical chain from the advance of scientific knowledge to increased professional and public knowledge, advocacy, and constituency building by activists at all levels, enactment of policy, programme implementation, changes in risk factors, and finally changes in disease incidence. However such sequences cannot be easily documented or

established in a convincing way. Early phases of development were often marked by the formation of physician-led voluntary organizations which sought to mobilize public support for research, prevention and treatment in many countries: The American Heart Association (founded 1924), the National Heart Foundation in Australia (founded 1959), and the British Heart Foundation (founded 1961). The European Heart Network now unites 29 member organizations from 24 European countries. In Australia the National Heart Foundation, with extensive public support, was rich enough to actually lead in establishing national risk factor surveys during the 1980s. In Finland the local government of North Karelia, a district with extremely high risks of premature death from vascular disease, petitioned the national government for support in reducing their disease burden. From this emerged the North Karelia project which spanned epidemiological investigation and mass mobilization to promote dietary change. Finland became an exemplar in the implementation, at community level, of science-based policies to reduce vascular risk (Puska 2005).

Few countries have long time series of data on public knowledge of risk factors or of risk factor levels. However, there are indications—not least in the salience of health claims in television advertising for foods—that public awareness of the determinants of cardiovascular risk has increased over time. By the mid-1990s a vast majority (96%) of men and women in England correctly identified at least three ways of achieving a healthier diet (Bridgwood et al. 1996, p. 62).

One assessment of associations between risk factor changes and declines in mortality from cardiovascular diseases plus diabetes in 26 ‘industrialized’ countries (including Chile, Argentina, and Brazil but no post-Communist country in Europe) found that ‘trends in body mass index (BMI), blood pressure, serum cholesterol, and smoking together may have alone accounted for nearly one-half of the decline in men and about 40% in women’ (Di Cesare et al. 2013, p. 847). However, many of the risk factor estimates in this study were imputed using hierarchical Bayesian imputation.

Capewell and colleagues have shown, in a series of studies, that favourable shifts in risk factors contributed much more than clinical measures (pharmacoprevention plus treatment) to the reductions in years of life lost from ischaemic heart disease in Scotland (1975–1994), USA (1980–2000), Ireland (1985–2000), and England and Wales (1981–2000) (Gouda et al. 2012).

To all of these activities intended to reduce disease risk must be added other background factors. As early as the 1960s, results from the study of 'Coronary Heart Disease in Seven Countries' confirmed that local food cultures varied markedly in their association with coronary risk—protective in the Mediterranean area and in Japan (once traditional high sodium intakes were abandoned) while being adverse, for example, in Northern Europe (Keys 1980). Favourable trajectories in mortality from circulatory diseases in countries such as France, Spain and Japan may owe a good deal to protective food cultures as well as deliberate efforts to reduce risk. While the effects of rising incomes on animal fat and cigarette consumption may have been adverse in the early decades of the twentieth century, once the associated risks were understood, rising incomes will have brought benefits, for example via the ability to support further scientific research the dissemination of findings and by the diversification of food supplies—reduced fat dairy products, margarines with low trans-fat content, year-round supplies of fresh fruit and vegetables, etc. Richer populations can also afford the relatively expensive strategy of controlling blood pressure and high cholesterol by long-term medication. These changing patterns of association between vascular risk factors and national income levels have recently been described (Danaei et al. 2013).

Large scale failures of public health policy

Eastern Europe 1965–1990: a pervasive failure of institutional adaptation

In the period up to the mid-1960s premature adult mortality fell at least as fast in Eastern as in Western Europe (Fig. 3.6.2). The Soviet model of health service organization (named after Semashko, the first commissar of health) dealt well with the main public health problems it was designed for—notably communicable diseases and maternal and child health (Field 1967). But when faced with the challenge of tackling non-communicable disease not only did things go badly wrong but these regimes appeared incapable of changing course in the face of manifest failure.

The communist-ruled countries with transplanted Soviet public health institutions, failed to build modern scientific foundations for responding to noncommunicable diseases: they did not invest in developing modern

epidemiology and they failed to make good use of international findings. The early years of the Soviet regime had been characterized by an enthusiasm for science but subsequent developments were not favourable to population based research in medicine. According to Terris, the policy of Semashko, of integrating prevention and treatment led, in practice, to the domination of Soviet medicine by clinicians rather than by public health experts (Terris 1988). Later, the legacies of the 'Lysenko affair'[1] weakened all biological sciences (Krementsov 1997) and the control of science through hierarchical systems of research institutes limited innovation. Soviet economic philosophy also saw the route to human betterment through the physical domination of nature, not by attending to biologically based needs. The physical sciences and engineering (related, in Soviet economic theory, to the 'productive sector') were valued and rewarded more than medicine (seen as part of the 'unproductive sector'). The idea that 'dietary affluence' could harm health was denounced as a capitalist ruse. These adverse internal characteristics were augmented by the isolation of the cold war years; at late as 1988, Terris observed that there were no Soviet members of the International Epidemiological Association. The transformative potential of the quantitative revolution in medicine and, in particular, of the randomized trial, was not realized (McKee 2007). Finally, there were problems of the supply of healthy foods: systems for counter-seasonal supplies of fresh fruit and vegetables were not developed (as in the west) and these foods tended to disappear from the shops in winter and spring. Such circumstances would not have provided a propitious background for programmes promoting dietary fruit and vegetable consumption.

In some of the 'laggard' post-communist countries, public awareness of risk factors has been notably low. When survey respondents in north-east Bulgaria were asked an open-ended question (in 2000 and 2003), why they thought stroke was so common in their region, high blood pressure and cigarette smoking accounted for less than 3% of all reasons offered (Dokova et al. 2005).

As shown in Figure 3.6.2, the East/West gap in Europe was mainly created by favourable changes in the West that were not matched in the East. This point has been somewhat obscured by the focus on 'things going wrong in the east', for example, the catastrophic effects of binge drinking in Russian men in the 1990s. But if we take for illustration, say, cardiovascular mortality in middle-aged women in Hungary and Spain (among whom binge

drinking and heavy smoking were not especially prevalent): all of the increased difference between Spanish and Hungarian mortality levels from the 1950s and the 1990s—from Hungarian levels being 1.5 times Spanish levels to being 4.5 times—was created by decreases in mortality in Spain that were unmatched in Hungary (personal calculations using data from the World Health Organization (WHO) mortality database (WHO n.d.)). What mainly needs to be explained is the absence of favourable changes in the East.

The surprising underperformance of the USA

The public health failure of the United States is extraordinary on many counts. It is contrary to expectations based on national income and schooling levels. It is occurring in the face of national strengths in medical and public health science. It is occurring in spite of prodigious levels of spending on clinical medicine. And, far from showing any early effects of corrective action, US mortality trajectories continue to deviate further from trends elsewhere in the developed world.

There is also surprising ethnic selectivity. Hispanics are largely spared, with midlife death rates 30–40% below those for non-Hispanic whites (Woolf et al. 2018)—despite median household wealth just one-tenth that of non-Hispanic whites (Taylor et al. 2011). Mortality among US blacks, though substantially higher than in whites has also been declining faster, with reducing black white differences. It is thus among non-Hispanic whites that trends deviate furthest from expectation. The USA has also experienced large increases in economic inequality, and while these are important in themselves (and dealt with elsewhere in this textbook, in Chapter 6.3), it is not clear that they can explain the unfavourable trends in average health levels (Panel on Understanding Cross-National Health Differences Among High-Income Countries 2013) (Lynch et al. 2004).

The relative failure of health improvement in the USA since the 1980s constitutes an enormous loss. Writing in 2002, the US economist Nordhaus asked his readers to consider the following choice (Nordhaus 2002):

> You must forgo either the health improvements over the last half-century or the non-health improvements. That is, you must choose either (a) 1948 health conditions and 1998 non-health living standards or (b) 1998 health conditions and 1948 non-health living standards. Which would you choose?

Respondents to his straw poll tended slightly to favour b), a result consistent with his more formal analyses:

> The major result that comes through using all techniques is that the value of improvements in life expectancy is about as large as the value of all other consumption goods and services put together.

Applied to the present, the implication is that the USA is losing the equivalent of a substantial fraction of its national income. It cannot be ‘great again’ without finding the means of closing the health gap with the generality of developed countries.

Although the US failure is generally pervasive across child and adult health and across major disease groups and injuries, it does appear that a higher prevalence of obesity and increased consumption of opioids, are making notable contributions (Jeffery et al. 2018; Preston et al. 2018). In this case, the puzzle is pushed one link further back in the causal chain: why are excess adiposity and the liberal prescription of opioids more prevalent in the USA than in Other Developed countries?

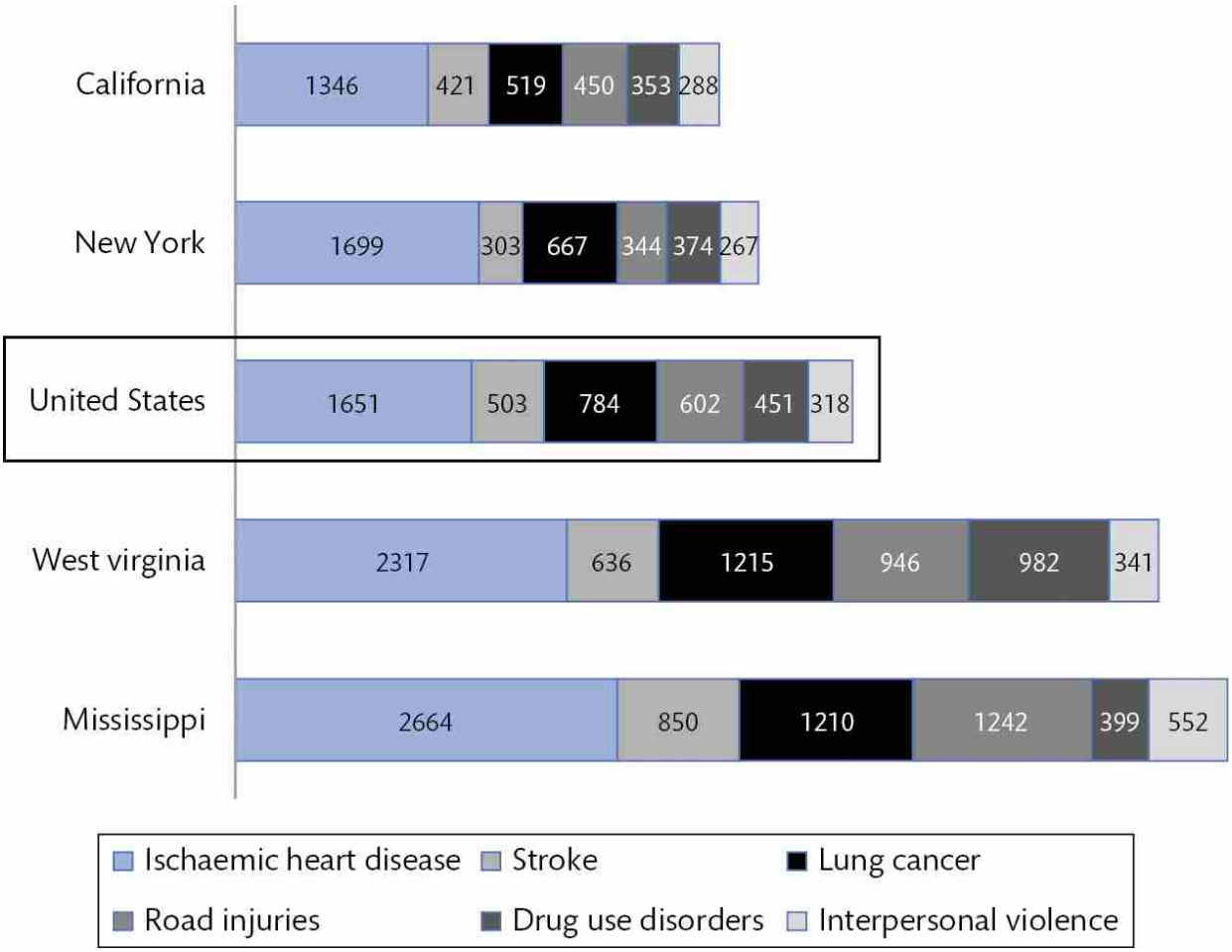


Fig. 3.6.3 Years of life lost in the US and four constituent states from six causes, age-standardized estimates for both sexes, 2016.

Source: data from U. S. Burden of Disease Collaborators (2018), 'The State of US Health, 1990–2016: Burden of Diseases, Injuries, and Risk Factors Among US States. *JAMA*, 319 (14), 1444–72.

So far we have been discussing the USA as though it were spatially homogeneous. In fact it is a very large country with marked internal heterogeneity in health experience across states and counties (see Fig. 3.6.3). Primary responsibility for public health programmes resides with the states (Institute of Medicine 2003, p. 102). The success of Californians in reducing health losses from tobacco have already been noted. This is now reflected in lung cancer mortality, at all ages, being one-third below the national rates (U.S. Burden of Disease Collaborators 2018, Fig. 6)—with even greater differences at the youngest ages. Between 1990 and 2016 California moved from twenty-fourth to second rank for life expectancy at birth. New York state, containing New York City, noted for its vigorous chronic disease prevention programmes (Ong et al. 2017), moved from thirty-ninth to fifth.

(The magnitude of these movements will have been exaggerated by the larger than average effects of the HIV epidemic in these two states in 1990.) By contrast, in 21 ‘downwardly mobile’ states, mortality at ages 20 to 55 actually increased over this period: In Kentucky, Oklahoma, New Mexico, West Virginia, and Wyoming, the increases in mortality exceeded 10%. Comparing lost life years by cause across the states of the USA in 2016 it is apparent that levels of the major causes tend to vary together. Thus rates in California and New York for 16 and 15, respectively, out of 20 causes considered in the GBD study are significantly below national rates. Similarly, states with above average rates of years of life lost tend to have high losses from most causes. The risk factors for premature mortality thus appear to vary together. Figure 3.6.3 illustrates the contribution of major preventable causes to differences in premature mortality between two low mortality states—California and New York—and two high mortality states—West Virginia and Mississippi.

In Eastern Europe there had been state-imposed embargoes on both knowledge and action to counter chronic disease risks; but in the USA relevant knowledge-based institutions are strong. The block in not on knowledge; it is on action. Characterizing institutional differences between countries is not straightforward, but some cross-national comparisons are suggestive.

Programmes to reduce traffic injury are relatively easy to identify and observe, as are their effects. The authors of the US National Academies report, *Shorter lives, poorer health* (Panel on Understanding Cross-National Health Differences Among High-Income Countries 2013) noted with some enthusiasm a study of traffic injury control in the USA and five comparison countries (Transportation Research Board [TRB] 2011): ‘ … the panel was struck by how closely the [Transport Research Board] committee’s overall findings parallel those documented in Part 1 of this report … ’ (p. 226). The TRB report noted how jurisdictions in the USA resist the application of measures proven effective elsewhere for reducing drink driving, increasing seatbelt and motorcycle helmet use, and controlling speed. It provides a detailed comparison of driving speed management in France, the United Kingdom, and Australia with the USA, using as subheadings ‘management and planning’, ‘technical implementation of countermeasures’, and ‘political and public support’. The USA came out poorly under all three heads. The TRB report concluded that ‘Most U.S. state and local transportation safety

agencies lack the institutional and technical capacities required to apply the management techniques observed in the benchmark countries' (p. 4).

Scientific evidence and expert opinion often carry less weight in political deliberations in the USA than in Western Europe. Nathanson contrasts the situation in Britain where '[c]redible scientific knowledge is monopolized and deployed by closely linked experts based within and outside the government—medical civil servants, the British Medical Association, the Royal Colleges, and government appointed advisory bodies' with the situation in the USA where experts 'enjoy no comparable authority' (p. 223).

One of the two dimensions on which the World Values Survey ranks countries from traditional/religious values at one end to secular-rational values at the other. Among developed countries, the USA is placed towards the traditional/religious end of this scale (World Values Survey 2017). Substantial proportions favour creation over evolution as an explanation of human origins (Swift 2017). Arguments based on orthodox interpretations of scientific evidence may face more competition than they would elsewhere. The much delayed uptake in the USA of syringe exchange programmes to help reduce HIV transmission in injecting drug users is often cited as an example of conservative morality blocking effective evidence based measures. It was not until December 2015 that a federal funding ban on these programmes was lifted. This happened after the prescription opioid epidemic led to new epidemics of injecting drug use and HIV transmissions in Republican Party constituencies that happened to be poor and rural, and not in Democrat voting cosmopolitan coastal cities (Showalter 2018).

Commercial interests are powerful and politically well-organized. Lobbying is a $4b 'industry' with one study identifying 37,000 (sic) lobbyists active at federal level in the decade to 2008 (Bertrand et al. 2014). The chronicles of California's tobacco wars provide insight into the challenges of operating in such an environment. The tobacco industry's carefully planned strategy for 'neutralizing' the participation of the California Medical Association to the Proposition 99 campaign in 1987–88 is very sobering (Glanz and Balbach 2000, pp. 46–9).

Responsibilities for public health programmes (like those already noted for traffic injury control) rest at state level. It is likely that the weaknesses noted in traffic injury control apply in other areas of disease and injury prevention at state level including: failure to marshall strong political and community support for preventive programmes; limited managerial capacity

and accountability for achieving programme goals; and lack of longer-term planning horizons. In 2003 an Institute of Medicine report on *The Future of the Public's Health in the 21st Century* concluded that 'the nation's entire governmental public health infrastructure—its human resources, information systems, and organizational capacity—must be revitalized and strengthened' (Institute of Medicine 2003, p. 169). Evidence that this is underway is awaited.

Emulation and globalization of public health strategies

In the previous section, 'Large scale failures of public health policy', we discussed the early, creative phases of societal responses to hazards such as tobacco and traffic injury. However, institutions evolve. For example, opportunities for discussion and comparison of early national experiences in tobacco control were created by the founding of world conferences: the first in New York in 1967, the second in London in 1971—a series that continues to the present with the seventeenth held in Capetown in 2018. The journal *Tobacco Control* was founded in 1992 and has since attained one of the highest impact factors among all public health journals. Knowledge of 'what works' in tobacco control has evolved in parallel with tobacco control activity.

Over time the number and range of civil, professional, and scientific organizations actively supportive of stronger measures to reduce tobacco consumption has grown as has support at the level of national governments and intergovernmental bodies. In the late 1990s these trends coalesced in negotiations which led to the unanimous adoption of the Framework Convention on Tobacco Control at the World Health Assembly of 2003 (Studlar 2006). This has been ratified by all but two of our included countries (USA and Switzerland).

The WHO website now presents a distillation of expert consensus on 'best practice' in tobacco control, summarized in the acronym MPOWER:

Monitor tobacco use and prevention policies

Protect people from tobacco use

Offer help to quit tobacco use

Warn about the dangers of tobacco

Enforce bans on tobacco advertising, promotion, and sponsorship

Raise taxes on tobacco.

The website includes 12 page profiles on each of over 190 countries, tracking their compliance with the MPOWER framework—conveniently and conspicuously summarized in a colour-coded bar (WHO n.d., Tobacco).

Global processes of emulation and policy convergence are not confined to tobacco. Also to be found on the WHO site are a 'Global Status Reports' on Alcohol and Health and on Road Safety (each with single page profiles for all countries) and 'More Active People for a Healthier World' (without country profiles).

Of our group of 39 developed countries, 28 are (currently) members of the European Union. From the late 1980s the European Commission began its engagement with public health by working on a directive to regulate tobacco advertising. This faltered due to member state resistance (especially from Germany) and adverse judicial rulings (Duina and Kurzer 2004). However such a directive was finally passed in 2002, and since then the EU has expanded its engagement in public health, extending from tobacco to alcohol, nutrition, and physical activity along with its previously established role in communicable disease control (European Commission n.d.). It is also active in promoting traffic safety—through the transport portfolio.

For the post-communist countries of Eastern Europe and for 'latecomers' such as the Asian Developed countries, the products of these supranational institutions provide practical guidance and models of 'best practice' for national emulation. They will be contributing, for example, to the recent above average declines in deaths from smoking in the younger cohorts in Eastern Europe—presaging further substantial declines in tobacco attributable mortality as these cohorts age.

The Asian Developed group—Japan, Singapore, and South Korea—have been late starters (with the partial exception of Japan) and rapid improvers for child and adult survival (Fig. 3.6.1 and Table 3.6.1). They have followed technology-intensive development strategies with high investments in schooling and in research. They have been moderate to good adopters of international best practices in tobacco control and traffic safety (based on

WHO score cards). They too can be assumed to be benefitting from supranational developments.

Recognizing that healthy populations are not the result of hospitals and health systems alone, other initiatives which address social determinants of health also have had increasing influence in developed countries. Health in All Policies (HiAP) was developed by the European Union and subsequently shaped by the Commission on the Social Determinants of Health and has been adopted across developed, and more recently, developing countries Kickbusch 2013). Furthermore, the Healthy Cities programme, which was launched over 30 years ago, was a WHO strategy designed to bring Health for All to the local level and recognized the importance of developing enabling environments for healthy lives (Tsouros 2015).

Reflections and prospects

In the period following the Second World War, the developed world faced a continuation of the challenges of communicable disease against which it had enjoyed substantial success earlier in the century and for which it now had new weapons (antibiotics and new vaccines). The momentum of post-war reconstruction provided a favourable political context. In the following decades continuing marked success in reducing deaths from communicable disease in children and adults has been universal.

But in this post-war period new challenges were to emerge for which existing knowledge and institutional forms were not adequate—conditions such as lung cancer, road traffic injuries, and ischaemic heart disease. Strategies that were ultimately to succeed included the building of new scientific bases for tacking non-communicable disease and injuries combined with what turned out to be cumulative piecemeal changes to social institutions related to health damaging behaviour patterns. Especially in their early phases, the actual form and content of the latter varied markedly between countries. Over time, as knowledge deepened and extended, and practical experiences in areas such as tobacco control, traffic injury reduction, and heart disease prevention were compared, it became possible to identify ‘best practice’ strategies and to codify these at an international level. Within countries, such institutions evolved over time, increasingly subject to scientific and public scrutiny. Since the mid-twentieth century risks of premature death have more than halved.

Where political regimes blocked these developments—as in the communist-ruled countries of Eastern Europe—the health of adults either failed to improve, or, especially in the case of men, deteriorated. With the end of communist party rule, these countries have, to varying extents, resumed improving adult mortality trajectories. Countries such as the Czech Republic and Slovenia have been rapid improvers and now have adult male mortality levels that lie within the range of other developed countries and that are only half of those of laggards such as Latvia and Lithuania.

The poor performance of the USA is noteworthy and puzzling. Adult mortality levels (at ages 15–60) are now around twice those of the best performing countries and the gap has continued to widen. Given the several background conditions favourable to the attainment of higher levels of health, suspicion falls on institutional obstacles to more effective action. There is now evidence that some other countries may be joining the US in experiencing slowdowns and reversals of previous mortality declines (Ho and Hendi 2018).

Respect for science, objectivity, and expertise have, in combination with a sense of human solidarity, provided the foundation for major health gains over the past several decades. They are currently under challenge from the upsurge of populism, so the prospects for successfully reversing recent adverse trends remain unclear.

REFERENCES

Action on Smoking and Health (2017). *Key Dates in the History of Anti-Tobacco Campaigning*. London: Action on Smoking and Health.

Berridge, V. (2007). *Marketing Health: Smoking and the Discourse of Public Health in Britain, 1945–2000*. Oxford; New York: Oxford University Press.

Bertrand, M., Bombardini, M., and Trebbi, F. (2014). Is it whom you know or what you know? An empirical assessment of the lobbying process. *American Economic Review*, 104(12), 3885–920.

Borland, R., Winstanley, M., and Reading, D. (2009). Legislation to institutionalize resources for tobacco control: the 1987 Victorian Tobacco Act. *Addiction*, 104(10), 1623–9.

Bornmann, L. and Mutz, R. (2015). Growth rates of modern science: a bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology*, 66(11), 2215–22.

Bridgwood, A., Malbon, G., Lader, D., and Matheson, J. (1996). *Health in England, 1995. What People Know, What People Think, What People Do*. London: Office for National Statistics.

Clarke, A., et al. (2007). A bibliometric overview of public health research in Europe. *The European Journal of Public Health*, 17(Suppl 1), 43–9.

Conference of the Parties to the WHO FCTC (2003). *WHO Framework Convention on Tobacco Control*. Geneva: World Health Organization.

Cowan, R.S. (1983). *More Work for Mother: The Ironies of Household Technology from the Open Hearth to the Microwave*. New York: Basic Books.
Danaei, G., et al. (2013). The global cardiovascular risk transition: associations of four metabolic risk factors with national income, urbanization, and Western diet in 1980 and 2008. *Circulation*, 127(14), 1493–502.
Davison, G. and Yelland, S. (2004). *Car Wars: How the Car Won Our Hearts and Conquered Our Cities*. Crows Nest, Australia: Allen & Unwin.
Deaton, A. (2013). *The Great Escape: Health, Wealth, and the Origins of Inequality*. Princeton, NJ and Oxford: Princeton University Press.
Di Cesare, M., et al. (2013). The contributions of risk factor trends to cardiometabolic mortality decline in 26 industrialized countries. *International Journal of Epidemiology*, 42(3), 838–48.
Dokova, K.G., et al. (2005). Public understanding of the causes of high stroke risk in northeast Bulgaria. *The European Journal of Public Health*, 15(3), 313–16.
Doll, R. and Hill, A.B. (1950). Smoking and carcinoma of the lung; preliminary report. *British Medical Journal*, 2(4682), 739–48.
Duina, F. and Kurzer, P. (2004). Smoke in your eyes: the struggle over tobacco control in the European Union. *Journal of European Public Policy*, 11(1), 57–77.
European Commission (n.d.). *Public Health*. [Online] Available at: https://ec.europa.eu/health
Field, M.G. (1967). *Soviet Socialized Medicine: An Introduction*. New York: Free Press.
GBD 2016 Risk Factors Collaborators (2017). Global, regional, and national comparative risk assessment of 84 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *The Lancet*, 390(10100), 1345–422.
Glanz, S.A. and Balbach, E.D. (2000). *Tobacco War: Inside the California Battles.* Berkeley, CA: University of California Press.
Global Burden of Disease Collaborative Network (2017). *Global Burden of Disease Study 2018 (GBD 2018) Results.* Seattle, WA: Institute for Health Metrics and Evaluation (IHME).
Gouda, H.N. and Powles, J. (2012). Why choice of metric matters in public health analyses: a case study of the attribution of credit for the decline in coronary heart disease mortality in the US and other populations. *BMC Public Health*, **12**, 88.
Ho, J.Y. and Hendi, A.S. (2018). Recent trends in life expectancy across high income countries: retrospective observational study. *British Medical Journal*, **362**, k2562.
Institute of Medicine (2003). *The Future of the Public's Health in the 21st Century*. Washington, DC: The National Academies Press.
Jeffery, M.M., et al. (2018). Trends in opioid use in commercially insured and Medicare Advantage populations in 2007–16: retrospective cohort study. *British Medical Journal*, **362**, k2833.
Keys, A. (1980). *Seven Countries: A Multivariate Analysis of Death and Coronary Heart Disease*. Cambridge, MA: Harvard University Press.
Kickbusch, I. (2013). Health in all policies. *British Medical Journal*, **2013**, 347.
Kmietowicz, Z. (2018). 'Tragedy' of 35 deaths from measles in Europe last year is unacceptable, says WHO. *British Medical Journal*, **360**, k795.
Komlos, J. (1993). The secular trend in the biological standard of living in the United Kingdom, 1730–1860. *Economic History Review*, **46**, 115–44.
Krementsov, N.L. (1997). *Stalinist Science*. Princeton, NJ: Princeton University Press.
Kuznets, S. (1973). Modern economic growth: findings and reflections. *American Economic Review*, 63(3), 247–58.
Lewis, J. (1980). *The Politics of Motherhood. Child and Maternal Welfare in England, 1900–1939*. London: Croom Helm.
Lock, S., Reynolds, L.A., and Tansey, E.M. (1998). *Ashes to Ashes: The History of Smoking and Health*—Symposium and Witness Seminar organized by the Wellcome Institute for the History

of Medicine and the History of Twentieth Century Medicine Group on 26–27 April 1995. Amsterdam: Rodopi.
Lynch, J., et al. (2004). Is income inequality a determinant of population health? Part 1. A systematic review. *Milbank Q*, 82(1), 5–99.
McKee, Martin (2007). Cochrane on Communism: the influence of ideology on the search for evidence. *International Journal of Epidemiology*, 36(2), 269–73.
McKeown, T. (1988). *The Origins of Human Disease*. Oxford: Basil Blackwell.
Mokyr, J. (2002). *The Gifts of Athena: Historical Origins of the Knowledge Economy*, xiii, p. 359. Princeton, NJ: Princeton University Press.
Mokyr, J. and Stein, R. (1997). Science, health and household technology: the effect of the Pasteur Revolution on consumer demand. In: Bresnahan, T.F., et al. (eds.) *The Economics of New Goods*: *Studies in Income and Wealth*. Chicago, IL: University of Chicago Press), pp. 143–205.
Nader, R. (1965). *Unsafe at Any Speed*. New York: Grossman Publishers.
Nathanson, C.A. (2009). *Disease Prevention as Social Change: The State, Society, and Public Health in the United States, France, Great Britain, and Canada*. New York: Russell Sage Foundation, p. 328.
National Institutes of Health (NIH) (n.d.). *The NIH Almanac*. [Online] Available at: http://www.nih.gov/about/almanac/appropriations/index.htm
Nordhaus, W.D. (2002). *The Health of Nations: The Contribution of Improved Health to Living Standards* (Working Paper 8818). Boston, MA: National Bureau of Economic Research.
Ong, P., et al. (2017). Evaluating the effectiveness of New York City health policy initiatives in reducing cardiovascular disease mortality, 1990–2011. *American Journal of Epidemiology*, 186(5), 555–63.
Organisation for Economic Co-operation and Development (OECD) (n.d.). *Member Countries*. [Online] Available at: http://www.oecd.org/about/members-and-partners/
Panel on Understanding Cross-National Health Differences Among High-Income Countries (2013). Woolf, S.H. and Aron, L. (eds.). *U.S. Health in International Perspective: Shorter Lives, Poorer Health*. Washington, DC: National Academies Press.
Peto, R., Watt, J., and Boreham, J. (2013). Deaths from smoking. *International Union Against Cancer*. [Online] Available at: http://www.ctsu.ox.ac.uk/deathsfromsmoking
Pierce, J.P. and Gilpin, E.A. (2001). News media coverage of smoking and health is associated with changes in population rates of smoking cessation but not initiation. *Tobacco Control*, 10(2), 145–53.
Pierce, J.P., et al. (2010). Forty years of faster decline in cigarette smoking in California explains current lower lung cancer rates. *Cancer Epidemiology and Biomarkers Preview*, 19(11), 2801–10.
Pollock, D. (1999). *Denial and Delay: The Political History of Smoking and Health, 1951–1964*. London: Action on Smoking and Health.
Preston, S.H., Vierboom, Y.C., and Stokes, A. (2018). The role of obesity in exceptionally slow US mortality improvement. *Proceeding of the National Academy of Sciences U S A*, 115(5), 957–61.
Preston, S.H. and Haines, M.R. (1991). *Fatal Years. Child Mortality in Late Nineteenth-Century America*. Princeton, NJ: Princeton University Press.
Proctor, Robert (2011). *Golden Holocaust: Origins of the Cigarette Catastrophe and the Case for Abolition*. Berkeley, CA: University of California Press.
Puska, P. (2005). Community change and the role of public health. In: Marmot, M. and Elliott, P. (eds.) *Coronary Heart Disease Epidemiology: From Aetiology to Public Health* (2nd ed.). Oxford: Oxford University Press, pp. 893–907.
Robertson, L.S. (2010). Motor vehicle industry. In: Wiist, W.H. (ed.) *The Bottom Line or Public Health: Tactics Corporations Use to Influence Health and Health Policy and what we can do to counter them*. Oxford: Oxford University Press, pp. 225–48.

Schroeder, S.A. (2004). Tobacco control in the wake of the 1998 master settlement agreement. *The New England Journal of Medicine*, 350(3), 293–301.
Searle, J.R. (2005). What is an institution?. *Journal of Institutional Economics*, 1(1), 1–22.
Searle, J.R. (2010). *Making the Social World: The Structure of Human Civilization*. Oxford: Oxford University Press.
Secretary of State for Social Services (1988). *Public Health in England: The Report of the Committee of Inquiry into the Future Development of the Public Health Function (D Acheson, Chairman)*. Cm 289; London: HMSO.
Showalter, D. (2018). Federal funding for syringe exchange in the US: explaining a long-term policy failure. *International Journal of Drug Policy*, **55**, 95–104.
Smeed, R.J. (1972). The usefulness of formulae in traffic engineering and road safety. *Accident Analysis & Prevention*, **4**, 303–12.
Studlar D.T. (2006). Tobacco control policy instruments in a shrinking world: how much policy learning?. *International Journal of Public Administration*, 29(4–6), 367–6.
Swift, A. (2017). In U.S., belief in creationist view of humans at new low. [Online] Available at: https://news.gallup.com/poll/210956/belief-creationist-view-humans-new-low.aspx
Szreter, S. (1988). The importance of social intervention in Britain's mortality decline c.1850–1914: a re-interpretation of the role of public health. *Journal of the Society for the Social History of Medicine*, **1**, 1–37.
Taylor, P., et al. (2011). Wealth gaps rise to record highs between whites, blacks and Hispanics. Pew Social & Demographic Trends; Washington, DC: Pew Research Center. [Online] Available at: https://www.pewsocialtrends.org/2011/07/26/wealth-gaps-rise-to-record-highs-between-whites-blacks-hispanics/
Terris, M. (1988). Restructuring and accelerating the development of the Soviet health service: preliminary observations and recommendations. *Journal of Public Health Policy*, 9(4), 537–43.
Thelen, K.A. (2004). *How Institutions Evolve: The Political Economy of Skills in Germany, Britain, the United States, and Japan*. Cambridge: Cambridge University Press, p. 333.
Tomes, N. (1998). *The Gospel of Germs: Men, Women, and the Microbe in American Life* Cambridge, MA: Harvard University Press.
Transport Accident Commission (2013). Road safety statistical summary (yearly). *Transport Accident Commission*. [Online] Available at: http://reporting.tacsafety.com.au
Transportation Research Board (2011). *Achieving Traffic Safety Goals in the United States: Lessons from Other Nations*. *TRB Special Report 300*. Washington, DC [Online] Available at: http://onlinepubs.trb.org/onlinepubs/sr/sr300.pdf
Tsouros, A.D., (2015). Twenty-seven years of the WHOP European Healthy Cities movement: a sustainable movement for change and innovation at the local level. *Health Promotion International*, **30**, i3–i7.
U.S. Burden of Disease Collaborators (2018). The State of US Health, 1990–2016: burden of diseases, injuries, and risk factors among US states. *JAMA*, 319(14), 1444–72.
V. Lance Tarrance and Associates (1987). *California Constitutional Amendment Survey: A Confidential Report Prepared for the Tobacco Institute*. Houston, TX: Tarrance Hill Newport and Ryan.
Williamson, Oliver E. (2000). The new institutional economics: taking stock, looking ahead. *Journal of Economic Literature*, 38(3), 595–613.
Woolf, S. H., et al. (2018). Changes in midlife death rates across racial and ethnic groups in the United States: systematic analysis of vital statistics. *British Medical Journal*, **362**, k3096.
World Health Organization (WHO) (n.d.). *WHO Mortality Database*. [Online] Available at: http://www.who.int/healthinfo/mortality_data/en/
World Health Organization (WHO) (n.d.). *Tobacco*. [Online] Available at: http://www.who.int/tobacco/en/

World Values Survey (2017). World values survey, findings and insights. [Online] Available at: http://www.worldvaluessurvey.org/WVSContents.jsp

Wynder, E.L. and Graham, E.A. (1950). Tobacco smoking as a possible etiologic factor in bronchiogenic carcinoma: a study of six hundred and eighty-four proved cases. *JAMA*, **143**, 329–36.

Zylbersztejn, A., et al. (2018). Child mortality in England compared with Sweden: a birth cohort study. *Lancet*, 391(10134), 2008–18.

---

[1] Trofim Denisovich Lesynko (1898–1976) was a Soviet agronomist who rejected Mendelian genetics in favour of theories of environmentally acquired inheritance. He and his theories were supported by Joseph Stalin and in 1940 he was appointed as Director of the Institute of Genetics. Scientists who refused to denounce genetics were dismissed form their posts, imprisoned, and even sentenced to death.

SECTION 4

Information systems and sources of intelligence

4.1 Information systems in support of public health in high-income countries
Tjeerd-Pieter van Staa and Liam Smeeth

4.2 Community diagnosis and health information systems in low- and middle-income countries
Zunyou Wu and Jennifer McGoogan

4.3 New communication technologies, social media, and public health
Patrick S. Sullivan, Aaron J. Siegler, and Lisa Hightow-Weidman

4.1

Information systems in support of public health in high-income countries

Tjeerd-Pieter van Staa and Liam Smeeth

Introduction to information systems

Public health sciences have been described by a Wellcome Trust working group as follows: ‘effective public health actions are based on scientifically derived information about factors influencing health and disease and about effective interventions to change behaviour at the level of the individual, the family, the community or wider society’ (Public Health Sciences Working Group et al. 2004, p. 5). This field of science has made major contributions to the improvement of health such as the effects of sanitary reforms on infant mortality and, more recently, the reductions in cardiovascular disease incidence due to lipid lowering and blood pressure control. However, many challenges remain, such as how to tackle the increased prevalence of obesity and diabetes and how to relate the costs of interventions to their benefits (Public Health Sciences Working Group et al. 2004). This chapter will describe how information systems containing healthcare data can support public health research. We will first describe examples of public health information systems, followed by a description of various developments in these systems, including increased public health use of routinely collected electronic healthcare records (EHRs). More and more clinicians and healthcare professionals are using computers to store information: indeed the ‘meaningful use’ of electronic health records for patient care has recently been mandated in the United States (Blumenthal and Tavenner 2010). These data sources provide opportunities, among others, to test the effectiveness and impact of public health interventions. We discuss the role of randomization in public health research and conclude

this chapter by highlighting the importance of an impact assessment of public health interventions.

Sources of public health information

Vital statistics

Death certificates are routinely collected in many countries in order to report on specific causes of death including, for example, alcohol-related deaths, suicides, and drug-related deaths, deaths involving methicillin-resistant *Staphylococcus aureus* (MRSA) and *Clostridium difficile*, and estimates of excess winter mortality. Trends in, for example, all-age all-cause mortality, deaths from cancer, circulatory diseases, suicide and injury of undetermined intent, and accidents are reviewed periodically as are the association of mortality rates with socioeconomic deprivation. Live and stillbirth figures are also routinely collected in many countries. The birth counts may be stratified by occurrence within or outside marriage, multiple births, mother's area of residence and country of birth, place of confinement, and father's social class (examples are given at http://www.statistics.gov.uk).

Population surveys

Population surveys are frequently used information resources. One example is the US National Health Interview Survey, which targets annually an interview with 40,000 households. Questions include health status and health services utilization and activity limitations due to ill health. The Behavioural Risk Surveillance System is a telephone survey in the US involving 150,000 people. Data are collected on health risks and behaviours, exercise, and diet. The National Health and Nutrition Examination Survey examines diet, nutrition, health behaviours, and risk factors in 5,000 randomly selected persons.

A population census provides information about the characteristics of the population and facilitates understanding of the similarities and differences in populations locally, regionally, and nationally. The results are then used to allocate public money for services including healthcare services. They

are also used to identify areas with the greatest public health needs. As an example, national census data from the Strasbourg metropolitan area in eastern France were used to develop a small-area index of socioeconomic deprivation. It showed an association between increased rates of myocardial infarction and worse neighbourhood deprivation (Havard et al. 2008).

Disease surveillance

In many countries, there are systems for reporting and notification of cases of certain diseases to the public health authorities. Infectious diseases such as yellow fever and diphtheria are notifiable on diagnosis of a suspected case and should not wait for laboratory confirmation. All notifiable cases are then reviewed and appropriate action may be taken. Analyses of local and national trends are typically published on a regular basis. Despite legal requirements to report certain infectious diseases, reporting rates for these conditions are not always optimal. To address this, for example, a New York public health institute facilitated public health reporting by creating alerts within EHRs to remind clinicians at the point of care that a particular diagnosis is reportable and providing a link to the reporting form. Patient demographics are automatically populated in the form (Calman et al. 2012).

Administrative databases

There are many administrative databases that contain information recorded for the purposes of managing the healthcare system. Several of these systems can be used for public health research. The Hospital Episode Statistics is a data warehouse containing details of all admissions to National Health Service hospitals in England. It contains clinical information about diagnoses and operations on individual patients, demographic information such as age, gender, and ethnicity, administrative information such as time waited and date of admission, and geographical information on where the patient was treated and the area in which they lived. Main procedures (such as surgical operations) are also recorded. It contains admitted patient care data from 1989 onwards, with more than 12 million new records added each year, and outpatient attendance data from 2003 onwards, with more than 40 million new records added each year (Health and Social Care Information System n.d.).

Registries

A registry is an organized system that uses observational study methods to collect uniform data (clinical and other) to evaluate specified outcomes for a population defined by a particular disease, condition, or exposure. There are many different types of registries. An example of a product registry is the UK national registry of patients on biological therapy: patients starting such therapy are registered and followed, for example, for long-term safety issues. Healthcare service registries include patient-based individual clinical encounters, such as office visits or hospitalizations, and procedures. Examples include registries enrolling patients undergoing a procedure (e.g. carotid endarterectomy, appendectomy, or primary coronary intervention) or admitted to a hospital for a particular diagnosis (e.g. community-acquired pneumonia). These registries may be used to evaluate the processes and outcomes of care for quality measurement purposes. Disease registries use the state of a particular disease or condition as the inclusion criterion. These registries typically enrol the patient at the time of diagnosis at a routine healthcare service. Examples include cancer registries which are widely used for research (Edwards and Bell 2000).

Immunization registries have been created in order to assemble in one site a record of all immunizations and to provide reminder and recall notices when an immunization is due or late. Vaccination schedules are becoming complex due to the number of different vaccinations a child should receive. These may be further complicated by frequent changes of healthcare provider or insurance company. Immunization registries can provide a central repository of all the vaccinations given to children (Linkins 2001).

Electronic health record databases

EHRs are increasingly being used for research and public health purposes. There are currently over 300 EHR databases in 45 countries (http://www.ispor.org/intl_databases). This section describes a few examples of EHR databases that have been widely used for research. The Clinical Practice Research Datalink, previously known as the General Practice Research Database, collates the anonymized EHR information for over 5 million patients currently registered at participating general

practices. General Practitioners (GPs) play a key role in the UK healthcare system, as they are responsible for primary healthcare and specialist referrals. If a patient is treated in secondary care or by a specialist, the GP is informed about major clinical outcomes, and long-term treatments are frequently handed back to the GP. Almost all GPs in the UK use computers for maintaining health records, and communications between different health providers are increasingly sent electronically (Williams et al. 2012). The Dutch PHARMO Record Linkage System collects information on patient demographics, drug dispensing, hospital morbidity, clinical laboratory and pathology results, and GP information for more than 3 million community-dwelling inhabitants of 48 geographic areas in the Netherlands (Herings and Pedersen 2012). A recent development in the Netherlands is the Mondriaan project that has developed an IT and governance infrastructure for enrichment and linkage of EHR. Privacy enhancing technology such as linkage through a trusted third party (TTP) is applied and currently data from GPs (one million), community pharmacists (12 million), and various other EHR sources are retrieved for research purposes and as feedback information to healthcare providers (www.projectmondriaan.nl). In Denmark and Sweden, each national healthcare system provides universal coverage to all residents (5.5 million inhabitants in Denmark and 9.2 million inhabitants in Sweden). Healthcare coverage includes visits to GPs, specialists, hospital admissions, and outpatient visits; drug costs are either partially or completely covered. A centralized civil registration system has been in place in each country for many years, allowing for personal identification of each person in the entire population and for the possibility of linkage to all national registries containing civil registration numbers, for example, patient registry, cancer registry, prescription databases, and registry of causes of death (Furu et al. 2010). The Rochester Epidemiology Project medical records-linkage system captures healthcare information for the entire population of Olmsted County, Minnesota in the United States. It includes a dynamic cohort of over half a million patients who received healthcare for any reason. The data available electronically include demographic characteristics, medical diagnostic codes, surgical procedure codes, and death information (including causes of death). The system covers residents of all ages and both sexes, regardless of socioeconomic status, ethnicity, or insurance status (St Sauver et al. 2012).

Recent developments in information systems

The electronic information systems in the healthcare system are evolving and increasing (Table 4.1.1). Initial use of electronic healthcare data mostly consisted of aggregate analyses of administrative data such as hospital admission data. When clinicians started to use computers for record keeping, the first research databases collating anonymized EHRs were created. One of the first of these was the VAMP research database that started in 1987. This database eventually developed into the Clinical Practice Research Datalink. The richness and completeness of many EHR databases have been increasing over time as more information is being shared electronically between different parts of the healthcare system. Laboratory data provide an example in which test results are increasingly being communicated electronically and loaded automatically into the patient's EHR.

Table 4.1.1 Stages in the development of EHR databases

Time period (approximate)	Development of EHR databases
1980s onwards	Data collected for administrative purposes (such as hospital admission data and death certificates); mainly used for aggregate analyses
1990s onwards	Clinicians starting to use computers for record keeping (replacing paper records) and data collated into research database. These data initially mainly used for drug safety monitoring
2000s onwards	Monitoring of clinic encounters for symptoms that may represent infectious diseases and other conditions of public health concern
2000s onwards	Linkages between various EHR databases; mainly used to obtain complementary information or to validate outcomes
2010s onwards	Enrichment of routinely collected data by prospective data collection within EHR databases (e.g. collection of blood samples for genetic analyses or patient questionnaires)
2010s onwards	Development and implementation of triage tools to guide clinicians in hospital referral of patients with, e.g. influenza; electronic alerts for eligible patients or patients prescribed unsafe combinations of medications
2010s onwards	Introduction of randomization at the point of care using the EHR database to identify potentially eligible patients and for follow-up collection of major clinical outcomes (i.e. pragmatic and cluster trials); mainly used to evaluate the effects of medicines in routine clinical practice

An important development is the increased linkage between different healthcare databases. Typically, linkages are done by a TTP that collects from each data source the patient's identifiers (such as registration number in the healthcare system, patient's date of birth, gender, and postcode), and research numbers. The TTP then matches the records of the various databases retaining the linked records with the various research numbers without any patient identifiers. This approach allows the combination of different healthcare databases. Linkages between different EHR datasets are being done more frequently, benefiting both the quality and completeness of the EHR data. The UK Clinical Practice Research Datalink has been linked individually and anonymously to other healthcare datasets, including the national registry of hospital admissions, the national death certificates (with primary and secondary cause of death), and prospective disease registries,

such as the cancer and cardiovascular disease registries (Williams et al. 2012).

Comparisons of the information in linked EHR databases can highlight incomplete data records and inform efforts to improve data recording in the healthcare system. In a cohort study, primary care records were linked with those from cancer registries in the UK National Cancer Data Repository. Comparison of the two datasets showed a concordance rate of 83.3%, which varied by cancer type. Cancer registries recorded larger numbers of patients with lung, colorectal, and pancreatic cancers, whereas GPs recorded more haematological cancers and melanomas (Boggon et al. 2013).

Anonymized linkages between local air pollution data, records from primary care, hospital admissions, and death certificates have been used recently to evaluate the relationships between ambient outdoor air pollution and incident myocardial infarction, stroke, arrhythmia, and heart failure. While evidence was weak for relationships with myocardial infarction, stroke, or arrhythmia, consistent associations between pollutant concentrations and incident cases of heart failure were found (Atkinson et al. 2013).

An example of a public health resource that uses linked healthcare databases is the CALIBER programme. It is linking multiple data sources including the longitudinal primary care data from the Clinical Practice Research Datalink, the national disease registry of acute coronary syndrome, hospitalization, and procedure data from hospital admission records, cause-specific mortality from death certificates, and postcode-based social deprivation. Current cohort analyses involve a million people in initially healthy populations and disease registries (Denaxas et al. 2012). A recent study that used this resource was an analysis of the effect of influenza and influenza-like illnesses on triggering acute myocardial infarction (AMI). A self-controlled case series (i.e. only cases were evaluated) identified cases of myocardial infarction from the national registry of acute coronary syndromes and looked for associations with exposure to influenza from the consultation records of GPs. It found that influenza and other acute respiratory infections can act as a trigger for AMI (Warren-Gash et al. 2012).

The Massachusetts Department of Public Health in the United States has developed a system that loads EHRs every 24 hours from clinicians' proprietary software systems. These extracts are organized into separate

databases for patient demographics, vital signs, diagnosis codes, test orders and results, medication prescriptions, allergies, social histories, and provider contact details. These data provide clinically detailed surveillance data. Sensitive and specific disease-detection algorithms have been developed to overcome the limited accuracy of code diagnostic data by considering all available data. Diabetes surveillance is done using these algorithms and can provide information on, for example, the frequency of referrals for medical nutrition counselling stratified by race and locality. A comparison of this system and conventional reporting by clinicians showed a significant increase in the quantity and quality of case reports of notifiable conditions (Klompas et al. 2012).

The minimization of risk of preventable harm to patients is a priority for many healthcare organizations. In Scotland, a system has been implemented in routine primary care that provides a rapid audit method of screening electronic patient records to detect patient harm. The objective of this tool is to support clinical improvement efforts (De Wet and Bowie 2011).

The combination of routinely collected EHRs with prospectively collected data is an important development in improving the quality of information resources. A study in the Clinical Practice Research Datalink recruited about 8,000 persons aged 50–64 year for a survey about working life. Study participants completed questionnaires about their work and home circumstances and the EHRs will be used to analyse health outcomes over time. The inter-relationship between changes in employment (with reasons) and changes in health (e.g. major new illnesses, new treatments, mortality) will be examined. This study will contribute to the evidence base as to whether working beyond the traditional retirement age is feasible for those with major health problems associated with ageing and the effect of occupational and personal circumstances (e.g. savings, retirement intentions, domestic responsibilities, whether work is arduous or rewarding) (Palmer et al. 2015; D'Angelo et al. 2016).

Syndromic surveillance is the practice of monitoring clinic encounters for symptoms that may represent infectious diseases and other conditions of public health concern. There are several examples in which EHRs are used to conduct this surveillance. An example is QSurveillance which provides weekly information on the consultation rates of flu-like illness in general practices covering a total population of almost 22 million patients (QSurveillance n.d.). A comparable system has been developed in New

York State, United States, in which routinely collected EHRs from health centres are transferred daily to the research database. Respiratory illness, fever, diarrhoea, and vomiting are the key symptoms monitored, with analysis to determine when the incidence of these syndromes exceeds expected thresholds (Calman et al. 2012).

The EHRs can also be used to guide clinicians in hospital referral of patients with influenza-like illness during a pandemic. During pandemics of novel influenza and outbreaks of emerging infections, surge in healthcare demand can exceed capacity to provide normal standards of care. In such exceptional circumstances, triage tools may aid decisions in identifying people who are most likely to benefit from higher levels of care. A recent study (called FLU-CATs) developed a system for real-time capture of symptoms in patients presenting in primary care with influenza-like illness. GPs were prompted by a pop-up box to provide symptoms when a diagnosis of influenza-like illness was entered into the EHR. The collected data can then be used to refine and update the triage tools for predicting hospital admission and measure clinical outcomes such as hospital admission or death and GPs' clinical assessments and management decisions. Such a system is also potentially adaptive, since with ongoing data collection, triage tools can be regularly adapted to changes in virus, human behaviour, and models of healthcare provision in the community (Venkatesan et al. 2015).

Pandemic preparedness is a major public health activity in many countries, including the United Kingdom. The major objectives of the UK strategy are, in case of an influenza pandemic: to identify key clinical, epidemiological, and virological features of the new influenza virus; to count severe cases, and identify risk groups affected, to describe the evolving pandemic and its impact at the population level (e.g. by age-group) particularly in relation to hospitalizations and mortality; and to measure the uptake and safety of various pharmaceutical countermeasures. Several information systems would be used to collect the data to be used for this, including primary care consultations, records of calls to telephone help-lines and web-based advisory services relating to influenza-like illness, virological 'sentinel' surveillance schemes in primary care, laboratory analysis of a sample of cases to identify the genetic features of the virus, and analysis of death records (Department of Health 2011).

Future developments in information systems

There is an increasing need to perform studies across different EHR systems and across different countries mainly because of the need for a larger sample size. The healthcare systems in most countries consist of multiple healthcare providers who often use different systems to store data, either on paper or electronically. Furthermore, physicians often record data differently and inconsistently, both on paper and electronically. Various international initiatives have used different approaches to combine heterogeneous EHR databases. One approach focuses on IT aspects with the aim to develop EHR systems that are interoperable and allow seamless transfer of data (http://www.transformproject.eu/; http://www.ehr4cr.eu/). All the information in the various databases is mapped together with a detailed understanding of the content of each of the data elements and use of a single coding dictionary across the different databases. An alternative approach is to maintain the EHR data structure as collected by the participating health professionals but to develop a common protocol across the different databases. The operational definitions of how the data are classified will vary by individual EHR database but the research questions will be kept similar. This model is currently used by the Innovative Medicines Initiative PROTECT project (http://www.imi-protect.eu/). The third approach is used by the OMOP initiative in the United States: all EHR data from the different databases are integrated into a central research database according to a common data model (http://omop.fnih.org/). A distributed network model where basic analyses are run on federated datasets generating common input data and subsequent local aggregation and central pooling of results, constitutes the fourth approach to dealing with heterogeneous EHR databases. The EU-ADR project uses this approach. In one of their studies, data from eight European healthcare databases (administrative claims, medical records) were combined. This data set included over 1.9 million individuals (59,594,132 person-years follow-up) who used 2,289 different drugs. It found for a frequent event such as AMI, there were 531 drugs (23% of total) for which an association with relative risk ≥2, if present, can be investigated. For a rare event such as rhabdomyolysis, there are 19 drugs (1%) for which an association of same magnitude can be investigated. It concluded that even larger databases

would be required to detect signals for less frequent exposures and outcomes (Coloma et al. 2012).

A recent development is the use of EHRs to measure the quality of clinical care and provide financial incentives through 'pay-for-performance' programmes (Campbell et al. 2009). EHRs are used to evaluate clinician and system performance with the goal of making healthcare safer and more efficient. The adoption of computerized clinical records itself may be associated with improved care and outcomes (Cebul et al. 2011) although some doubts remain (Classen and Bates 2011). There are several different dimensions in which quality of healthcare can be measured. Examples of different quality measures include the number of patients with diabetes seeing an ophthalmologist, clinical outcomes in patients with hypertension, percentage of clinicians reviewing out-of-range laboratory results with a certain number of hours, and the percentage of patients receiving incorrect medications (Weiner et al. 2012). The quality and completeness of the EHRs are of course critical in achieving the goals of the 'pay-for-performance' programmes.

Internet postings, blogs, and social media postings may provide another source of information for the early detection of new infectious disease epidemics. A recent study analysed the queries to online search engines, which are used by millions of people around the world each day. It concluded that this approach accurately estimated the level of weekly influenza activity in each region of the United States, with a reporting lag of about 1 day. The use of search queries could detect influenza epidemics in areas with a large population of web search users (Ginsberg et al. 2009). This potential use of Internet searches was tested in a US study that compared Internet searches with rates of outpatient visits for influenza-like illness and of confirmed laboratory tests. It was found that Internet searches were highly correlated with rates of outpatient visits for influenza-like illness but less correlated with rates of laboratory-confirmed influenza (Ortiz et al. 2011). The main limitation of these analyses of unstructured Internet data is that it may be difficult to separate activities by concerned healthy people from those with the disease of interest. Healthcare seeking behaviour, physician testing practices, and Internet search behaviour can be influenced by high levels of media coverage (Ortiz et al. 2011).

The information on the Internet can also be evaluated for the monitoring of side effects of medicines. A study evaluated the text on Internet message

boards dedicated to drug abuse in order to compare different prescription opioids. Over 48,000 posts were analysed and the unstructured text was coded. It was found that the number of posts on these drug abuse message boards varied between the different opioids (Butler et al. 2007). While this approach is interesting, one cannot conclude that these differences in drug abuse messages were caused by the pharmacological differences between the opioids. Different levels of information provision by, for example, the clinicians or in the information sheet or prescribing to different patient populations could also lead to different levels of Internet activities.

Randomization in public health research

The present allocation of healthcare interventions is often inconsistent. Clinicians will be influenced in their prescribing behaviour, not only by clinical considerations, but also by other issues including drug preferences, exposure to marketing materials, and guidance from the local healthcare funders, which will vary considerably between practices and clinicians (Adamson et al. 2012). Clinicians will respond differently to uncertainties in the evidence base. The prescribing of antibiotics to patients with mild to moderate exacerbations of chronic obstructive pulmonary disease provides an example of this. A review of EHRs found major differences in the rate of antibiotics prescribing to these patients. Very few patients received an antibiotic in some clinics while antibiotics were routinely prescribed in other clinics. This variability in care (due to a lack of randomized trials conducted in these patient groups) produces a potentially unfair lottery for the receipt of healthcare interventions. Treating patients in this inconsistent manner generates no new evidence to improve clinical practice. On the other hand, randomization with structured data collection would provide the evidence to guide clinicians in selecting treatments. A possible model would be to offer randomization to all willing patients as part of routine clinical care in every situation where there is genuine uncertainty about which of two or more widely accepted treatments is best. The EHRs could be used to measure major clinical outcome and follow progress (van Staa et al. 2012). An analysis of the trial evidence for medicines used by millions of patients and with blockbuster sales showed that only one of the 24 top blockbusters in 2011 had a randomized trial including more than 10,000

participants and few of the blockbusters had evidence of beneficial effects on mortality (Ioannidis 2013). A mega-trial with mortality as an outcome could address concerns about side effects and also provide evidence of effectiveness of medicines in routine clinical practice.

EHR databases could offer an ideal platform to undertake large pragmatic trials, with randomization at the point of care and collection of follow-up data using the EHR (van Staa et al. 2012). Such an approach would allow the assessment of effectiveness of healthcare interventions in everyday clinical practice among representative populations. A recent study developed and evaluated the methods to implement simple pragmatic trials in primary care, identified the barriers and facilitators for clinicians and patients, as well as the experiences of trial participants. As part of this study, patients with high cardiovascular risk were randomized between simvastatin and atorvastatin (van Staa et al. 2012). Potentially eligible patients were identified in the EHRs using risk prediction tools. Outcome data such as continuation of statin treatment and occurrence of heart attacks were collected using the routinely collected data from primary care, linked data from hospital admissions, a prospective disease registry, and death certificates. This study concluded that EHR point-of-care trials are feasible, although the recruitment of clinicians is a major challenge owing to the complexity of trial approvals. These trials will provide substantial evidence on clinical effectiveness only if trial interventions and participating clinicians and patients are typical of usual clinical care and trials are simple to initiate and conduct (van Staa et al. 2014).

Discontinuation and non-compliance with medication over time can be an important outcome measurement in pragmatic trials. For example, if patients preferentially use e.g. atorvastatin for a longer period of time, this could mean a substantially lower number of patients suffering a heart attack. In most trials, study participants are regularly monitored and instructed how to take their medicines. But the rates of persistence are often considerably lower in routine clinical practice, as stringent monitoring procedures do not apply. Pragmatic trials that collect follow-up information unobtrusively (e.g. from EHR databases) would provide the answers that decision makers need: that is, will this intervention make a difference in routine clinical practice compared to alternative strategies (Tunis et al. 2003).

The TRANSFoRm project developed an approach to embed trial functionalities within the EHRs including the pre-population of electronic Case Report Forms directly from the EHRs and the potential to record trial-specific data at the point of care. In addition, it also implemented data standards which allowed users to query the EHRs without a detailed understanding of how data were originally recorded and stored in the EHRs and what coding system was used (i.e. interoperable data systems) (Ethier et al. 2017). The (TRANSFoRm) clinical trial tools were then used for automated identification, recruitment, and follow-up. A small feasibility study was conducted in three general practices in Poland. Participants were physicians and patients with gastro-oesophageal reflux disease. It found that physicians were satisfied with the usefulness of the system, as it enabled easier and faster identification, recruitment and follow-up of patients compared with existing methods (Mastellos et al. 2016).

Cluster trials randomly allocate entire areas or health service organizational units to intervention or control groups, with outcomes evaluated for individuals within each cluster. They facilitate pragmatic evaluation of the effectiveness of interventions delivered in routine practice settings. Public health interventions could be tested in cluster trials as one can, for example, implement the novel intervention in one set of randomly selected clinics and the old intervention in the remaining clinics. The effectiveness of screening programmes could also be tested in cluster trials, with half of the clinics conducting the screening and the other half not. An example of a study that used EHR data is a study that tested the effectiveness of providing clinicians with the guideline recommendations on antibiotic prescribing in respiratory illness. Electronic prompts are activated during a patient's consultation for upper respiratory illness. The EHRs are used to measure the outcomes of interest, which include the rate of antibiotic prescribing (Gulliford et al. 2011).

Phased access to new interventions has been proposed as a method to measure effectiveness and safety. Practices or regions would be randomized to early or late access and the EHR data would be used to measure the outcomes of interest. Because practices are randomly chosen for each new intervention, a practice will have access to some of the new interventions (Adamson et al. 2012).

Impact assessment of public health interventions

The measurement of outcomes of public health interventions in routine clinical practice should be considered a major public health activity. This impact assessment can inform the extent of uptake but also identify any issues with the implementation of the public health activity. However, this impact assessment is not consistently done. For example, measurements of outcomes are lacking in a major UK programme that targeted the use of statins in patients with high cardiovascular risk. Following a detailed review of evidence, the UK National Health Service introduced a population-wide vascular risk assessment programme. This consists of a systematic approach to assessing risk of vascular diseases for everyone between 40 and 74 years who is not yet diagnosed with cardiovascular disease or treated for risk factors. Statin treatment should be initiated if a patient has a high risk (20% or greater 10-year risk) of cardiovascular disease (National Health Service Health Check Programme 2008). However, a recent analysis of the EHRs found that many healthy patients were prescribed a statin despite having a below-threshold cardiovascular risk and that there was wide variation between practices in the extent of statin prescribing to patients at high risk.

Another example of the difference between intended use according to guidelines and actual use in clinical practice concerns the selective cyclooxygenase-2 inhibitors (coxibs). These drugs ranked, before September 2004, among the most commonly used medications in the world. They were developed to minimize the upper gastrointestinal side effects of conventional non-steroidal anti-inflammatory drugs. A large number of cost-effectiveness analyses were conducted in order to provide clinicians with guidance on which patient groups should be treated with coxibs. These analyses were based on mathematical models that took into account the higher prescription costs and lower incidence of gastrointestinal side effects of coxibs. The assumptions that were used in these models were obtained from the main coxibs randomized trials. However, an analysis of EHRs found that the vast majority of coxib users in routine clinical practice would not have been eligible for the main coxibs randomized trials, as they did not have osteoarthritis or rheumatoid arthritis and only used coxibs intermittently or short term. Thus, the many cost-effectiveness analyses that were conducted for coxibs lacked external validity and were of limited value in guiding clinicians on how to treat patients in routine clinical

practice. The authors of the EHR analyses concluded that the field of health technology assessments should move from evaluating cost-*efficacy* in ideal (hypothetical) populations with ideal interventions, to cost-*effectiveness* in real populations with pragmatic interventions (van Staa et al. 2009). The clinicians were not provided with any guidance on how to use coxibs in the vast majority of patients who could be treated with coxibs. Public health interventions should be developed with a focus on who could be targeted in routine clinical practice rather than on who had been included into randomized trials. It has been proposed that public health programmes should be preceded by a systematic review of the research evidence assessing the likely effects of these programmes followed by an impact evaluation after launch (Oxman et al. 2010).

Scientific challenges in research with electronic healthcare data

Data quality is of course very important for research that uses EHR data. There are several dimensions of data quality. The accuracy and validity of the information in EHR databases will depend on the level and specificity of the coding of medical data at the clinics. A clinic that mostly records data using unstructured free-text or using non-specific codes will provide data that are less useful for research. Data quality also depends on the completeness of information. A clinic that is not routinely informed about, for example, hospitalizations of their patients will not provide complete data, even if the clinic codes the medical data to highest standards. Reliability of information includes the level of changes in data collection over time. As an example of secular changes, the Quality and Outcomes Framework introduced in England in 2004 resulted in a substantive increase of the data included in this framework. Relevance of the collected data for the desired use is another component of data quality. An EHR database that is based on records from GPs may not contain all the data considered relevant by a specialist for making a diagnosis. Timeliness of information processing is also critical for data quality (Audit Commission 2007). Studies that use EHR data for public health research will need to consider these different aspects of data quality.

There are many different EHR systems and similar information may be recorded and stored differently, making it difficult to extract information in a standardized manner or to share information between different partners. Complex systems may need to be built in order to make computer systems interoperable and to exchange EHR data with public health agencies. An example of such a system is the US National Electronic Disease Surveillance System. This is a web-based infrastructure for public health surveillance data exchange between the Centers of Disease Control and the 50 US states. Forty-seven states had in 2010 fully operational general communicable disease electronic surveillance systems, of which 39 states had systems that were interoperable and 42 states had the capacity to receive electronic laboratory reports (Centers for Disease Control and Prevention 2011).

Ethical aspects of research with electronic healthcare data

Healthcare data contain sensitive information about individuals. Adequate protection of individuals' privacy and data security are very important requirements for researchers to adhere to. EHR databases typically remove patients' identifiers (such as name, address, and postcode) from the research database. In some databases, researchers will never be able to contact the clinician or patient; these data are considered fully anonymized. In other databases, researchers will not know the patients' identifiers but they are able to contact the clinician or patients through a gatekeeper; these data are considered pseudo-anonymized. In the Clinical Practice Research Datalink, the patient's clinician can be contacted (following approval by an ethics committee), who can then review the request and contact the patient, if appropriate.

There is considerable debate about whether (pseudo)anonymized healthcare data should be made available for research. In an opt-in system, data can only be made available to researchers dependent on their informed consent. In an opt-out system, patients can refuse to have their data used for research. In other systems, anonymized data are considered exempt from consent or dissent requirements and can be used for research as long as the data do not contain patient identifiers. The use of health data for research raises complex ethical questions of privacy. One interesting approach to

evaluating public acceptability of different approaches to challenging ethical questions like this is the use of citizens' juries. They are based on the premise that, given time, opportunity, support and resources, ordinary people can make decisions about complex matters. A project in the North of England explored what control informed citizens would seek over the use of EHRs for research after participating in a deliberative process using citizens' juries. They addressed the question of to what extent should patients control access to EHRs for research use? At the end of several days of deliberation, 33 out of 34 jurors voted in support of the secondary use of data for research, with 24 wanting individuals to be able to opt out, 6 favouring opt in, and 3 voting that all records should be available without any consent process. All jurors thought that public benefit was a key justification for access (Tully et al. 2018).

The critical question is whether the right of data privacy trumps all other rights and duties or whether there is a balance between different considerations. There is the right of patients to receive proper treatments and the duty of the healthcare to, for example, monitor treatments for effectiveness and safety, and be cost-effective. A healthcare system that generates and applies the best evidence for collaborative healthcare choices of each patient and provider has been defined as a learning healthcare system (McGinnis 2010). Such a system would continuously test interventions and collect data on the outcomes and then use the results to inform and improve clinical practice. Scientific research including public health requires high-quality data, and a learning healthcare system that aims to continuously improve cannot be achieved without such data. It has been proposed that data governance should be viewed as a matter of weighing up the value accruing to individuals, such as privacy and consent, against the value that the research may generate for the public (Rumbold et al. 2011).

The discussion about rights of data privacy and the research use of EHRs should not be restricted to abstract legal notions but also consider the likelihood of data privacy breaches and how these can be minimized. Staff training and standard procedures are essential, but the skills and attitudes of staff are also very important to ensure that data are treated with appropriate care. Regular audits of data security by external experts can also help to maintain a culture of continuous improvement (MacKenzie et al. 2011). With appropriate data security procedures, the risk of breaches of data privacy could be minimized. Also, the standards and quality of research are

important considerations in the balancing of rights of data privacy and the research use of EHRs. Registration of the study prior to the start of the analyses and external access to protocols after completion of the analyses have been advocated strongly for randomized trials (Chan et al. 2006). External access to protocols and the possibility for independent researchers to replicate the study findings could help to improve research standards. A public health system that uses EHRs for high-quality and transparent research and with minimal risk of privacy breaches is clearly ethically superior to a system that applies few standards.

Conclusion

Information systems are critical for public health activities with information being exchanged between public health agencies, clinicians/healthcare providers, individuals, and communities. The EHRs can provide information on diseases and healthcare activities and electronic alerts within the EHRs can be used to remind clinicians of necessary or required public health activities. The effectiveness of some public health interventions can be monitored and also tested using the EHRs. Better data on risks and benefits of healthcare interventions can also help to improve decision-making and informed consent by individuals (Calman et al. 2012). The increasing computerization of the healthcare system will offer many important opportunities to improve public health activities.

REFERENCES

Adamson, J., van Staa, T., and Torgerson, D. (2012). Assuring the safety and effectiveness of new drugs: rigorous phase IV trials randomizing general practices to delayed access to new drugs. *Journal of Health Services Research & Policy*, 17(1), 56–69.

Atkinson, R.W., Carey, I.M., Kent, A.J., van Staa, T.P., Anderson, H.R., and Cook, D.G. (2013). Long-term exposure to outdoor air pollution and incidence of cardiovascular diseases. *Epidemiology*, 24(1), 44–53.

Audit Commission (2007). *Improving Information to Support Decision Making: Standards for Better Quality Data*. Report. London: Audit Commission.

Blumenthal, D. and Tavenner, M. (2010). The 'meaningful use' regulation for electronic health records. *The New England Journal of Medicine*, 363(6), 501–4.

Boggon, R., van Staa, T.P., Chapman, M., Gallagher, A.M., Hammad, T.A., and Richards, M.A. (2013). Cancer recording and mortality in the General Practice Research Database and linked cancer registries. *Pharmacoepidemiology and Drug Safety*, 22(2), 168–75.

Butler, S.F., Venuti, S.W., Benoit, C., Beaulaurier, R.L., Houle, B., and Katz, N. (2007). Internet surveillance: content analysis and monitoring of product-specific internet prescription opioid abuse-related postings. *The Clinical Journal of Pain*, 23(7), 619–28.

Calman, N., Hauser, D., Lurio, J., Wu, W.Y., and Pichardo, M. (2012). Strengthening public health and primary care collaboration through electronic health records. *American Journal of Public Health*, 102(11), e13–18.

Campbell, S.M., Reeves, D., Kontopantelis, E., Sibbald, B., and Roland, M. (2009). Effects of pay for performance on the quality of primary care in England. *The New England Journal of Medicine*, 361(4), 368–78.

Cebul, R.D., Love, T.E., Jain, A.K., and Hebert, C.J. (2011). Electronic health records and quality of diabetes care. *The New England Journal of Medicine*, 365(9), 825–33.

Centers for Disease Control and Prevention (CDC) (2011). State electronic disease surveillance systems—United States, 2007 and 2010. *MMWR Morbidity and Mortality Weekly Report*, 60(41), 1421–3.

Chan, A.W., Upshur, R., Singh, J.A., Ghersi, D., Chapuis, F., and Altman, D.G. (2006). Research protocols: waiving confidentiality for the greater good. *British Medical Journal*, 332(7549), 1086–9.

Classen, D.C. and Bates, D.W. (2011). Finding the meaning in meaningful use. *The New England Journal of Medicine*, 365(9), 855–8.

Coloma, P.M., et al. (2012). Electronic healthcare databases for active drug safety surveillance: is there enough leverage? *Pharmacoepidemiology and Drug Safety*, 21(6), 611–21.

D'Angelo, S., et al. (2016). Job dissatisfaction and the older worker: baseline findings from the Health and Employment After Fifty study. *Occupational & Environmental Medicine*, 73(8), 512–19.

Denaxas, S.C., et al. (2012). Data resource profile: cardiovascular disease research using linked bespoke studies and electronic health records (CALIBER). *International Journal of Epidemiology*, 41(6), 1625–38.

Department of Health (2011). *UK Influenza Pandemic Preparedness Strategy 2011*. London: Department of Health. [Online] Available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/213717/dh_131040.pdf

De Wet, C. and Bowie, P. (2011). Screening electronic patient records to detect preventable harm: a trigger tool for primary care. *Quality in Primary Care*, 19(2), 115–25.

Edwards, D. and Bell, J. (2000). Cancer registries—future development and uses in Britain. *Journal of Public Health Medicine*, 22(2), 216–19.

Ethier, J.F., et al. (2017). eSource for clinical trials: implementation and evaluation of a standards-based approach in a real world trial. *International Journal of Medical Informatics*, 106, 17–24.

Furu, K., Wettermark, B., Andersen, M., Martikainen, J.E., Almarsdottir, A.B., and Sørensen, H.T. (2010). The Nordic countries as a cohort for pharmacoepidemiological research. *Basic & Clinical Pharmacology & Toxicology*, 106(2), 86–94.

Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S., and Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, 457(7232), 1012–14.

Gulliford, M.C., et al. (2011). Cluster randomised trial in the General Practice Research Database: 1. Electronic decision support to reduce antibiotic prescribing in primary care (eCRT study). *Trials*, 12, 115.

Havard, S., Deguen, S., Bodin, J., Louis, K., Laurent, O., and Bard, D. (2008). A small-area index of socioeconomic deprivation to capture health inequalities in France. *Social Science & Medicine*, 67(12), 2007–16.

Health and Social Care Information System (n.d.). *Hospital Episode Statistics.* [Online] Available at: http://www.hscic.gov.uk/hes

Herings, R.C. and Pedersen, L. (2012). Pharmacy-based medical record linkage systems. In: Strom, B.L., Kimmel, S.E., and Hennessy, S. (eds.) *Pharmacoepidemiology* (5th ed.). New York: Wiley-Blackwell, pp. 270–86.

Ioannidis, J.P.A. (2013). Mega-trials for blockbusters. *The Journal of the American Medical Association*, 309(3), 239–40.

Klompas, M., et al. (2012). Integrating clinical practice and public health surveillance using electronic medical record systems. *American Journal of Preventive Medicine*, 42(6 Suppl. 2), S154–62.

Linkins, R.W. (2001). Immunization registries: progress and challenges in reaching the 2010 national objective. *Journal of Public Health Management and Practice*, 7(6), 67–74.

Mackenzie, I.S., Mantay, B.J., McDonnell, P.G., Wei, L., and MacDonald, T.M. (2011). Managing security and privacy concerns over data storage in healthcare research. *Pharmacoepidemiology and Drug Safety*, 20(8), 885–93.

Mastellos N, et al. (2016). Feasibility and acceptability of TRANSFoRm to improve clinical trial recruitment in primary care. *Family Practice*, 33(2), 186–91.

McGinnis, J.M. (2010). Evidence-based medicine—engineering the Learning Healthcare System. *Studies in Health Technology and Informatics*, 153, 145–57.

National Health Service Health Check Programme (2008). *Putting Prevention First—Vascular Checks: Risk Assessment and Management.* London: Department of Health. [Online] Available at: http://www.healthcheck.nhs.uk/Library/F26E1C01d01.pdf

Ortiz, J.R., Zhou, H., Shay, D.K., Neuzil, K.M., Fowlkes, A.L., and Goss, C.H. (2011). Monitoring influenza activity in the United States: a comparison of traditional surveillance systems with Google Flu Trends. *PLoS One*, 6(4), e18687.

Oxman, A.D., et al. (2010). A framework for mandatory impact evaluation to ensure well informed public policy decisions. *The Lancet*, 375(9712), 427–31.

Palmer, K.T., et al. (2015). Health and Employment after Fifty (HEAF): a new prospective cohort study. *BMC Public Health*, 15, 1071.

Public Health Sciences Working Group convened by the Wellcome Trust (2004). *Public Health Sciences: Challenges and Opportunities.* London: The Wellcome Trust. [Online] Available at: http://www.wellcome.ac.uk/stellent/groups/corporatesite/@policy_communications/documents/web_document/wtd003191.pdf

QSurveillance (n.d.). Website. [Online] Available at: http://www.qsurveillance.org

Rumbold, B., Lewis, G., and Bardsley, M. (2011). *Access to Person-Level Data in Health Care: Understanding Information Governance.* London: Nuffield Trust. [Online] Available at: http://www.nuffieldtrust.org.uk/publications/access-person-level-data-health-care-understanding-information-governance

St Sauver, J.L., et al. (2012). Data resource profile: the Rochester Epidemiology Project (REP) medical records-linkage system. *International Journal of Epidemiology*, 41(6), 1614–24.

Tully MP, et al. (2018). Investigating the extent to which patients should control access to patient records for research: a deliberative process using citizens' juries. *Journal of Medical Internet Research*, 20(3), e112.

Tunis, S.R., Stryer, D.B., and Clancy, C.M. (2003). Practical clinical trials: increasing the value of clinical research for decision making in clinical and health policy. *The Journal of the American Medical Association*, 290(12), 1624–32.

Van Staa, T.P., Leufkens, H.G., Zhang, B., and Smeeth, L. (2009). A comparison of cost effectiveness using data from randomized trials or actual clinical practice: selective cox-2 inhibitors as an example. *PLoS Medicine*, 6(12), e1000194.

Van Staa, T.P., et al. (2012). Randomised Evaluations of Accepted Choices in Treatment (REACT) trials: large-scale pragmatic trials within databases of routinely collected electronic healthcare records. *British Medical Journal*, 344, e55.

van Staa, T.P., et al. (2014). The opportunities and challenges of pragmatic point-of-care randomised trials using routinely collected electronic records: evaluations of two exemplar trials. *Health Technology Assessment*, 18(43), 1–146.

Venkatesan, S., et al. (2015). Development of processes allowing near real-time refinement and validation of triage tools during the early stage of an outbreak in readiness for surge: the FLU-CATs Study. *Health Technology Assessment*, 19(89), 1–132.

Warren-Gash, C., et al. (2012). Influenza infection and risk of acute myocardial infarction in England and Wales: a CALIBER self-controlled case series study. *The Journal of Infectious Diseases*, 206(11), 1652–9.

Weiner, J.P., Fowles, J.B., and Chan, K.S. (2012). New paradigms for measuring clinical performance using electronic health records. *International Journal for Quality in Health Care*, 24(3), 200–5.

Williams, T., van Staa, T.P., Puri, S., and Eaton, S. (2012). Recent advances in the utility and use of the General Practice Research Database as an example of a UK Primary Care Data resource. *Therapeutic Advances in Drug Safety*, 3, 89–99.

4.2

Community diagnosis and health information systems in low- and middle-income countries

Zunyou Wu and Jennifer McGoogan

Introduction

Broad recognition of health as a basic human right began with the formation of the World Health Organization (WHO) in 1946 (see Box 4.2.1), as a specialized agency within the United Nations (UN) (WHO 1946). Twenty years later in 1966, the International Covenant on Economic, Social and Cultural Rights was ratified by the UN General Assembly (GA). Article 12 of the Covenant reinforces international commitment to protecting health as a natural human right, recognizing *'the right of everyone to the enjoyment of the highest attainable standard of physical and mental health'* (UN GA 1966). Since then, these agreements have been followed by other similar commitments; for example, the Declaration of Alma-Ata (1978), the Health For All goals (1981), the World Medical Association Declaration of Ottawa (1998), and the Abuja Declaration (2001).

Box 4.2.1 Constitution of the World Health Organization (WHO)

WHO founding principles (WHO 1946):

- Health is a state of complete physical, mental, and social well-being and not merely the absence of disease or infirmity.
- The enjoyment of the highest attainable standard of health is one of the fundamental rights of every human being without distinction of race, religion, political belief, economic or social condition.
- The health of all peoples is fundamental to the attainment of peace and security and is dependent on the fullest co-operation of individuals and states.
- The achievement of any State in the promotion and protection of health is of value to all.
- Unequal development in different countries in the promotion of health and control of diseases, especially communicable disease, is a common danger.
- Healthy development of the child is of basic importance; the ability to live harmoniously in a changing total environment is essential to such development.
- The extension to all peoples of the benefits of medical, psychological, and related knowledge is essential to the fullest attainment of health.
- Informed opinion and active co-operation on the part of the public are of the utmost importance in the improvement of the health of the people.
- Governments have a responsibility for the health of their peoples which can be fulfilled only by the provision of adequate health and social measures.

In 2000, the UN Millennium Declaration, with the support of 189 of its member states as well as 22 international organizations, launched the Millennium Development Goals (MDGs)—eight international development goals meant to be achieved by 2015 (UN GA 2000). MDG 4 (Reduce Child Mortality), MDG 5 (Improve Maternal Health), and MDG 6 (Combat HIV/AIDS, Malaria and Other Diseases) were meant to directly influence health. This helped to galvanize national and international health organizations by bringing awareness to these issues and pushing them to the forefront of political agendas worldwide. It also helped to provide renewed momentum, as considerable funds, earmarked for initiatives aimed at achieving the MDGs, became available through international aid organizations such as the United States President's Emergency Plan for AIDS Relief (PEPFAR), the President's Malaria Initiative (PMI), the GAVI Vaccine Alliance, The Bill & Melinda Gates Foundation, and The Global Fund to Fight AIDS, Tuberculosis and Malaria (The Global Fund).

The aspirations that the MDGs embodied were ambitious, but the tactical work that would be required to achieve them fit neatly into the usual purview of public health, which had already long been doing the work of elevating average health status and reducing burden of disease, combatting health inequity and inequality, ensuring provision of basic services, and improving health system efficiency, all at the community level. Thus, one of the near-immediate positive outcomes of the MDGs was the recognition of the human

race itself as a community and hence, ‘global health’ or ‘global public health’ as a concept (Marten 2019), a movement which had first begun in the mid-1990s with a refocusing of efforts at the WHO (Brown et al. 2006).

While overall, the MDGs were generally considered successful—development was accelerated, lives were saved, people were lifted out of poverty (UN 2015)—nearly every aspect of the MDGs were also heavily criticized (Fehling et al. 2013; Marten 2019). For example, the MDGs were written primarily by the United States, Europe, and Japan, with the heavy influence of large international aid organizations. Neither the citizens nor the governments of the developing nations the MDGs were meant to help were involved in their creation. Moreover, although the MDGs were written as global goals, assessments of performance against the goals were repeatedly applied at the individual country level. The mismatch between global-level goal setting and country-level performance measurement meant that many countries struggled and some were, arguably unfairly, labelled as ‘off-track’. Furthermore, the sense of futility felt in countries where the MDGs were not realistically feasible was harmful and, in some cases, counterproductive to growth and development of local public health infrastructure, governance, and accountability (Cohen et al. 2014; Easterly 2009; Easterly 2013; Fehling et al. 2013; Vandemoortle 2009).

This problem was further exacerbated by insufficiencies in data availability, quality, timeliness, and potential for aggregation and disaggregation, in statistical methodologies and capacity, and in national (and international) information systems. Although over the course of the MDG era (2000–2015), more and better data became available, coordination with statistical systems improved, and new statistical methodologies were developed, these issues were serious obstacles to progress in the MDG era (2000–2015) (UN 2015), and remain so today.

Not surprisingly, it was low- and middle-income countries (LMIC; see Box 4.2.2), in particular, that faced unreachable standards for success. These nations generally did not have the data collection infrastructure and information management systems to support measurement of current state, much less evaluation of achievement over time. For instance, some countries’ civil registration systems were inefficient and ineffective, leading to large gaps in vital statistics and related data. Consequently, governing bodies at local, subnational, and national levels relied on incomplete, outdated, and/or low-quality evidence for policy development and planning purposes. Thus, experience gained in monitoring progress toward the MDGs

underscored the power of effective use of data for setting appropriate targets, focusing efforts, successfully implementing interventions, assessing performance, and increasing accountability (UN 2015).

Box 4.2.2 Low- and middle-income countries (LMIC)

List of LMIC defined by the World Bank as having a gross national income per capita of less than 12,236 United States dollars (USD) in 2018 (World Bank 2018). Studies conducted in countries listed in bold font highlighted as examples within the chapter.

Afghanistan
Albania
Algeria
American Samoa
Angola
Argentina
Armenia
Azerbaijan
Bangladesh
Belarus
Belize
Benin
Bhutan
Bolivia
Bosnia & Herzegovina
Botswana
Brazil
Bulgaria
Burkina Faso
Burundi
Cabo Verde
Cambodia
Cameroon
Central African Rep.
Chad
China
Colombia
Comoros
Congo, Dem. Rep.
Congo, Rep.
Costa Rica
Cote D'Ivoire
Cuba
Djibouti
Dominica
Dominican Rep.
Ecuador
Egypt, Arab Rep.
El Salvador
Equatorial Guinea
Eritrea
Ethiopia
Fiji
Gabon
Gambia, The
Georgia
Ghana
Grenada
Guatemala
Guinea
Guinea-Bissau
Guyana
Haiti
Honduras
India
Indonesia
Iran, Islamic Rep.
Iraq
Jamaica
Jordan
Kazakhstan
Kenya
Kiribati
Korea, Dem. People's Rep.
Kosovo
Kyrgyz Republic
Lao PDR
Lebanon
Lesotho
Liberia
Libya
Macedonia, FYR
Madagascar
Malawi
Malaysia
Maldives
Mali
Marshall Islands
Mauritania
Mauritius
Mexico
Micronesia, Fed. Sts.
Moldova
Mongolia
Montenegro
Morocco
Mozambique
Myanmar
Namibia
Nepal
Nicaragua
Niger
Nigeria
Pakistan
Panama
Papua New Guinea
Paraguay
Peru
Philippines
Romania
Russian Federation
Rwanda
Samoa
Sao Tome & Principe
Senegal
Siberia
Sierra Leone
Solomon Islands
Somalia
South Africa
South Sudan
Sri Lanka
St. Lucia
St. Vincent & the Grenadines
Sudan
Suriname
Swaziland
Syrian Arab Republic
Tajikistan
Tanzania
Thailand
Timor-Leste
Togo
Tonga
Tunisia
Turkey
Turkmenistan
Tuvalu
Uganda
Ukraine
Uzbekistan
Vanuatu
Venezuela, RB
Vietnam
West Bank & Gaza
Yemen, Rep.
Zambia
Zimbabwe

Fundamentally, this activity of describing the current state, setting goals for a future state, and measuring progress against those goals is the basis of community diagnosis. Careful and deliberate selection of health indicators, or metrics, is important to the planning, execution, and evaluation of any community diagnosis project. However, the effectiveness of health metrics selected and used relies heavily on the data available. Thus, robust information systems that can support timely collection, analysis, and distribution of high-quality data are central to any effective community diagnosis project. Recognition of the interconnectedness of community diagnosis, health metrics, and health information systems is important for supporting the iterative improvement of all three over time, especially in LMIC where sustained development is at its most challenging, yet the need for overall improvement in health status is greatest.

Community diagnosis

In contrast to clinical diagnosis, where health is assessed at an individual level through patient–physician interaction, community diagnosis is an assessment of health at a population level through community engagement in public health practice. As such, *community diagnosis* (sometimes also called a community health profile or health needs assessment) is defined by the WHO as *'a quantitative and qualitative description of the health of citizens and the factors which influence their health. It identifies problems, proposes areas for improvement and stimulates action'* (WHO 1995).

Community diagnosis projects can have one or more overarching goals, such as:

Description of a community's health status, health problems, and their determinants
Assessment of a community's health resources, services, and systems
Evaluation of community members' attitudes toward health and health services
Identification of priority issues and development of action plans for improvement
Definition of an epidemiologic baseline for measurement of change over time

A *community*, in the sense of community diagnosis, can be any cluster of people within the broader society who share at least one common characteristic; for example, geographic location, ethnicity, occupation, exposure to a specific environmental factor, or presence of a particular risk factor. However, it is not only the characteristics of the people (e.g. vital statistics, sociodemographics, religious and cultural beliefs, educational attainment) that define the community. Other factors—environment, safety, physical, and mental health and social services, economics, communication, transportation, and government—must also be included as the community is defined and as the current state of health in the community is described. Notably, any one individual may be a member of several different communities by this definition—citizen of Village A, but practitioner of Religion B, and worker in Factory C, etc. (Tindana et al. 2007).

The process of community diagnosis

Community diagnosis, whether performed in high-income settings or in LMIC, generally follows a fairly well-defined process that includes five main steps (see for example Box 4.2.3). However, the process is not only customizable, it is meant to be customized. Ideally, the base process of community diagnosis outlined here should be tailored to match the intent of the community diagnosis initiative, the resources and time available to execute it, and the unique attributes of the community itself. The process is meant to be pragmatic and its result is meant to be informative (i.e. improve understanding of the issue) and actionable (i.e. identify actions to be taken to improve the situation).

Box 4.2.3 Community diagnosis in a small village in Iran

Summary of methods used and results obtained in a community diagnosis effort in Qala-Sayed village, Kazeroun, Fars Province, Iran in 2013 (Basseej et al. 2015).

Step 1: Community assessment team created
Step 2: Primary and secondary data collected
Step 3 and 4: Data analysed and combined
Step 5: Community informed
Step 6. Health priorities selected
Step 7: Community assessment document created
Step 8: Action plan developed

A total of 649 households comprised of a mostly-rural population of 2,514, 25% of which were under the age of 11. For many years already, the village has had working water, electricity, gas, and telephone utilities. It had one 'health house', four schools, five mosques, one nursery school, four butchers, and 21 grocery stores.

Insufficient health knowledge was identified as the highest priority problem, followed by diabetes, oral health, addiction, and hypertension.

The plan of action focused on non-academic training designed to improve villagers' knowledge (regardless of educational attainment) of diabetes, with particular emphasis on healthy eating habits.

Step 1: Initiation. It is important that any community diagnosis effort is managed and coordinated by a dedicated team, committee, or working group. This team should be cross-functional in make-up, comprised of individuals from relevant government agencies, health bureaus, non-governmental organizations (NGOs), and representatives from the community. Once formed, this team should begin the process with refining the scope of the project. Taking into consideration the goals of the initiative, the financial budget and other resources available, and the time allotted to conduct the work, the team should have a detailed understanding of the work ahead, a plan for the methods of data collection and analysis to be used, and at least a high-level strategy for the ultimate output of the community diagnosis project (i.e. development and dissemination of the community diagnosis report). Development of a strategy for engagement of, and communication with, the community the project is intended to serve should be a critical element of the project's up-front planning. Another important component of this early planning stage is the selection of health indicators that ideally would be used, although this may be modified during the project based on data availability and analyses conducted. Furthermore, the team should anticipate obstacles and risks to the project and draft plans for overcoming barriers and mitigations for risks realized.

Step 2: Collection. Both primary (e.g. surveys, interviews, focus group discussions, mapping, observations) and secondary (e.g. records review) data collection methods are critical for gaining a thorough understanding of the community being studied. For the same reason, both quantitative and

qualitative data must be collected since neither provide a complete picture without the other (see for example Box 4.2.4).

Care must be taken to ensure representativeness of the data collected. For example, teams should be cautions of potentially causing over-representation of subgroups from which data are easier to obtain. This and other types of bias could lead to inaccurate findings and interpretations. Even more dangerous is the possibility that, as a result of these biases, inappropriate actions are taken, time and resources are wasted, opportunities for health intervention are lost, appropriate actions are delayed, and trust and belief in this process by the community is eroded. To help prevent these issues, it is crucial that different types of data be collected from multiple sources. While each source may be important individually, each will also have strengths and weaknesses. Taken together, however, data collected from a broad variety of sources are likely to be complimentary, more comprehensive, and less susceptible to bias.

Box 4.2.4 Malaria community diagnosis in Lomahasha, Swaziland

The importance of qualitative data is highlighted in a mixed-methods community diagnosis project aimed at better understanding social and behavioural factors influencing the success of malaria control measures in a rural area of Swaziland (Dlamini et al. 2017).

Individual interviews
Multiple researchers were strategically stationed around the village of Lomahasha on walkways and roadsides. They randomly stopped individuals passing by and interviewed them using a semi-structured questionnaire meant to collect both quantitative and qualitative information.

Focus group discussions
Focus group discussions were spontaneously convened by researchers who approached groups of people in Lomahasha already collected for some social reason (e.g. boy soccer players loitering in the street, women socializing near a water source, people waiting at bus stops).

Researcher observations
Six, one-day visits were made each year in 2004, 2006, and 2010 during which researchers drove around different parts of the Lomahasha community or stayed inside individual homesteads from early morning to late evening noting behaviours of individuals and social interactions among groups.

Several key findings important to future malaria control efforts were gleaned from qualitative data:

- Knowledge of malaria was relatively high, yet social norms and behaviours remained contradictory to control programme intentions—people stayed outside late into the evening, maintained pools of stagnant water, and sought care for symptoms late and in the wrong places.
- Severe poverty, drought, and the persistent presence of malaria in the community have desensitized people Lomahasha such that they regard malaria as a less serious problem and a normal part of life.

The study design most commonly employed for community diagnosis projects is cross-sectional. *Cross-sectional studies*, which consist of retrospectively evaluating a representative sample of a population at a single point in time, have the advantage of being relatively inexpensive and fast. They also facilitate the simultaneous investigation of multiple exposures and outcomes while providing the ability to control for confounding. Although these types of studies to not allow for assessment of temporality or causality (Thiese 2014), repeated cross-sectional measurement nevertheless facilitates observations of changes over time for the purpose of evaluating action plans executed or interventions implemented.

Step 3: Analysis. Raw data should be compiled, tabulated, and summarized using descriptive statistics. Statistical tests may be run and inferences made. Rates and ratios may be further analysed to assess correlates or determinants of health, injury, disease, etc. Historical trends in certain indicators may be accessible for comparison to current data, and projections may be made to help predict future changes in health indicators for planning purposes. Re-analysis (by the same methods) of selected health indicators after action plans are carried out or interventions are implemented is important for evaluating their effectiveness. Additionally, community data may be compared to that of other similar communities (i.e. Village A compared to Village B) and to larger geographical or administrative areas (i.e. village to county/district to province/state). Ultimately, the analyses selected by the team must relate to the objectives of the project, but again, caution must be exercised to ensure that bias is not introduced at this stage as well. Additionally, in selecting the analyses to be conducted, the team should be mindful of the need to communicate results of the project. The output of the analyses should be reasonably accessible and understandable to a broad range of audiences—government officials, public health and healthcare workers, and the general public.

Step 4: Diagnosis. Interpretations and conclusions drawn from the results of the analyses form the diagnosis of the community. This includes for

example, the current state of health, determinants of health, and health problems prioritized for intervention. This diagnosis must relate back to the initial intent of the community diagnosis project, and the diagnosis itself must be informative and actionable—suggestions for actions to be taken to improve health in the community should follow easily from the diagnosis. Furthermore, it must be specific and relevant to the community it is meant to serve.

Step 5: Dissemination. At its most basic, this step is the generation of a community diagnosis report, describing in detail the scope of the project, the methods used to collect and analyse data for the project, the results observed (i.e. the diagnosis), and the team's recommendations for further action. Emphasis should be placed on transparency. The report should be detailed, yet also include easily understood summaries of key points.

However, generation of the report itself is not enough. The report must be written in such a way that the information it contains is accessible not only to policymakers, public health authorities, or officials, other government agencies, and health workers, but also to the community members themselves. Ideally, a communication plan should be generated in parallel with the report. The results of the community diagnosis project should be shared with the general public in a variety of ways; for example, through presentations at meetings of local government committees, NGOs, and community-based organizations, through press-releases and news stories, and through community events such as health fairs or kick-off celebrations for new health interventions. Taking an active role in helping the community understand the project and how it will help improve health is important for building trust and belief in the process. Since the community diagnosis process is meant to be iterative, these kinds of activities may deliver the added benefit of improving participation in future community diagnosis projects.

Community engagement

Successfully conducting community diagnosis not only requires meticulous planning and execution, but also engagement and mobilization of the community to participate in the process.

Community diagnosis cannot be successfully performed or applied without involvement of community it serves. *Community engagement* (CE), the collaborative process of working with community members and groups

to address issues that impact their well-being, can take many forms. For instance, rates of participation by members of the community must be high in order for large-scale community-based studies or household surveys to be successful. Thus, involving community leaders, influencers, and champions in the planning process, communication plan, and launch, and ensuring that they are seen and heard by the community through all stages of the project can be important for encouraging participation from those who may be reluctant.

There are several important points during a community diagnosis initiative where focus should be placed on CE. A selection of these CE activities is listed here, but this list is meant to be neither exclusive nor comprehensive:

Creation of a project subteam whose mission is to lead CE
Identification of community leaders, influencers, and champions to be involved at each stage of the project
Review of plans to ensure the project is relevant to the community, culturally sensitive and pragmatic, not disruptive, designed to ensure fairness in community members' receipt of benefits from the project, and protective of vulnerable members
Identification of local community health workers who will participate in data collection
Development of a clear and consistent messaging strategy to ensure the community understands the project goals, processes, timelines, and expected outcomes and benefits
Sensitization of the community and the local media about the launch of the project
Execution of a high-profile launch, perhaps with a special community event where community leaders, influencers, and champions show their support for the effort
Provision of regular updates and feedback to the community during the project
Execution of a similarly high-profile project closure, with gratitude for community participation, transparent reporting on project outcomes, and details of action plans as well as funding and execution timelines for those plans

The benefit of collaboratively working with the community itself through truly authentic partnerships before, during, and after a community diagnosis

project cannot be overstated (see for example Box 4.2.5). It is not only palpable in the near-term on the current project, but also carries over in the longer term on future projects if community members feel their voices are heard, their contributions are valued, and their lives or the lives of those in their families or in their communities are bettered through this iterative and focused effort on improving health and well-being.

Box 4.2.5 Community engagement in rural Vietnam

A descriptive study of rural communes in Quang Tri Province in north central Vietnam underscores the value of community engagement (CE) in the community diagnosis process (Cho et al. 2018).

CE embedded in study methods

- **Study teams:** Study teams were comprised of nurse researchers, physicians, programme officers, programme evaluators, and local Vietnamese medical school faculty members.
- **Instruments review:** Local experts reviewed all questions included in all instruments used in the study for appropriateness relative to the local context.
- **RPA approach:** The study employed a rapid participatory appraisal (RPA) approach, which is a method for quickly collecting information about a set of problems and their causes that places strong emphasis on involvement of community members (Annett and Rifkin 1995).

Value of CE reflected in study findings

- **Multiple data sources:** Inclusion of multiple sources and types of data, made possible by CE, facilitated a better and more comprehensive understanding of community health needs.
- **Qualitative data:** Use of qualitative data, obtained through CE, helped compensate for relatively weak quantitative data (small data set from a small convenience sample).
- **Collaboration with local experts:** Many of the study communes were approached for the first time in this study and collaboration with local experts was essential for understanding the sociocultural context.
- **Community involvement:** Involvement of community members in the collection of information was critical to creating *'an acceptable and sustainable environment for successfully undertaking a new study'*.

Finally, CE is increasingly considered an ethical requirement of research involving human subjects, and thus, is no longer really considered optional (Tindana et al. 2007).

Community diagnosis in low- and middle-income countries

In LMIC, community diagnosis is particularly challenging for a wide variety of reasons. Primary data collection is fraught with difficulty in creating representative, non-biased samples and retaining participants in studies, as well as barriers to field work including low educational attainment and poor infrastructure. Secondary data collection efforts can be frustrating as the data required are often unavailable, incomplete, inaccurate, out-of-date, and difficult to obtain. These complications make for complex and labour-intensive data entry and analysis, which may reduce confidence in results obtained and interpretations made.

Challenges with primary data collection

It is nearly impossible to remove all bias when sampling a population. One significant source of bias is caused simply by community member participation in primary data collection efforts. Those who elect to participate are very likely to be more informed or more concerned about the issue being examined and addressed through the project and those who decline may, in fact, be the more vulnerable to the issue at hand. Obtaining high participation rates in LMIC is already challenging but obtaining any level of participation from those most in need of intervention is extremely difficult.

Dropout is an issue for projects that require repeated measurement or even just a seemingly-small time interval between study enrolment and primary data collection. Dropout can be caused by a wide range of issues, but in LMIC, predominant among these is the movement of people over time. Large proportions of populations in LMIC communities are either transient (i.e. characterized by frequent movement due to lack of stable housing and/or employment) or migrant (i.e. characterized by regular movement over large distances between family residence and employment).

Low education levels and literacy rates, more common in LMIC settings, mean the target population for a community diagnosis project is often ill-informed or unable to understand the importance of the health research project. This can render less expensive primary data collection techniques, such as self-administered questionnaires, impossible. Rather, home visits by study workers with high-level language skills and cultural knowledge may be required. This drives up the cost of the study, thereby driving down the feasible sample size, and increasing the risk of non-representative, biased sampling.

There are innumerable challenges with this kind of house-to-house field work in LMIC. Remote rural communities as well as poorer urban communities lacking or having poor-quality basic infrastructure that is susceptible to seasonal or severe weather effects mean transportation into and out of communities by field workers can be difficult, time-consuming, and sometimes dangerous. Lack of geospatial mapping, or simply street names and house numbers, creates obstacles for ensuring the homes of enrolled participants are correctly located. Field work that requires electricity (e.g. for visual observations or inspections), water (e.g. for hand washing before and after physical examinations), internet access (e.g. for record keeping), or other utilities can be problematic.

Finally, some of the world's LMIC most in need of improvements to community health are in the midst of ongoing violent conflict. Accessing these regions is unsafe for field workers seeking to gather primary data, and moreover, engaging with foreigners may be unsafe for community members as well. Additionally, interest among community members in participating in a community health needs assessment in regions of conflict is likely to be low, since primary among their concerns is not falling victim to the violence around them.

Challenges with secondary data collection

Challenges with secondary data collection in LMIC are numerous and persistent and can be frustrating for researchers attempting to conduct a community diagnosis or evaluate the effects of health policy changes or health interventions. Secondary data collection centres around review of records from a wide variety of sources—vital statistics from registries and censuses, surveillance systems, routine information systems, electronic medical records, disease registries, and other forms of information repositories. These records can be a rich source of quantitative information for community diagnosis initiatives and are frequently the focus of study in the public health field.

However, one of the most frequently cited challenges during the MDG era was the poor availability and reliability of data in LMIC—both for baseline assessment and for evaluation of progress toward MDG targets. What information systems did exist at that time suffered from, for instance, large gaps in record keeping, incomplete records, and poor-quality data entry (see for example Box 4.2.6 and Box 4.2.7). Many LMIC lacked birth and death

registries, health records repositories, or health statistics at any administrative level. Methods for gathering information in the absence of these secondary data sources were often unreliable and sometimes based on assumptions. Furthermore, these methods were variable within and between countries and over time, making individual country progress difficult to interpret and nearly impossible to compare. These issues, caused by widespread lack of robust, thoughtfully-designed, and secure information systems in LMIC, persists today, presenting major challenges to the mission of global public health and global sustainable development (Fehling et al. 2013).

Box 4.2.6 Timeliness of spotted fever surveillance reporting in Brazil

Study of spotted fever (SF) surveillance records in Brazil highlighting challenges with timely data availability (de Oliveira and Angerami 2018).

Background: In 2001, SF, an infectious tick-borne disease endemic to southeastern Brazil, became a compulsorily notifiable disease and in 2014 became an immediately-notifiable disease (within 24 hours of suspicion). Two systems have tracked SF reports in Brazil: the digital disease detection (DDD) system launched in 1997 and the national surveillance (SINAN) system launched in 2007.

Aim: To examine the timeliness of SF notification in the DDD and SINAN systems in southeastern Brazil.

Method: Records from Brazil's DDD and SINAN systems were matched and dates compared.

Findings: Only 62 records contained sufficient information to be matched and included in the study. More than 90% of these cases were reported to SINAN a median of 20.5 days earlier than DDD. Both systems' records contained different information and the data that could be extracted from each system were valuable and complementary.

Interpretation: The authors noted several critical limitations of these surveillance systems. For instance, although reporting to SINAN was faster due to immediate mandatory reporting of suspicion, reports were not available until diagnoses were confirmed generally 30 days later. Furthermore, SINAN did not contain location information, forcing reliance on the DDD system for identification of new transmission areas, clusters, and outbreaks. Delayed access to fragmented information prevents rapid and informed responses to health needs.

Box 4.2.7 Continuity of cancer registry reporting in Kosovo

Study of cancer registry records in Kosovo highlighting challenges with data quality and quantity (Berisha et al. 2018).

Background: Although malignant disease is a leading cause of death worldwide and an important contributor to morbidity and disability, public health efforts continue to be hindered in LMIC like

Kosovo by lack of non-communicable disease registry data, which prevents informed allocation of limited resources.

Aim: To investigate the incidence, prevalence, and types of malignancies reported in Kosovo.

Method: Records from the Kosova Cancer Registry (KCR), which was reinstated in 2011 after an extended gap in reporting, were reviewed. Data were collected and tabulated and incidence and prevalence calculated.

Findings: Incidence was 93.4 (per 100,000) in 2012, 83.3 in 2013, and 167.9 in 2014, while prevalence was 135.8 (per 100,000) in 2012, 101.6 in 2013, and 186.5 in 2014. Most common among men were malignancies of the skin and respiratory and intrathoracic organs, and among women were malignancies of the breast and reproductive tract.

Interpretation: Although the authors concluded that cancer incidence and prevalence are increasing in Kosovo, and that investments must be made in prevention, screening, diagnostics, and therapeutics, they also emphasized several critical limitations of their study. They stated that they were *'not satisfied with the quantity and quality of the reporting'* and highlighted the importance of better quality data for better public health planning and the critical need for political commitment to support and expand coverage of existing reporting efforts, such as the KCR.

Challenges with data management and analysis

Aforementioned issues with the data and data collection methods all have ramifications for the management of datasets as well as statistical analyses, which can be difficult to overcome. For example, collection of handwritten data on paper forms—common in LMIC where use of laptops or tablets is often precluded by financial constraints and suboptimal internet access—creates the momentous task of keeping track of all the paper generated by the study, accurately deciphering and entering all the data, and then quality-checking the data entry. In terms of analysis, for instance, standard practice to address issues of representativeness or bias in sampling is to attempt to control for confounding. However, in LMIC, confounding may be caused by a multitude of factors, many of which researchers, due to their own unconscious biases and lack of knowledge, may be unaware. This can easily result in confounding by unmeasured factors having a substantial effect on results and interpretation. As another example, in the case of high dropout-rate, teams will often add 'replacements'. However, these replacement participants, often added late in the study, may have different characteristics and perhaps different exposures, which can be difficult to handle in the analysis phase.

The future of community diagnosis in low- and middle-income countries

Without question, properly executing community diagnosis in LMIC is difficult work. However, despite all of the barriers and challenges,

performing a community diagnosis initiative is still a worthwhile endeavour, and having a little data or lesser-quality data is nevertheless better than having zero data. After all, the work of global public health is both a science and an art, meant to achieve a noble purpose—to prevent disease, prolong life, and promote health. In LMIC, where health needs are greatest and resources are most constrained, effectively executing the methodical process of community diagnosis can be especially impactful.

The Sustainable Development Goals

The Sustainable Development Goals (SDGs), the successor to the MDGs, were launched by the UN GA in 2016 and were comprised of 17 goals and 169 associated targets (see Box 4.2.8) (UN GA 2015). In contrast to the MDGs, where three of the eight goals were focused on improving community health, the SDGs contained only one goal directly related to health—Goal 3: Ensure healthy lives and promote well-being for all at all ages—which contained 13 targets.

Box 4.2.8 The United Nations Sustainable Development Goals (SDGs)

Launched in 2016, the 17 SDGs to be achieved by 2030 contained one goal directly related to health (Goal 3), which itself had 13 targets (UN GA 2016) (see Fig. 4.2.1.)

Fig. 4.2.1 Targets.
Reprinted from United Nations (2017) 'The United Nations Sustainable Development Goals', https://unstats.un.org/sdgs/report/2017/overview/

Goal 3: Ensure healthy lives and promote well-being for all at all ages

3.1	By 2030, reduce the global maternal mortality ratio to less than 70 per 100,000 live births
3.2	By 2030, end preventable deaths of newborns and children under 5 years of age, with all countries aiming to reduce neonatal mortality to at least as low as 12 per 1,000 live births and under-5 mortality to at least as low as 25 per 1,000 live births
3.3	By 2030, end the epidemics of AIDS, tuberculosis, malaria, and neglected tropical diseases and combat hepatitis, water-borne diseases, and other communicable diseases
3.4	By 2030, reduce by one-third premature mortality from non-communicable diseases through prevention and treatment and promote mental health and well-being
3.5	Strengthen the prevention and treatment of substance abuse, including narcotic drug abuse and harmful use of alcohol
3.6	By 2020, halve the number of global deaths and injuries from road traffic accidents
3.7	By 2030, ensure universal access to sexual and reproductive healthcare services, including for family planning, information and education, and the integration of reproductive health into national strategies and programmes
3.8	Achieve universal health coverage, including financial risk protection, access to quality essential healthcare services and access to safe, effective, high-quality, and affordable essential medicines and vaccines for all
3.9	By 2030, substantially reduce the number of deaths and illnesses from hazardous chemicals and air, water and soil pollution and contamination
3.a	Strengthen the implementation of the World Health Organization Framework Convention on Tobacco Control in all countries, as appropriate
3.b	Support the research and development of vaccines and medicines for the communicable and non-communicable diseases that primarily affect developing countries, provide access to affordable essential medicines and vaccines
3.c	Substantially increase health financing and the recruitment, development, training, and retention of the health workforce in developing countries, especially in least developed countries and small island developing states
3.d	Strengthen the capacity of all countries, in particular developing countries, for early warning, risk reduction, and management of national and global health risks

A great many of the aforementioned criticisms of the MDGs have been attributed to the way in which they were developed—largely by the United States, Europe, and Japan together with the World Bank, the International Monetary Fund, and the Organization for Economic Cooperation and Development without consultation from, and some would say, even without consideration of, LMIC. Regardless of whether one subscribes to this view or not, the complaints by many of LMIC not being represented in the MDG development process and then being the recipient of its dictated priorities, targets, measures, and deadlines (Cohen et al. 2014; Easterly 2009; Fehling et al. 2013; Vandemoortle 2009), combined with the emerging notion of ‘global public health’ as well as other factors, drove a very different approach to the development of the SDGs leading up to the 2015 MDG deadline (Coonrod 2014; Marten 2019).

In many ways, the process used to develop the SDGs strongly resembled community diagnosis effort on a global scale. Cross-functional working groups were formed, the scope of the project was refined, and a plan for community engagement was developed (i.e. initiation step). Massive amounts of information were collected (i.e., collection step), compiled, and tabulated (i.e. analysis step). Interpretations were made, priorities were set, and metrics were selected (i.e. diagnosis step), and finally, the SDGs were communicated (i.e. dissemination step).

Engagement of the global community in preparation for the development of the SDGs began in 2012. It was an extraordinarily large effort—widely considered the most inclusive and consultive participatory processes the world has ever seen. Led by a 30-member Open Working Group of the UN GA, government officials from more than 100 countries were engaged in face-to-face discussions, recommendations of more than 5,000 civil society organizations were reviewed, executives from more than 250 companies were consulted, and feedback from more than a million individual citizens was collected via the UN's online Survey for a Better World, which asked people to select six issues that matter most to them (UN 2013).

The SDGs were the output of this massive, global-scale community diagnosis initiative. Without question, the new goals address a much broader range of issues, and the targets beneath the health and well-being goal are much more comprehensive (see Box 4.2.8) (UN GA 2015). However, it remains to be seen how successful the health and well-being SDG will be at driving forward a unified global public health agenda and improving health and well-being for all peoples of the world (UN 2017; Marten 2019).

The 'data revolution'

The crucial importance of information sources and data to global sustainable development was recognized in the MDG era, and the UN High-Level Panel, tasked with developing recommendations on the post-2015 development agenda, called for a 'data revolution'. The panel identified the need for substantial investment to build capacity worldwide in data collection and statistical analysis methods and tools to ensure that monitoring and evaluation at all stages is strengthened to support decision-making, guide prioritization, and ensure accountability (UN 2013). This theme extended into the 2015 UN GA resolution detailing the SDGs and its critical nature is

underscored by its inclusion as targets under SDG 17 (UN GA 2015) (see Box 4.2.9).

The need to 'harness the power of data for sustainable development' features prominently in the UN SDGs Report 2017 (UN 2017):

> To fully implement and monitor progress on the SDGs, decision makers need data and statistics that are accurate, timely, sufficiently disaggregated, relevant, accessible and easy to use. Data availability and quality have steadily improved over the years. However, statistical capacity still needs strengthening and data literacy must be enhanced at all levels of decision-making. This will require coordinated efforts on the part of data producers and users from multiple data systems. It will also demand innovative ways to produce and apply data and statistics in addressing the multifaceted challenges of sustainable development.

The report goes on to emphasize the importance of building capacity via innovative approaches, creating synergies across different data environments, ensuring data improve clarity and do not mask disparities, improving data presentation and literacy, and implementing best practices and standards related to data management.

Box 4.2.9 SDG targets for data, monitoring, and accountability

SDG Goal 17: Strengthen the means of implementation and revitalize the Global Partnership for Sustainable Development contains two targets important to the improvement of data, monitoring, and accountability (UN GA 2015).

17.18: By 2020, enhance capacity-building support to developing countries, including for least developed countries and small island developing states, to increase significantly the availability of high-quality, timely, and reliable data disaggregated by income, gender, age, race, ethnicity, migratory status, disability, geographic location, and other characteristics relevant in national contexts.

17.19: By 2030, build on existing initiatives to develop measurements of progress on sustainable development that complement gross domestic product, and support statistical capacity-building in developing countries

Sources of information

There are many sources of information and data that are employed to conduct community diagnoses and to evaluate changes in community health over time. Each individual source has its own advantages and disadvantages, yet in combination with other sources (that have different advantages and disadvantages), the complementary larger data set becomes more powerful.

Several of the most common sources of health information used in LMIC are detailed here.

Census

Generally conducted every 10 years and typically at the national level, a *census* systematically collects data on the population number and sociodemographic profile of citizens. A major advantage of census data is its provision of a population count and age profile that may be disaggregated by a range of geographical, social, economic, and demographic variables, which can contribute significantly to health planning (see for example Box 4.2.10).

Box 4.2.10 Census data identifies an ageing population in Jamaica

Jamaica's census data has provided concrete evidence of significant and meaningful population ageing, 2001 compared to 2011 (Eldemire-Shearer et al. 2014), and projected to 2021 (PopulationPyramid.net).

Since ageing is associated with increased disease burden, urgent action must be taken to ensure that health and social services are expanded to meet the demands not only of the current, but also the future, elderly population. To promote the health and well-being of Jamaica's ageing population, investments need to be made in combatting chronic disease and age discrimination, creating programmes that support healthy, active ageing, and expand services for diagnosing and treating acute and chronic mental and physical illness in the elderly population. See Figure 4.2.2.

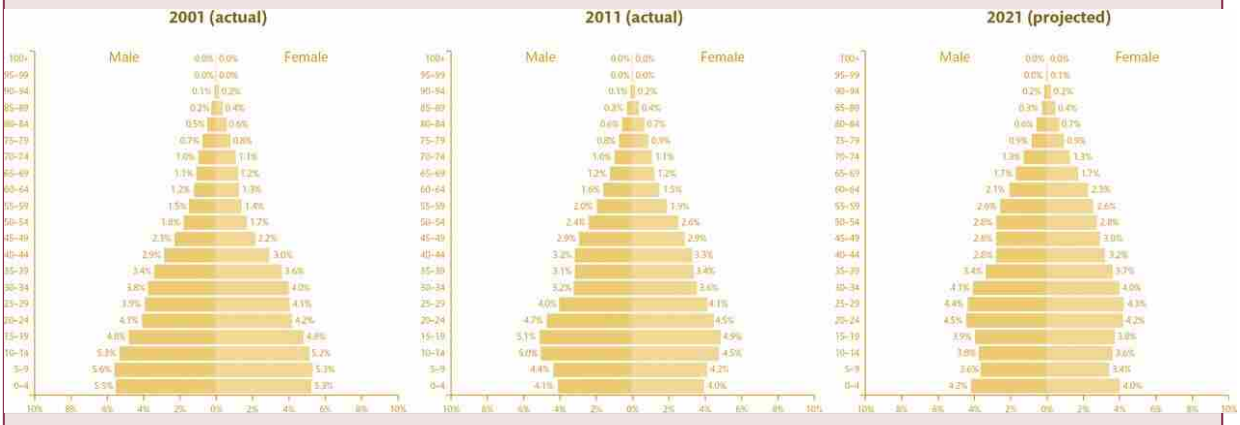


Fig. 4.2.2 Population pyramids.

Source: data from Eldemire-Shearer D, et al. (2014) 'Ageing of Jamaica's population—what are the implications for healthcare?' *West Indian Med J* 63(1): 3–8 and Statistical Institute of Jamaica (STATIN). Population and Housing Census 2011: General report. Volume 1. Kingston: STATIN; 2011.

However, while this form of data collection is somewhat successful in avoiding the pitfalls of sampling bias due to its focus on full participation by all citizens, bias still exists as many of those who fail to comply are perhaps the most vulnerable in society to health disparities. This is particularly true in LMIC. Furthermore, a census normally does not collect data that are directly related to the health of citizens and thus, cannot usually stand on its own for community health evaluation and monitoring.

Vital registry

Vital registration is the systematic recording of vital events including births, deaths, adoptions, marriages, and divorces. A country's vital registry (VR) can therefore provide important information on population dynamics (i.e. life expectancies, fertility rates, population growth rates) and health, via death records that can be used to evaluate morbidity and mortality, that a census cannot. Although VR is also intended to capture data from a nation's entire population, vital events are commonly under-reported in LMIC. According to a 2014 World Bank and WHO report, an estimated 85 million children under the age of five in Africa and 135 million in Asia and the Pacific were not registered. Likewise, an estimated 67% of deaths worldwide are not accounted for in VRs (WB and WHO 2014). Inadequacy of death reporting severely limits the use of VR alone to infer mortality rates from death records. Even deaths that are recorded sometimes have somewhat limited value since medical certification of deaths is often limited and thus, causes of death are not classified or miss-classified.

Verbal autopsy

VR death records can be augmented by *verbal autopsy* (VA), a process whereby a trained health worker interviews a person familiar with the deceased using a structured questionnaire for the purpose of gathering enough health information to determine a probable cause of death. However, VA has significant limitations beyond the obvious fact that it requires substantial time and resources. Although the WHO has developed VA standards as well as an instrument designed for routine use in cases were undocumented or poorly documented deaths require investigation (WHO 2016), considerable variability in VA procedures persist in many settings.

Furthermore, non-standard VA conducted by untrained or inexperienced individuals tends to be highly subjective, risking bias, or miss-classifications of causes of death.

Box 4.2.11 Household survey of essential obstetric and newborn care (ENOC) in Ecuador

A study of the impact of a 3-year intervention intended to improve ENOC availability, accessibility, demand, uptake, and quality in rural Cotopaxi province, Ecuador highlights the utility of household surveys, and the value of combining these data with data collected through other means (Broughton et al. 2016).

Method: Household survey

Women with a child under the age of 24 months were surveyed using a structured questionnaire meant to assess access to and uptake of postpartum and post-neonatal care, knowledge and practices, and patient-reported quality of care at study baseline (2010) and completion (2013). Women in geographic areas known (via census data) to have many households with children were selected by multistage sampling.

Method: Routine health information system

Clinical records at participating facilities were randomly sampled and extracted data were used to assess facility quality of care.

Method: Vital registry

Vital registry data were provided by the provincial government to support evaluation of neonatal mortality for the study.

Researchers found that receipt of postnatal visits and postpartum counselling on newborn care rose, health knowledge increased, quality of care improved, and newborn mortality declined.

Household surveys

Household surveys may be used to collect a broad range of data both as a means of regular surveillance and as a method for evaluating health interventions (see for example Box 4.2.11). *Household surveys* are a relatively common way of collecting demographic, social, wealth, and health data in LMIC as the size and complexity of the survey questionnaire and the size and complexity of the survey sample can be easily scaled up or down according to research needs and resource constraints.

Ideally, household surveys involve multistage or two-stage stratified sampling designs to facilitate generation of a sample that is representative on

a provincial or national level and to facilitate acquisition of data that may be aggregated and disaggregated by sampling stage/geographic area. However, single-stage cluster sampling plans may also be used if two-stage designs are not feasible, this may reduce costs and increase efficiencies in some cases.

To support the conduct of the household surveys, a standardized, structured questionnaire is developed and preferably pilot tested prior to use to ensure the data collected indeed answer the questions being asked in the study, to ensure the language used is easily understood, and to ensure the questions are culturally sensitive. However, although biomarker testing for clinical factors (e.g. anaemia, blood pressure, lead, vitamin A) and communicable diseases (e.g., diabetes malaria, measles, syphilis) has increasingly been included in household surveys in LMIC, information collected by household survey still relies heavily on self-reporting and is therefore subject to recall and social desirability biases.

Routine health information systems

The day-to-day collection of health information at public and private healthcare facilities and institutions is supported by *routine health information systems* (RHIS). A primary goal of RHIS is to provide reliable information that can be used to support health system improvement. As such, RHIS can be used to examine all manner of health indicators related to, for example, healthcare delivery, utilization, and policy implementation. Today, nearly all countries maintain these data repositories albeit in varying levels of completeness, complexity, and compliance. LMIC, especially, tend to have RHIS that operate on a suboptimal level, unable to provide complete and accurate information that is relevant and easily accessible in a timely fashion. Many types of interventions designed to strengthen RHIS have been evaluated, but evidence of effectiveness of these interventions in a range of settings has been mixed (Leon et al. 2015). Furthermore, even the most highly evolved RHIS still has limitations. For example, the creation of RHIS records and the augmentation of those records over time selects for those with more serious symptoms and those more likely to seek care. Many different types of RHIS exist (two examples described as follows), sometimes even within the same LMIC, often created for different purposes, and typically disconnected from each other, which causes fragmentation and reduced access to timely data.

Electronic medical records (EMR) are a system of digital patient records that replaces traditional, paper-based patient charts in clinical settings. This important source of information can shed light on all kinds of health indicators including facility-centric metrics (e.g. healthcare delivery, utilization, cost, and quality metrics) and patient-centric metrics (e.g. mortality, morbidity, and disability metrics). Having a facility-based or health system-based EMR system facilitates horizontal integration and greatly improves the speed with which health indicators can be evaluated. However, implementation of EMR systems in LMIC settings can be challenging and adopting practices for regular use of the data contained within the EMR system to evaluate and improve upon healthcare delivery is often slow and difficult due to constraints on human resource capacity.

Disease-specific information systems are generally developed, implemented, and maintained in response to a major emerging health issue. The structure and degree of complexity of these systems is highly variable as they are often constructed quickly with bare minimum functionality to meet the urgent needs of an acute situation, and then expanded over time.

One example of such a system is China's HIV/AIDS Comprehensive Response Information Management System (CRIMS). Early in China's HIV response, data collection was a highly-fragmented, paper-based, and labour-intensive exercise that was at once incomplete and duplicative—the many international aid organizations and donors who were funding much of China's initial efforts all required different reporting. Health workers, particularly at the local levels, were inundated with paperwork that did not directly serve their main mission, providing and expanding testing, treatment, and care services. This was exacerbated by fragmentation of information at the China Center for Disease Control (China CDC), the public health arm of the central government's Ministry of Health (now called the National Health and Family Planning Commission). The China CDC was maintaining eight different databases for sentinel surveillance, case reporting, treatment, prevention, methadone maintenance treatment (MMT), and other services and indicators. By 2005, the situation had become untenable. It was impossible to understand the epidemic and the effects of epidemic response efforts in a clear and holistic way.

Thus, in 2006, the National Center for AIDS/STD Control and Prevention (NCAIDS, a division of the China CDC) began developing an integrated, national HIV-specific information system, and in 2008, CRIMS was finally launched. CRIMS is a fully-secure, real-time, web-based system that

contains records for all individuals ever diagnosed with HIV in China and all individuals ever enrolled in MMT in China. It was built not only to optimize efficiency for health workers putting information into the system, but also to maximize performance for public health personnel and researchers extracting information out of the system.

The power of the new system in delivering value to health and public health workers and officials, researchers, and policymakers was felt nearly immediately. CRIMS has helped to generate an enormous body of evidence for the development and implementation of numerous important iterative improvements in HIV testing, treatment, care, and prevention programmes and has most certainly had an immeasurable positive impact on the trajectory of China's HIV epidemic (Mao et al. 2010) (see Box 4.2.12).

Box 4.2.12 The human immunodeficiency virus (HIV) care continuum in China

In 2014, the Joint United Nations Programme on HIV/AIDS (UNAIDS) set ambitious targets for the global HIV response to be achieved by the year 2020. These so-called 90-90-90 targets are 90% of those with HIV infection know their status, 90% of those with a confirmed diagnosis on antiretroviral therapy (ART), and 90% of those on ART virologically suppressed (UNAIDS 2014; UNAIDS 2017).

China was a partner in the development of these targets and is supportive of transparent and objective measurement of performance against the targets. In China, this would not be possible without the national HIV/AIDS Comprehensive Response Information Management System (CRIMS) (Ma et al. 2018).

Methods:
A nationwide, biennial cross-sectional study was conducted using CRIMS records from 2005 to 2015 (Ma et al. 2018), and 2017 (NCAIDS 2018). Individuals were categorized based on the definitions in the table. These definitions and the method by which proportions were calculated were consistent with the UNAIDS guidance on the evaluating progress toward the 90–90–90 targets (UNAIDS 2017).

Results:
Consistent improvement in the performance of China's HIV care continuum is shown in Figure 4.2.3. Bars represent the total number of people living with HIV and proportions are presented above the bars. As of the end of 2015, 68% of those estimated to be infected had been diagnosed, 67% of those diagnosed were on ART, and 65% of those on ART were virally suppressed (68–67–65) (Ma et al. 2018), and as of the end of 2017, performance against the targets had again improved to 79–80–91 (NCAIDS 2018).

Category	Number and proportion
Infected	Total number of individuals with HIV infection in China as estimated via a joint effort between UNAIDS, WHO, and China's Health and Family Planning Commission.
Diagnosed	Total number of individuals in CRIMS with diagnosed HIV infection who were still alive at the end of the year. Proportion diagnosed = number diagnosed/estimated number infected.
On ART	Total number of individuals in CRIMS who were receiving ART during the year and were still alive at the end of the year. Proportion on ART = number on ART/number diagnosed.
Suppressed	Total number of individuals in CRIMS who achieved viral suppression during the year and were still alive at the end of the year. Proportion suppressed = number suppressed/number on ART.

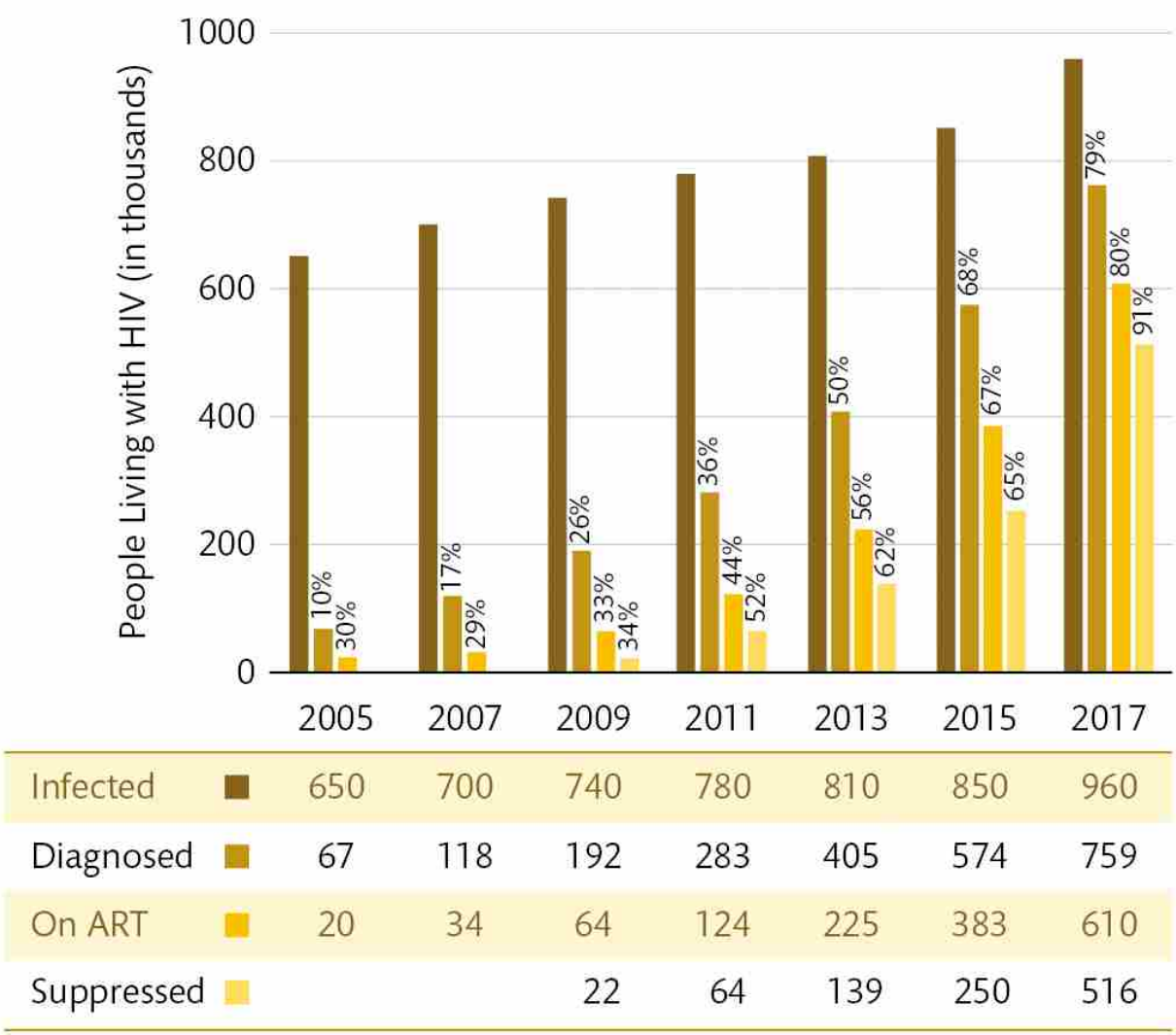


Fig. 4.2.3 Improvement in the performance of China's HIV care continuum.

Source: data from National Center for AIDS/STD Control and Prevention, China CDC. (2018) Annual progress of national HIV/HCV programs in 2017 and priorities and challenges in 2018. Presentation at the National Annual Provincial AIDS Directors Meeting. Ningbo, Zhejiang, China, 19 April 2018.

Key findings made possible by CRIMS

- The performance of China's HIV care continuum has consistently improved since 2005, and by the end of 2017, China had met 'The Third 90' target—90% of those on ART virally suppressed.
- However, China must urgently redouble its efforts in case finding and treatment initiation and retention, especially among those at greatest risk for attrition—adolescents, minorities, and injecting drug users.

HIV care policy implications

- Suboptimal care continuum performance observed from 2005 to 2013 spurred an interventional study and a cluster-randomized clinical trial (Wu et al. 2015, 2017), which provided ample, quality evidence for the individual (i.e. improved mortality rates) and community (i.e. improved viral suppression rates) benefit of streamlining and accelerating the care continuum.
- As a result, China began implementation of a new testing and treatment algorithm intended to drive faster diagnosis and treatment initiation, the effects of which are, at least in part, reflected in the performance of the care continuum in 2017.

Sentinel surveillance systems

When high-quality data on a disease or condition of interest are unavailable through passive channels (e.g. RHISs), other means of identifying health trends and disease outbreaks must be used. *Sentinel surveillance systems* can be a fairly economical method whereby a limited number of reporting units (e.g. hospitals) are carefully selected for their high likelihood of accurately identifying cases and their commitment to promptly report those cases. Sentinel surveillance data can provide estimates of symptom frequency, disease prevalence, incidence, and geographical distribution, and other key indicators, either among a broader population as a whole, or among subpopulations at particular risk.

Geographic information systems

Although spatial mapping of health problems is not new to public health, the development of new technologies and methods for rapidly manipulating enormous datasets has substantially increased our capacity for examining larger problems over larger geographic areas. *Geographic information systems* (GIS) are systems of computer hardware and software that are capable of combining high precision location information with health databases and analysis tools to visualize geographic distributions of health indicators such as morbidity, mortality, prevalence, and incidence related to a particular disease or condition of interest.

Other sources

There are many other sources of important information that can contribute to community diagnosis. These include lot quality assurance sampling, community key informants, risk factor surveillance mechanisms, morbidity

and mortality data, and health system data related to facilities and human resources, as well as service delivery, utilization, and cost.

Health metrics

Communities are diagnosed and changes in health status are measured over time using a wide range of objective health indicators or metrics. However, selecting the appropriate group of metrics is as important as properly measuring performance against them using quality data. Ideally, metrics for the assessment of community health should be (Bilheimer 2010):

Feasible—Performance against the indicator must be measurable. Good-quality, current data required to accurately assess performance against the metric must be available.
Valid—The indicator must be valid, meaning it must actually measure what it was designed to measure. Validation of the method of measurement should be prioritized.
Reliable—Measurement of performance against the metric must be robust enough to be unaffected by the individual doing the measuring. This reduces bias and facilitates comparison of repeated measurements.
Relevant—The indicator must be relevant. It should substantively contribute to improving understanding of the health issue being investigated.
Sensitive—The metric must be sensitive enough to allow detection of meaningful changes in performance against the metric.
Specific—The indicator must be specific enough to reflect changes only in the health issue being investigated.
Actionable—The metric should be actionable at some level—if not at the community level, then at the state/province or national level.
Accessible—The indicator should be accessible, meaning it should be quickly and easily understood by policymakers, health and public health workers, and community members.

Additionally, using metrics that have been used before and have recognized operational definitions can be advantageous. Finally, itis also important to consider the ability to aggregate or disaggregate selected metrics by geography, age group, sex, ethnicity, and perhaps other sociodemographic parameters.

Arguably, individual metrics that do not meet these criteria may be of questionable value. Thus, when considering selection of an indicator that lacks one or more of these characteristics, the benefit of using it should be weighed against the cost of measuring performance against it. However, it would be a rare situation where a single metric would provide sufficient information to adequately assess changes in the health of a community over time. Hence, groups of metrics should generally be selected—each individual metric providing meaningful information on its own and complementary information to other selected metrics, and together as a group of metrics providing a comprehensive assessment that is capable of meeting the needs of the project.

Community diagnosis projects have used a broad range of health metrics to investigate issues of interest (see Box 4.2.13). Since each individual community diagnosis project will have different research questions, each will spend considerable time carefully selecting the group of health metrics that best meet the aforementioned criteria alone, are complementary and comprehensive together, and meaningfully address the issues of interest (Bilheimer 2010).

Box 4.2.13 Health metrics

The following list of health metrics is intended to be neither exclusive nor comprehensive. Rather, it is meant to be a categorized selection of health indicators commonly used in community health projects.

Mortality metrics

- Case fatality rates
- Crude death rates
- Infant/child mortality rates
- Life expectancies
- Maternal mortality rates
- Mortality rates
- Proportionate mortality ratios
- Specific death rates

Morbidity metrics

- Admission/discharge rates
- Attendance rates
- Causes of death
- Disease incidence
- Disease prevalence
- Hospital stay duration rates
- Morbidity rates
- Notification rates

Disability metrics

- Disability-free life expectancy
- Disability rates
- Functional health status
- Lost work/school rates
- Restricted activity rates

Social/mental health metrics

- Accident/injury rates
- Alcohol dependence prevalence
- Mental illness incidence
- Mental illness prevalence
- Substance use prevalence
- Suicide/homicide rates
- Traffic accident rates

Nutritional metrics

- Anthropometrics measurements
- Clinical indicators of nutritional status (e.g. anaemia)
- Low birth weight rates
- Obesity rates

Healthcare delivery metrics

- Cost-effectiveness measures
- Hospital bed/population ratios
- Hospital facility/population ratios
- Physician/nurse ratios
- Physician/population ratios

Healthcare utilization metrics

- Antenatal care attendance rates
- Family planning uptake rates
- Hospital bed occupancy/ turnover rates
- Infant immunization rates
- Oral health attendance rates

Environmental metrics

- Clean water coverage rates
- Food safety indicators
- Pollution measures
- Safe sanitation coverage rates
- Tobacco use rates
- Vector density rates

Socioeconomic metrics

- Family sizes
- Homelessness rates
- Household income
- Housing condition measures
- Literacy rates
- Per capita GDP
- Population change rate
- Poverty rates
- Unemployment rates

Health policy metrics

- Healthcare expenditure per capita
- Proportion of GDP spent on healthcare
- Proportion of total health resources dedicated to primary care

Other metrics

- Basic needs indicators
- Quality of life indicators
- Risk factors for disease/injury

Since the effective use of health metrics to drive active, intentional efforts to improve community health is heavily dependent on the availability of good-quality, timely data, which is a serious challenge in most LMIC, many of the most helpful health metrics may be very difficult, if not impossible, to use in some settings. Thus, there must be some flexibility in the assessment of the value of particular health metrics. Above all, it is critical that a pragmatic and iterative approach to the selection of health metrics in LMIC settings is taken. This may first mean assessing what data are available that meet reasonable standards of completeness, quality, reliability, and timeliness, and using that data to measure what is possible to measure in the short term. Then, assessing what health indicators should be measured, and

developing the information system infrastructure to make it possible to measure what should be measured in the longer term. A few of the health metrics more commonly used LMIC are described in more detail here.

Mortality metrics

A broad range of mortality measures may be used to assess health in communities, all of which rely heavily on VRs, and where vital registration data are incomplete, on census data, household surveys, and health services records. However, since deaths commonly go unreported in LMIC, measures of mortality in these settings can generally be considered underestimates. *Mortality rates* may be calculated overall, or for a specific subgroup within the community (e.g. age group, ethnic group), as the numbers of deaths in a year per unit of population, usually per 1,000 or 10,000. This fairly simple metric is a valuable means of easily understanding health in the community as health must be improving if mortality is declining. Disaggregation of mortality measures by geography, demographic, and socioeconomic factors may offer clues as to how to help improve community health. However, there are many more specific methods of measuring mortality that are also commonly used and standardization of measurement aids comparison across communities and countries.

Maternal mortality ratio (MMR) is the number of maternal deaths per 100,000 live births in one year where a maternal death is defined as the death of a women due to pregnancy or childbirth, during pregnancy, or within 42 days after delivery. Since complications during pregnancy and childbirth are a leading cause of death and disability among women in LMIC, the MMR is an important indicator of women's health. MMR is the subject of the first of the 17 SDG health goal targets—reduce MMR to less than 70 per 100,000 live births by 2030.

Child mortality may be measured in a number of ways and is a good indicator of health among a community's most vulnerable members. Child mortality is usually measured based on specific age-based categories—probability of death within the first month of life (i.e. neonatal mortality), before the first birthday (i.e. infant mortality), before the fifth birthday (i.e. under-5 mortality), and between 5 and 14 years of age (i.e. mortality among children aged 5–14). Neonatal, infant, and under-5 mortality is generally expressed per 1,000 live births, while mortality among children 5–14 is usually per 1,000 children surviving to 12 months of age. Child mortality is

also emphasized in the health SDG as the second of 17 health targets—reduce neonatal mortality to less than 12 per 1,000 live births and under-5 mortality to less than 25 per 1,000 live births by 2030. Although many LMIC are still far from meeting these targets, measurement of child mortality is improving, which should help draw attention to the issue and facilitate development of meaningful interventions (see for example Box 4.2.14).

Box 4.2.14 Perinatal mortality in Kirakira, Solomon Islands

In a study conducted in the remote, provincial Kirakira Hospital, Solomon Islands, stillbirths, and early neonatal deaths in years 2014 to 2016 were examined to evaluate perinatal mortality rate (Jones et al. 2018).

Method: Birth registry and hospital records

A retrospective audit of birth registry and hospital records was conducted for 2014 to 2016. Perinatal deaths were classified as either stillbirths (i.e. infants that failed to breathe independently after birth) or early neonatal deaths (i.e. infants that breathed independently after birth and subsequently died).

Measure: Perinatal mortality rate

Perinatal mortality rate was calculated as the sum of stillbirths and early neonatal deaths divided by the total number of live births expressed per 1,000.

Result: Persistently very high

A mortality rate of 31 per 1,000 live births was observed, double the 12 per 1,000 SDG target rate, and unchanged over the past 6 years.

Although likely an underestimate since only an estimated 28% of deliveries in the province occurred in Kirakira Hospital, this study found that perinatal mortality was twice the figure officially reported previously for the Solomon Islands.

Morbidity metrics

Morbidity encompasses all conditions that are a departure from being healthy, or in a state of physiological and psychological well-being. There are many ways to measure morbidity in a community, many of which focus on data from RHISs. However, VR, VA, household surveys, and sentinel surveillance systems can also be important sources of data for evaluating morbidity metrics. Similar to mortality indicators, morbidity metrics are generally also easy to grasp as community health improves as morbidity

declines and disaggregation by a number of factors may yield clues as to how to help improve health in the community.

Prevalence, or the total number of people living with a particular disease, disability, or injury among a population at a given time (i.e. point prevalence) or within a given time interval (i.e. period prevalence), is a very common measure of morbidity in LMIC since it is fairly easy to measure or estimate. It is also accessible to a broad audience since it is simply expressed as a percentage of the entire population or per some unit of population. For example, in the Russian Federation, the prevalence of chronic hepatitis C Virus (HCV) infection was estimated to be 336 cases per 100,000 people in 2013 (Mukomolov et al. 2016), making it one of the top countries in the world for chronic HCV cases in the world. Prevalence measurement is one of the most straightforward means of evaluating the effect of interventions on burden of disease in a community (see for example Box 4.2.15).

Box 4.2.15 Mali trachoma prevalence

A study conducted the Kayes region of Mali underscores the value of using prevalence measures to evaluate the effects of interventions (Traoré et al. 2018).

Background: Trachoma, an eye disease caused by infection, is the leading cause of preventable blindness globally. Endemic in 41 countries, trachoma has been highly prevalent in Mali's southwestern Kayes region. In 2002, Mali implemented the WHO's comprehensive strategy to combat trachoma that focuses on surgery, antibiotics, cleanliness, and water/sanitation. Three rounds of mass drug administrations occurred prior to 2006, and since 2009, surgical, behavioural, and environmental interventions have been scaled up.

Aim: To evaluate the effectiveness of these interventions by measuring trachoma prevalence in Kayes, Mali in 2015.

Method: The 2015 population of the Kayes region was estimated, using census data, to be 2.5 million. A cross-sectional study using two-stage, cluster random sampling was performed in the four Kayes districts with the aid of village authorities to facilitate data collection via household survey with both questionnaire and medical examination components.

Result: A total of 11,620 people were examined. Prevalence of active trachoma among children aged 1–9 years declined below the 5% elimination threshold in all four districts and prevalence of trichiasis among adults >15 years decreased below the 0.1% elimination threshold in three districts.

Conclusion: Trachoma interventions implemented in Mali have successfully improved community health and Mali is nearing trachoma elimination.

Incidence is the occurrence of new cases of disease, injury, or disability in a population. Generally, it may be measured as either a proportion (i.e. the number of newly identified cases among a population) or a rate (i.e. the number of newly identified cases in a population during a measured quantity

of observed time). While prevalence measures the cumulative total number of cases regardless of the cases being new or old, incidence measures focus solely on new cases. Measurement of incidence tends to be considerably more difficult as it requires long-term observation of a cohort of people, which may be time and resource intensive.

Causes of death, commonly evaluated for a population as a whole, and for adults (age 15 and older) separately from children (age 14 and younger), may also be used as a good indicator of morbidity. VR is the standard source of information for evaluating causes of death. However, as these data are often incomplete in LMIC, they are supplemented with information obtained via VA and health services records. There are a number of ways to look at causes of death. For example, top causes may be ranked by crude death rates either overall or categorized into communicable diseases, non-communicable diseases, injury, and other causes (e.g. maternal, perinatal, nutritional).

Other metrics

There are a great many other indicators that can make important contributions to community health understanding (see Box 4.2.13), which cannot be covered in any depth here. However, careful consideration of metric selection, measurement, and communication is a critical task that must be prioritized early in any community diagnosis effort. Although the limitations in information sources and metrics measurement in LMIC must be recognized, it should not be merely accepted. It is imperative not only that health indicators are pragmatically selected (i.e. select from among metrics that are currently measurable), but also that health information sources are pragmatically developed to achieve the ability to measure health indicators that are important to the future improvement of health in the community they are meant to serve.

Looking forward

There are many good paths forward for community diagnosis and information systems in LMIC. However, to accelerate improvement in community health at any level beyond its current rate, new methods of solving problems and working together toward common goals must be both creative and realistic.

Private–public partnerships for health

Over the past roughly 20 years, some LMIC have engaged in what is now known as international private–public partnerships for health (PPPHs) (Kostyak et al. 2017). This is where a private, for-profit companies (e.g. pharmaceutical companies), typically based in other countries, enter into collaborative relationships with a public, non-profit entities (e.g. national governments) for the purpose of investigating and addressing larger, more expensive public health problems and providing humanitarian aid. These PPPHs can range from the simple, short-term, one-way arrangements (e.g. the pharmaceutical industry has spent billions of US dollars on cash and product donations to LMIC) to the complex, long-term, true partnerships where experience, expertise, and resources are pooled to achieve common health goals that would be out of reach by either party working independently, maximizing benefit to communities while minimizing risk to stakeholders.

The biggest benefits of PPPHs are the acceleration of innovation and research and development and the expansion of access to affordable health interventions. However, many challenges remain. Geopolitical instability and insecurity make private enterprises reluctant to invest in some countries, gains achieved through PPPHs may be unsustainable beyond the funded period, accountability and oversight is often lacking, and ethical questions abound. Profit-driven private companies may not be motivated purely by altruism, may not measure performance or success by unbiased means, and may not ensure conflicts of interest are properly investigated and adjudicated. Creation of guidelines or regulations governing PPPHs and mechanisms to audit compliance, may mean that LMIC entering into these partnerships can better ensure their people reap the full benefit of these relationships (Kostyak et al. 2017).

Multisectoral collaboration

The interconnectedness and indivisibility of the 17 SDGs was recognized by the UN and heavy emphasis has been placed on cross-sector collaboration and partnerships (UN GA 2015). Like many of the SDGs, ensuring healthy lives and promoting well-being for all at all ages (Goal 3) is unachievable by the healthcare sector alone. However, the healthcare sector has traditionally

operated in a very siloed fashion, separated from, for example, the financial sector, the public security sector, the education sector, and others.

All sectors have important, complimentary contributions to make toward achievement of health SDG targets, as well as health goals at regional, national, subnational, and community levels. The untapped potential of multisector collaboration could be a huge difference-maker, particularly for LMIC where this kind of cooperation has perhaps been less common. However, in order to attain full benefit of this kind of coordinated effort, strong leadership is required. Leadership capacity in LMIC must therefore be expanded and not only distributed across sectors, but connected across sectors, to ensure collaboration naturally yields solutions that are inclusive and implementation plans that are practical. Furthermore, it is critical that champions for health 'causes', or targets, are cultivated within the different sectors and they a have clear vision of how each of their pieces of the tactical and strategic work contribute to the larger common goal (Rasanathan et al. 2017).

New technologies

Promising new technologies have the potential to help the field of public health to make more than just incremental steps forward in LMIC. We are only just starting to embark on development of mHealth technologies, and the power of 'big data' is only beginning to be understood and applied. Together with new portable and point-of-care health technologies for clinical diagnosis and evaluation, the possibilities for rapidly collecting enormous data sets from the field and then applying newly emerging data analytics techniques mean global public health is on the precipice of making huge leaps forward.

Conclusion

As an international community, we must renew and redouble our focus on a human-rights based approach to improvement of global public health. One way to do this is through enhancement of community diagnosis in LMIC by helping in the further development of information systems that support the measurement of performance against important and actionable health metrics. Greater sharing across borders of experience and expertise, tools

and protocols, and lessons learned from both failures and successes can accelerate development of robust information systems in these settings. The resulting larger quantities of higher-quality data, properly analysed by reliable and validated methods, can yield invaluable information upon which health and public health authorities, funding agencies, international aid organizations, and policymakers can better base critical decisions related to allocation of limited resources, programme development, and implementation, and application of emergency services. This evidence-based, targeted approach will maximize improvements in health in individual LMIC, which will naturally drive the betterment global public health.

REFERENCES

Annett, H. and Rifkin, S.B. (1995). Guidelines for rapid participatory appraisals to assess community health needs: a focus on health improvements for low-income urban and rural areas. *World Health Organization (WHO).* [Online] Available at: http://apps.who.int/iris/handle/10665/59366

Basseej, M.H., et al. (2015). Community assessment for determining the health priority problems and community diagnosis: a case study of Qala-Sayed village, Kazeroun, Fars Province, Iran. *Iranian Journal of Public Health*, 44(9), 1303–5.

Berisha, M., et al. (2018). Impact of the national population register in improving the health information system of malignant diseases in Kosova. *Acta Informatica Medica*, 26(1), 62–6.

Bilheimer, L.T. (2010). Evaluating metrics to improve population health. *Preventing Chronic Disease*, 7(4), 1–4.

Broughton, E., et al. (2016). Evaluation of an intervention to improve essential obstetric and newborn care access and quality in Cotopaxi, Ecuador. *Frontier of Public Health*, 4, 247.

Brown, T.M., et al. (2006). The World Health Organization and the transition from 'international' to 'global' public health. *American Journal of Public Health*, 96(1), 62–72.

Cho, S., et al. (2018). Community health needs assessment: a nurses' global health project in Vietnam. *International Nursing Review*, 65(4), 505–14.

Cohen, R.L., et al. (2014). Country progress towards the Millennium Development Goals: adjusting for socioeconomic factors reveals greater progress and new challenges. *Global Health*, 10, 67.

Coonrod, J. (2014). MDGs to SDGs: top 10 differences. The Hunger Project. [Online] Available at: https://advocacy.thp.org/2014/08/08/mdgs-to-sdgs/

de Oliveira, S.V. and Angerami, R.N. (2018). Timeliness in the notification of spotted fever in Brazil: evaluating compulsory reporting strategies and digital disease detection. *International Journal of Infectious Diseases*, 72, 16–18.

Dlamini, S.V., et al. (2017). Knowledge of human social and behavioral factors essential for the success of community malaria control intervention programs: the case of Lomahasha in Swaziland. *Journal of Microbiology, Immunology and Infection*, 50, 245–53.

Easterly, W. (2009). How the millennium development goals are unfair to Africa. *World Development*, 37, 26–35.

Easterly, W. (2013). *The Tyranny of Experts: Economists, Dictators, and the Forgotten Rights of the Poor*. New York: Basic Books.

Eldemire-Shearer, D., et al. (2014). Ageing of Jamaica's population—what are the implications for healthcare? *West Indian Medical Journal*, 63(1), 3–8.

Fehling, M., et al. (2013). Limitations of the Millennium Development Goals: a literature review. *Global Public Health*, 8(10), 1109–22.

Joint United Nations Programme on HIV/AIDS (2014). 90-90-90–An ambitious treatment target to help end the AIDS epidemic. [Online] Available at: http://www.unaids.org/en/resources/documents/2017/90-90-90

Joint United Nations Programme on HIV/AIDS (2017). Ending AIDS: progress towards the 90-90-90 targets. [Online] Available at: http://www.unaids.org/en/resources/documents/2017/20170720_Global_AIDS_update_2017

Jones, P.D., et al. (2018). High perinatal mortality rates persist in Kirakira: the Sustainable Development Goals for health remain out of reach in the provinces of Solomon Islands. *Journal of Paediatric Child Health*, 54(8), 895–9.

Kostyak, L., et al. (2017). A means of improving public health in low- and middle-income countries? Benefits and challenges of international public-private partnerships. *Public Health*, 149, 120–9.

Leon, N., et al. (2015). Routine health information system (RHIS) interventions to improve health systems management (Protocol). *Cochrane Database of Systematic Reviews*, (12), CD012012.

Ma, Y., et al. (2018). The human immunodeficiency virus care continuum in China: 1985–2015. *Clinical Infectious Diseases*, 66(6), 833–9.

Mao, Y., et al. (2010). Development of a unified web-based national HIV/AIDS information system in China. *International Journal of Epidemiology*, 39(Suppl 2), ii79–89.

Marten, R. (2019). How states exerted power to create the Millennium Development Goals and how this shaped the global health agenda: lessons for the sustainable development goals and the future of global health. *Global Public Health*, 14(4), 584–99.

Mukomolov, S., et al. (2016). Hepatitis C in the Russian Federation: challenges and future directions. *Hepatic Medicine*, 8, 51–60.

National Center for AIDS/STD Control and Prevention, China CDC (2018). Annual progress of national HIV/HCV programs in 2017 and priorities and challenges in 2018. Presentation at the National Annual Provincial AIDS Directors Meeting. Ningbo, Zhejiang, China, 19 April 2018.

Rasanathan, K., et al. (2017). Governing multisectorial action for health in low- and middle-income countries. *PLoS Medicine*, 14(4), e1002285.

These, M.S. (2014). Observational and interventional study design types; an overview. *Biochemia Medica*, 24(2), 199–210.

Tindana, P.O., et al. (2007). Grand challenges in global health: community engagement in research in developing countries. *PLoS Medicine*, 4(9), e273.

Traoré, L., et al. (2018). Prevalence of trachoma in the Kayes region of Mali eight years after stopping mass drug administration. *PLoS Neglected Tropical Diseases*, 12(2), e0006289.

United Nations (UN) (2013). *A New Global Partnership: Eradicate Poverty and Transform Economies Through Sustainable Development. The Report of the High-Level Panel of Eminent Persons on the post-2015 Development Agenda*. [Online] Available at: https://sustainabledevelopment.un.org/index.php?page=view&type=400&nr=893&menu=1561

United Nations (UN) (2015). *Millennium Development Goals Report*. [Online] Available at: http://www.un.org/millenniumgoals/2015_MDG_Report/pdf/MDG%202015%20rev%20(July%201).pdf

United Nations (UN) (2017). *The Sustainable Development Goals Report 2017*. [Online] Available at: https://unstats.un.org/sdgs/report/2017/

United Nations General Assembly (UN GA) (1966). *International Covenant on Economic, Social and Cultural Rights*. [Online] Available at: http://www.ohchr.org/Documents/ProfessionalInterest/cescr.pdf.

United Nations General Assembly (UN GA) (2000). *The United Nations Millennium Declaration*. [Online] Available at: http://www.un.org/millennium/declaration/ares552e.pdf

United Nations General Assembly (UN GA) (2015). Transforming our world: the 2030 agenda for sustainable Development. [Online] Available at: https://www.un.org/sustainabledevelopment/sustainable-development-goals/

Vandemoortle, J. (2009). The MDG conundrum: meeting the targets without missing the point. *Development Policy Review*, 27, 355–71.

World Bank (2018). Data: low & middle income. [Online] Available at: https://data.worldbank.org/income-level/low-and-middle-income

World Bank and the World Health Organization (WB and WHO) (2014). Global civil registration and vital statistics: scaling up investment plan 2015–2024. [Online] Available at: http://www.who.int/healthinfo/civil_registration/WB-WHO_ScalingUp_InvestmentPlan_2015_2024.pdf

World Health Organization (WHO) (1946). *Constitution of the World Health Organization.* [Online] Available at: http://apps.who.int/gb/bd/PDF/bd47/EN/constitution-en.pdf?ua=1

World Health Organization (WHO) (1995). City health profiles: how to report on health in your city. [Online] Available at: http://www.euro.who.int/__data/assets/pdf_file/0009/101061/wa38094ci.pdf?ua=1

World Health Organization (WHO) (2016). Verbal autopsy standards: ascertaining and attributing causes of death. [Online] Available at: http://www.who.int/healthinfo/statistics/verbalautopsystandards/en/

Wu, Z., et al. (2015). Simplified HIV testing and treatment in China: analysis of mortality rates before and after a structural intervention. *PLoS Medicine*, 12, e1001874.

Wu, Z., et al. (2017). Testing and linkage to HIV care in China: a cluster-randomised trial. *Lancet HIV*, 4(12), e555–65.

4.3

New communication technologies, social media, and public health

Patrick S. Sullivan, Aaron J. Siegler, and Lisa Hightow-Weidman

Introduction

Communication technologies have evolved at a breathtaking pace, and the resulting opportunities to disseminate information and to engage in interpersonal communication have critical implications for public health. These conduits for dissemination of public health content—whether prevention messages, public health advisories or alerts, or clinical or public health guidelines—offer opportunities to reach the right people with the right information more quickly than ever. Considering the public health framework for capitalizing on new communication technologies requires us to consider the technologies themselves, the kinds of content that can be shared, the audiences for such information, and theories of behavioural and systems change. This discussion must also acknowledge that the technology platforms, data speeds, and predominant communications software are heterogenous around the world. In this chapter, we articulate the major dimensions and considerations of the use of new communication technologies for public health, and reflect on important differences in different global public health settings.

Global versus developed country trends in the United States

Two-thirds of the global population are now mobile subscribers with smartphones, accounting for over half of the total connections globally (GSMA Intelligence 2017). According to the International Telecommunication Union, in 2015 there were more than 7 billion mobile

telephone subscriptions across the world, over 70% of which were in low-or middle-income countries (International Telecommunications Union 2015; International Telecommunications Union 2017). However, even though smartphone adoption is near saturated in many developed countries, approximately half the world is not yet online. India and sub-Saharan Africa account for 42% of the world's unconnected, with more than 60% of their respective populations not yet on the internet.

Trends in the United States follow a similar pattern, with the digital divide still seen related to income and geographic locations. Roughly three in ten adults with household incomes below $30,000 a year (29%) don't own a smartphone. More than four in ten with low incomes don't have home broadband services (44%) or a traditional computer (46%) (Pew Research Center 2019). By comparison, each of these technologies is nearly ubiquitous among adults in households earning $100,000 or more a year. Rural Americans also remain less likely than non-rural adults to have home broadband, smartphones, and other devices (Pew Research Center 2019).

eHealth has been a priority for the World Health Organization (WHO) since 2005, when the World Health Assembly resolution WHA58.28 was adopted: 'eHealth is the cost-effective and secure use of information communication technologies (ICT) in support of health and health-related fields, including healthcare services, health surveillance, health literature, and health education, knowledge and research'. The WHO Global Observatory for eHealth (GOe) 2015 survey of Member States documented the surge in adoption of eHealth in countries. All 194 WHO Member States were surveyed, of which 125 responded. Slightly more than half of the countries (n = 73; 58%) reported having an eHealth strategy in place (WHO 2016). Of those countries with an eHealth strategy, 69 (94%) reported that they had special funding allocated for the implementation of their strategy

Phone technology, data speed, data replacing cellular, cost structures

Although fixed-telephone subscriptions continue to decline with a penetration rate of 12.4% in 2018, the number of mobile cellular telephone subscriptions is greater than the global population. In 2018, there were more fixed-broadband connections (1.1 billion) than fixed-telephone connections

(942 million). The number of active mobile broadband subscriptions have increased from 268 million in 2007 to 5.3 billion in 2018. Mobile broadband is more affordable than fixed-broadband services in most developing countries. However, mobile broadband prices represent more than 5% of gross national income (GNI) per capita in most least developed countries (LDCs) and are therefore unaffordable for most of the population. Nearly the entire world population (96%) now lives within reach of a mobile cellular network, and most of the global population (90%) can access the internet through a 3G or higher speed network. Declining cost of smartphones and data plans globally is likely to continue and allow for increased access.

Since the inception of mobile phone service, there has been an ongoing trend towards increasing data speed. Data speeds have increased three orders of magnitude since the introduction of 2G service in 1991; the maximum speed of 5G networks (10 Gbps) are nearly 200,000 times faster than the 2G speeds of three decades earlier. These trends are important, because some public health applications of technology (e.g. streaming videos, transmitting high resolution images, video-based counselling, or situational assessments) require faster data speeds. As is the case for coverage of smartphone technologies, access to higher data speeds has substantial variability along dimensions of urbanicity versus rurality, and between low-/middle-income countries and high-income countries.

Multiple social networking channels, time on social network, generational changes

Social media platforms are ubiquitous globally, with substantial differences in use based on geography and demographics (race, age, gender). Within the United States, those ages 18 to 24 are substantially more likely than those ages 25 to 29 to say they use Snapchat (73% vs. 47%) and Instagram (75% vs. 57%). Roughly three-quarters of Facebook users (74%) visit the site daily, including about half who do so several times a day. Roughly eight in ten Snapchat users ages 18 to 29 (77%) say they use the app every day, including 68% who say they do so multiple times day. Similarly, 76% of Instagram users in this age group visit the site daily, with 60% reporting that they do so several times per day.

Popular social media and messaging platforms like Facebook and WhatsApp have drawn attention for their potential role in spreading

misinformation, facilitating political manipulation, and increasing violence and hate crimes. However, social media may also have a positive impact on the lives of youth. Most teens aged 13–17 (81%) say social media makes them feel more connected to what's going on in their friends' lives, while around two-thirds say these platforms make them feel as if they have people who will support them through tough times.

Need to anticipate technologies that work for youth

According to a 2017 report by ICT, in developed countries, 94% of young people aged 15–24 use the internet, compared with 67% of young people in developing countries and only 30% in LDCs. The proportion of young people aged 15–24 using the internet (71%) is significantly higher than the proportion of the total population using the internet (48%).

Types of social media

Social media channels constitute a dynamic set of applications that are designed to facilitate the circulation and promotion of information through social connections. Because the specific apps delivering these channels evolve over time, we here describe types of channels and their features, and cite current-day examples of these communication types. They vary largely based on the format of information sharing and length of messages, and on the means of sharing through social networks. Subsequently, we will discuss public health applications of these types of communication channels and provide examples of such applications.

Interaction of technology and public health

Technology-based public health services may interact with members of communities, healthcare providers, and policymakers in multiple ways.

Providing information: Websites, including mobile optimized websites, and apps provide static information on a wide range of topics. For example, public health agencies may provide information and guidelines on topics ranging from vaccine recommendations,[1] local and regional HIV prevention and care plans,[2] and Malaria control.[3] Twitter is used by many public agencies to distribute public health alerts (see Fig. 4.3.1a). Formats include

text about health programmes or conditions, infographics (which illustrate data through simple illustrations; see Fig. 4.3.1b).

Providing counselling or navigation: Chat functions and service locators are used by public health providers to distribute information, make referrals for appropriate services, and provide recommendations for service providers nearest to users. For example, PleasePrEPME: Connect is an online programme to promote appropriate use of pre-exposure prophylaxis (PrEP) that offers chat-based navigation for PrEP to US residents.[4] Services include health systems navigation, identifying appropriate providers, and HIV information.

Behavioural change interventions: The concept of providing behaviour change interventions through eHealth or mHealth channels is appealing for many reasons, including the opportunity to reach people in need of behaviour change services who do not live near public health centres, and the opportunity to provide such services at minimal marginal cost through computer-delivered interventions (Sullivan et al. 2013). In general, excitement about the promise of such interventions has been more consistent than rigorous evidence of the interventions themselves (Evers 2006). In fields as diverse as smoking cessation (Augustson et al. 2016; Ponciano-Rodríguez et al. 2018), diabetes prevention (Joiner et al. 2017), medication adherence (Linn et al. 2011), and engagement in health research (Khosropour et al. 2013), results on the efficacy of eHealth interventions have been mixed. However, newer generations of eHealth interventions feature increasingly interactive technologies (Muessig et al. 2015), and recent reports suggest efficacy of theory-based interventions that include multiple sessions, rich media, and interactive components (Mustanski et al. 2018). Meta-analyses suggest that certain types of interventions may be especially well suited to certain health behaviours (e.g. text messaging for smoking cessation and physical activity (Head et al. 2013)).

Identifying service locations: Other service leverage the geolocation features of computers and smartphones to provide tailored recommendations of service locations near to technology users. For example, an online service, PrEPLocator.org, provides tailored referrals to PrEP providers and driving directions to nearest providers (Siegler et al. 2017). Other locator apps or websites include referrals to blood pressure check points (HealthMapNYC.org; see Figure 4.3.1c), and syringe exchange services

(www.nasen.org). AIDSVu.org provides locators for HIV testing, PrEP providers, and HIV treatment (Valdiserri and Sullivan 2018).

(a)

Public Health England @PHE_uk · May 1

Do you or your child have a vaccination appointment coming up?

It's still vital to attend to get protected against preventable diseases.

Only reschedule your appointment if you or a member of your household has #coronavirus symptoms.

Find out more: nhs.uk/conditions/vac...

4 128 102

(b)

World Health Organization (WHO) @WHO · Apr 29

FACT:
Spraying or introducing bleach or another disinfectant into your body WILL NOT protect you against #COVID19 and can be dangerous.

More: bit.ly/COVID19Mythbus...

#coronavirus

#KnowTheFacts

PAHO/WHO and 5 others

295 1.4K 2.8K

Fig. 4.3.1 Examples of social media and apps for promoting public health. (a) Twitter public health alert; (b) Twitter public health infographic (c) Geolocation for nearby health service locations.

Commodity ordering: As people increasingly use technology to order goods and services, public health programmes also have opportunities to use communication technologies to deliver commodities to promote health and prevent new health problems. For example, programmes to deliver home-testing kits for sexually transmitted infections have been highly acceptable. A randomized study of online ordering and mail distribution of STI test kits in London showed a doubling of the rate of STI testing and in the rate of STI diagnoses (Wilson et al. 2017). Similar mailout STI kits have been shown to have high acceptability and completion rate of specimen return among men who have sex with men (MSM) (Mustanski et al. 2017; Siegler et al. 2019a) and women (Gaydos et al. 2006) in the United States.

Data collection/outbreak identification: New communication technologies provide opportunities to collect information for public health action, both through active engagement of users of social media, through data mining of public social media content, and through use of social media content to tailor health interventions. Accessing people to participate in health studies, including surveys, offers opportunities to collect data episodically on geographically diverse populations. For example, the American Men's Internet Survey has collected data on over 10,000 internet-recruited MSM

annually since 2012, and publishes reports on trends in risk and preventive behaviours (Sanchez et al. 2015; Sanchez et al. 2018). The systematic analysis of public postings or search strings can also serve as a crowd-sourced surveillance system. Searches for influenza-related topics have been reported to correlate closely with traditional influenza surveillance systems (Cook et al. 2011), but cannot currently replace other traditional surveillance systems (Olson et al. 2013). Twitter-based systems for detecting outbreaks at large public gatherings (Yom-Tov et al. 2014) and a Twitter-based system for pharmacovigilance have also been evaluated (Freifeld et al. 2014).

Data visualization and access to services: The visualization of data—whether the data arise from social media, new media, or other sources—represents a powerful opportunity to depict the relationship between public health needs and current or future resources. For example, electronic mapping of surveillance data on HIV concentration relative to existing HIV testing facilities has informed the development and implementation of targeted community-based HIV testing programmes (Valdiserri and Sullivan 2018). In a different public health application, geolocation data on the whereabouts of users of a sex-seeking mobile phone app were plotted to identify areas of a city in which sexual health promotion programmes should be targeted (Delaney et al. 2014). In Slovenia, data visualization methods were used to depict the accessibility and availability of public healthcare resources (Lavrač et al. 2007); in the United States, data on public and private PrEP providers were used to illustrate ‘service deserts’, in which residents did not have meaningful geographic access to PrEP services (Weiss et al. 2018).

Drawbacks and public health concerns

Despite the substantial opportunities for public health presented by new communications media, there are also important drawbacks, and it is important to weigh the opportunities and concerns about new media. For example, the anonymous and continual opportunities to comment to and about others has given risk to an epidemic of cyberbullying: by 2008, nearly one in eight teens reported cyberbullying others, and nearly a quarter of teens in the Netherlands reported being cyberbullied (DeHue et al. 2008); being a victim of cyberbullying is reported more often by young women than by young men (Brochado et al. 2017). Internet addiction is also a public

health concern (Young 1998), and has been hypothesized to be associated with depression and obesity (Young 1998; Flisher 2010). WHO has organized a series of meetings on the public health implications of addictive behaviours; gaming addiction has been listed as a mental health condition by the WHO in the ICD-11 revision in 2018 (Reed et al. 2019).

Some mobile apps might also increase behaviours that undermine public health priorities. For example, the rise of dating apps for gay and bisexual men has occurred concurrently with rises in HIV infections (Chapin-Bardales et al. 2018), condomless sex (Hess et al. 2017), and diagnoses of sexually transmitted infections (Abara et al. 2016) among these men. Finally, there are privacy concerns associated with using new communication technologies to facilitate delivery of public health information and services. The transmission of health-related information must be conducted in compliance with recognized standards of security and encryption to avoid disclosure of health-related information. Further, providers of public health services through electronic modes of communication should establish and adhere to policies to limit disclosure of users' health information and secondary or collateral uses of data. For reporting of programme results with geolocating data, suppression schemes should be adopted to prevent indirect identification of users of new media health resources (Sullivan 2020).

Theory relevant to social media

The Diffusion of Innovations theory explores patterns and rates of new technology adoption. The theory posits, and has demonstrated empirical support for, five key attributes of new technologies that enhance their adoption. The *relative advantage of the innovation*, including economic, convenience, or satisfaction, influences it rate of uptake. The *compatibility of a new technology with social values or prior experiences*, and the *perceived complexity* of using it, also influence adoption. Interventions that can be easily tried (*trialability*) and whose impact can be observed (*observability*) are also more likely to have rapid uptake. These intervention traits do not exist in a vacuum, but instead are based upon each individual's values, and therefore are subject to change over time as societal norms and customs shift. For instance, the relative advantage of a social media intervention is likely to shift over time because social media platforms and their usage norms shift. New technology-based interventions to provide

health information may not be competing with calls to a clinician's office, but instead with 'google it up' strategies (Sundstrom 2016). Unlike many intervention theories, Diffusion of Innovations also describes characteristics of those adopting an innovation in order to facilitate targeting of programmes in the most relevant way. Innovators and early adopters (those most likely to first partake of a new technology or opportunity) are characterized by high ability and motivation to adopt. Stable traits of innovators have not been consistently identified, and therefore staging of individuals should be applied to a specific domain of the area of interest, where motivation and ability can be more granularly assessed and targeted (Greenhalgh et al. 2008).

Engagement in the research field, public health engaging with technologies

Public health adoption of interventions and programmes based on new communication technologies calls for engagement between the developers of technologies and public health programmes. Often, the new technologies that will shape public health opportunities are developed by commercial interests. Engagement between technology developers and public health officials is critical. For example, Building Healthy Online Communities[5] is a partnership between the developers of mobile apps for sex seeking and public health constituencies. App owners and developers receive feedback from public health about how to support the health of their users, and researchers gain access to best practices for advertising and targeting app users for public health interventions.

Increasingly, questions emerge about the best ways to implement public health interventions, and this field of inquiry is especially focused on how to implement eHealth interventions—an emerging area for which there are not a large number of historical examples of implementation. The implementation science field of inquiry ideally starts in the stage of design of eHealth interventions, with researchers discussing possible implementation approaches if the intervention proves successful. After efficacy trials or demonstrations of programmes, a second phase of study of implementation practices may occur. For example, after randomized trial evidence that an online intervention reduced rates of new sexually transmitted infections in young gay men (Mustanski et al. 2018), researchers

developed a research-focused protocol to ask about the best ways to implement the intervention in authentic public health settings.[6] The study will compare costs and participant engagement in two delivery models: one model in which the intervention is delivered by a local community-based organization, and a second model in which a single national provider delivers the intervention. These studies also provide further evidence about the performance of interventions in 'real world' public health settings.

Future trends in social media and public health

Current trends suggest that the future will be characterized by decreasing costs of mobile devices, increasing smartphone ownership rates, increasing data speeds, and increasing sophistication of platforms to disseminate content through social media. In one sense, this suggests an opportunity to democratize public health resources using social media as coverage and data speed increase. Simultaneously, computing platforms and artificial intelligence systems will offer new opportunities to analyse large quantities of data and use automated systems to identify emerging public health issues, predict where public health resources might be needed, and provide a broader contextual understanding of health behaviours (Neill 2012).

Another exciting trend is the growth of wearable technologies that can record certain biometric data (e.g. body temperature, heart rate), location, and movement (Patel et al. 2015). These devices offer several avenues to the improvement and evaluation of public health programmes. First, the devices document levels of physical activity, objective measures of which are key to understanding the risks of populations and the impacts of interventions to increase activity. Second, the devices may directly detect certain health conditions (e.g. potential atrial fibrillation), which can both promote early medical intervention and serve as a data source for planning health resources. Third, these devices provide data on the location of individuals. Such data are valuable in considering the locations in which certain health risks might occur, and the related deployment of health interventions. Preliminary studies suggest that people are willing to wear devices and share data to improve the health services offered to them (Duncan 2013). However, these new data streams should prompt careful consideration of how to protect the privacy and medical data of people who wear devices and share data for public health purposes.

Implementation of new communication and social media-based interventions

There are important questions about how public health programmes or interventions based on new communication technologies or social media would be implemented. Current public health delivery systems vary by country and operate in diverse structures, with varying levels of centralized or disseminated public health authority. Regardless, implementation of technology-based programmes or public health systems will require a workforce with specialized skills and facility with both technologies used for social media and public health intervention. The Combined Framework for Implementation Research (CFIR) suggests that we should consider how new technologies will impact public health delivery systems by considering multiple levels of implementation impacts (Damschroder et al. 2009). For example, public health programmes or interventions should be developed in consideration of relative advantage to current solutions, adaptability to diverse public health settings, and cost. Similarly, patient needs and resources and external policies (for example, on data security and data sharing) must be considered. A trained workforce with skills and comfort with requisite technology tools will be required, and champions within all levels of public health systems will be needed to overcome barriers to implementation.

Literacy and preference

There are multiple aspects of literacy and preference that should be considered as technology- and social media-based public health tools are developed and implemented. For example, full benefit from a mobile app to improve behaviours to reduce obesity will be realized only if users possess the technology literacy, health literacy, and linguistic literacy to be able to interact with the app and translate its information and suggestions to appropriate health actions. In turn, these aspects of literacy vary across elements of geography, age, race/ethnicity, and social determinants of health. Not coincidentally, some communities with highest needs for health interventions are also communities that are impacted by poor health literacy, low access to health facilities, poverty, and lack of meaning access to

technologies (e.g. sufficient bandwidth for multimedia tools, monetary resources to pay for data plans). There may be similar 'mismatches' of need and willingness to adopt eHealth tools for older people, who are at increased risk for many chronic health conditions and who, as a group, have lower eHealth literacy (Chesser et al. 2016). There might also be cultural and normative values favouring in-person modes of receiving health advice or services. For example, African American and Latinx patients in a Health Maintenance organization expressed concerns about using electronic modes of communication for health services, citing the lack of technical support for using eHealth tools and concerns about losing personal relationships with providers (Ansari et al. 2010). Such preferences for in-person public health services should both be qualitatively understood in the process of designing new eHealth tools, and should be respected by ensuring that comparable health services are available in traditional settings for those with such preferences.

Opportunities for the use of technology in the conduct of public health studies

Technology offers new opportunities to make the conduct of clinical trials and implementation research more efficient and to allow the collection of rich, granular, and time-, or location-informed data. For example the use of mobile devices is ideally suited to the collection of data as ecological momentary assessments, which aim to assess phenomena as they occur in the environments in which they occur (Stone and Shiffman 1994). Geofencing technology allows for the triggering of messaging or short surveys when the research participant is near a location of interest (for example, triggering data collection when a research subject is near a location of a health service or a risk environment) (Biello et al. 2019). By offering the opportunity to complete short data collections on mobile devices, recall bias can be minimized: for example, an app-based assessment approach to collecting data on condom usage and failure allowed collection of granular data on rates of condom breakage in a study of a new condom type (Siegler et al. 2019b). Other unique opportunities to improve the conduct of trials and granularity of data include incorporating data on activity spaces of participants (Duncan et al. 2014) and the use of paradata—data on the

patterns of technology uses—to enhance understanding of the mediators of effect in evaluating app-based interventions (Bauermeister et al. 2017).

Conclusion

It is tempting to cast the emergence of new communication technologies and social media as unprecedented disruptors of public health strategies. Although it is true that the speed with which these tools and communication channels have emerged is remarkable, public health and medicine have transitioned through many disruptive phases of medical practice and technology. At their core, new communication technologies and social media are new approaches to disseminating information, collecting information, and engaging in dialogue with the people that public health seeks to serve. They will change public health in ways that are incremental and the process of realizing their potential will evince many benefits of using these approaches in public health, and likely also some unintended outcomes.

It is critical to monitor and anticipate that, as with many new technologies, there exists a real possibility that the introduction of new technologies might initially exacerbate health inequities in access to services or in outcomes, because of iniquities in the component elements of access—access to suitable hardware, access to adequate data speed, and access to resources to pay for data access. These inequities must be anticipated and mitigated or avoided. Yet, these technologies also offer special potential to increase the availability of health services to populations who might have less access for reasons of lack of proximity to health facilities or lack of resources to commute within cities.

The growth of new communications media and social media are remarkable challenges and opportunities for public health. Getting the most out of these opportunities will require a willingness for technology developers and public health agencies to work together to leverage emerging capacities. Evaluation of public health new programmes and a commitment to systematically studying implementation methods will yield data to inform rational decisions about how, where, and when technology-based public health programmes and interventions will do the most good.

REFERENCES

Abara, W.E., Hess, K.L., Fanfair, R.N., Bernstein, K.T., and Paz-Bailey, G. (2016). Syphilis trends among men who have sex with men in the United States and Western Europe: a systematic review of trend studies published between 2004 and 2015. *PLoS One*, 11, e0159309.

Ansari, A.A., et al. (2010). Blocking of alpha4beta7 gut-homing integrin during acute infection leads to decreased plasma and gastrointestinal tissue viral loads in simian immunodeficiency virus-infected rhesus macaques. *Journal of Immunology*, 186, 1044–59.

Augustson, E., et al. (2016). Text to Quit China: an mHealth smoking cessation trial. *American Journal of Health Promotion*, 31, 217–25.

Bauermeister, J.A., Golinkoff, J.M., Muessig, K.E., Horvath, K.J., and Hightow-Weidman, L.B. (2017). Addressing engagement in technology-based behavioural HIV interventions through paradata metrics. *Current Opinion in HIV and AIDS*, 12, 442–6.

Biello, K.B., Marrow, E., Mimiaga, M.J., Sullivan, P., Hightow-Weidman, L., and Mayer, K.H. (2019). A mobile-based app (MyChoices) to increase uptake of HIV testing and pre-exposure prophylaxis by young men who have sex with men: protocol for a pilot randomized controlled trial. *JMIR Research Protocols*, 8, e10694.

Brochado, S., Soares, S., and Fraga, S. (2017). A scoping review on studies of cyberbullying prevalence among adolescents. *Trauma, Violence, & Abuse*, 18, 523–31.

Chapin-Bardales, J., et al. (2018). Trends in human immunodeficiency virus diagnoses among men who have sex with men in North America, Western Europe, and Australia, 2000–2014. *Annals of Epidemiology*, 28, 874–80.

Chesser, A., Burke, A., Reyes, J., and Rohrberg, T. (2016). Navigating the digital divide: a systematic review of eHealth literacy in underserved populations in the United States. *Informatics for Health and Social Care*, 41, 1–19.

Cook, S., Conrad, C., Fowlkes, A.L., and Mohebbi, M.H. (2011). Assessing Google flu trends performance in the United States during the 2009 influenza virus A (H1N1) pandemic. *PLoS One*, 6, e23610.

Damschroder, L.J., Aron, D.C., Keith, R.E., Kirsh, S.R., Alexander, J.A., and Lowery, J.C. (2009). Fostering implementation of health services research findings into practice: a consolidated framework for advancing implementation science. *Implementation Science: IS*, 4, 50–50.

Dehue, F., Bolman, C., and Völlink, T. (2008). Cyberbullying: youngsters' experiences and parental perception. *CyberPsychology & Behavior*, 11, 217–23.

Delaney, K.P., Kramer, M.R., Waller, L.A., Flanders, W.D., and Sullivan, P.S. (2014). Using a geolocation social networking application to calculate the population density of sex-seeking gay men for research and prevention services. *Journal of Medical Internet Research*, 16, e249.

Duncan, D.T. (2013). What's your Walk Score®?: web-based neighborhood walkability assessment for health promotion and disease prevention. *American Journal of Preventive Medicine*, 45, 244–5.

Duncan, D.T., et al. (2014). Application of global positioning system methods for the study of obesity and hypertension risk among low-income housing residents in New York City: a spatial feasibility study. *Geospatial Health*, 9, 57.

Evers, K.E. (2006). eHealth promotion: the use of the Internet for health promotion. *American Journal of Health Promotion*, 20, 1–14.

Flisher, C. (2010). Getting plugged in: an overview of internet addiction. *Journal of Paediatrics and Child Health*, 46, 557–9.

Freifeld, C.C., et al. (2014). Digital drug safety surveillance: monitoring pharmaceutical products in twitter. *Drug Safety*, 37, 343–50.

Gaydos, C.A., et al. (2006). Internet-based screening for chlamydia trachomatis to reach nonclinic populations with mailed self-administered vaginal swabs. *Sexually Transmitted Diseases*, 33, 451–7.

Greenhalgh, T., Robert, G., Bate, P., Macfarlane, F., and Kyriakidou, O. (2008). *Diffusion of Innovations in Health Service Organisations: A Systematic Literature Review*. Hoboken, NJ:

John Wiley & Sons.
GSMA Intelligence (2017). Global mobile trends 2017. [Online] Available at: https://www.gsmaintelligence.com/research/?file=3df1b7d57b1e63a0cbc3d585feb82dc2&download
Head, K.J., Noar, S.M., Iannarino, N.T., and Harrington, N.G. (2013). Efficacy of text messaging-based interventions for health promotion: a meta-analysis. *Social Science & Medicine*, 97, 41–8.
Hess, K.L., Crepaz, N., Rose, C., Purcell, D., and Paz-Bailey, G. (2017). Trends in sexual behavior among men who have sex with men (MSM) in high-income countries, 1990–2013: a systematic review. *AIDS and Behavior*, 21, 2811–34.
International Telecommunications Union (2015). Measuring the Information Society Report. [Online] Available at: https://www.itu.int/en/ITU-D/Statistics/Documents/publications/misr2015/MISR2015-w5.pdf
International Telecommunications Union (2017). Mobile-cellular telephone subscriptions In: *Key ICT Indicators for Developed and Developing Countries and the World (Totals and Penetration Rates)*. [Online] Available at: https://www.itu.int/en/ITU-D/Statistics/Documents/statistics/2017/ITU_Key_2005-2017_ICT_data.xls
Joiner, K.L., Nam, S., and Whittemore, R. (2017). Lifestyle interventions based on the diabetes prevention program delivered via eHealth: a systematic review and meta-analysis. *Preventive Medicine*, 100, 194–207.
Khosropour, C.M., Johnson, B.A., Ricca, A.V., and Sullivan, P.S. (2013). Enhancing retention of an Internet-based cohort study of men who have sex with men (MSM) via text messaging: randomized controlled trial. *Journal of Medical Internet Research*, 15(8), e194.
Lavrač, N., Bohanec, M., Pur, A., Cestnik, B., Debeljak, M., and Kobler, A. (2007). Data mining and visualization for decision support and modeling of public health-care resources. *Journal of Biomedical Informatics*, 40, 438–47.
Linn, A.J., Vervloet, M., Van Dijk, L., Smit, E.G., and Van Weert, J.C. (2011). Effects of eHealth interventions on medication adherence: a systematic review of the literature. *Journal of Medical Internet Research*, 13, e103.
Muessig, K.E., Nekkanti, M., Bauermeister, J., Bull, S., and Hightow-Weidman, L.B. (2015). A systematic review of recent smartphone, internet and web 2.0 interventions to address the HIV continuum of care. *Current HIV/AIDS Reports*, 1–18.
Mustanski, B., et al. (2017). Internet-based HIV prevention with at-home sexually transmitted infection testing for young men having sex with men: study protocol of a randomized controlled trial of keep it up! 2.0. *JMIR Research Protocols*, 6, e1.
Mustanski, B., Parsons, J.T., Sullivan, P.S., Madkins, K., Rosenberg, E., and Swann, G. (2018). Biomedical and behavioral outcomes of Keep It Up!: an eHealth HIV prevention program RCT. *American Journal of Preventive Medicone*, 55, 151–8.
Neill, D.B. (2012). New directions in artificial intelligence for public health surveillance. *IEEE Intelligent Systems*, 27, 56–9.
Olson, D.R., Konty, K.J., Paladini, M., Viboud, C., and Simonsen, L. (2013). Reassessing Google flu trends data for detection of seasonal and pandemic influenza: a comparative epidemiological study at three geographic scales. *PLoS Computational Biology*, 9, e1003256.
Patel, M.S., Asch, D.A., and Volpp, K.G. (2015). Wearable devices as facilitators, not drivers, of health behavior change. *The Journal of the American Medical Association*, 313, 459–60.
Pew Research Center (2019). Digital gap between rural and nonrural America persists. [Online] Available at: https://www.pewresearch.org/fact-tank/2019/05/31/digital-gap-between-rural-and-nonrural-america-persists/
Ponciano-Rodríguez, G., Reynales-Shigematsu, L.M., Rodríguez-Bolaños, R., Pruñonosa-Santana, J., Cartujano-Barrera, F., and Cupertino, A.P. (2018). Enhancing smoking cessation in Mexico using an e-Health tool in primary healthcare. *Salud Pública de México*, 60, 549–58.

Reed, G.M., et al. (2019). Innovations and changes in the Icd-11 classification of mental, behavioural and neurodevelopmental disorders. *World Psychiatry*, 18, 3–19.
Sanchez, T.H., Sineath, R.C., Kahle, E.M., Tregear, S.J., and Sullivan, P.S. (2015). The annual American men's internet survey of behaviors of men who have sex with men in the United States: protocol and key indicators report 2013. *JMIR Public Health and Surveillance*, 1, e3.
Sanchez, T.H., Zlotorzynska, M., Sineath, R.C., Kahle, E., Tregear, S., and Sullivan, P.S. (2018). National trends in sexual behavior, substance use and HIV testing among United States men who have sex with men recruited online, 2013 through 2017. *AIDS Behaviour*, 22, 2413–25.
Siegler, A.J., et al. (2019a). Developing and assessing the feasibility of a home-based PrEP monitoring and support program. *Clinical Infectious Diseases*, 68(3), 501–4.
Siegler, A.J., et al. (2019b). Levels of clinical condom failure for anal sex: a randomized cross-over trial. *EClinicalMedicine*, 17, 100199.
Siegler, A.J., Wirtz, S., Weber, S., and Sullivan, P.S. (2017). Developing a web-based geolocated directory of HIV pre-exposure prophylaxis-providing clinics: the PrEP locator protocol and operating procedures. *JMIR Public Health Surveillance*, 3, e58.
Stone, A.A. and Shiffman, S. (1994). Ecological momentary assessment (EMA) in behavorial medicine. *Annals of Behavioral Medicine*, 16(3), 199–202.
Sullivan, P.S., et al., (2020). A data visualization and dissemination resource to support HIV prevention and care at the local level: analysis and uses of the AIDSVu Public Data Resource. *Journal of Medical Internet Research*, 22(10), e23173.
Sullivan, P.S., Grey, J.A., and Simon Rosser, B.R. (2013). Emerging technologies for HIV prevention for MSM: what we have learned, and ways forward. *Journal of Acquired Immune Deficiency Syndrome*, 63 Suppl 1, S102–7.
Sundstrom, B. (2016). Mothers 'Google It Up:' Extending communication channel behavior in diffusion of innovations theory. *Health Communication*, 31, 91–101.
Valdiserri, R.O. and Sullivan, P.S. (2018). Data visualization promotes sound public health practice: the AIDSVu Example. *AIDS Educational Preview*, 30, 26–34.
Weiss, K., Bratcher, A., Sullivan, P.S., and Siegler, A. (2018). *Geographic Access To PrEP Clinics Among Us MSM: Documenting PrEP Deserts*. Abstract 1006. Presented at CROI 2018, 4–7 March 2018, Boston.
Wilson, E., et al. (2017). Internet-accessed sexually transmitted infection (e-STI) testing and results service: a randomised, single-blind, controlled trial. *PLoS Medicine*, 14, e1002479.
World Health Organization (WHO) (2016). *Global Diffusion of eHealth: Making Universal Health Coverage Achievable*. [Online]. Available at: https://www.who.int/goe/publications/global_diffusion/en/
Yom-Tov, E., Borsa, D., Cox, I.J., and McKendry, R.A. (2014). Detecting disease outbreaks in mass gatherings using internet data. *Journal of Medical Internet Research*, 16, e154.
Young, K.S. (1998). Internet addiction: the emergence of a new clinical disorder. *Cyberpsychology & Behavior*, 1, 237–44.

---

[1] https://www.canada.ca/en/public-health/services/provincial-territorial-immunization-information/provincial-territorial-routine-vaccination-programs-infants-children.html

[2] https://sanac.org.za/

[3] https://health.go.ug/programs/national-malaria-control-program

[4] https://www.pleaseprepme.org/sites/default/files/file-attachments/PleasePrEPMe%20Connect%20AIDS%202018_0.pdf

[5] www.bhocpartners.org

[6] https://news.northwestern.edu/stories/2018/august/keep-it-up-grant/

SECTION 5

Epidemiological and biostatistical approaches

5.1 Epidemiology: the foundation of public health
Roger Detels

5.2 Cross-sectional studies
Manolis Kogevinas and Leda Chatzi

5.3 Principles of outbreak investigation
Sopon Iamsirithaworn, Panithee Thammawijaya, and Kumnuan Ungchusak

5.4 Case–control studies
Noel S. Weiss

5.5 Cohort studies
Alvaro Muñoz and F. Javier Nieto

5.6 Qualitative research imagination
Jennie Popay and Fran Baum

5.7 Methodological issues in the design and analysis of cluster randomized trials
Kathy J. Baisley, Richard J. Hayes, and Lawrence H. Moulton

5.8 Community intervention trials in high-income countries
John W. Farquhar and Lawrence W. Green

5.9 Natural and quasi-experiments
Peter Craig

5.10 Clinical epidemiology
Fiona F. Stanaway, Naomi Noguchi, Clement Loy, Sharon Reid, and Jonathan C. Craig

5.11 Validity and bias in epidemiological research
Sander Greenland and Tyler J. VanderWeele

5.12 Causation and causal inference
Katherine J. Hoggatt, Tyler J. VanderWeele, and Sander Greenland

5.13 Systematic reviews and meta-analysis
Nandi Siegfried and Lawrence Mbuagbaw

5.14 Statistical methods
Gail Williams and Robert S. Ware

5.15 Measuring the health of populations: the Global Burden of Disease study methods
Theo Vos, Christopher J.L. Murray and Alan D. Lopez

5.16 Mathematical models of transmission and control of infectious agents
Alex Welte and Cari van Schalkwyk

5.17 Public health surveillance
Nguyen Tran Hien, James W. Buehler, and Ann Marie Kimball

5.18 Life course epidemiology and analysis
Elizabeth Rose Mayeda, Alexandra M. Binder, and Lindsay C. Kobayashi

5.1

Epidemiology

The foundation of public health

Roger Detels

Introduction to epidemiology

Epidemiology is the basic science of public health, because it is the science that describes the relationship of health and/or disease with other health-related factors such as human pathogens in human populations. Furthermore, epidemiology has been used to generate much of the information required by public health professionals to develop, implement, and evaluate effective intervention programmes for the prevention of disease and promotion of health, such as the eradication of smallpox, the anticipated eradication of polio and guinea worm disease, and the prevention of heart disease and cancer. Unlike pathology, which constitutes a basic area of knowledge, and cardiology, which is the study of a specific organ, epidemiology is a philosophy and methodology that can be applied to learning about and resolving a very broad range of health problems. It is not enough to know what the various study designs and statistical methodologies are. The ‘art’ of epidemiology is knowing when and how to apply the various epidemiological strategies creatively to answer specific health questions.

The uses and limitations of the various epidemiological study designs are presented in this chapter to illustrate and underscore the fact that the successful application of epidemiology requires more than a knowledge of study designs and epidemiological methods. These designs and methods must be applied appropriately, creatively, and innovatively if they are to yield the desired information. The field of epidemiology has been expanding dramatically over the last three decades, as epidemiologists have demonstrated new uses and variations of traditional study designs and

methods (Pearce 2012). We can anticipate that the scope of epidemiology will expand even more in the future as increasing numbers of creative epidemiologists develop innovative new strategies and techniques.

The chapters in this section present detailed discussions of the principles and methods of epidemiology. In this introductory chapter, I will attempt to define epidemiology, to present ways in which epidemiology is used in the advancement of public health, and finally, to discuss the range of applications of epidemiological methodologies.

What is epidemiology?

There are many definitions of epidemiology, but every epidemiologist will know exactly what it is that he or she does. Defining epidemiology is difficult primarily because it does not represent a body of knowledge, as does, for example, anatomy, nor does it target a specific organ system, as does cardiology. Epidemiology is a scientific strategy/method for studying a health problem that can be applied to a wide range of problems, from transmission of an infectious disease agent to the design of a new strategy for healthcare delivery. Furthermore, this methodology is continually changing as it is adapted to a greater range of health problems and more techniques are borrowed and adapted from other disciplines such as mathematics and statistics.

Maxcy, one of the pioneer epidemiologists of the past century, offered the following definition:

> Epidemiology is that field of medical science which is concerned with the relationship of various factors and conditions which determine the frequencies and distributions of an infectious process, a disease, or a physiologic state in a human community. (Lilienfeld 1978, p. 88)

The word itself comes from the Greek *epi*, *demos*, and *logos*; literally translated, it means the study (*logos*) of what is upon (*epi*) the people (*demos*). John Last, in the *Dictionary of Epidemiology* (Last 2001), has defined epidemiology as follows:

> The study of the distribution and determinants of health-related states or events in specified populations, and the application of this study to the control of health problems.

Last's definition underscores that epidemiologists are concerned not only with disease but also with 'health-related events', and that ultimately epidemiology is committed to control of disease. All epidemiologists, however, will agree that epidemiology concerns itself with populations rather than individuals, thereby separating it from the rest of medicine and constituting the basic science of public health. Following from this, therefore, is the need to describe health and disease in terms of frequencies and distributions in the population. The epidemiologist relates these frequencies and distributions of specific health parameters to the frequencies of other factors to which populations are exposed in order to identify those that may be causes of ill health or promoters of good health. Inherent in the philosophy of epidemiology is the idea that ill health is not randomly distributed in populations, and that elucidating the reasons for this non-random distribution will provide clues regarding the risk factors for disease and the biological mechanisms that result in loss of health.

Because epidemiology usually focuses on health in *human* populations, it is rarely able to provide experimental proof in the sense of Koch's postulates, as can often be done in the laboratory sciences. Epidemiology more often provides an accumulation of increasingly convincing indirect evidence of a relationship between health or disease and other factors. This process, referred to as causal inference (see Chapter 5.14), includes considering an observed relationship in terms of its strength, consistency, specificity, temporality, biological gradient, plausibility, coherence, and experimental evidence (Hill 1965).

Although they will differ on the exact definitions of epidemiology, most epidemiologists will agree that they try to characterize the relationships among the *agent*, the *environment*, and the *host* (usually human). The epidemiologist considers health to represent a balance among these three forces, as shown in Figure 5.1.1.

Changes in any one of these three factors may result in loss of health. For example, the host may be compromised as a result of treatment with steroids, making him or her more susceptible to agents that do not ordinarily cause disease. On the other hand, a breakdown in the water supply system may result in an increased exposure of people to agents such as cryptosporidium, as happened in 1993 in Milwaukee, Wisconsin (MacKenzie et al. 1994).

Finally, some agents may become more or less virulent over time—often because of the promiscuous use of antibiotics—thereby disturbing the

dynamic balance among agent, host, and environment. Two examples are the cases of acute necrotizing fasciitis caused by group A streptococcus (Communicable Disease Surveillance Centre 1994) and the development of multidrug-resistant tuberculosis (Chapman and Henderson 1994).

The epidemiologist uses another triad to study the relationship of agent, host, and environment: *time–place–person*. Using various epidemiological techniques, described in subsequent chapters, the epidemiologist describes disease or disease factors occurring in the population in terms of the characteristics of time (e.g. trends, outbreaks, etc.), place (the geographic area in which the disease is occurring), and person (the characteristics of the affected individuals; e.g. age, gender, etc.) to elucidate the causative agent, the natural history of the disease, and the environmental factors that increase the likelihood of the host acquiring the disease. With this information, the epidemiologist is able to suggest ways to intervene in the disease process to prevent either disease or death.

Epidemiology has been described as the ‘art of the possible’. Because epidemiologists work with human populations, they are rarely able to manipulate events or the environment as can the laboratory scientist. They must, therefore, exploit situations as they exist naturally to advance knowledge. They must be both pragmatic and realistic; they must realize both the capabilities and limitations of the discipline. Morris has said that the ‘epidemiologic method is the only way to ask some questions … one way of asking others and no way at all to ask many’ (Morris 1975, p. 96). The art of epidemiology is to know both when epidemiology is the method of choice and when it is not, and how to use it to answer the question.

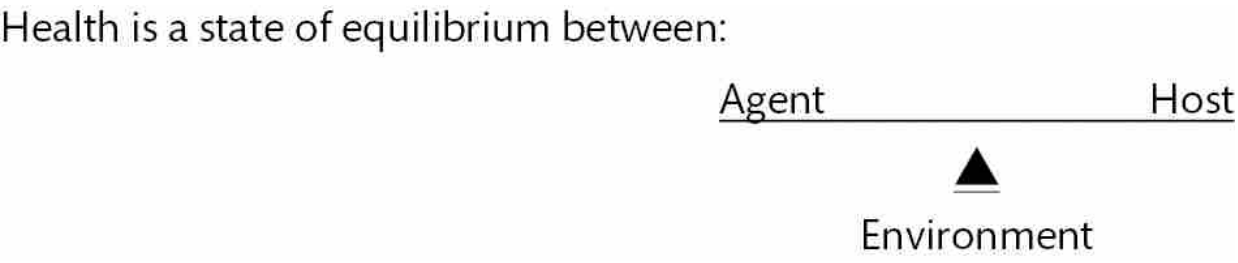


Fig. 5.1.1 The dynamic relationship between agent, host, and environment in epidemiology.

Applying the epidemiological method to resolve a health question successfully can be compared to constructing a memorable Chinese banquet. It is not enough to have the best ingredients and to know the various Chinese cooking methods. The truly great chefs must be able to select the appropriate ingredients and cooking methods to bring out the flavours of each dish, and further, must know how to construct the correct sequence of dishes to excite the palate without overwhelming it. They create a memorable banquet by adding their creative genius to the raw ingredients and the established cooking methods. Similarly, it is not enough for the epidemiologist to know the various strategies and methods of epidemiology; the innovative epidemiologist must be able to apply them creatively to obtain the information needed to understand the natural history of the disease. It is not enough to know what a cohort study is; the epidemiologist must know when the cohort design is the appropriate design for the question at hand, and then must apply that design appropriately and creatively. These skills make epidemiology more than a methodology. It is this opportunity for creativity and innovation that provides excitement for the practitioner and makes the successful practice of epidemiology an art.

For example, Imagawa et al. (1989) identified probable transient HIV-1 infection in men, implying clearance of the virus by the immune system of the men, by focusing their viral isolation studies on the relatively few HIV-1-antibody-negative homosexual men who had the highest numbers of sexual partners; a simple cohort study of antibody-negative individuals would have required a cohort of thousands of men rather than the 133 high-risk men studied. the effects of passive smoking were demonstrated by cohort studies of non-smoking family members of smokers and in nursing students by comparing the reported symptoms in room-mates of smokers and non-smokers who kept diaries of their symptoms: the room-mates of the smokers had a 1.8 greater risk of episodes of phlegm than room-mates of non-smokers (Colley et al. 1974; Tager et al. 1979; Schwartz and Zeger 1990). Tashkin et al. (1984) demonstrated that children of smokers had lower levels of lung function than children of non-smokers. All of these investigators used traditional study designs, but demonstrated their creativity by applying that design to those specific populations which were most likely to reveal a relationship if it existed.

Epidemiological studies rarely provide ‘proof’ of a causal relationship. Thus, there is continuing debate among epidemiologists about what

constitutes adequate criteria for inferring a causal relationship from epidemiological studies (Rothman 1988). Hill (1965) suggested the following minimum criteria for establishing a causal relationship between exposure and consequence/outcome: strength of association (statistical probability and risk ratio), consistency of findings across multiple studies, specificity of the relationship, temporality (outcome follows causation not heuristic), biological gradient (a dose–response relationship), plausibility, coherence (consistency with prior knowledge), experimental evidence, and analogy (relationship hypothesized is similar to that in known relationships). Susser has added to these criteria the ability of the observed relationship to correctly predict other relationships (Rothman 1988). Rothman and Greenland define a cause as 'an event, condition or characteristic that preceded the disease onset and that, if the event, condition or characteristic has been different in a specified way, the disease would not have occurred at all or would not have occurred until sometime later' (Rothman and Greenland 1998, p. 8). The sufficient component cause model defined a sufficient cause as 'a complete causal mechanism' that inevitably produces disease (Rothman and Greenland 1998, p. 8). A minimal set of factors results in an outcome, that is, each factor in a sufficient cause is a component cause and most 'causes' that are the 'causes of interest' in studies are component causes of a sufficient cause. For example, HIV infection can only occur following exposure to the HIV virus (sufficient cause). The exposure to the virus occurs through component causes (unprotected heterosexual intercourse, intravenous drug use). Therefore, a sufficient cause is a minimal set of factors that will cause the outcome. The debate goes on, but the principle is the same: epidemiological studies seldom provide 'proof' of a causal relationship in the sense of Koch's postulates, but may be used to reveal a possible relationship and build a convincing case that this relationship is causal. However, the presence of a relationship does not imply a causal relationship. The 'art' of causal inference assists in distinguishing causal from non-causal relationships (Aschengrau and Seage 2008).

More recently there is increased recognition that disease is not a product of individual circumstances at a given point of time, but rather a product of circumstances over a life course, possibly beginning *in utero* and continuing during a life course (Galea et al. 2010). This includes the complex interactions over time that result in a dynamic up- or down-regulation of

different factors that closely approximates the temporal nature of causality (Galea et al. 2010).

Uses of epidemiology in support of public health

Epidemiology is the basic science of public health because it is the health science that describes health and disease associations in *populations* rather than in individuals, information essential for the formulation of effective public health initiatives to prevent disease and promote health in the community. I have taken the liberty of updating the 'Functions of Epidemiology', first expounded by Morris (1957) and Holland et al. (2007). They are as follows:

Describe the spectrum of the disease: disease represents the end point of a process of alteration of the host's biological systems. Although many disease agents are limited in the range of alterations they can initiate, others, such as measles, can cause a variety of disease end points. For example, the majority of infections with rubeola (measles) virus result in the classical febrile, blotchy-rash disease, but the rubeola virus can also cause generalized haemorrhagic rash and acute encephalitis. Years after initial infection, rubeola can also cause subacute sclerosing panencephalitis (SSPE), a fatal disease of the central nervous system.

Various types of epidemiological studies have been used to elucidate the spectrum of disease resulting from many agents and conditions; for example, cohort studies have been used to document the role of high blood pressure as a major cause of stroke, myocardial infarct, and chronic kidney disease. For rare diseases such as SSPE and multiple sclerosis, case–control studies have been useful to identify the role of the rubeola virus (Detels et al. 1973; Alter 1976). Knowing the spectrum of disease that can result from specific infections and conditions allows the public health professional to design more effective intervention strategies; for example, education, screening, and treatment programmes to reduce the prevalence of high blood pressure will also reduce the incidence of myocardial infarct, stroke, and chronic kidney disease (Hypertension Detection and Follow-up Program Cooperative Group 1979).

Describe the natural history of disease: epidemiological studies can be used to describe the natural history of disease, to elucidate the specific alterations in the biological system in the host, and to improve diagnostic accuracy. For example, cohort studies of individuals who were infected with HIV, the 'AIDS virus', revealed that a drop in the level of T lymphocytes having the CD4 marker was associated with being infected with HIV (Polk et al. 1987), and that a further decline in CD4 cells was associated with developing clinical symptoms and AIDS (Detels et al. 1987). This observation stimulated immunologists to focus their research on the interaction of the immune system and HIV. From a clinical perspective, clinicians can target HIV-antibody-positive individuals who have declining CD4 cells for viral suppressive antiretroviral treatment.

Epidemiology can also be used to describe the impact of treatment on the natural history of disease. For example, a cohort study design was used to demonstrate the public health effectiveness of combined highly active

antiretroviral therapy on reducing the incidence of AIDS and extending survival of those who already had the disease (Detels et al. 1998). Thus, describing the natural history of AIDS among both treated and untreated individuals has assisted researchers to focus their studies and clinicians to use the treatment modalities available more effectively (Phair et al. 1992). The field of 'clinical epidemiology' applies research on the natural history of disease to improving the diagnostic accuracy of physicians in their clinical practice (Sackett et al. 1991).

Community diagnosis: epidemiological surveys are often used to establish the morbidity and mortality from specific diseases, allowing efficient use of limited public health funds for control of those diseases having the greatest negative impact on the health of the community. For example, an epidemiological survey in one area of China identified the epidemic of HIV due to unsafe plasma donations in villages (Wu and Detels 1995; Wu et al. 2001; Ji et al. 2006). The use of disability-adjusted life years (DALYs) has allowed quantification of the importance of non-lethal conditions on the public's health (Murray 1994).

Describe the clinical picture of a disease: epidemiological strategies can identify who is likely to get a disease such as capillariasis, the characteristic symptoms and signs, the extent of the epidemic, the risk factors, and the causative agent, and can help to determine the effectiveness of treatment and control efforts (Detels et al. 1969b).

Identify factors that increase or decrease the risk of acquiring disease: having specific characteristics increases the probability that individuals will or will not develop disease. These 'risk factors' may be social (smoking, drinking), genetic (ethnicity), dietary (saturated fats, vitamin deficiencies), and so on. Knowing these risk factors can often provide public health professionals with the necessary tools to design effective programmes to intervene before disease occurs. For example, descriptive, cross-sectional, case–control, cohort, and intervention studies have all shown that smoking is the biggest single risk factor for ill health, because it is a major risk factor for cardiovascular disease, chronic respiratory disease, and many cancers (e.g. of the lung, nasopharynx, and bladder). Thus, smoking is the leading cause of disability and death in developed countries, if not the world. Health education campaigns and other strategies to stop or reduce smoking, based on these epidemiological studies, are now a major public health activity in most countries of the world.

Identify precursors of disease and syndromes: high blood pressure, a treatable condition, has been identified through case–control and cohort studies as a precursor to heart disease, stroke, and kidney disease (Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure 1997).

Test the efficacy of intervention strategies: a primary objective of public health is to prevent disease through intervention in the disease process. But a vaccine or other intervention programme must be proven effective and safe before it is used in the community. Double-blind placebo-controlled trials are a necessary step in developing an effective intervention programme, whether that programme is administration of a new vaccine, a behavioural intervention strategy to stop smoking, or a community intervention study to lower heart disease. Although it may be argued that injection of a saline placebo is no longer considered ethical, a proven vaccine, such as polio, can often be used as a placebo for a trial of a new vaccine for a different disease, as was used for trials of rubella vaccines in Taiwan (Detels et al. 1969a). Widespread use of an intervention not subjected to epidemiological studies of efficacy may result in implementation of an ineffective intervention programme at great public expense, and may actually result in greater morbidity and mortality because of an increased reliance on the favoured but unproven intervention and a reduced use of other strategies that are thought to be less effective but which are actually more effective.

Although an intervention such as a vaccine may have been demonstrated to have efficacy in double-blind trials, it may fail to provide protection when used in the community. Double-blind trials may demonstrate the ‘biological efficacy’ of the vaccine; but if the vaccine is not acceptable to the majority of the public, they will refuse to be vaccinated, and the ‘public health effectiveness’ of the vaccine will be very low. For example, the early typhoid vaccine provided some protection against small infecting inocula, but the frequency of unpleasant side effects with the whole-cell vaccines and the need for multiple injections in the past influenced many people against being vaccinated (Chin 2000).

Another problem of inferring public health efficacy from small vaccine trials is that volunteers for these trials may not be representative of the general public, which needs to be protected against a specific disease. Thus, broad-based intervention trials also need to be carried out, to demonstrate the

acceptability and public health effectiveness of a vaccine or other intervention to the entire population in need of protection.

As there are adverse effects associated with any vaccine, ongoing evaluations of the cost–benefit relationship of specific vaccines are important. By comparing the incidence of smallpox with the incidence of adverse effects from the smallpox vaccine, Lane et al. (1969) demonstrated that more disease resulted from routine use of the vaccine in the United States than by transmission from imported cases.

There are several epidemiological strategies that can be used for ongoing evaluation of intervention programmes. Serial cross-sectional studies can be used to determine if there has been a change in the prevalence of disease or of indicators of health status over time. The cohort design can be used to compare incidence of disease in comparable populations receiving and not receiving the prevention programme. The case–control design can be used to determine if there are differences in the proportion of cases and non-cases who received the intervention programme.

Investigate epidemics of unknown aetiology: epidemiological strategies were used to establish the extent, cause, modes of transmission, and risk factors for Ebola haemorrhagic fever, which first occurred in the Congo in 1976 (Feldmann et al. 1996; King 2013), and capillariasis in the Philippines in the mid-1960s (Detels et al. 1969b).

Evaluate public health programmes: departments of health are engaged in a variety of activities to promote the health of the community, ranging from surveillance and vaccination programmes to clinics for the treatment of specific diseases. Ongoing evaluation of such programmes is necessary to ensure that they continue to be cost-effective. Periodic review of routinely collected health statistics can provide information about the effectiveness of many programmes. For programmes for which relevant statistics are not routinely available, cohort studies and serial cross-sectional studies of the incidence and changing prevalence of the targeted disease in the populations which are the intended recipients of these programmes can measure whether the programmes have had an impact and are cost-effective. For example, most countries have established sexually transmitted infections (STI) clinics, but studies in Thailand (Prempree et al. 2007) and Beijing (Zhao et al. 2007) have demonstrated that the majority of the patients with STIs do not attend the government STI clinics.

Elucidate mechanisms of disease transmission: understanding the mechanisms of disease transmission can suggest ways in which public health professionals can protect the public by stopping transmission of the disease agent. Epidemiological studies of the various arboviral encephalitides have incriminated certain species of mosquitoes as the vectors of disease and specific animals as the reservoirs for the viruses. For example, public health efforts in California to prevent western equine encephalitis have concentrated on control of the mosquito vector and vaccination of horses, which are a reservoir of the virus. Although an effective vaccine for smallpox had been available for almost 200 years, eradication of the disease was not achieved until the recognition that the low infectivity of varicella virus and the relatively long incubation for development of smallpox could be used to develop a strategy of surveillance for cases, with identification and immediate vaccination of all susceptible contacts (containment). Using this containment strategy based on epidemiological principles, smallpox was eradicated through a worldwide effort in less than 10 years (Fenner et al. 1988).

Elucidate the molecular and genetic determinants of disease progression: epidemiological strategies helped to elucidate the changes in the human immune response by CD4 and CD8 cells that accompany infection and disease progression by disease agents (Detels et al. 1983; Fahey et al. 1984; Ho et al. 1995), and genetic factors (such as CCR5 absence or heterozygosity) that prevent HIV infection and slow progression of HIV disease (Liu et al. 2004).

Implementation science: recently the scientific community and policymakers have recognized the slow rate at which scientific advances are translated into public health policies and medical and public health practice. Thus, there is growing support for identifying and testing strategies that will speed up the process of translating science into action and practice. Epidemiological methods will play an important role in identifying and testing these strategies.

From the examples given in the preceding list, it should be clear that epidemiology functions as the backbone or core of evidence-based public health practices, as well as a key strategy for evaluating the effectiveness of both clinical and public health interventions. As such it is a particularly pertinent instrument for evidence-based policy formulation as well as a tool

for building advocacy and resource justification for rational policies and programmes.

Applications of epidemiology

Specific epidemiological study designs are used to achieve specific public health goals. These goals range from identifying a suspected exposure–disease relationship to establishing that relationship, to designing an intervention to prevent it, and finally, to assessing the effectiveness of that intervention. The usual sequence of study designs in the identification and resolution of a disease problem is as follows:

Ecological studies
Cross-sectional (prevalence) surveys
Case–control studies
Cohort studies
Experimental studies
Adaptive trial design.

There are, however, many exceptions to the application of this sequence of study designs, depending on such factors as the prevalence and virulence of the agent and the nature of the human response to the agent.

The earliest suspicion that a relationship exists between a disease and a possible causative factor is frequently obtained by observing correlations between exposure and disease from existing data such as mortality statistics and surveys of personal or national characteristics. These can be correlations observed across geographical areas (ecological studies) or over time, or a combination of both. Many of the initial epidemiological investigations on chronic bronchitis used vital statistics data, particularly data on mortality. Case–control studies identified smoking as a possible causal factor for chronic bronchitis. Subsequent cohort studies confirmed the relationship. A decline in respiratory symptoms of chronic bronchitis and a concurrent but slower decline in lung function have been observed in individuals who cease smoking (Colley 1991). Using the information obtained from these studies, public health has implemented vigorous antismoking campaigns and has used epidemiology to assess their effectiveness.

Although this is the usual sequence in which the various epidemiological study designs are applied, there are exceptions to this sequence. Furthermore, all study designs are not appropriate to answer all health questions. The usual applications of each of the epidemiological study designs and their limitations are, therefore, presented briefly in this subsection and in greater depth in subsequent chapters.

Ecological studies

The use of existing statistics to correlate the prevalence or incidence of disease in groups or populations to the frequency or trends over time of suspected causal factors in specific localities has often provided the first clues that a particular factor may cause a specific disease. These epidemiological strategies, however, document only the co-occurrence of disease and other factors in a population; the risk factors and the disease may not be occurring in the same people within the population. These types of descriptive studies are inexpensive and relatively easy to perform, but the co-occurrence observed may be merely due to chance. For example, the incidence of both heart disease and lung cancer has increased concurrently with the increased use of automatic washing machines in the United States. Few people, however, would attribute the increase in these two diseases to the use of automatic washing machines. Thus, ecological studies must be interpreted with caution. Nonetheless, they often reveal important relationships and can provide a strong rationale for undertaking more expensive analytic studies (see Chapter 5.2).

Cross-sectional/prevalence surveys

Cross-sectional/prevalence surveys establish the frequency of disease and other factors in a community. Because they require the collection of data, however, they can be expensive. They are useful to estimate the number of people in a population who have disease and can also identify the difference in frequency of disease in different subpopulations. This descriptive information is particularly useful to health administrators who are responsible for developing appropriate and effective public health programmes.

Cross-sectional studies can also be used to document the co-occurrence of disease and suspected risk factors not only in the population but also in specific individuals within the population. The cross-sectional study design is useful to study chronic diseases such as multiple sclerosis, which has a reasonably high prevalence, but an incidence that is too low to make a cohort study feasible (Detels et al. 1978). On the other hand, they are not useful for studying diseases that have a very low prevalence, such as SSPE or variant Creutzfeldt–Jakob disease. Cross-sectional studies are subject to problems of respondent bias, recall bias, and undocumented confounders. Further, unless historical information is obtained from all the individuals surveyed, the time-relationship between the factor and the disease is not known. Also, prevalence surveys identify people who have survived to that time point with disease, and thus, under-represent people with a short course of disease.

The cross-sectional study design is used in two special types of studies: field studies and surveillance. Field studies are usually investigations of acute outbreaks, which require immediate identification of the causative factors if effective public health interventions are to be implemented in a timely fashion. Surveillance is the monitoring of disease- or health-related factors over time and uses serial cross-sectional surveys to observe trends in populations. Surveillance is important to identify diseases that are becoming an increasing public health problem, to assure that diseases already brought under control remain under control, and to evaluate the impact of public health intervention strategies (see Chapters 5.3, 5.4, and 5.17).

Case–control studies

The case–control study compares the prevalence of suspected causal factors between individuals with disease and controls who do not have the disease. If the prevalence of the factor is significantly different in cases than it is in controls, this factor may be associated with the disease. Although case–control studies can identify associations, they do not measure risk. An estimate of relative risk, however, can be derived by calculating the odds ratio. Case–control studies are often the analytic study design used initially to investigate a suspected association. Compared to cohort and experimental studies, they are usually relatively cheap and easy to perform. Cases can often be selected from hospital patients and controls either from hospitalized patients with other diseases or by using algorithms or formulas for selecting

community (neighbourhood) or other types of controls. Selection bias, however, is often a problem, especially when using either hospitalized cases or controls. The participants are seen only once, and no follow-up is necessary. Although time sequences can often be established retrospectively for factors elicited by interview, they usually cannot be for laboratory test results. For example, an elevation in factor B may either be causally related, or it may be a result of the disease process and not a cause. Furthermore, factors elicited from interview are subject to recall bias; for example, patients are often better motivated to recall events than controls because they are concerned about their disease.

The case–control study is particularly useful for exploring relationships noted in observational studies. A hypothesis, however, is necessary for case–control studies. Relationships will be observed only for those factors studied. Therefore, case–control studies are not useful for determining the spectrum of health outcomes resulting from specific exposures, because a definition of a case is required in order to do a case–control study. On the other hand, these studies are the method of choice for studying rare diseases, and are often indicated when a specific health question needs to be answered quickly (see Chapter 5.5).

Cohort studies

Cohort studies follow defined groups of people without disease to identify risk factors associated with disease occurrence. They have the advantage of establishing the temporal relationship between an exposure and a health outcome, and thus measure risk directly. Because the population studied is often defined on the basis of its known likelihood of exposure to suspected factors, cohort studies are particularly suitable for investigating health hazards associated with environmental or occupational exposures. Further, these studies will measure more than one outcome of a given exposure, and therefore, are useful for defining the spectrum of disease resulting from exposure to a given factor. Occasionally, a cohort study is done to elucidate the natural history of a disease when a group that has a high incidence of disease, but in which specific risk factors not previously known, can be identified. Although this cohort is not defined based on a known exposure, questions are asked, and biological specimens are collected from which exposure variables can be identified concurrently or in the future.

Unfortunately, cohort studies are both expensive and time consuming. Unless the investigator can define a cohort in which risk factors were measured at some time in the past and has the assurance that the cohort has been completely followed up for disease outcomes in the interim (historical cohort), this design can take years to decades to yield information about the risks of disease resulting from exposure to specific factors. Ensuring that participants remain in a cohort study for such long periods of time is both difficult and expensive. Further, the impact of those who drop out of follow-up must be taken into account in the analysis and interpretation of these studies. Finally, exposures may vary over time, complicating the analysis of their impact.

Because of the cost and complexity of cohort studies, they are usually done only after descriptive, cross-sectional, and/or case–control studies have suggested a causal relationship. The size of the cohort to be studied is dependent, in part, on the anticipated incidence of the disease resulting from the exposure. For diseases with a very low incidence, population-based cohort studies are usually not feasible, either in terms of the logistics or of the expense of following very large numbers of people, or both. Cohort studies establish the risk of disease associated with exposure to a factor, but do not 'prove' that the factor is causal; the observed factor may merely be very closely correlated with the real causative factor or may even be related to the participants' choice to be exposed.

A variant of the cohort study that has become popular is the 'nested case–control study' (Gange et al. 1997). Cases arising from a cohort study are compared with individuals followed in the cohort who have not developed disease using the usual case–control analytic strategies. The advantage of this type of study is that the exposure variables are collected before knowledge of the outcome, and therefore, are unlikely to be tainted by recall bias (see Chapter 5.6).

Experimental studies

Experimental studies differ from cohort studies because it is the investigator who makes the decision about who will be exposed to the factor based on the specific design factors to be employed (e.g. randomization, matching, etc.). Therefore, confounding factors such as choice that may have led to the subjects being exposed in the cohort studies are usually not a problem in

experimental studies. Because epidemiologists usually study human populations, there are few opportunities for an investigator to deliberately expose participants to a suspected factor. On the other hand, intervention studies of randomly assigned individuals (see Chapter 5.7) or communities (see Chapters 5.8–5.10) to receive or not receive an intervention programme that demonstrates a subsequent reduction in a specific health outcome in the intervention group do provide strong evidence, if not proof, of a causal relationship. Because of the serious implications of applying an intervention that may alter the biological status of an individual or the sociopolitical behaviour of a community, intervention studies should not be undertaken until the probability of a causal or risk relationship has been well established using the other types of study designs. A randomized clinical trial or vaccine trial is a typical example of an experimental study (see Chapters 5.7–5.10).

Meta-analysis

Because individual epidemiological studies rarely provide proof of causation and results of different studies can vary for a number of reasons, including small sample size, a recent trend has been to combine similar studies to increase the power of the analysis. This strategy for data synthesis is known as *meta-analysis*. It has been especially helpful in studying diseases with a low incidence or where similar studies have given conflicting results. But meta-analysis can yield opposite conclusions, especially if different sets of studies are used in the analysis. Meta-analysis investigates and identifies all available studies both quantitatively and qualitatively. The qualitative analysis excludes poor-quality studies that may be considered flawed by the investigators (Aschengrau and Seage 2008). Meta-analysis is prone to heterogeneity problems as different case definitions and different sources of controls cannot be adjusted for in the analysis (Rothman et al. 2008) (see Chapter 5.15).

Adaptive trial designs

Adaptive trial designs provide a way to address uncertainty about choices made during the planning of a study and allow for planned modifications after initiation of the study at pre-specified time points or end points as data

is accumulated within a study (Kairalla et al. 2012). This design provides flexibility and efficiency (smaller sample size) of study designs and increases the likelihood of a 'successful' trial that will yield the outcome of interest. However, improper adaptations can result in study bias. This type of study design can be applied in comparative effectiveness studies (Kairalla et al. 2012).

Methodological issues

Epidemiological studies, because they deal with humans, are subject to problems such as bias (deviation of results from truth), due to the strategies of recruiting participants or to differential recall among persons with and without disease, and confounding, due to factors which are associated with both the exposure variables and outcome variable under study. A cause may combine with another cause to produce an outcome (Susser et al. 2006). For example, a genetic abnormality can be modified by prenatal exposures, resulting in an adverse outcome (e.g. autism) or methylation of key genes. This process is called interaction or effect modification and attempts are made to capture these with statistical measures (Susser et al. 2006). In the last several decades, many new techniques have been developed to reduce the effect of these factors, which can influence the outcome of a study and, in some instances, can cause apparent relationships to be observed which are in fact false (see Chapters 5.12 and 5.13). Because epidemiologic studies rarely meet the criteria of Koch's postulates, care must be taken to infer causality from the results of a study. Recognizing this problem, the field of epidemiology has expanded to develop strategies, methodologies, and guidelines to improve the ability to make causal inferences from epidemiologic studies (Pearl 2016).

Summary

Epidemiology is the core science of public health because it defines health and disease in human populations, describes disease aetiology, and evaluates public health control efforts. It achieves these goals through a variety of strategies and methods. Epidemiology is a dynamic science that is

continually evolving new strategies and methods in support of public health goals.

REFERENCES

Alter, M. (1976). Is multiple sclerosis an age-dependent host response to measles? *The Lancet*, 1(7957), 456–7.

Aschengrau, A. and Seage, G.R. (2008). *Essentials of Epidemiology in Public Health*. Sudbury, MA: Jones and Bartlett.

Chapman, S.W. and Henderson, H.M. (1994). New and emerging pathogens multiply resistant *Mycobacterium tuberculosis*. *Current Opinion in Infectious Diseases*, 7, 231–7.

Chin, J. (2000). *Control of Communicable Diseases Manual*. Washington, DC: American Public Health Association.

Colley, J.R.T. (1991). Major public health problems: respiratory system. In: Holland, W.W., Detels, R., and Know, G. (eds.) *Oxford Textbook of Public Health* (2nd ed.). Oxford: Oxford University Press, pp. 227–48.

Colley, J.R.T., et al. (1974). Influence of passive smoking and parental phlegm on pneumonia and bronchitis in early childhood. *The Lancet*, 2(7888), 1031–4.

Communicable Disease Surveillance Centre (1994). Invasive group A streptococcal infections in Gloucestershire (England/Wales). *Communicable Disease Report*, 4, 97–100.

Detels, R., et al. (1969a). Prevention of clinical and subclinical rubella infection: efficacy of three HPV-77 derivative vaccines. *American Journal of Diseases of Children*, 118, 295–300.

Detels, R., et al. (1969b). An epidemic of intestinal capillariasis in man: a study in a barrio in northern Luzon. *American Journal of Tropical Medicine and Hygiene*, 18(5), 676–82.

Detels, R., et al. (1973). Further epidemiological studies of subacute sclerosing panencephalitis. *The Lancet*, 819, 11–14.

Detels, R., et al. (1978). Multiple sclerosis and age at migration. *American Journal of Epidemiology*, 108, 386–93.

Detels, R., et al. (1983). Relation between sexual practices and T-cell subsets in homosexually active men. *The Lancet*, 1(8325), 609–11.

Detels, R., et al. (1987). Predictors of clinical AIDS in young homosexual men in a high-risk area. *International Journal of Epidemiology*, 16, 271–6.

Detels, R., et al. (1998). Effectiveness of potent antiretroviral therapy on time to AIDS and death in men with known HIV infection duration. *Journal of the American Medical Association*, 280(17), 1497–503.

Fahey, J., et al. (1984). Quantitative changes in T-helper or T-suppressor/cytotoxic lymphocyte subsets that distinguish acquired immune deficiency syndrome from other immune subset disorders. *American Journal of Medicine*, 76, 95–100.

Feldmann, H., Slenczka, W., and Klenk, H.D. (1996). Emerging and reemerging of filoviruses. *Archives of Virology*, 11(Suppl.), 77–100.

Fenner, F., et al. (1988). *Smallpox and its Eradication*. Geneva: World Health Organization.

Galea, S., Riddle, M., and Kaplan, G.A. (2010). Causal thinking and complex system approached in epidemiology. *Internal Journal of Epidemiology*, 39, 97–106.

Gange, S., et al. (1997). Design of nested studies to identify factors related to late progression of HIV infection. *Journal of Acquired Immune Deficiency Syndromes and Human Retroviruses*, 15(Suppl.), S5–S9.

Hill, A.B. (1965). The environment and disease: association or causation? *Proceedings of the Royal Society of Medicine*, 58, 295–300.

Ho, D.D., et al. (1995). Rapid turnover of plasma virion and CD4 lymphocytes in HIV-1 infection. *Nature*, 373, 123–6.
Holland, W.W., et al. (eds.) (2007). *The Development of Modern Epidemiology: Personal Reports from Those Who Were There*. Oxford: Oxford University Press.
Hypertension Detection and Follow-up Program Cooperative Group (1979). Five-year findings of the hypertension detection and follow-up program. I: reduction in mortality of persons with high blood pressure, including mild hypertension. *Journal of the American Medical Association*, 242, 2562–71.
Imagawa, D.T., et al. (1989). Human immunodeficiency virus type 1 infection in homosexual men who remain seronegative for prolonged periods. *The New England Journal of Medicine*, 320, 1458–62.
Ji, G., et al. (2006). Correlates of HIV infection among former blood/plasma donors in rural China. *AIDS*, 20(4), 585–91.
Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure (1997). The sixth report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure. *Archives of Internal Medicine*, 157, 2413–46.
Kairalla, J.A., Coffey, C.S., Thomann, M.A., and Muller, K.E. (2012). Adaptive trial designs: a review of barriers and opportunities. *Trials*, 13, 145.
King, J.W. (2013). *Ebola Virus Infection*. Emedicine.com, Inc. [Online] Available at: http://www.emedicine.com/MED/topic626.htm
Lane, J.M., et al. (1969). Complications of smallpox vaccination, 1968: national surveillance in the United States. *The New England Journal of Medicine*, 281, 1201–8.
Last, J. (ed.) (2001). *Dictionary of Epidemiology* (4th ed.). New York: Oxford University Press.
Lilienfeld, D.E. (1978). Definitions of epidemiology. *American Journal of Epidemiology*, 107, 87–90.
Liu, C., et al. (2004). Lack of associations between HLA class II alleles and resistance to HIV-1 infection among white, non-Hispanic homosexual men. *Journal of Acquired Immune Deficiency Syndromes*, 37(2), 1313–17.
MacKenzie, W.R., et al. (1994). A massive outbreak in Milwaukee of cryptosporidium infection transmitted through the public water supply. *The New England Journal of Medicine*, 331, 161–7.
Morris, J.N. (1957). *Uses of Epidemiology*. London: Churchill Livingstone.
Morris, J.N. (1975). *Uses of Epidemiology* (3rd ed.). London: Churchill Livingstone.
Murray, C.J. (1994). Quantifying the burden of disease: the technical basis for disability-adjusted life years. *Bulletin of the World Health Organization*, 72, 429–45.
Pearce, N. (2012). Classification of epidemiologic study designs. *International Journal of Epidemiology*, 41, 393–7.
Pearl, J. (2016). *Causal Inference in Statistics: A Primer* (1st ed.). West Sussex England: John Wiley and Sons.
Phair, J., et al. (1992). Acquired immune deficiency syndrome occurring within 5 years of infection with human immunodeficiency virus type-1: the Multicenter AIDS Cohort Study. *Journal of Acquired Immune Deficiency Syndromes*, 5, 490–6.
Polk, B.F., et al. (1987). Predictors of the acquired immunodeficiency syndrome developing in a cohort of seropositive homosexual men. *New England Journal of Medicine*, 316, 61–6.
Prempree, P., et al. (2007). The sources of treatment of sexually transmissible infections in a rural community in central Thailand. *Sex Health*, 4(1), 17–19.
Rothman, K.J. (ed.) (1988). *Causal Inference.* Chestnut Hill, MA: Epidemiology Resources.
Rothman, K.J. and Greenland, S. (1998). *Modern Epidemiology* (2nd ed.). Philadelphia, PA: Lippincott-Raven.
Rothman, K.J., Greenland, S., and Lash, T.L. (eds.) (2008). *Modern Epidemiology* (3rd ed.). Philadelphia, PA: Lippincott Williams & Wilkins.

Sackett, D.L., et al. (1991). *Clinical Epidemiology: A Basic Science for Clinical Medicine*. London: Little, Brown.

Schwartz, J. and Zeger, S. (1990). Passive smoking, air pollution, and acute respiratory symptoms in a diary of student nurses. *American Review of Respiratory Disease*, 141, 62–7.

Susser, E., Schwartz, S., Morabia, A., and Bromet, E.J. (2006). *Psychiatric Epidemiology*. New York: Oxford University Press.

Tager, I.B., et al. (1979). Effect of parental cigarette smoking on the pulmonary function of children. *American Journal of Epidemiology*, 110, 15–26.

Tashkin, D.P., et al. (1984). The UCLA population studies of chronic obstructive respiratory disease: VII. Relationship between parental smoking and children's lung function. *American Review of Respiratory Disease*, 129, 891–7.

Wu, Z. and Detels, R. (1995). HIV-1 infection in commercial plasma donors in China. *The Lancet*, 346, 61–2.

Wu, Z., et al. (2001). Prevalence of HIV infection among former commercial plasma donors in rural Eastern China. *Health Policy and Planning*, 16(2), 41–6.

Zhao, G., et al. (2007). The distribution of people seeking STD services in the various types of healthcare facilities in Chao Yang District, Beijing, China. *Sexually Transmitted Diseases*, 35, 65–7.

5.2

Cross-sectional studies

Manolis Kogevinas and Leda Chatzi

Introduction to cross-sectional studies

Cross-sectional studies examine the relationship between diseases (or other health-related characteristics) and other variables of interest as they exist in a defined population at a particular point in time. They could be defined as studies taking a snapshot of a society. Synonyms used for cross-sectional studies include prevalence studies and disease frequency studies.

The principal characteristic of cross-sectional studies is that they provide information on the *prevalence* of disease, that is, they include prevalent cases. In cross-sectional studies, exposure and disease are measured at the same point in time but this characteristic is shared by other epidemiological designs, for example, case–control studies. In many cross-sectional studies information on past exposures is not collected but this should not be regarded as a characteristic defining these studies. The outcome measured in cross-sectional studies can be a continuous variable such as blood pressure or forced expiratory volume in 1 second (FEV_1), as compared to a dichotomous outcome measured in case–control studies and, in most occasions, in cohort studies.

As is frequently the case in epidemiology, studies may use mixed designs. For example, a cross-sectional study may measure a biomarker referring to current exposure (e.g. vitamin E), may request information for the past (e.g. use of health services the last year), may identify older and recent cases (e.g. subjects having had asthma in childhood or those having their first attack of asthma in the last few months), and may convert into a cohort study if subjects included in the cross-sectional studies are followed up. The statistical analysis of cross-sectional studies depends on their hybrid design and is frequently similar to that of a case–control study using logistic

regression and calculating (prevalence) odds ratios. Cross-sectional studies are extensively used to measure the prevalence of disease, exposures, or other health-related variables in a population. In these cases, the sample must be representative of the total population.

In this chapter we will first describe the uses of cross-sectional studies in epidemiological and public health research and then discuss methodological issues concerning the design, the main biases of these studies including response rates, and how to improve participation. We will finally discuss issues related to the statistical analysis of cross-sectional studies. Many of these issues are also relevant to other epidemiological designs.

Uses of cross-sectional studies

Cross-sectional studies have been used to evaluate the prevalence of diseases and of health-related variables and also to evaluate the aetiology of diseases. Large, multipurpose, national cross-sectional studies are conducted in several countries for administrative reasons and to provide background documentation on the health status of a population.

National multipurpose surveys

National (or regional) multipurpose surveys are carried out routinely or ad hoc in several countries. The purpose and themes of these surveys are multiple and they may cover different health problems in successive years. The Health Survey in England collects data from a representative sample of approximately 8,000 adults aged 16 and over and of 2,000 children in England (Department of Health 2012). Each year the survey consists of a series of core elements and also includes special topics. In 2012 the core survey covered general health, smoking, drinking, fruit and vegetable consumption, height, weight, blood pressure measurements, and collected blood and saliva samples. Special topics included were cardiovascular disease, hypertension, diabetes, social care, drinking, chronic pain, adult and child obesity, and information on well-being and dental health.

The National Health and Nutrition Examination Survey (NHANES) in the United States is the largest national multipurpose survey and is conducted on a routine basis (Centers for Disease Control and Prevention 2014). The NHANES is a programme of studies designed to assess the health and

nutritional status of the civilian, non-institutionalized population of the United States. The survey examines a nationally representative sample of about 5,000 persons each year and combines interviews and physical examinations.

Results of national health surveys can be used for several purposes, such as the evaluation of the prevalence of health-related behaviours or of symptoms, time trends in risk factors or health symptoms, the identification of causes of disease, and evaluation of health needs. Changes in health problems and risk factors in the population over time can be identified through repeated surveys. Repeated national multipurpose surveys may examine a sample of the same population, but not intentionally the same individuals. Thus, they are particularly useful for assessing changes in the same community, state, or nationally which may have undergone changes due to other factors, such as access to healthcare, population mixes and migration into and out of the community. This type of information may allow health planning through the identification of needs and effectiveness of existing policies and interventions. Time trends in obesity and in levels of toxic substances such as lead in blood are examples of the use of these types of surveys. The prevalence of overweight and obesity has increased in recent years in both children and adults in many countries. Time trends and the effect of socioeconomic status on childhood obesity in the United States were examined using data from two large, nationally representative health surveys and data systems: the 1976–2008 NHANES and the 2003 and 2007 National Survey of Children's Health (NSCH) (Singh and Kogan 2010). The prevalence of overweight and obese children increased significantly in relation to decreased levels of household education and income in both 2003 and 2007 (Fig. 5.2.1). The prevalence of obesity for children with parents having fewer than 12 years of education was 30% in 2007 compared to around 10% for children whose parents had a college degree. In addition the socioeconomic health disparities became wider in the period examined (Fig. 5.2.1).

The existence of a large national database has several advantages for the identification or confirmation of risk factors of a disease. Problems of cross-sectional studies that evaluate risk factors of diseases are discussed in the next section and particularly potential biases from the inclusion of prevalent cases and from frequent lack of information on past exposures. However, the evaluation of a national representative population sample, as is the case in

several multipurpose surveys, has the advantage of capturing a global image of the society including population groups that may be omitted in other types of design. An example provided here is of a study of cleaners based on NHANES. The importance of this study lies, among others, in the fact that domestic cleaners are a population group that is very hard to identify through the usual design strategies.

Information on the population distribution of health problems and risk factors, provide important clues to researchers on the causes of disease and also may confirm previous hypotheses in large population samples. Arif and colleagues used data from the Third National Health and Nutrition Examination Survey (NHANES III) 1988–1994, to evaluate associations between occupation and work-related asthma and work-related wheezing among US workers (Arif et al. 2003). They identified several occupations that were at risk of developing work-related asthma with domestic cleaners and equipment cleaners showing the highest risks. The population attributable risk for work-related asthma was 26%. The study confirmed a previous hypothesis on asthma in cleaners (Kogevinas et al. 1999), provided an evaluation of asthma risk in other occupations in the US population, and also gave an estimate of the burden of occupational asthma.

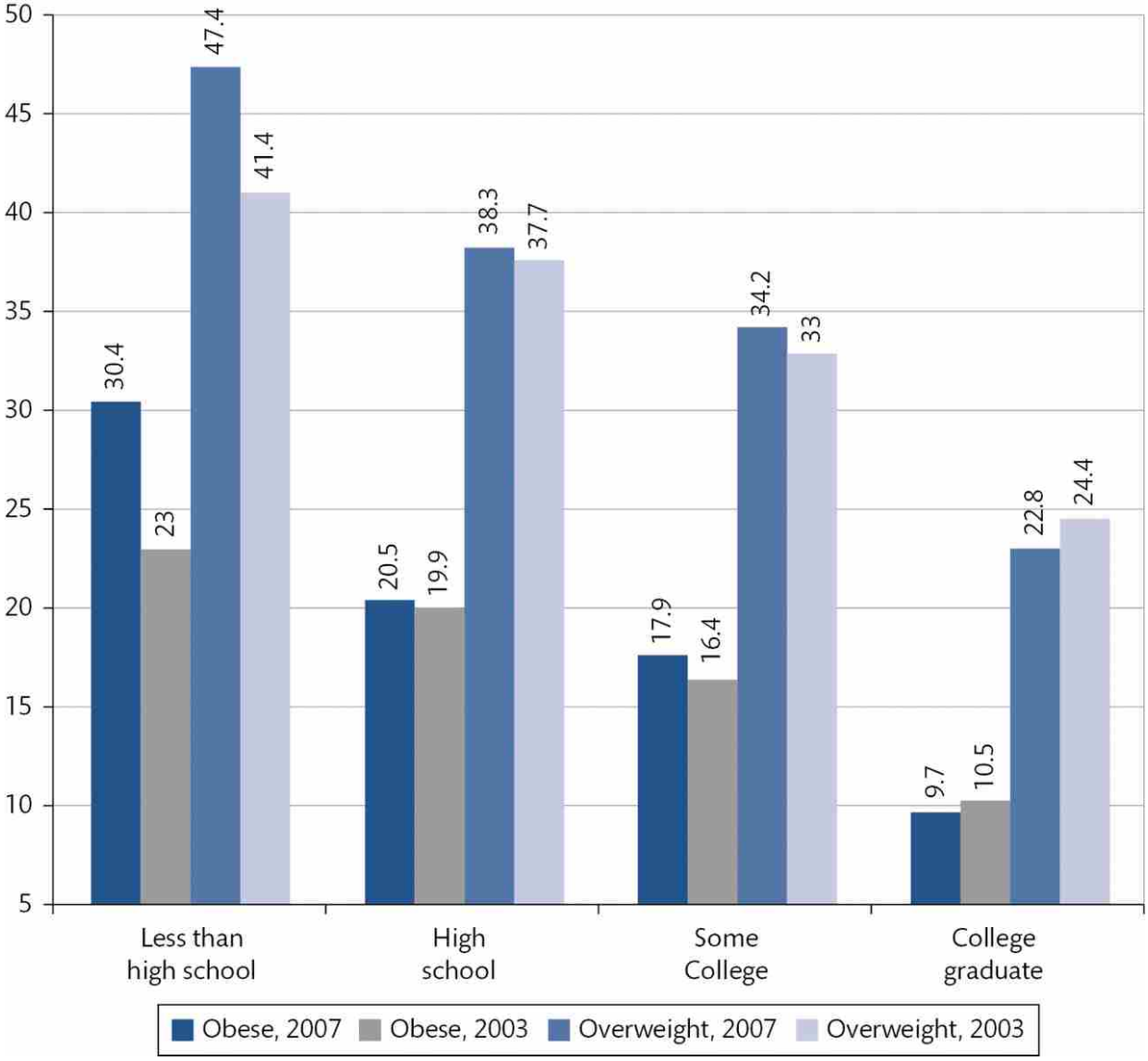


Fig. 5.2.1 Trends in the prevalence of overweight and obese children aged 10–17 years by parental education, United States, 2003–2007.

Reproduced from Singh, G.K. and Kogan, M.D., *Childhood Obesity in the United States, 1976–2008: Trends and Current Racial/Ethnic, Socioeconomic, and Geographic Disparities: A 75th Anniversary Publication*, Health Resources and Services Administration, Maternal and Child Health Bureau, Department of Health and Human Services, Rockville, Maryland, USA, 2010, available at: http://www.hrsa.gov/healthit/images/mchb_obesity_pub.pdf.

By identifying the healthcare needs of the population, government agencies and private sector organizations can establish policies and plan research, education, and health promotion programmes that will help improve present health status and prevent future health problems.

Studies examining the prevalence of disease

Cross-sectional studies are, by definition, the appropriate design to evaluate the prevalence of diseases. An evaluation of the prevalence of a disease may be performed for administrative purposes or for reasons related to the evaluation of aetiological factors. The main issues regarding the conduct of these studies include the representativeness of the sample, the size of the sample, and the measurement of outcomes and exposures. The International Study of Asthma and Allergies in Childhood (ISAAC) is among the largest cross-sectional studies ever done.

The ISAAC study was designed to allow comparisons of the prevalence of allergic disorders in childhood between populations in different countries and their trends over time (Asher et al. 1995). In the early 1990s although it was generally perceived that asthma prevalence was increasing there existed very limited population-based estimates. The study included simple 'core' instruments for measuring the prevalence of allergic disorders, suitable for different geographical locations and languages: (1) written questionnaires on the prevalence and severity of asthma, rhinitis, and eczema for self-completion in 13–14-year-olds, or for completion by parents of 6–7-year-olds; and (2) video questionnaires on the prevalence and severity of asthma for self-completion by 13–14-year-old children. In phase I, children aged 13–14 years were studied in 155 centres in 56 countries (n = 463,801) and children aged 6–7 years were studied in 91 centres in 38 countries (n = 257,800) (Anonymous 1998). Up to 20-fold variations in the prevalence of 'current wheeze' (in the last 12 months) were observed between centres worldwide (range 1.8–36.7%), with a sevenfold variation observed between the tenth and ninetieth percentiles (4.4%; 30.9%). The highest 12-month period prevalences were from centres in the United Kingdom, Australia, New Zealand, and the Republic of Ireland; the lowest prevalences were from centres in Eastern Europe, Albania, Greece, China, Taiwan, Uzbekistan, India, Indonesia, and Ethiopia.

Aetiological research

Cross-sectional studies are particularly valuable for the investigation of the aetiology of non-fatal diseases, degenerative diseases, diseases with no clear point of onset, for example, chronic bronchitis, and effects on physiological variables, as well as for low forced vital capacity (FVC) or elevated liver

enzyme levels. In diseases with no clear point of onset it is difficult to identify incident cases and conduct cohort or incident case–control studies. The prevalence of a disease will be highest if it occurs relatively frequently and if the disease persists in time. A disease could be described as persistent if it is not rapidly fatal and if it does not usually manifest as a transitory condition. Diseases such as asthma and osteoarthritis that have a relatively high prevalence have frequently been examined in cross-sectional studies.

The European Community Respiratory Health Survey (ECRHS) is in many aspects similar in design to ISAAC and was initially conducted as an international cross-sectional study enrolling young adults mainly from European centres to evaluate the prevalence and aetiology of asthma and related diseases (Burney et al. 1994). In an initial screening short questionnaire, the ECRHS included information from around 140,000 individuals from 22 countries. A subsample of the study incorporated extensive questionnaires and measurement of atopic markers, spirometry, and methacholine challenge. The key findings of the study are the large geographical differences in the prevalence of asthma, atopy, and bronchial responsiveness in adults, with high prevalence rates in English-speaking countries and low prevalence rates in the Mediterranean region and Eastern Europe. Analyses of risk factors have highlighted the importance of occupational exposure for asthma in adulthood (Kogevinas et al. 1999). The association between sensitization to individual allergens and bronchial responsiveness was strongest for indoor allergens (mite and cat). Analysis of treatment practices has confirmed that the treatment of asthma varies widely between countries and that asthma is often undertreated (Janson et al. 2001). Subjects were subsequently followed up in a prospective study which included subjects with and without asthma at the initial contact.

Design of cross-sectional studies

Main design issues in cross-sectional studies

The simplest form of a survey in a population is a one-time measurement of the prevalence of a disease (Checkoway et al. 2004). On some occasions, repeated surveys (panel studies) are conducted on the same or different individuals.

When evaluating the prevalence of a disease the sample studied should be representative of the base population from where subjects are recruited (internal validity). The random error of an estimate depends directly on the size of the study and standard techniques exist to calculate this error (e.g. free software such as Epi Info™ (Centers for Disease Control and Prevention (2013)). The measurement of outcomes may be particularly complicated since frequently prevalence studies are conducted for diseases that have a gradual onset and that are not fatal. In addition, problems may occur when comparisons of prevalence are done between countries that may have different systems for the identification of the disease or even different perceptions and oral descriptions of a disease. An example of this is the large prevalence studies conducted for the evaluation of the prevalence of children's asthma such as the ISAAC study (see earlier description). Wheezing is a key symptom defining asthma in childhood. However the perception of what is wheezing is not necessarily the same in different countries and, further, some languages do not have a single word to describe what in English is called 'wheezing'.

Generalizability is another aspect of representativeness and refers to whether the findings in one population can be extrapolated to other populations (external validity). Generalizability is less of a problem when examining risk factors of disease since biological mechanisms through which an agent may provoke disease tend to be similar between populations. Generalizability may be, however, more difficult to ensure when examining prevalence since this may vary considerably within and between populations and extrapolations from one population to another can be problematic.

Panel studies apply a hybrid design that combines cross-sectional and cohort study methods. They are series of cross-sectional studies performed over time on the same group of individuals (measuring, however, prevalence rather than incidence), or in different groups of a population in subsequent time periods. The objective is to evaluate change in health status in relation to changes in exposure. Small groups, or panels, of individuals are followed over short time intervals, and health outcomes, exposure, and potential confounders are ascertained for each subject on one or more occasions. This design is especially useful for the study of physiological variables (i.e. pulmonary function, blood pressure), for which changes over several years may indicate early stages of disease processes. It is also useful to evaluate time trends in the prevalence of a disease.

The third phase of the ISAAC study (ISAAC, phase III) could be regarded as one form of a panel study or serial cross-sectional study. ISAAC-III was a repetition of the phase I survey after 5–10 years to examine time trends in the prevalence of allergic disorders in the same ages in centres and countries that participated in ISAAC phase I. The findings from ISAAC phase III indicate that international differences in asthma symptom prevalence have reduced with decreases in prevalence in English-speaking countries and Western Europe and increases in prevalence in regions where prevalence was previously low. Although there was little change in the overall prevalence of current wheeze, the percentage of children reported to have had asthma increased significantly, possibly reflecting greater awareness of this condition and/or changes in diagnostic practice. The increases in asthma symptom prevalence in Africa, Latin America, and parts of Asia indicate that the global burden of asthma is continuing to rise, but the global prevalence differences are lessening (Pearce et al. 2007).

Panel studies may also evaluate the same individuals before and after an intervention or an exposure. These types of studies have been extensively used in occupational and environmental epidemiology to evaluate changes in symptoms or physiological parameters following an intervention in a workplace or in the general environment such as air pollution. An example is a study that examined the effects of short-term exposure to diesel traffic in 60 adults with asthma (McCreanor et al. 2007). Each participant walked for 2 hours along Oxford Street in London (United Kingdom) and, on a separate occasion, through Hyde Park, London. Physiological and immunological measurements were performed after each occasion. It was found that walking for 2 hours on Oxford Street induced consistent reductions of up to 6.1% in FEV_1 and up to 5.4% in FVC that were significantly larger than the reductions in FEV_1 and FVC after exposure in Hyde Park. Similar changes were also observed for markers of inflammation in the lungs.

Panel studies, particularly when examining the same individuals, could be regarded as prospective studies and frequently this distinction between study designs is blurred. The study in Oxford Street is a crossover study that includes aspects of a prospective experimental design and aspects of a cross-sectional study. The main difference between cross-sectional panel studies and prospective designs is that panel studies examine the prevalence of a disease or a physiological parameter rather than incidence.

Sampling and response rates

Cross-sectional studies aiming to identify the prevalence of a specific factor in a community should select a representative sample of that community. The same is, in principle, desirable in cohort and case–control studies but in these studies the main issue is internal validity rather than external validity (generalizability representativeness). In cohort studies the most important issue concerning subject selection and validity of results is completeness of follow-up. In case–control studies the crucial issue concerning selection of subjects is that cases and controls should be selected from the same study base; whether cases are representative of the general population is, by contrast, not crucial for the internal validity of a case–control study.

In a cross-sectional study, however, representative sampling is the only procedure that allows selection of subjects in a population that can be used as the basis of generalization. A sampling frame should be identified that allows the identification of any subject in a population that could be included in a sample. Different methods can then be applied to select the subjects. Sampling error estimation has been exhaustively addressed in statistics. This depends on the number of subjects included in a study and can be estimated. Non-sampling error refers to bias, and may be of considerably higher importance than sampling error. Enlarging the sample does not reduce non-sampling error. Non-response may be non-random and may, therefore, be related to factors of interest for a study such as education, ethnicity, or obesity.

Response rates and dealing with non-response

A response rate of 60% has been frequently accepted as an adequate response in population surveys. However, the issue of non-response is complex and the acceptance of a fixed percentage as an adequate and probably unbiased response rate is an oversimplification. To some extent, this seems similar to the arbitrary acceptance of $p < 0.05$ as the level for statistical significance (Timothy et al. 2012). There is a general perception, not only among epidemiologists, that it is becoming increasingly difficult in many industrialized countries to achieve high response rates in surveys (Hartge 2006). In an evaluation of 355 original epidemiological papers published in ten high-impact journals, it was found that average participation in epidemiological studies has fallen in the last 30 years and this has

particularly affected controls in population-based studies (Morton et al. 2006). At least some information regarding participation was provided in 59% of cross-sectional studies, 44% of case–control studies, and 32% of cohort studies. In 51 out of 86 cross-sectional studies that reported response rates, the mean participation was 74%, ranging from 28% to 100%. In recent years, response rates among population controls of around 50% are not uncommon. Participation in cross-sectional studies decreased by approximately 1% per year from 1970 to 2003 (Morton et al. 2006).

The use of telephones as a means for defining a sampling frame has been rapidly decreasing. The main assumption for the use of random digit dialling (RDD) was that there was one working residential phone number per household. This assumption is no longer valid due to rapid changes in telecommunication, particularly the use of mobile phones. The rapid increase in the use of mobile phones throughout the world has made nearly impossible the identification of a roster of phones identified to a specific area through which a sample could be derived. In addition, in several countries the use of telephone contacts for commercial reasons has resulted in a decrease of response rates to any such type of contact. Computer-assisted telephone interviews can be a valid and efficient means of collecting information among subjects who have been contacted and agreed to participate.

Studies collecting biological samples

In recent years an increasing proportion of epidemiological studies collect biological samples, following the advances in molecular and genetic techniques and their application in large-scale studies. A recent review of 355 studies identified very high participation rates (above 90%) in cohort and cross-sectional studies collecting biospecimens. However, only about a third of the studies reporting collection of biospecimens also reported response rates for the biological samples. A total of 134 of 355 (38%) articles reported the collection of biological specimens to measure exposure or disease and the proportion of cohort and case–control studies that collected biological specimens increased over time while that of cross-sectional studies remained fairly constant (Morton et al. 2006). Large cross-sectional studies, such as the NHANES, are regularly collecting biological

samples and the inclusion of the collection of biospecimens does not appear to influence the representativeness of the sample.

An increasing number of studies are collecting saliva or other biospecimens by mail. Several studies that have collected saliva samples by mail have obtained response rates that are not very different from those observed for paper-and-pencil questionnaires. In a study on 2,994 subjects at Geneva University, Switzerland, the response rate for subjects requested to provide a mail questionnaire and saliva was 52% while for those requested to complete only the mail questionnaire, it was 63%; using financial incentives increased the response rates by 6–11% (Etter et al. 2005). Sociodemographic factors may affect response rates in such a type of sampling. Self-collection of saliva was evaluated, concerning both response and quality of DNA retrieved in a random sample of 611 men (ages 53–87 years) in Sweden (Rylander-Rudqvist et al. 2006). The response rate was 80% and varied from 89% for those aged 67–71 years to 71% in those aged 77–87 years. Similar to other studies, the DNA extracted was of high quality and could be used as an alternative to blood DNA in molecular epidemiological studies. There is little empirical evidence in several cultures to evaluate general tendencies concerning participation when biospecimens are collected. Whether non-response may differentially affect the analysis of biomarkers is unknown. One study examined a series of genetic data (polymorphisms, haplotypes, and short tandem repeats) from 2,955 individuals of three studies and found no evidence of differential results by participation (Bhatti et al. 2005). A note of caution, however, has been made against assuming that willingness to participate should not vary with a particular polymorphism or other forms of genetic variation (Hartge 2006).

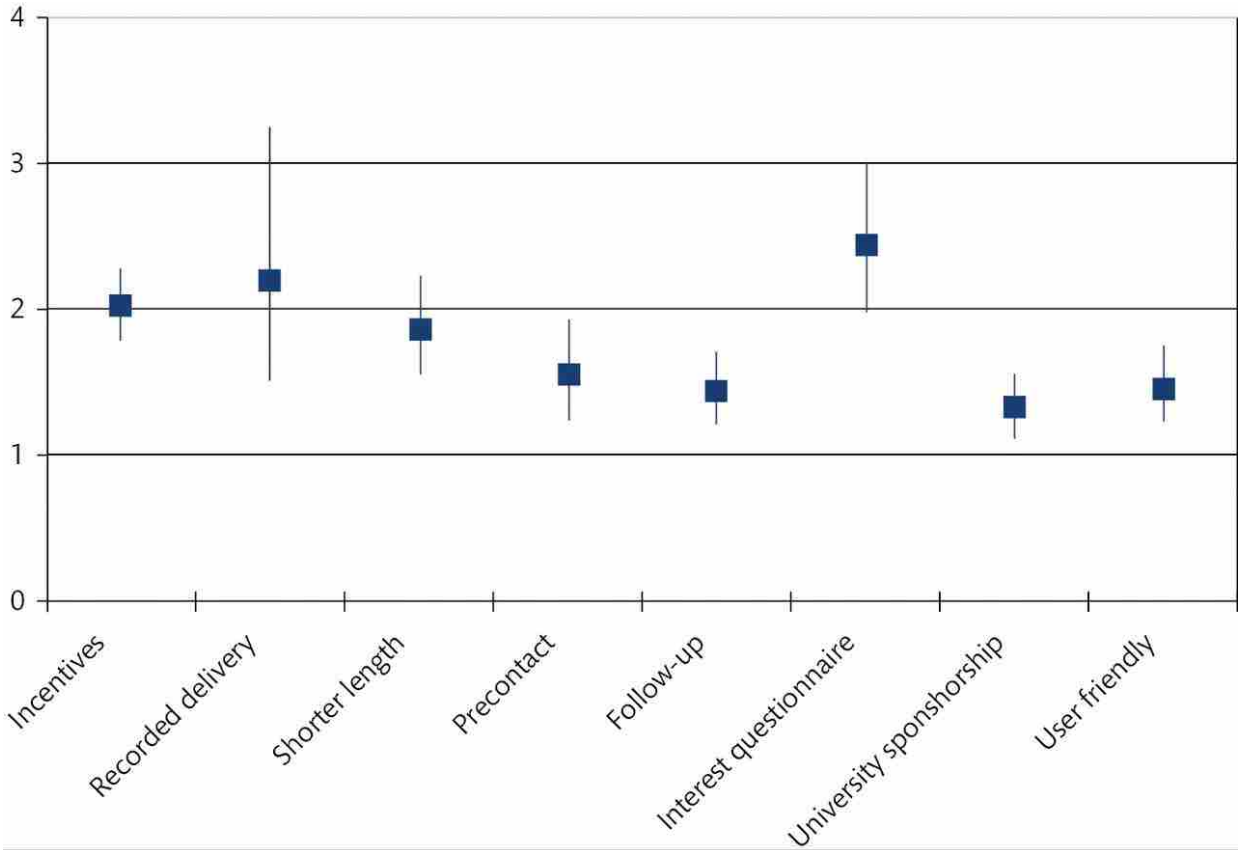


Fig. 5.2.2 Effects on questionnaire response (odds ratios and 95% confidence intervals) of eight strategies where combined trials included over 1,000 participants.

Source: data from Edwards, P. et al., Increasing response rates to postal questionnaires: systematic review, *British Medical Journal*, Volume 324, p. 1183.

Improving non-response

The difficulties in achieving high response rates and the potential biases resulting from self-selection makes indispensable the understanding of the reasons for non-response and the application of methods for limiting non-response. Numerous studies have evaluated factors affecting non-response in epidemiological and other types of studies. Most of the factors frequently identified such as age, sex, education, exposure status, method of recruitment, and type of questionnaire used may also be culture specific. This indicates that, in addition to some general guidelines, there is a need to verify main approaches to increase response in different sociocultural settings, and these may change in different time periods. Nevertheless, some general guidelines can be drawn up based on numerous observational and on randomized trials.

A comprehensive systematic review of available randomized trials from any relevant discipline (not restricted to medicine or epidemiology) evaluated factors that influence response to postal questionnaires (Edwards et al. 2002). Information from 292 randomized controlled trials that included 258,315 participants was reviewed. A total of 75 strategies that could influence response to postal questionnaires were reviewed. The review identified questionnaire length, use of monetary incentives particularly when not conditional on response, appearance of the package, layout of the questionnaire, type of follow-up, and registered delivery as the most influential factors in terms of participation rates. Figure 5.2.2 shows effect estimates for the most influential factors. Contacting participants before sending questionnaires increased response as did follow-up contact. Questionnaires designed to be of more interest to participants were more likely to be returned but questionnaires containing questions of a sensitive nature were less likely to be returned. Questionnaires originating from universities were more likely to be returned than were questionnaires from other sources, such as commercial organizations. Overall, response to postal questionnaires can be improved by following strategies shown to be effective in other research.

Adjusting for non-response

The most important measure to deal with potential biases from non-response is to minimize, as much as possible, the number of non-responders. As already discussed, this is done through selecting adequate ways of contact and through repeated contacts with subjects. Several approaches have been described to evaluate and adjust estimates of prevalence taking into account non-response.

An evaluation of basic characteristics of non-responders is important and may give clues concerning the representativeness of the sample. Such characteristics may be derived from information known for the target population from other surveys, for example, percentage of smokers. They can also be derived from a subsample of non-responders followed up intensively. Estimates of prevalence can then be weighed using the data of respondents but taking into account information on variables known for responders and non-responders such as age, sex, residence, etc.

A usual procedure in prevalence surveys is to perform repeated contacts of subjects to increase response rates. The final estimates are derived from the total number of responders. In an alternative procedure it can be assumed that responders at different mailings or contacts are not representative of the population. Under this assumption, the overall prevalence estimates are modelled using the prevalence of each mailing. A cumulative response at the time of each mailing can be calculated taking into account different covariates (e.g. age, sex) of the samples responding in each mailing, and estimates of the prevalence of non-responders are derived through extrapolation.

Such modelling was conducted analysing responses of 13,007 subjects in the three English centres of the ECRHS (Chinn et al. 1995). Modelling responses to take into account centre, age, and sex had only small effects on estimated prevalence. For example, in Norwich, out of 440 eligible individuals aged 20–24 years, 141 responded to the first mailing (32%), 58 to the second (cumulative response 45.2%), and 45 to the third mailing (cumulative response 55.5%). Prevalence of wheezing was modelled taking into account differences among responders and non-responders by age and sex and this gave an estimated prevalence of wheezing of 25.7%. This estimate was further examined through modelling. In a model that included a parameter for a linear trend in prevalence with each mailing by age and sex group, an overall prevalence of 23.9% was estimated. Overall in this study although several models gave statistically significant differences, in absolute terms the estimates of prevalence derived were rather similar. Such type of modelling is useful to explore the sensitivity of estimated prevalence to non-response bias. However, frequently changes due to such modelling are small and, in general, caution should be exerted when applying such models. In conclusion, the best way of adjusting for non-response is not having much of it.

Item non-response

Apart from the overall response rate the quality of the answers of participants deserves attention as well. The issue of item non-response has been little discussed and usually not reported. What is usually discussed and reported as ‘non-response’ refers to the non-contact with a selected person or the refusal of this person to participate (unit non-response). Failure of a

responder to answer an individual question or provide a sample (item non-response) should also be considered and can equally affect the representativeness of specific analyses in a study. In a study on bladder cancer in Spain, the overall response rate to the main questionnaire was 85%. However, the response to specific items was lower. For example, the questions on lifetime consumption of water and complete information on water quality in each of the residences of the subjects were available for 1,479 (59%) out of a total of 2,490 subjects (Villanueva et al. 2007). An evaluation of differences in sociodemographic characteristics of subjects included or excluded from the water analyses and also differences in the prevalence of known risk factors such as smoking indicated that on this occasion item non-response was not likely to bias results. Exclusion of subjects reduced, however, the statistical power of the study.

The inclusion of several variables in an analysis may result in a large amount of missing data due to item non-response. The pattern of missing data should be examined and, if possible, the causes of missing information determined. Multiple imputations methods are increasingly used to improve efficiency and minimize selection bias. It is frequently assumed that data are missing at random although this is not always easy to identify. Missing values are imputed and a set of complete data sets is generated that are further analysed individually to obtain a set of parameters and then combined into overall estimates (Rubin 1987). Several methods have been proposed to impute missing data and modern software has simplified the analysis of multiply imputed data (White et al. 2011; Cummings 2013).

Sampling methods

Different types of sampling may be applied in cross-sectional studies depending on the aims of the study, characteristics of the population studied, the available background information, and the means available in each study.

In a simple random sample, all subsets of the sampling frame are given an equal probability of being selected. What is used most frequently in large surveys is stratified sampling: this involves a first-stage selection of different categories of the population and subsequent selections at random within each stratum. The main reason for doing stratified sampling is efficiency, so as to secure adequate sampling of small population groups and avoid oversampling of larger groups. Several other probabilistic or non-

probabilistic types of sampling have been applied such as ‘cluster sampling’ in which the unit selected is a group of persons (e.g. a city block) rather than individuals, and ‘snowball sampling’ and ‘respondent-driven sampling’ where a first set of study subjects (‘seed participants’) is used to recruit more subjects into the sample. These latter types are convenience sampling methods that are applied when a complete identification of the primary sampling population is not possible or hard to find (Magnani et al. 2005). Respondent-driven sampling is a variation of snowball sampling that allows a better control of the selection of the population examined, mainly by collecting information from recruiters on the size of their social network and on also on persons approached but who did not participate. The availability of such information, however, does not secure the selection of a random sample from members of a network (Heimer 2005). In addition, respondent-driven sampling is believed to have higher external validity than other convenience sampling methods because it may extend better to all potential members of a network (Magnani et al. 2005). The same approach has been used for the identification of samples for interventions in hard-to-reach populations. For example, respondent-driven sampling was used in a cluster-randomized trial in India to assess the effectiveness of HIV counselling and testing among men who have sex with men (Solomon et al. 2013).

Snowball sampling was applied in a study on the prevalence of depression among female sex workers of eastern Nepal (Sagtani et al. 2013). Snowball sampling is especially useful when the objective is to reach populations that are inaccessible or hard to find. The participants were 210 female restaurant-based and street-based sex workers in five cities of eastern Nepal. In countries like Nepal where commercial sex is illegal, the criminalized status of their work means that commercial sex workers are prone to harassment and violence. Under such conditions, identifying a base population and sampling from this population is impossible. Sex workers were contacted through snowball sampling. The first few respondents were traced with the help of a non-governmental organization working on HIV prevention in the study area. Face-to-face interviews were conducted and additional respondents were nominated. The prevalence of depression among respondents was 82%. A major limitation of this type of sampling is that external validity of the study is questionable due to the unknown sampling frame.

Biased samples are samples where some members of the population are less likely to be included than others. (Note that some stratified samples are deliberately biased but the degree of bias can be estimated.) There are several classic examples of biased sampling. Among the most well-known is a large opinion poll in the US presidential election in 1936 that erroneously predicted that the incumbent president, Franklin Roosevelt, would lose by a large margin. This poll of approximately 2 million persons was based on people who were readers of a magazine, supplemented by records of registered automobile owners and telephone users. The error was produced because the sample overrepresented more affluent people who were less likely to vote for the Democratic candidate.

Examples of different sampling procedures are listed in the following paragraphs, namely sampling in a large multipurpose survey (NHANES) in the United States, a large international prevalence survey (ISAAC), together with a local survey of a difficult-to-identify population of drug addicts in Spain.

The NHANES survey in the United States applies a multistage (four stages) probability sampling to select participants representative of the civilian, non-institutionalized US population (Centers for Disease Control and Prevention 2014). In stage 1, primary sampling units (mostly single counties) are selected. In the second stage, these units are divided up into segments (generally city blocks) and sample segments are selected taking size into account. In stage 3, households within each segment are listed and a sample is randomly drawn. In the final stage, individuals are randomly chosen from a list of all persons residing in the selected households within designated age–sex–race strata. NHANES is designed to sample larger numbers of certain subgroups of particular public health interest including African Americans, adolescents, and others.

In ISAAC, a combination of non-random and random methods was used. The areas participating within each country were not selected at random but instead qualitative criteria and convenience sampling were used. Stratified random sampling was then used within a given geographical area to select the school children. All schools in each geographical area centre were listed and a random sample was selected. The sampling unit was the school and non-participating schools were replaced. All children on a selected school within the two specified age groups (6–7 and 13–14 years of age) were included.

Web-based research and the use of wireless technologies

Web-based research has been expanding in recent years. The Internet has been used in research and clinical settings to enhance contact with participants. Prevalence surveys have traditionally been based on self-administered paper questionnaires, face-to-face or telephone interviews. Internet surveys have been proposed as having advantages compared to traditional approaches, for example, face-to-face or telephone interviews, regarding turnaround time and expenses. Internet-based surveys similar to computer-assisted personal or telephone interviews have advantages concerning data management as compared to paper-and-pencil questionnaires.

The two main issues regarding the validity of use of Internet surveys are response rates and quality of the data collected. Several studies that focused on specific population groups with high indexes of Internet use have found comparable response rates between internet-based questionnaires and more traditional approaches (Pealer et al. 2001). However, many of these studies report response rates for populations recruited through the Internet rather than for population-based samples. These results are therefore difficult to generalize, although they indicate that in specific populations, internet-based surveys provide high-quality data.

Few studies exist that compare population-based data using both Internet and traditional type approaches. Results differ and depend crucially on the degree of education and computer literacy in different cultures, ages, and sociodemographic groups. Women referred for mammography at a Danish public hospital aged less than 67 years and without a history of breast cancer, were randomized to either receive a paper questionnaire, with a prepaid return envelope, or a guideline on how to fill in the internet-based version online (Kongsved et al. 2007). The response rate was 18% for the Internet group compared to 73% for the paper questionnaire. An interesting aspect of this study was that after sending a reminder and offering the possibility of using the alternative questionnaire form (i.e. a paper questionnaire in the Internet group), the response rate for the Internet group became comparable to that of the group assigned the paper questionnaire. In the Millennium cohort (Smith et al. 2007) of military personnel in the United States, a similar proportion of the approximately 70,000 responders chose to respond using the Internet. Web responders were more likely to be male, to be younger, have a higher education, work in technical occupations, to be

obese, and to smoke more cigarettes. They were less likely to be problem alcohol drinkers and to report occupational exposures. Several studies have evaluated quality of information from web-based studies and have found that it is similar to, if not better than, self-administered questionnaires. In the Danish mammography study mentioned earlier (Kongsved et al. 2007), the quality of the information was better in the Internet version with 98% having no missing data compared to 63% filling in a complete paper questionnaire. A study in Sweden (Bonn et al. 2013) compared web-based self-reports of bodyweight with weight measured at the study centre in a convenience sample of 149 individuals. As frequently happens with self-reported weight, overweight subjects (body mass index ≥25 kg/m^2) tended to underreport their weight and the overall mean difference between self-reported and measured weight was −1.2 kg. The Spearman correlation coefficient, however, between self-reported and measured weight was 0.98 ($p < 0.001$) indicating that self-reported weight and perhaps other anthropometric measures via the web can be a valid method of data collection.

The use of wireless technologies and devices is rapidly increasing and it is expected that smartphone technology has become a frequent way to access information in the World Wide Web. The ubiquity of mobile device ownership and the development applications allows new ways of contact and recording of health-related information in population studies. A study in Barcelona used smartphone technology to improve environmental exposure assessment applying novel sensor technologies integrated with mobile phones (de Nazelle et al. 2013). Geographic location and physical activity patterns were tracked through a smartphone application in 36 subjects and were then linked to space–time air pollution mapping for improved air pollution exposure assessment. This information provides an accurate account of mobility of the subjects during the day and of physical activity. Mobile technologies also allow individuals to record and communicate behaviour in real time from any location. A study on 50 trained subjects evaluated the feasibility of using Twitter to capture dietary behaviour and reasons for eating over three consecutive days (Hingle et al. 2013). A list of hashtags representing food groups, for example, #fruit, and reasons for eating, for example, #mood, were provided to participants who were encouraged to annotate hashtags with photos, text, and links. A total of 773 tweets including 2,862 hashtags were reported and eating behaviours were

visualized through ‘maps’ of hashtag co-occurrences suggesting time-varying diet and behaviour patterns.

On the basis of existing evidence it appears that internet-based surveys may provide high response rates in selected populations. In most general populations, however, internet-based surveys do not yet seem to provide adequate response rates and this may lead to selected sampling. Several studies indicate that Internet surveys may be efficiently combined with traditional techniques to save resources and perhaps provide more complete answers. Clearly, more research is needed in the future in population-based surveys to evaluate changes in response in internet-based research as more of the population becomes more technologically savvy. While sensor-based applications in environmental epidemiology have been widely used and have been shown to improve exposure assessment, there is a lack of empirical evidence in public health research concerning the use of popular software applications such as Twitter.

Bias in cross-sectional studies

Cross-sectional studies may suffer from the same potential biases as other epidemiological designs: misclassification, selection bias, and confounding. Issues of particular concern in cross-sectional studies are the potential biases from the mixing of incident and prevalent cases, the frequent lack of information on past exposures and consequently on the time sequence of events, and the lack of representativeness in studies evaluating the prevalence of disease or of other health-related variables. Because of these difficulties, it has been recommended that cross-sectional studies examining the risk factors of disease are appropriate for diseases that produce little disability, and are also appropriate to the presymptomatic phases of more serious disorders (Coggon et al. 1997).

Incidence–prevalence bias (known as Neyman bias; also discussed as length-bias) is a type of selection bias due to the mix of incident and prevalent cases in cross-sectional studies. Incidence– prevalence bias refers specifically to the selection of a case group that is not representative of cases in the community. Figure 5.2.3 (left graph) shows the hypothetical time of occurrence of disease and the duration of disease. A cross-sectional study would include a disproportionately high percentage of cases with long duration while those with more aggressive disease and hence short duration

would have a lower probability of being sampled. The latter is exemplified in Figure 5.2.3 (right graph). A bias would occur if survivors of a condition studied are atypical with respect to the exposure evaluated. If the exposure examined is not directly or indirectly related to survival, inclusion of prevalent cases would not bias any association. In addition, as noted by Hill et al. (2003), for diseases with extremely low case fatality (e.g. rheumatoid arthritis), studies based on prevalent cases are not subject to Neyman bias even if the risk factor under study produces increased mortality from other causes.

The association of smoking with Alzheimer's disease has been discussed as a potential effect of incidence–prevalence bias. Several cross-sectional studies have suggested that smoking exerts a protective effect on Alzheimer's disease. Evidence, however, from a few cohort studies is controversial. A cross-sectional study that included a prospective component suggested that part of the protective effect could be attributed to Neyman bias, specifically because smokers with dementia could have worse survival than smokers without dementia. The study was a population-based cohort of 668 people aged 75–101 years in Sweden (Wang et al. 1999). Smoking was negatively associated with prevalent Alzheimer's disease (adjusted odds ratio = 0.6, 95% confidence interval (CI) 0.4–1.1). Over 3-year follow-up (1989–1992), the hazard ratio of incident Alzheimer's disease due to smoking was 1.1 (95% CI 0.5–2.4). In addition, mortality at follow-up was greater among smokers in demented (hazard ratio = 3.4) than non-demented subjects (hazard ratio = 0.8). The authors concluded that smoking does not protect against Alzheimer's disease and that the cross-sectional association might be due to differential mortality.

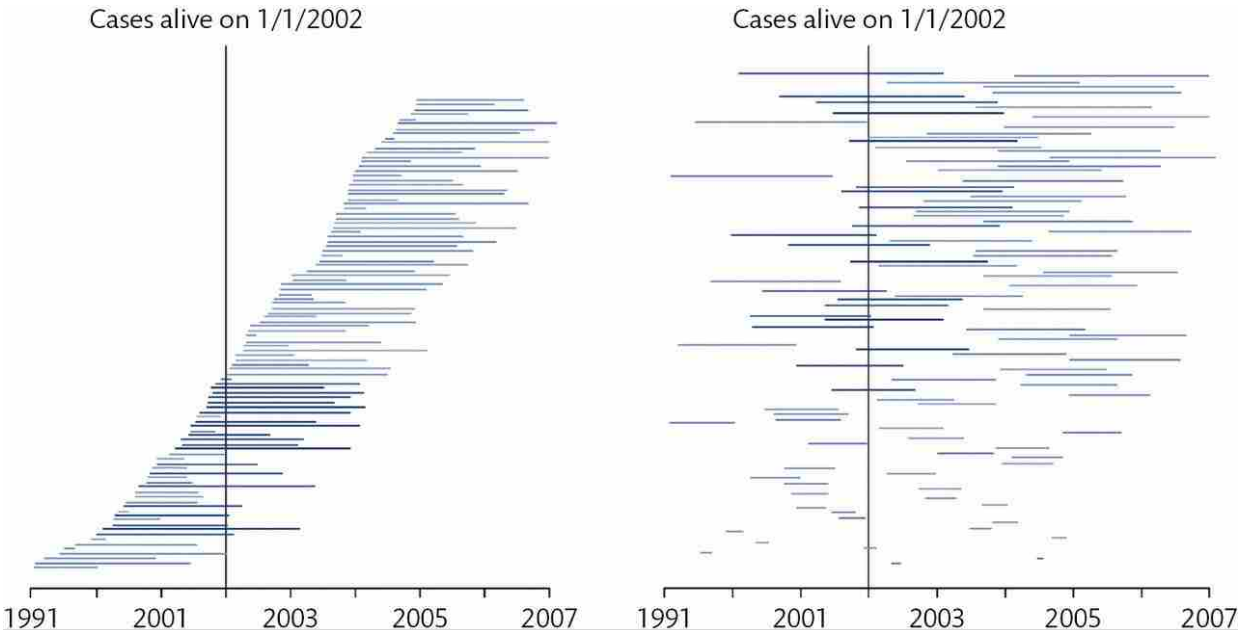


Fig. 5.2.3 Incidence–prevalence bias in cross-sectional studies. Note: the lines on this graph represent the duration of disease with the left end point representing the date that the disease was first diagnosed and the right end point representing the date that the patient died. In the left graph, subjects are ordered by time of initial diagnosis. A cross-sectional study in January 2002 selected those patients who were alive at that time (indicated with vertical line). Notice that patients diagnosed before 2002 with a short survival would have less probability of being selected compared to subjects diagnosed the same period but with a longer survival. In the graph on the right, patients with the shortest survival times appear at the bottom of the graph and the patients with the longest survival times appear at the top. Notice how rarely the patients with short survival times appear among the prevalent cases.
Reproduced from *StATS: Neyman bias* (December 15, 2004), available from http://www.pmean.com/04/NeymanBias.html, licensed under a Creative Commons Attribution 3.0 United States License.

When examining prevalent cases an additional problem may occur if diseased persons modify their exposure or if disease affects measures of exposure, for example, the presence of disease distorting biomarker levels. Treatment with insulin of diabetes mellitus type 1 is an example. Before insulin was available, patients had a very poor survival of less than 1 year. At that time, a cross-sectional study would simply not identify many diabetics. With insulin treatment, survival has spectacularly increased to more than 20 years. A cross-sectional study evaluating diabetes and use of insulin would therefore find an association, simply because diabetics take insulin while non-diabetics do not. This bias has also been described as a type of Neyman bias (incidence– prevalence bias) but in reality it is rather a case of reverse causation provoked by misclassification due to an erroneous evaluation of timing of exposure.

In recent years there has been an increase in the number of studies using data from large surveys such as the NHANES. Experimental data suggest that phthalates, a widespread contaminant, could produce metabolic derangements that may lead to obesity. The association between urinary phthalate metabolites and body mass was examined in a cross-sectional analysis of 2,884 children 6–19 years of age who participated in the 2003–2008 NHANES. Phthalates were measured at one time point and monoesters of phthalates typically have half-lives of around 24 hours. An association of phthalates with childhood obesity was identified among non-Hispanic black children while no significant associations were identified in any other ethnic groups. Making maximum use of a resource such as the NHANES is understandable and advisable given its cost and the population sampling and these findings should lead to additional research. However, making causal inferences from this type of cross-sectional data may be complex. Reverse causation may be a problem in cross-sectional analyses since both physiological aspects and behavioural pathways, for example, restricted mobility of subjects with disease, may influence measured exposures. The London bus drivers' study in the 1950s is a classic example of potential reverse causation.

The study by Jerry Morris on the incidence and prevalence of cardiovascular disease among London bus drivers and conductors (Morris and Dale 1955) is an example of selection bias in a cross-sectional study. Bus drivers in London were shown to have very different lipid profiles, blood pressure patterns, and eventually cardiovascular disease risk compared to bus conductors. Bus drivers were less physically active on the job than conductors, but could the difference in cardiovascular disease risk be ascribed to physical activity? One potential alternative explanation was that some characteristic of the persons such as obesity, could have led to self-selection in a job that was more (bus conductors) or less (bus drivers) active. The studies in London busmen were eventually the first to convincingly demonstrate the importance of physical activity for health.

The healthy worker effect may lead to an underestimation of an association in case–control studies that evaluate occupation and disease at a specific point in time (Checkoway et al. 2004). This happens because workers with health problems at work are more likely to change jobs than subjects without symptoms. In this case the selection is out of the job, contrary to what was a potential problem in busmen. Any cross-sectional

evaluation will, therefore, not capture cases that have occurred in the past and will over-represent healthy individuals who have stayed in the job. An example of this was observed in a cross-sectional study on occupational asthma in Spain (Kogevinas et al. 1996). Subjects were asked about their current occupation and about current respiratory symptoms. The prevalence odds ratios (PORs) for asthma in subjects employed in high-risk occupations was 1.4 with 95% CI of 0.6–3.3. Subjects were also asked whether they had suffered respiratory symptoms in the past and, if yes, what type of job were they doing. The POR for employment in a high-risk occupation at the time of respiratory problems was 2.8 (95% CI 1.5–5.2) indicating that, indeed, the cross-sectional analysis failed to identify health risks due to the selective migration of subjects with respiratory symptoms out of high-risk jobs.

This type of bias is less of a problem if current exposure is a reliable surrogate for past exposure and is not a problem for non-changeable characteristics, for example, race, gender, and genetic variability. Case–control or cross-sectional studies are currently the most widely applied designs evaluating the effects of genetic susceptibility, such as the effects of SNPs (single nucleotide polymorphisms), on disease risk.

Social desirability bias has long been recognized and falls within the wider category of misclassification (Armstrong et al. 1994). Requesting information on behaviours that are regarded as sensitive by respondents may lead to refusal to participate or may lead to a tendency to over-report behaviours regarded as socially desirable (or, what is the same, under-report what is regarded as socially undesirable). This has become particularly evident in recent years in large surveys that inquire into health-related behaviours such as sexual activity and drug use.

There is extensive research showing that self-administered questionnaires lead to more valid information on socially sensitive issues compared to face-to-face interviews (Schuman and Presser 1981; Armstrong et al. 1994). In recent years the use of computers in soliciting information has considerably expanded. In addition to CAPIs (computer-assisted personal interviews) the technique of audio-computer-assisted self-interviewing (ACASI) has been proposed as an efficient method to enhance validity when evaluating behaviours that could be perceived as socially undesirable (Turner et al. 1998). In ACASI (also quoted as audio-CASI), respondents hear the questions through headphones and enter their responses directly into a computer. An experimental comparison of sexual behaviours in adolescent

males using a paper-and-pencil self-administered questionnaire and ACASI identified an almost fourfold increase in the number of respondents reporting male–male sexual contact in the ACASI (5.5%) as compared to 1.5% in the paper-and-pencil questionnaire (Turner et al. 1998). Among the reported advantages of ACASI is that it reduces the literacy requirements of respondents compared to other forms of self-administered questionnaires (paper and pencil or computer assisted). ACASI has been tested in multiple settings and has generally been shown to provide higher estimates of prevalence of socially undesirable behaviours compared to face-to-face interviews (Ghanem et al. 2005). Results are less consistent when comparing ACASI with self-administered questionnaires (Morrison-Beedy et al. 2006). Different computer-assisted methods (audio or text) may have little impact on reported socially undesirable behaviours and the choice of one method over another should be based on an evaluation of wider considerations on the study settings.

Analysis of cross-sectional studies

Prevalence is the ratio/proportion of the number of cases of a disease in a population at a particular time, to the size of the population at that time:

$$\text{Prevalence} = \frac{\text{number of existing cases of disease}}{\text{total population at a point in time}}$$

Point prevalence is estimated at one point in time, while period prevalence denotes the number of cases in a time interval (e.g. 1 year).

For example, in a given state in June 2007, there were 14,000 women with a diagnosis of breast cancer. The total female population in June 2007 was 1,558,000 women. The prevalence of breast cancer in this population on this specific date is 14,000/1,558,000, which equals 0.009 (or 0.9%). This means that nearly 1% of the women in June 2007 had breast cancer, a figure that corresponds to the actual prevalence of this disease in several industrialized countries. This estimate does not show how many new cases of breast cancer occur each year or when the 14,000 women had their disease diagnosed. In these types of estimates it is usually assumed that none of the persons

diagnosed with a chronic disease is entirely cured, although such calculations could be incorporated into the estimation of prevalence.

Prevalence is, strictly speaking, a proportion although it is sometimes indicated, erroneously, as a rate. However, the term ‘prevalence rate’ should be regarded theoretically as an impossible concept (Elandt-Johnson 1975).

Prevalence reflects determinants of incidence as well as survival (duration of a disease). The prevalence of bladder cancer in the United States in January 2004 (377,523 persons alive with a history of bladder cancer) was higher than that of lung cancer (174,880 persons), even though the incidence of lung cancer (81.2 per 100,000) was clearly higher than that of bladder cancer (37.3/100,000). This is a reflection of the much better prognosis of bladder cancer (5-year relative survival rate of 79.5%) compared to 15% for lung cancer (National Cancer Institute n.d.).

In its most simple expression prevalence odds are equal to incidence rate times the duration of the disease:

$$\frac{\mathrm{P}}{1-\mathrm{P}} = 1 \times \mathrm{D},$$

where P is prevalence, I is incidence, and D is duration (survival). This, however, is only true under stable conditions, that is, unchanging rates and population structure that are unlikely in most occasions in true populations. If two populations are compared, say 1 indicating the exposed and 0 the non-exposed, then the POR can be calculated as:

$$\mathrm{POR} = \frac{P_1(1-P_1)I_1D_1}{P_0(1-P_0)I_0D_0}.$$

If the duration of the disease is expected to be equal in the exposed and unexposed, then the POR is equal to the incidence rate ratio (under steady conditions in the population). Under the same conditions, the ratio of prevalences (as compared to the ratio of prevalence odds) will only estimate the incidence rate ratio if the prevalence of the disease is low. In that occasion, 1 − P would be roughly equal to 1.

There has been some controversy whether cross-sectional studies should be calculating POR or prevalence ratios (Thompson et al. 1998; Pearce

2004). The choice of one or the other depends essentially on the purposes of the research, with the POR having advantages compared to the prevalence ratio if aetiological research is conducted. However, the prevalence ratio or probably the absolute difference in prevalences may be preferred if two populations are compared regarding the prevalence of disease.

Statistical analysis in cross-sectional studies is frequently limited to descriptive statistics through the use of standard statistical measures for proportions or a comparison of distributions (e.g. blood pressure) between an exposed and an unexposed group. Standard statistical tests comparing means of a distribution such as Student's *t*-test can be used. For the evaluation of multiple variables, standard techniques such as multiple linear regression can be used. For dichotomous outcomes, the POR is the most common effect measure estimated in cross-sectional studies comparing groups, using logistic regression or equivalent techniques. A wide range of other analytical regression techniques is readily available for the modelling of prevalence ratios (Thompson et al. 1998; Barros and Hirakata 2003). The selection of the effect measure, that is, POR versus the prevalence ratio as the preferred effect measure, should, therefore, not depend on the availability of statistical tools (Pearce 2004), but rather on the objectives of the research.

Summary

Cross-sectional studies are among the most frequently applied designs. The principal characteristic of cross-sectional studies is that they provide information on the prevalence of disease. In many cross-sectional studies information on past exposures is not collected but this should not be regarded as a characteristic defining these studies. On many occasions, cross-sectional studies tend to follow a hybrid design, an analysis incorporating aspects of cohort and of case–control studies. Repeated cross-sectional studies (panel studies) offer the possibility of assessing short-term and long-term physiological changes. Cross-sectional studies are extensively used to evaluate the prevalence of disease. Large cross-sectional studies are conducted in several countries to provide background information on health-related variables including health services use. In recent years these surveys have also incorporated the collection of biospecimens. In aetiological research, cross-sectional studies are suitable epidemiological means for studying non-fatal diseases and effects on physiological variables that do not

have a clear time of onset. However, the design of cross-sectional studies makes them less appropriate than other study designs for investigating causal associations. Among the major limitations of cross-sectional studies are the inclusion of prevalent rather than incident cases and the frequent lack of information on past exposures and on the time sequence of events. The representativeness of the sample is crucial when evaluating prevalence of a disease and several policies have been developed to increase participation in studies, reduce non-response, and improve the quality of the recorded information.

REFERENCES

Anonymous (1998). Worldwide variation in prevalence of symptoms of asthma, allergic rhinoconjunctivitis, and atopic eczema: ISAAC. The International Study of Asthma and Allergies in Childhood (ISAAC) Steering Committee. *The Lancet*, 351, 1225–32.

Arif, A.A., Delclos, G.L., Whitehead, L.W., Tortolero, S.R., and Lee, E.S. (2003). Occupational exposures associated with work-related asthma and work-related wheezing among U.S. workers. *American Journal of Industrial Medicine*, 44, 368–76.

Armstrong, B.K., White, E., and Saracci, R. (1994). *Principles of Exposure Measurement in Epidemiology*. New York: Oxford University Press.

Asher, M.I., et al. (1995). International Study of Asthma and Allergies in Childhood (ISAAC): rationale and methods. *European Respiratory Journal*, 8, 483–91.

Barros, A.J. and Hirakata, V.N. (2003). Alternatives for logistic regression in cross-sectional studies: an empirical comparison of models that directly estimate the prevalence ratio. *BMC Medical Research Methodology*, 3, 21.

Bhatti, P., et al. (2005). Genetic variation and willingness to participate in epidemiologic research: data from three studies. *Cancer Epidemiology, Biomarkers & Prevention*, 14, 2449–53.

Bonn, S.E., Trolle Lagerros, Y., and Bälter, K. (2013). How valid are web-based self-reports of weight? *Journal of Medical Internet Research*, 15(4), e52.

Burney, P.G., Luczynska, C., Chinn, S., and Jarvis, D. (1994). The European Community Respiratory Health Survey. *European Respiratory Journal*, 7, 954–60.

Centers for Disease Control and Prevention (2013). *Epi Info*™. [Online] Available at: http://wwwn.cdc.gov/epiinfo/

Centers for Disease Control and Prevention (2014). National Health and Nutrition Examination Survey. [Online] Available at: http://www.cdc.gov/nchs/nhanes.htm

Checkoway, H., Pearce, N., and Kriebel, D. (2004). *Research Methods in Occupational Epidemiology*. Oxford: Oxford University Press.

Chinn, S., Zanolin, E., Lai, E., Jarvis, D., Luczynska, C.M., and Burney, P.G. (1995). Adjustment of reported prevalence of respiratory symptoms for non-response in a multi-centre health survey. *International Journal of Epidemiology*, 24, 603–11.

Coggon, D., Rose, G., and Barker, D.J.P. (1997). *Epidemiology for the Uninitiated*. London: BMJ Publishing Group.

Cummings, P. (2013). Missing data and multiple imputation. *JAMA Pediatrics*, 13, 1–7.

De Nazelle, A., Seto, E., Donaire-Gonzalez, D., et al. (2013). Improving estimates of air pollution exposure through ubiquitous sensing technologies. *Environmental Pollution*, 176, 92–99.

Department of Health (2012). *Health Survey for England 2011*. [Online] Available at: https://www.gov.uk/government/publications/health-survey-for-england-2011

Edwards, P., et al. (2002). Increasing response rates to postal questionnaires: systematic review. *British Medical Journal*, 324, 1183.

Elandt-Johnson, R.C. (1975). Definition of rates: some remarks on their use and misuse. *American Journal of Epidemiology*, 102, 267–71.

Etter, J.F., Neidhart, E., Bertrand, S., Malafosse, A., and Bertrand, D. (2005). Collecting saliva by mail for genetic and cotinine analyses in participants recruited through the Internet. *European Journal of Epidemiology*, 20, 833–8.

Ghanem, K.G., Hutton, H.E., Zenilman, J.M., Zimba, R., and Erbelding, E.J. (2005). Audio computer assisted self-interview and face to face interview modes in assessing response bias among STD clinic patients. *Sexually Transmitted Infections*, 81, 421–5.

Hartge, P. (2006). Participation in population studies. *Epidemiology*, 17, 252–4.

Heimer, R. (2005). Critical issues and further questions about respondent-driven sampling: comment on Ramirez-Valles, et al. (2005). *AIDS and Behavior*, 9, 403–8.

Hill, G., Connelly, J., Hebert, R., Lindsay, J., and Millar, W. (2003). Neyman's bias re-visited. *Journal of Clinical Epidemiology*, 56, 293–6.

Hingle, M., Yoon, D., Fowler, J., et al. (2013). Collection and visualization of dietary behavior and reasons for eating using Twitter. *Journal of Medical Internet Research*, 15(6), e125.

Janson, C., et al. (2001). The European Community Respiratory Health Survey: what are the main results so far? European Community Respiratory Health Survey II. *European Respiratory Journal*, 18, 598–611.

Kogevinas, M., Anto, J.M., Soriano, J.B., Tobias, A., and Burney, P. (1996). The risk of asthma attributable to occupational exposures. A population-based study in Spain. Spanish Group of the European Asthma Study. *American Journal of Respiratory and Critical Care Medicine*, 154, 137–43.

Kogevinas, M., Anto, J.M., Sunyer, J., Tobias, A., Kromhout, H., and Burney, P. (1999). Occupational asthma in Europe and other industrialised areas: a population-based study. European Community Respiratory Health Survey Study Group. *The Lancet*, 353, 1750–4.

Kongsved, S.M., Basnov, M., Holm-Christensen, K., and Hjollund, N.H. (2007). Response rate and completeness of questionnaires: a randomized study of Internet versus paper-and-pencil versions. *Journal of Medical Internet Research*, 9, e25.

Magnani, R., Sabin, K., Saidel, T., and Heckathorn, D. (2005). Review of sampling hard-to-reach and hidden populations for HIV surveillance. *AIDS*, 19(Suppl. 2), S67–72.

McCreanor, J., et al. (2007). Respiratory effects of exposure to diesel traffic in persons with asthma. *The New England Journal of Medicine*, 357, 2348–58.

Morris, J.N. and Dale, R.A. (1955). Epidemiology of coronary atherosclerosis. *Proceedings of the Royal Society of Medicine*, 48, 667–72.

Morrison-Beedy, D., Carey, M.P., and Tu, T. (2006) Accuracy of audio computer-assisted self-interviewing (ACASI) and self-administered questionnaires for the assessment of sexual behavior. *AIDS and Behavior*, 10, 541–52.

Morton, L.M., Cahill, J., and Hartge, P. (2006). Reporting participation in epidemiologic studies: a survey of practice. *American Journal of Epidemiology*, 163, 197–203.

National Cancer Institute (n.d.). *Cancer Stat Fact Sheets*. [Online] Available at: http://seer.cancer.gov/statfacts/

Pealer, L.N., Weiler, R.M., Pigg, R.M., Jr., Miller, D., and Dorman, S.M. (2001). The feasibility of a web-based surveillance system to collect health risk behavior data from college students. *Health Education & Behavior*, 28, 547–59.

Pearce, N. (2004). Effect measures in prevalence studies. *Environmental Health Sciences*, 112, 1047–50.

Pearce, N., et al. (2007). Worldwide trends in the prevalence of asthma symptoms: phase III of the International Study of Asthma and Allergies in Childhood (ISAAC). *Thorax*, 62, 758–66.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Rylander-Rudqvist, T., Hakansson, N., Tybring, G., and Wolk, A. (2006). Quality and quantity of saliva DNA obtained from the self-administrated oragene method—a pilot study on the cohort of Swedish men. *Cancer Epidemiology, Biomarkers & Prevention*, 15, 1742–5.

Sagtani, R.A., Bhattarai, S., Adhikari, B.R., Baral, D., Yadav, D.K., and Pokharel, P.K. (2013). Violence, HIV risk behaviour and depression among female sex workers of eastern Nepal. *BMJ Open*, 3(6), e002763.

Schuman, H. and Presser, S. (1981). *Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording, and Context*. San Diego, CA: Academic Press.

Singh, G.K. and Kogan, M.D. (2010). *Childhood Obesity in the United States, 1976–2008: Trends and Current Racial/Ethnic, Socioeconomic, and Geographic Disparities. A 75th Anniversary Publication. Health Resources and Services Administration, Maternal and Child Health Bureau*. Rockville, MD: U.S. Department of Health and Human Services.

Smith, B., Smith, T.C., Gray, G.C., and Ryan, M.A. (2007). When epidemiology meets the Internet: web-based surveys in the Millennium Cohort Study. *American Journal of Epidemiology*, 166, 1345–54.

Solomon, S.S., Lucas, G.M., Celentano, D.D., Sifakis, F., and Mehta, S.H. (2013). Beyond surveillance: a role for respondent-driven sampling in implementation science. *American Journal of Epidemiology*, 178(2), 260–7.

Thompson, M.L., Myers, J.E. and Kriebel, D. (1998). Prevalence odds ratio or prevalence ratio in the analysis of cross-sectional data: what is to be done? *Occupational and Environmental Medicine*, 55, 272–7.

Timothy, P., Johnson, T.P., and Wislar, J.S. (2012). Response rates and nonresponse errors in surveys. *JAMA*, 307(17), 1805–806.

Turner, C.F., Ku, L., Rogers, S.M., Lindberg, L.D., Pleck, J.H., and Sonenstein, F.L. (1998). Adolescent sexual behavior, drug use and violence: increased reporting with computer survey technology, *Science*, 280, 867–73.

Villanueva, C.M., et al. (2007). Bladder cancer and exposure to water disinfection by-products through ingestion, bathing, showering, and swimming in pools, *American Journal of Epidemiology*, 165, 148–56.

Wang, H.X., Fratiglioni, L., Frisoni, G.B., Viitanen, M., and Winblad, B. (1999). Smoking and the occurrence of Alzheimer's disease: cross-sectional and longitudinal data in a population-based study, *American Journal of Epidemiology*, 149, 640–4.

White, I.R., Royston, P., and Wood, A.M. (2011). Multiple imputation using chained equations: issues and guidance for practice. *Statistics in Medicine*, 30, 377–99.

5.3

Principles of outbreak investigation

Sopon Iamsirithaworn, Panithee Thammawijaya, and Kumnuan Ungchusak

Introduction to principles of outbreak investigation

Knowledge of medicine and diseases has increased enormously over the last few decades. With the advance of knowledge, public health services in many countries can implement effective prevention programmes, and are able to protect people from many avoidable illnesses and death. However, people around the world still suffer and die from various known and unknown outbreaks of disease. Some outbreaks are old diseases that have re-emerged, some are newly identified or emerging, and some have been deliberately started. Outbreaks can occur anywhere, from a very remote area where no health facility exists to nosocomial outbreaks in a very sophisticated hospital where hundreds of health personnel are employed. Moreover, outbreaks may even spread across borders thereby threatening international health security. It is a challenge for governments and public health professionals of all countries to detect and control these outbreaks as early as possible. Outbreak investigations provide the opportunity to discover new aetiological agents, to understand factors that promote the spread of the diseases, and to identify the weaknesses of existing prevention and health programmes. For these reasons, all public health professionals should have the ability to detect, conduct, and play essential roles in supporting outbreak investigations.

This chapter provides a definition of an outbreak and describes the objectives of outbreak investigation, the methods for planning and conducting an investigation, and what needs to be done after the investigation has been completed. For simplicity, mainly examples of communicable disease investigations are discussed; however, the concepts and principles can also be applied to non-communicable diseases and injuries as well.

What is an outbreak?

The terms *outbreak* and *epidemic* can be used interchangeably. However, most people understand the term *outbreak*, which coveys a greater sense of urgency. Some epidemiologists prefer to use the term *epidemic* only in a situation that covers a very wide geographical area and involves large populations. For example, it is possible to use the term ‘outbreak of HIV’ to describe a sharply increasing human immunodeficiency virus (HIV) prevalence rate among commercial sex workers in a city where the normal rate was low in the previous year. However, the term *HIV epidemic* can be used when an abnormally high HIV prevalence is found among sex workers in many cities of the country.

In general, infectious diseases can be newly emerging but once the disease is able to sustain its transmission in the population of a geographical area we call it an endemic disease. Outbreak is used for a new emerging disease or situation when an endemic disease or health events occur at a greater frequency than normally expected within a specified population, time period, and place. There is often a misunderstanding that only communicable diseases can cause outbreaks. Non-communicable outbreaks such as mass sociogenic illnesses are sometimes reported as acute outbreaks of unexplained illness, especially in school settings (CDC 1990, 1996). Another example of an event that can be called an outbreak and needs investigation is mass casualty. Following a fire in a pub on New Year’s Eve 2009 in Thailand, an unprecedented number of 67 deaths and 153 hospitalizations resulted (Jongcherdchootrakul et al. 2011).

Because the criteria for judging an outbreak can be very subjective, it is useful to define ‘outbreak’ in more measurable terms. The criteria for judging that an outbreak has occurred can be one of the following:

The occurrence of a greater number of cases or events than normally occurs in the same place and during the same period as in past years. An example is detection of influenza outbreaks in the United States. The surveillance data from the 122 sentinel cities provides the seasonal baseline mortality attributed to influenza and pneumonia on a weekly basis, and an influenza outbreak is signalled by a rise of the mortality rates above the epidemic threshold. Another example is the community outbreak of HIV infection among injecting drug users in a rural county in Indiana, USA when 135 persons were tested positive for HIV during January-April 2015, whereas fewer than five cases of HIV infection had been reported annually in this county in previous years (Conrad et al. 2015).

A cluster of cases of the same disease occurs that can be linked to the same exposure. The term *cluster* means aggregations of relatively uncommon events or diseases in space and/or time in amounts that are believed or perceived to be greater than could be expected by chance (Porta 2008). For example, in Denmark during August 2007, the Danish clinical notification system identified many cases of *Shigella sonnei* among employees of two companies. The patients had eaten in their workplace canteen, which were served by the same catering company (Lewis et al. 2009).

A single case of disease that has never occurred before or might have a significant implication for public health policy and practice can be judged as an outbreak that merits investigation. The first documented case of avian flu (H5Nl) in the Hong Kong Special Administrative Region in a 3-year-old boy in May 1997 alerted the local authorities and scientists around the world to start a full-scale investigation (Lee et al. 1999).

How can an outbreak be detected?

Public health professionals need to maintain monitoring or surveillance of the disease situation in their local area or country, and also at the international level. It is possible to identify an outbreak by monitoring many sources of information, which will help to detect the abnormal occurrence of disease or symptom clusters. Some useful sources are listed as follows.

Health personnel

Doctors and nurses in hospitals have a good opportunity to observe an abnormal increase in the number of patients with a particular disease or syndrome. A cluster of food poisoning in a northern province of Thailand was reported to an epidemiologist during a business telephone conversation with his colleague. The epidemiologist started an investigation and identified the first confirmed outbreak of *Clostridium botulinum* food poisoning associated with home-canned bamboo shoots in the country (CDC 1999a). Without this personal contact, the outbreak would not have been investigated. Thus, public health authorities should maintain a cordial relationship with doctors and hospital staff, both in the governmental and private sectors. Conversely, doctors should report all suspected outbreaks to local public health authorities.

Laboratories

Every laboratory or network can serve as a source of support for an outbreak notification. However, it is important that the laboratory be qualified and capable of using state of the art methods. The avian flu outbreak in the Hong Kong Special Administrative Region was first discovered by the Influenza Surveillance Network, which reported an abnormal influenza type A (H5N1) (Lee et al. 1999). Without the necessary laboratory capacity, avian flu might have been overlooked and not have triggered a field investigation. A public health professional should communicate regularly with laboratory technicians and vice versa. Laboratory scientists can prevent further spread if they report abnormal findings to public health authorities regularly and without delay.

Official disease notification systems

Most countries have official systems for notification of cases and deaths of epidemic-prone diseases in hospitals. The systems are designed to detect outbreaks by comparing cases occurring in the current week or month with the average number of cases in the same area during the same period in past years.

For some diseases such as HIV/AIDS, a sentinel surveillance system was designed to monitor and detect abnormal trends in particular sentinel populations and sites. The first HIV sentinel serosurveillance in Thailand,

which started in June 1989, detected HIV prevalence of 44% among commercial sex workers in a popular northern tourist province. This finding was very alarming, and prompted a field investigation to confirm the high prevalence and to look for risk factors of HIV infection among sex workers (Siraprapasiri et al. 1991). The investigation confirmed the outbreak and revealed low levels of condom use, which led to a recommendation for promoting condom use in this high-risk population.

One of the most important functions that epidemiologists and public health professionals perform is regular analyses of reported disease data. Unfortunately, this task has been neglected, and the usefulness of disease reporting systems has been downgraded and often serves only as vital statistics reports. If this neglect of the reporting system can be overcome, the public health system will regain this powerful tool to detect and control outbreaks.

Newspapers and media

Public health professionals learn of outbreaks from the media more often than from the official surveillance system. Newspapers receive outbreak news directly from their reporters or people in the community, and are able to report them immediately. The Programme for Monitoring Emerging Diseases (ProMED), operated as an official programme of the International Society for Infectious Diseases was the prototype for a communication system that monitors emerging infectious diseases globally, which obtains much of its outbreak news from the domestic or international media reports, official reports, online summaries, local observers, and its subscribers. Although timeliness is the strength of the media, the validity of the information is often of concern, so media reports require verification.

Internet and social media

In addition to conventional newspaper and media, people increasingly use internet and social media (such as blogs, Facebook, Twitter, MySpace, etc.) as new ways to connect themselves with their families, friends, colleagues, and the world. Massive amounts of information on those internet blogs and social media are produced and shared extremely quickly, virtually real-time, and therefore create opportunities to epidemiologists and data scientists to

apply advanced data analytics or data mining techniques for detecting and tracking disease outbreaks or hidden health-related phenomena. However, sorting through this massive amount of data presents a real challenge. Although analytical techniques for utilizing internet and social media reports have not been widely utilized by public health organizations or government disease surveillance units, several studies have shown the potential of social media for outbreak detection (Corley et al. 2010; Charles-Smith et al. 2015).

Village leaders, health volunteers, and concerned citizens

In rural areas where there are no health facilities and communication is limited, village leaders or health volunteers can often help to recognize an abnormal increase in the numbers of some clinical diseases, such as diarrhoea, chickenpox, fever with rash, deaths from unknown aetiology both in humans and animals. For example, between 2004 and 2005 there was a nationwide outbreak of avian influenza (H5N1) in poultry. The Thai Ministry of Public Health asked the health volunteers to report any abnormal deaths of poultry in their village to local health officers and the animal health authorities. With these notifications from villagers and a policy to provide compensation to households that lost chickens through culling of the suspected infected flock, Thailand was able to control the highly pathogenic avian influenza outbreak in poultry that then resulted in only 25 human cases.

In many situations, public health officials received outbreak news from concerned citizens and key persons in community networks. Those include teachers, parents, religious leaders, safety officers of factories, etc. who observe a cluster of illnesses in their facilities. Good relationships and regular communications between surveillance officers and the community networks can be great portal to rapidly detect an outbreak in a community.

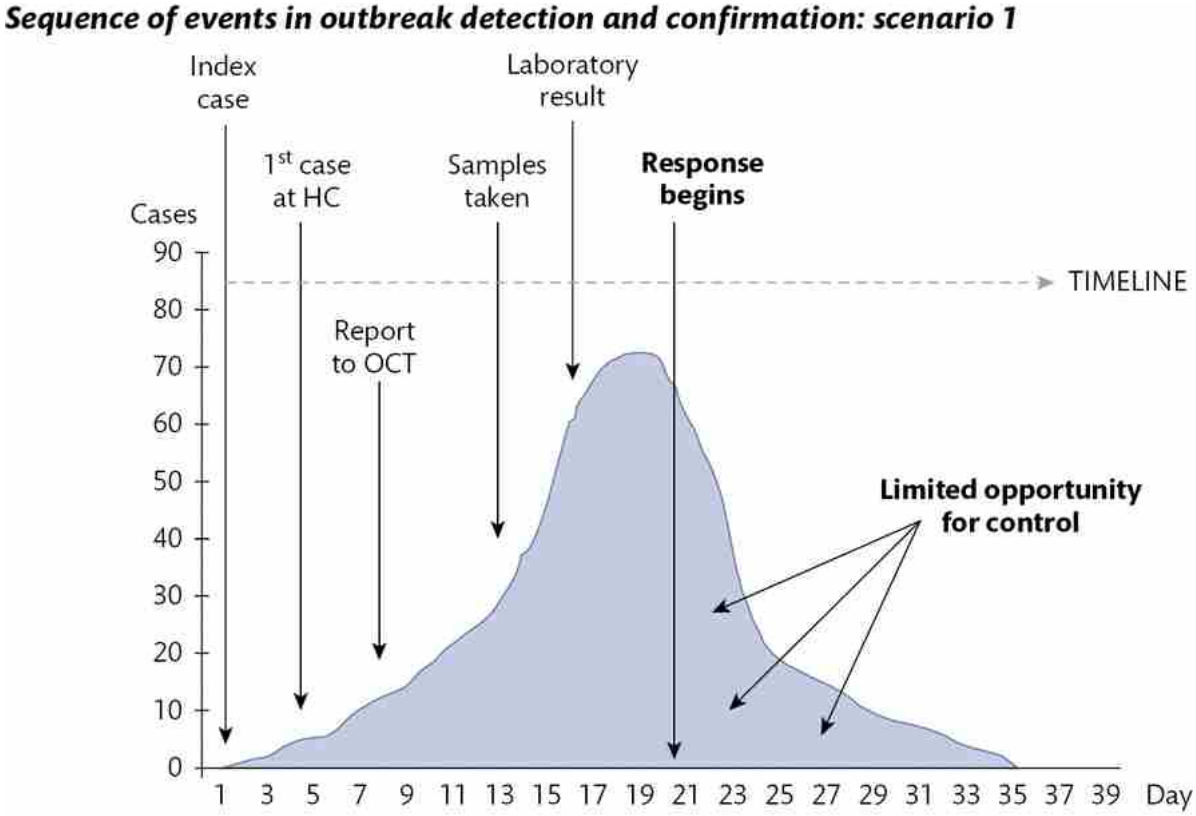
Purposes of outbreak investigation

An outbreak investigation can have many purposes, but the critical ones are as follows.

Controlling the current outbreak

This should be the primary ultimate goal. If the investigation can start early, the findings can guide implementation of appropriate control measures to stop further spread. The avian influenza (H5N1) outbreak investigation found a link between infection and illness in chickens and humans. The same virus was found in both. There were a total of 18 cases and six deaths before the Hong Kong Special Administrative Region decided to kill all of the 1.5 million chickens in the islands within 3 days to end the outbreak (Lee et al. 1999). There have been no human cases in Hong Kong since then. To achieve the goal of controlling the current outbreak, it is necessary to eliminate the delay in detecting the outbreak, to start an investigation as soon as possible, and to immediately implement appropriate preventive and control measures indicated by the investigation results. With early case detection and timely reporting to an outbreak control team, an opportunity of reducing a disease burden is potentially greater following the prompt response (Fig. 5.3.1).

Sequence of events in outbreak detection and confirmation: scenario 1



Sequence of events in outbreak detection and confirmation: scenario 2

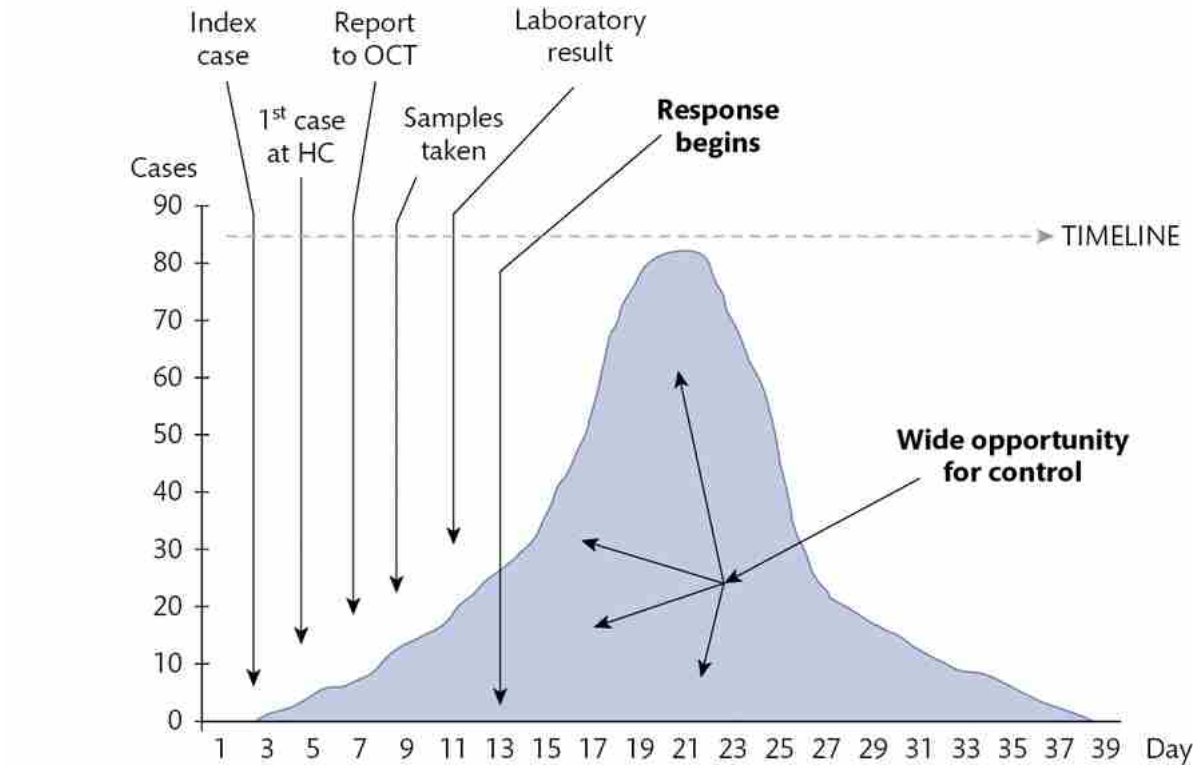


Fig. 5.3.1 The impact of early detection of and response to outbreak in reducing a disease burden. Reproduced with permission from Connolly M.A (ed.), *Communicable Disease Control in Emergencies: A Field Manual*, World Health Organization, Geneva, Switzerland, Copyright © World Health Organization 2005, http://www.who.int/hac/techguidance/pht/communicable_diseases/field_manual/en/.

Prevention of future outbreaks

Not all investigations start at the beginning or before the peak of the outbreak. The findings or lessons learned from the investigation may be too late to help fully control the current outbreak, but they can still contribute to prevention of future outbreaks. Through sound investigations, the weaknesses of the prevention programmes and/or necessary measures can be identified. If recommendations are taken seriously, the chance of recurrence of the same outbreak or other diseases that share common risk factors can be reduced. There was a thorough investigation of a reported case of a neurological syndrome after eating a common local dish of fish-porridge in Bangkok during October 2002. The investigation revealed that the meat was from puffer fish, which is widely used as a substitute because it is less expensive and easy to obtain from large markets. The result of the investigation was presented to the authorities of the Thai Food and Drug Administration. A new regulation to prohibit selling of puffer fish was enacted to minimize a risk of tetrodotoxin poisoning from puffer fish consumption in the country (Samitsuwan et al. 2005).

Research to provide knowledge of the disease

Information about new diseases and their natural history, clinical spectrum, incubation periods, etc. can often be best learned during an outbreak investigation. Outbreak investigation can help ascertain new information, for example, modes of transmission, population at risk, and so on, and so provide essential information which can significantly contribute to the medical/public health literature. The outbreak of encephalitis in Malaysia, which continued until the end of April 1999 and resulted in 257 cases and 100 deaths, prompted an international outbreak investigation, resulting in the discovery of a new virus, Nipah (CDC 1999b, 1999c). The mode of transmission was close contact with infected pigs through their blood and secretions.

Evaluation of the effectiveness of prevention programmes

Investigation of an outbreak of disease, which is the target of a public health programme, may reveal weaknesses in that programme. Investigation of an outbreak of vaccine-preventable diseases often identifies populations that have not received the vaccine. For example, investigation of a measles outbreak that occurred in 1993 in Espindola, a rural community in the Peruvian Andes, revealed that more than a quarter of the 553 residents were affected, and that more than 3% of those who contracted measles had died. One year before the outbreak, a national measles campaign targeting children younger than 15 years had been conducted. Although national data reported the coverage to be 78%, the investigation revealed that only 4% of the children in Espindola had actually been vaccinated (Sniadack et al. 1999).

Evaluation of the effectiveness of the existing surveillance system

An effective surveillance system should be able to detect a cluster of patients with the same diseases or syndromes. However, outbreaks of diseases under surveillance may emerge without the notice of health officers, especially when there is a shortage of surveillance staff or a change in the surveillance procedure. Some attributes of the surveillance system can be evaluated during an outbreak investigation, such as the timeliness, validity, sensitivity, appropriateness of case definitions, and use of the surveillance information for outbreak detection.

Training of health professionals

The Epidemic Intelligence Service of the US Centers for Disease Control and Prevention (CDC) and 69 field epidemiology training programmes around the world use real outbreaks as an opportunity for training health professionals, as well as to provide emergency services by investigating the causes and determinants of outbreaks. Recommendations and actions taken by these applied epidemiology programmes contributed to a limited scale of outbreaks and improved the health of the public.

Responding to public, political, and legal concerns

In many situations, an investigation must be conducted because the media has publicized the complaints of people to politicians, or even when there are widespread rumours. The main objective for this kind of investigation is to verify the outbreak and diagnosis. If it is groundless, the investigator can supply the media with new information that can end the rumours. Conversely, if it is a true outbreak, investigators must decide what steps need to be taken.

In general, for a real outbreak, many objectives can be achieved fully or at least partially. However, the ultimate goal is to control the current outbreak and to prevent future ones. It is unethical for investigators to compromise this ultimate goal with other goals such as training or non-essential research that does not directly contribute to control activities.

Components of an investigation team

In this chapter, the term *investigators* represents people who are directly involved in planning and conducting outbreak investigations from start to finish. In principle, local health professionals at the district or provincial level should assume the role of investigators and start work as soon as possible. For complicated or difficult field investigations, additional disciplines or international experts can provide assistance. It is best to form an investigation team with a team leader in charge of the operation. A good investigation team should include the following:

A field epidemiologist who is technically competent to conduct field investigations systematically. The field epidemiologist usually serves as one of the primary investigators, and should be involved in all the investigative steps.
Disease control specialists who are experienced in implementing basic disease control measures such as food and environmental sanitation, vector control, and vaccination. If available, an educator who can provide essential knowledge to villagers in clear terms is also very useful for disease control implementation.
Laboratory technicians who are able to provide basic and advanced laboratory support to the investigation team. They might not need to travel to the field and collect the specimens themselves, except when a special collection procedure is required.
Specialists in particular areas (e.g. veterinary medicine) would be very helpful for an outbreak investigation of zoonotic diseases. An entomologist is a key team member for an outbreak investigation of vector-borne diseases.
A social scientist with expertise in qualitative methods can help identify risk behaviours among affected populations and assess the acceptability of the recommended interventions.
Public health administrators who are good at providing logistic support, mobilizing resources, and providing administrative expertise for the team.
In certain situations, when the outbreak has caused panic or has gained the intense attention of the public, the investigation team should recruit or appoint a person to be in charge of public relations and press releases. This person should appropriately reassure and not unduly alarm the public.

In practice, all of these team members are not always available at the subdistrict or district level, due to limited human resources. Public health professionals and field epidemiologists need to have basic knowledge of all these relevant disciplines and be able to assume these tasks if needed. In the event of a technical problem or the need for consultation from the field, telecommunication equipment should be provided to field investigators.

Due to the sudden nature of a field investigation, it is best to establish in advance a list of people who will be on call and are ready to join an investigation team once an outbreak has occurred. Most countries including Thailand developed ‘Rapid Response Teams’ at their appropriate administrative unit for responding to local epidemics and/or public health emergencies. Public health officers of those teams received practical

epidemiology training, and subsequently became frontline armies tracking outbreaks in their local area.

Issues to be considered before implementing an investigation

The team leader should consider all of the following issues before initiating a field investigation.

Assessing the existence of the outbreak

No matter how outbreak news is obtained, an investigator should confirm the validity of the information. The best way is to have direct communication with the responsible local health authority or field staff. It is not unusual for the information to be groundless. Sometimes, the outbreak has in fact occurred, but the media incorrectly quoted the name of the location. The investigator should carefully check with all other possible local health authorities so as not to overlook the outbreak.

Gathering available basic information

If the local health authority or field staff confirm the existence of the outbreak, the investigator should ask for additional information related to the situation and control measures being implemented. It is not necessary to gather all of the following listed information before leaving for the field, but having it will help investigators to plan an effective investigation.

Information related to the disease situation

What are the main symptoms and signs of the patients?
By whom and how was the diagnosis made; for example, using only clinical or also laboratory evidence?
How many patients were seen and how many died?
What was the average age of the patients and were there any differences in gender distribution?
Where did the patients come from? Are they clustered or scattered?
When was the increased number first observed and what is the trend at the moment?

Information related to control and response activities

What has already been done in terms of the field investigation and implementation of control activities?
Are there any serious constraints that compromise the field investigation and/or implementing control measures?
Who are the key people responsible for the investigation and control activities?

Ensuring that clinical specimens and suspected materials are collected in a timely and appropriate manner

It is absolutely critical to contact the doctors who saw the cases and made the diagnosis to obtain relevant clinical specimens such as serum and blood for future laboratory tests. The items to which the cases were exposed, such as food and water, should be collected immediately before anything is unintentionally destroyed or untestable. The investigator should contact the local and reference laboratories so that the necessary supplies and equipment can be obtained immediately.

Obtaining permission and adequate support from local and national authorities

The investigation team should ask the permission of the local health authorities, which will create a sense of shared responsibility and partnership. Usually, local authorities are pleased to receive assistance. In a few situations, the local authority might be unhappy about having outsiders conduct the field investigation because of the sensitive nature of the problem. The investigator needs to convince local authorities that a thorough and sound investigation will incur more benefits than harm. The investigator should also request field support from the key authorities, such as field staff who will facilitate the fieldwork and provide transportation, medication, etc. In some situations, if the outbreak is possibly of international concern spelled out by the International Health Regulations (IHR) (Connolly 2005), the principal investigator should inform national health authorities so they are aware and can plan further necessary steps to report to the World Health

Organization (WHO) and deal with the media and international communities.

Planning the field operation

The investigator needs to have a short meeting with team members to summarize the situation, set up the objectives of the investigations, divide responsibility among team members, and check the readiness of laboratory and logistical support.

It is also important to plan the duration of the field investigation. The investigation team should stay in the field until all investigation processes such as data collection, analysis, interpretation, and the executive summary have been completed. Leaving the field without accomplishing all these objectives could delay the implementation of control measures. Most outbreak investigations should plan to obtain preliminary results and recommendations within a week of the start of the investigation to ensure that the findings will be timely enough for control of the current outbreak. Additional studies and subsequent investigations can be conducted later.

The investigation team should not spend too much time preparing a perfect plan, because of the urgency of the outbreak; rather, they should obtain what is most necessary and start the investigation as soon as possible.

Steps of outbreak investigation

An outbreak investigation is an observational study by nature, because the events have already occurred. Every outbreak investigation needs to start with a good descriptive study, followed by analytical studies whenever possible and necessary. Conclusions about the causes, sources, mechanisms, and determinants of the outbreak need to be drawn, based on sound epidemiological, clinical, laboratory, and environmental evidence.

A descriptive study can help to identify the risk population(s) and risk area(s) so that immediate interventions can be directed to those who need them the most. A good descriptive study can also generate hypotheses about how the outbreak has spread and what factors contributed to the abnormal occurrence of the disease. In theory, hypotheses derived from a descriptive study should be confirmed by further analytical study; in reality, this is not always possible, due to many constraints.

It is preferable to translate the methodology for outbreak investigation into steps of action. Gregg (1996) has divided the outbreak investigation process into ten steps; with a slight modification, this ten-step investigation process has been summarized in Box 5.3.1. Steps 1–4 use descriptive epidemiology to generate basic facts and hypotheses, steps 5–7 are processes to test the hypotheses and draw conclusions, and steps 8–10 emphasize the importance of communication of the results and follow-up of the recommendations.

Box 5.3.1 Ten steps to take in an outbreak investigation

1. Confirm the existence of the outbreak.
2. Verify the diagnosis and determine the aetiology of the disease.
3. Develop a case definition, start case finding, and collect information on cases.
4. Describe persons, places, and times, and generate hypotheses.
5. Test the hypotheses using an analytic study.
6. Carry out necessary environmental or other studies to supplement the epidemiological study.
7. Draw conclusions to explain the causes or the determinants of the outbreak, based on clinical, laboratory, epidemiological, and environmental studies.
8. Report and recommend appropriate control measures to concerned authorities at the local, national, and if appropriate, international levels.
9. Communicate risk to mobilize cooperation from other public health professionals and the general public.
10. Follow-up the recommendations to ensure implementation of control measures.

Source: data from Gregg, M.B. (ed.), Conducting a field investigation, p. 44–59, in *Field Epidemiology*, Oxford University Press, New York, USA, Copyright © Oxford University Press, Inc. 1996.

This outline does not imply a strict course of action. In real outbreak investigations, many steps may be undertaken concurrently, depending on the situation.

Step 1: confirm the existence of an outbreak

The main question is: ‘Is this a true outbreak?’ Applying the definition of an outbreak outlined in this chapter, the investigator should be able to establish or refute the existence of the outbreak. Investigators should review the number of cases with the local health officers or hospital staff and compare it with the number found during the same period recorded in past years.

For example, the outbreak of trichinosis in North Rhine-Westphalia, Germany, was confirmed because there were 52 cases in a 3-month period between November 1998 and January 1999, compared with no more than ten cases annually during the same time period in the past 10 years (CDC 1999d).

Step 2: verify the diagnosis and aetiology of the disease

If the number of cases fits the case criteria for the outbreak, the next related questions are:

What is the correct diagnosis and aetiology of the disease?
What can be done immediately to prevent new cases from occurring?

Knowing the exact diagnosis and aetiology of the disease will help to establish appropriate preventive measures immediately. This will protect susceptible people and allow the team to start educating villagers to avoid the risk factors. For example, many adults in a remote village were ill with fever, muscle and joint pain, rashes, etc. If the diagnosis and aetiology of the disease is unknown, the local public health officials will find it very difficult to educate the public or implement effective preventive programmes. Until the serology of some patients showed sharply rising immunoglobulin M (IgM) antibodies to dengue virus, control measures to destroy the larvae of the *Aedes* mosquito, the vector of dengue which breeds in water containers, may not be started.

In many situations, an unknown or unclear diagnosis can cause panic due to rumours. This was demonstrated by an outbreak of pneumonic plague in Surat, India in 1994, and of encephalitis in Malaysia in 1999, which resulted in severe panic among local people and foreign tourists. Thus, it is very important to obtain the exact diagnosis as rapidly as possible.

Investigators should have basic knowledge of the clinical diagnosis and how to confirm the aetiology of suspected disease by using well-established laboratory techniques. It is recommended that investigators visit and talk with some patients, review and visualize the signs and symptoms, and hold discussions with the attending doctors. The information from this step will help to develop a case definition to facilitate active case finding. The information on aetiology will also help to interpret the findings from the later descriptive study and establish a causal association.

The investigators should also visit the laboratory facilities and ask for either positive or negative results of the testing of specimens. It is not necessary for all cases to be laboratory confirmed, but at least some of the apparent clinical cases or deaths should be confirmed. Once there are laboratory-confirmed cases, the investigator will find it more reasonable to assume that other cases with the same clinical manifestations in the same period and location are the same disease.

It is unfortunate that specimens from patients are often thrown away when the primary results (either positive or negative) are obtained. The investigators should plan with the doctors and laboratory scientists to perform further investigations on the specific strain, and to establish drug sensitivities, genetic markers, etc. Many new laboratory technologies, such as serological testing, culture and isolation, and molecular techniques, are very powerful for diagnosing and tracking the connections between cases, even in different geographical areas. For example, outbreaks of *Shigella sonnei* infections in Denmark and Australia were detected during the same week in August 2007, and 215 laboratory-confirmed cases were reported in Denmark, with an additional 12 cases in Australia. Antibiotic resistance and pulsed field gel electrophoresis in isolates identified in those two outbreaks were indistinguishable. A hypothesis that the patients were exposed to the same implicated food item from the same original source was soon proven (Lewis et al. 2009).

Step 3: develop an appropriate case definition, start case finding, and collect information on cases

At this stage, the investigator needs to answer at least three questions:

Who should be counted as a case?
Are there more undetected cases in the hospitals and in the community?
What are the characteristics of cases?

To answer these three questions, the investigators must follow three steps.

Develop an appropriate case definition

It is important that the investigator develop a case definition, which will be applied consistently during the investigation. The definition should be

sensitive or adequate at the beginning, in order to capture actual cases. A good case definition for investigative purposes should be specific to time and location. The case definition should not include any specific suspected exposure that the investigator plans to verify. This would create selection bias when the investigator generates the hypothesis or tests the hypothesis in the subsequent steps. Using the information obtained from the previous steps, the investigator can divide the case definition into different levels.

For example, in September 2012, when two patients with a novel coronavirus from Saudi Arabia and Qatar were found, WHO requested member states to start surveillance of the disease by using the interim case definitions as shown in Box 5.3.2 (WHO 2012). Note, there was a further update in 2014 to reflect the MERS-CoV outbreak, and doubtless an interim for SARS-CoV-2 is forthcoming.

Box 5.3.2 2012 WHO interim case definitions for a novel coronavirus

1. Patients to be investigated (referred to as ‘patient under investigation’):
    - A person with an acute respiratory infection, which may include fever (≥38°C, 100.4°F) and cough; *and*
    - Suspicion of pulmonary parenchymal disease (e.g. pneumonia or acute respiratory distress syndrome (ARDS)), based on clinical or radiological evidence of consolidation; *and*
    - Travel to or residence in an area where infection with novel coronavirus has recently been reported or where transmission could have occurred (see updated area in WHO website); *and*
    - Not already explained by any other infection or aetiology, including all clinically indicated tests for community-acquired pneumonia according to local management guidelines.
2. Probable case
    - A person fitting the definition given in (1) of a ‘patient under investigation’ with clinical, radiological, or histopathological evidence of pulmonary parenchyma disease (e.g. pneumonia or ARDS) but no possibility of laboratory confirmation either because the patient or samples are not available or there is no testing available for other respiratory infections, *and*
    - Close contact with a laboratory-confirmed case, *and*
    - Not already explained by any other infection or aetiology, including all clinically indicated tests for community-acquired pneumonia according to local management guidelines.
3. Confirmed case
    - A person with laboratory confirmation of infection with the novel coronavirus.

Source: data from World Health Organization, Global Alert and Response (GAR): WHO Interim Case Definition—Novel Coronavirus, as of 29 September 2012, Copyright © World Health Organization 2012, available from http://www.who.int/csr/disease/coronavirus_infections/case_definition_29092012/en/index.html.

Active case finding

In places where there is good surveillance, cases from all hospitals can be reported to the epidemiology unit at the district or provincial level. The investigator can apply the case definition and count the number of cases, and review data that are collected on the reporting form. If enough cases and basic information on cases have already been gathered, the investigator can start descriptive epidemiology. Conversely, if only a few cases have been seen at the health facility, the investigator should plan to search for more cases in the community. The investigator must start a process called *active case finding*. The objective of active case finding is to gather enough cases to analyse. At the same time, this case finding will give a better picture of the magnitude of the outbreak. The investigator must frequently visit many

hospitals in the outbreak area and review medical records themselves. Sometimes, they must search for cases in each village by interviewing house to house, which is the true nature of outbreak field investigations and the way people learn epidemiology. This active case finding in the community also provides two more benefits:

Control measures can be implemented if the aetiology of the disease is known and treatable. During an investigation of influenza A (H1N1) 2009 outbreak in a prison in August 2011, the investigator conducted active case finding and laboratory investigation to confirm an aetiology. A total of 421 case-patients with an attack rate of 19% were identified. A mobile clinic from the provincial hospital was deployed into the prison to provide early diagnosis and oseltamivir was given to 17% of cases to prevent complication. Meanwhile, multiple intensive control measures including promoting hand washing, mask use, and case isolation were implemented to reduce the transmission in the prison. (Karnjanapiboonwong et al. 2011)
Rapid environmental assessment can be started during the visit to the affected families or villages. From the direct interview with the cases, the investigator can develop hypotheses and implement necessary interventions immediately, such as sanitation improvement among food handlers and treatment of water to prevent a waterborne outbreak.

In situations in which the outbreak is not localized but widespread, the investigator might need to use the media to alert the public about the outbreak. People can then avoid suspected exposures and see a healthcare provider if they have developed symptoms compatible with the case definition.

Collecting information on cases

For each case, the investigator should collect at least four types of information:

Identifying information: name, hospital number, contact person, and address of contact. This additional information will help to avoid duplication of enumerated cases. The investigator can also maintain communication with these cases when more information is needed.
Demographic information: age, gender, occupation, religion, ethnicity, area of residence, place of employment, etc. This important information may help to determine the characteristics and distribution of cases.
Clinical information: symptoms and signs, date of onset, duration of illness, and results of diagnostic procedures. These data will help to confirm a true case, provide the pattern of clinical manifestations, and also the distribution of cases by time.
Suspected risk factors: investigators can ask for a history of exposure to factors before the disease developed. The timing of interest is usually one incubation period if the aetiology is known or suspected. Questions about contact with other patients who have similar clinical symptoms are also very helpful. If the diagnosis is not known, the investigator should collect this information in a qualitative manner.

The investigator should develop a questionnaire as a tool to collect the relevant information from the hospitals or during active case finding in the community. In practice, some of the information will not be available or not of good quality. The investigator should validate all data in doubt.

In addition to using a traditional paper-based questionnaire, development of information and communication technology allows epidemiologists to use mobile devices, such as smartphones or tablet computers, to collect data during outbreak investigation. The power of mobile technology will enable outbreak response teams to collect and transfer data rapidly and smoothly. For example, during the 2014–2016 epidemics of Ebola virus disease in West Africa, mobile devices and related applications were used to collect data when conducting contact follow-up and identifying new cases (Tom-Aba et al. 2015). Using the right technology in the right situations will therefore effectively facilitate outbreak situation awareness and prompt response.

Step 4: describe the outbreak in person, place, and time, and hypotheses formation

In this step, investigators need to answer the following questions:

What are the main clinical features?
What population(s) is at risk?
What are the risk factors?
What are the most likely explanations of how the outbreak began?

A simple approach is to analyse clinical information from each case and see the distribution of cases in terms of person, place, and time. The analysis should be done using rates rather than absolute numbers; the investigator needs to obtain the denominators from an available source or to estimate them. Using rates, the investigator can compare and determine the populations and areas of highest risk. With the advent of computers, many software programs are available to analyse these data. A popular one is EpiInfo™, which is public domain software from the US CDC that is designed specifically for field investigations. In the absence of a computer, investigators can still analyse data manually. Individual questionnaires can be compiled into a line listing, which includes important variables of all cases. With this line listing, simple counts can be made. In this way, investigators will gain knowledge about populations and areas of risk. Resource and control measures can then be directed to the risk populations and risk areas. Enumeration will also produce information for hypothesis formation to explain how and why the outbreak occurred.

Clinical manifestations of cases

Signs and symptoms of cases can be analysed in percentages and presented in a summary table. In an outbreak of unknown aetiology, clinical information will help to establish the diagnosis. For an outbreak in which the aetiology is already known, the investigators still need to compare the clinical information found in the investigation with previous knowledge. Any discrepancy between the investigation and previous knowledge such as the attack rate, mortality rate, severity, and so on should be carefully examined, because this might indicate that a new strain or different specific host response is occurring. The high mortality rate of approximately 40% in the Nipah encephalitis outbreak in Malaysia 1999 (CDC 1999c) indicated that this outbreak might not be due to the usual endemic encephalitis.

Populations at risk

Investigators should analyse the characteristics of cases by gender, age, occupation, ethnicity, and so on. Initially this can be carried out by examining the proportion of all cases, but the specific attack rate by age and gender will be more useful for comparisons and hypothesis formation. The rates will provide useful indicators of the possible aetiology of the outbreak. In the Ebola haemorrhagic fever outbreak in Zaire in 1976, all ages and both genders were affected, but females slightly predominated. Age- and gender-specific attack rates indicated that adult females had the highest attack rate. This finding suggested that parenteral injection with unsterilized syringes and needles given during antenatal care in the local hospital was the means of transmission (WHO 1978).

The outbreak of Nipah encephalitis in Malaysia in early 1999 was initially thought to be due to Japanese encephalitis alone. However, after careful analysis of the descriptive information, it was clear that the cases were mainly male, adult, involved in pig farming, and of Chinese ethnicity. This descriptive information did not fit the pattern for Japanese encephalitis, which mainly affects children of both genders, with no preference for a particular ethnicity. The investigator then began to suspect other organisms and to hypothesize that pig farming increased the risk of becoming ill.

Primary, index, and outlier cases

In infectious diseases, the first case on the epidemic curve is called the primary case. The primary case is important, because of the possibility that he/she brought the disease into the community either from the local area or from other locales. In investigations, sometimes the first case that draws the attention of the investigator is not the primary case. We call this the ‘index case’. Index cases may be detected at the hospital or during the active case finding. The investigator may find the primary case among those who were sick in the community but did not seek medical care. Cases that appear at the very beginning or at the end of the epidemic who are markedly different from the rest of cases should also be given careful attention. These cases are called *outlier cases*, and can provide important information about the source and the way in which the disease is spread. A tourist who happened to be only a short time in the village during the outbreak and got sick is an example of an outlier.

Typically, the very first and very last cases in the epidemic curve should be critically appraised. The first case may not be the true index case because of misdiagnosis, a case unrelated to the epidemic, etc. The late cases may be due to misdiagnosis, a case unrelated to the epidemic, or a secondary case that had a different exposure than the majority of the cases in the epidemic.

During the outbreak of severe acute respiratory syndrome (SARS) in February 2003, the index case was identified as a 65-year-old medical doctor from Guangdong, mainland China, who stayed on the ninth floor of a hotel in Hong Kong (Fig. 5.3.2). He had treated patients with atypical pneumonia prior to his departure, and he was symptomatic upon arrival in Hong Kong. He infected over a dozen guests and international visitors. During the outbreak investigation, the investigators tried to identify contacts of this index case to prevent further transmission. A history and information of this index case obtained by the investigation team traced these cases back to previous atypical pneumonia outbreaks or undiagnosed SARS since late 2002 (World Health Organization Regional Office for the Western Pacific 2006).

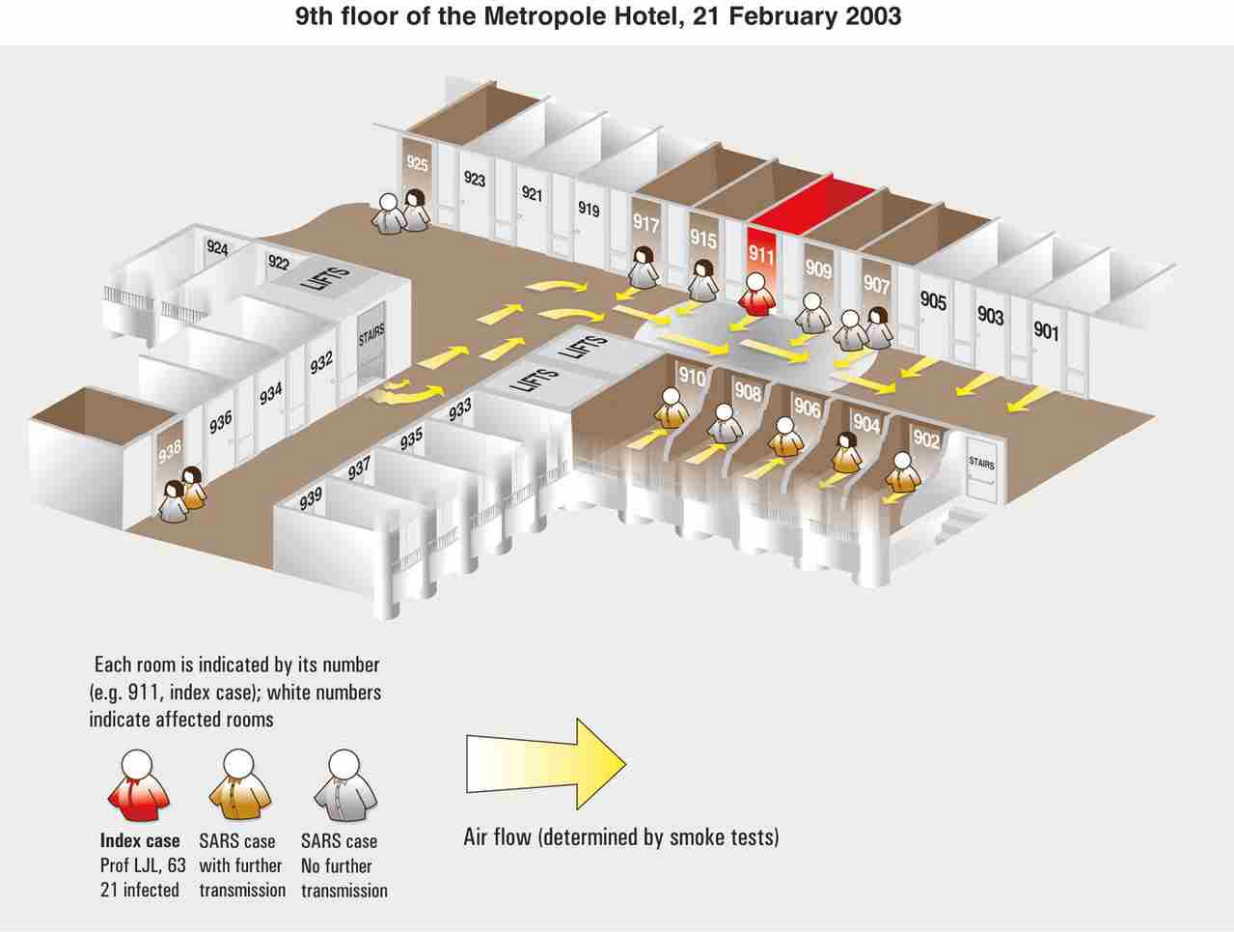


Fig. 5.3.2 Index case of SARS—Hong Kong, 2003.
Reproduced with permission from World Health Organization Regional Office for the Western Pacific, *SARS: How a Global Epidemic was Stopped*, World Health Organization, Western Pacific Region, Manila, Philippines, Copyright © 2006.

Location

Investigators can calculate attack rates of cases in different locations. These can be places of residence, places of employment, sites of exposure, etc. Locations with high attack rates often indicate the sources of infection or contamination. A spot map showing the locations of cases can give a very good idea of the source, as demonstrated by Snow in his classical investigation of a cholera outbreak in the Golden Square area of London between August and September 1854 (Frost 1936). In that investigation, Snow found most cases clustered around the Broad Street pump. From this information, Snow deduced that the Broad Street pump was probably the primary source of the outbreak.

If cases are scattered in many locations, investigators should explore the secular pattern of the cases over time, which will indicate whether the outbreak started in one area and spread to other areas, or whether people living in different places had a common exposure. The SARS outbreak is a good example of when cases are distributed worldwide originally from a single source of infection.

Time

The objective is to show the occurrence of cases over time and look for a pattern of occurrence. In general, there are two major types of outbreaks: a common source and a propagated source. The way to differentiate between these two patterns of outbreak is to draw an epidemic curve that shows the number of cases (on the *y*-axis) over time of onset (on the *x*-axis). The epidemic curve of each outbreak will suggest whether the mode of spread is by a common source or from person to person.

Common source outbreak

This kind of outbreak occurs when people are infected by exposure to the same source of infection. For example, a group of people contract viral hepatitis A because they ate the same contaminated food served during a wedding party. A common source outbreak can be divided into a point common source and a continuous common source:

A point common source outbreak occurs when there is a single source that exists for a very short time and all cases have common exposure to it at that particular time.
A continuous common source outbreak occurs when there is only one source that provides continuous or intermittent exposure over a longer period of time.

Epidemic curve of a point common source outbreak: the epidemic curve shows a sharp increase of many cases suddenly followed by a rapid decline, although not as rapid as at the beginning of the epidemic. Another criterion to judge a point common source is that the first and the last cases usually occur within one incubation period.

If the aetiology and the data of the incubation period of the disease are known, the investigators can then roughly estimate the probable time of the initial exposure. This can be done by identifying the peak of the epidemic from the curve and counting back on the *x*-axis the equivalent of the incubation period. The investigator can also use the first case and count back on the *x*-axis the duration of the minimum incubation period, which will also give a rough estimate of the time of exposure. In the opposite way, if the information on aetiologic agent and incubation period is unknown but the common exposure time is known, the mean incubation is the duration from exposure to the peak of the curve. The shortest incubation period is the duration from the time of exposure to the onset of the first case and the longest incubation period is the duration from the time of exposure to the onset of the last case.

If the aetiology of the disease is not known, but the epidemic curve fits well with a point common source outbreak, the investigator can estimate the average incubation period if he or she knows the times of common exposure.

Epidemic curve for a continuous common source outbreak: the epidemic curve shows an abrupt increase in the number of cases, but instead of having a peak and a decline within one incubation period, new cases persist for a longer time, with a plateau shape instead of a peak before decreasing. In the outbreak of food poisoning caused by *Shigella sonnei* infections linked to consumption of imported raw baby corn in Denmark and Australia in August 2007, the peak of the outbreak was observed 6 days after the onset date of the first case and gradually declined (Fig. 5.3.3). All of the cases developed symptoms within a 2-week period, and the outbreak subsided soon after a recall of the implicated batches of baby corn on 17 August 2007 (Lewis et al. 2009). However, if there are many peaks or irregular jagged curves, it suggests an intermittent common source.

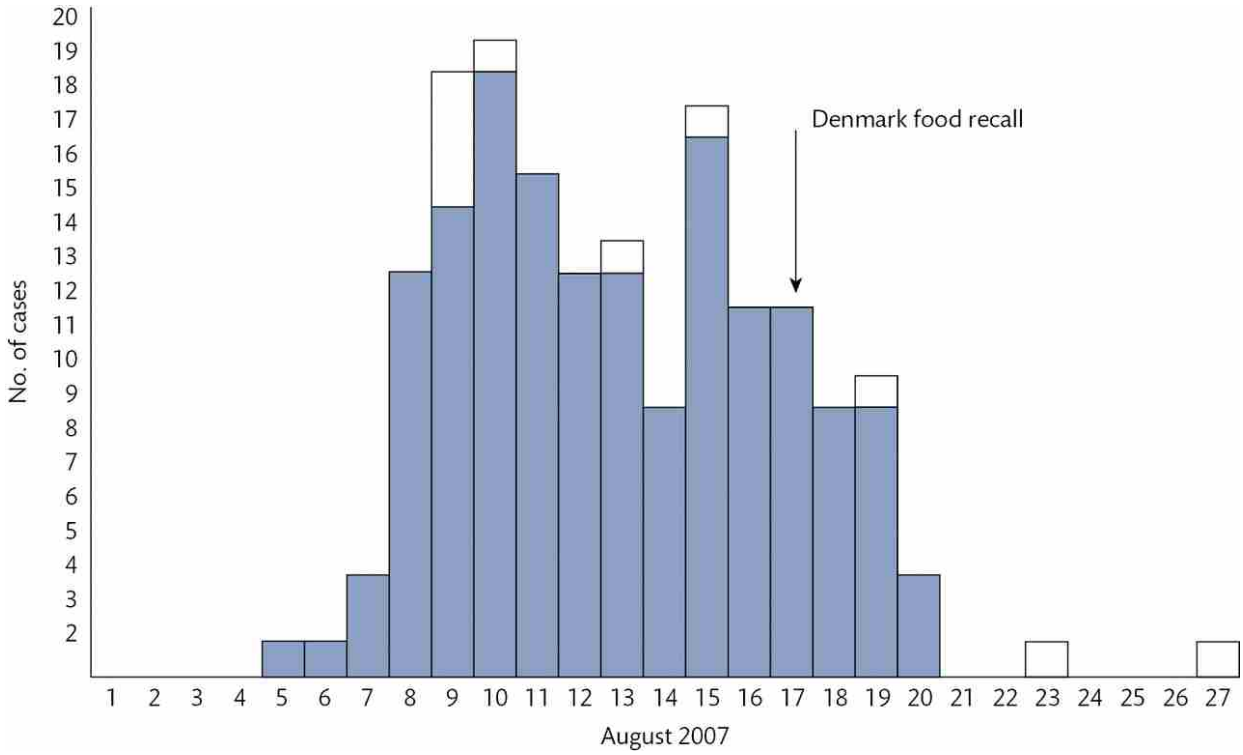


Fig. 5.3.3 Dates of illness onset for 163 cases in an outbreak of *Shigella sonnei* infections linked to consumption of imported raw baby corn in Denmark (■) and Australia (□) in August 2007.

From Lewis HC et al., Outbreaks of Shigella sonnei infections in Denmark and Australia linked to consumption of imported raw baby corn, *Epidemiology and Infection*, Volume 137, Special Issue 03, pp. 326–34, Copyright © 2009 Cambridge University Press, reproduced with permission, DOI: http://dx.doi.org/10.1017/S0950268808001829.

Propagated source outbreak

A propagated outbreak is caused by transmission from one person to another, and requires direct contact such as touching, biting, kissing, or sexual activity, or via vector such as in vector-borne disease outbreak.

Epidemic curve of propagated source outbreak: the epidemic curve shows a slow increase in the number of cases, with progressive peaks approximately one incubation period apart. The investigators might observe an abrupt decrease in new cases, because everyone has already been infected. The span from the first to the last case will also last longer than several incubation periods.

The outbreaks of SARS in Singapore demonstrated the propagated source of transmission beginning from a young female who returned home from Hong Kong. This patient had stayed on the same hotel floor as the index

case. She had initially infected many hospital staff before the disease spread for further generations (Kamps and Hoffmann 2003). The outbreak resulted in 238 cases and 33 deaths in a few months (Fig. 5.3.4).

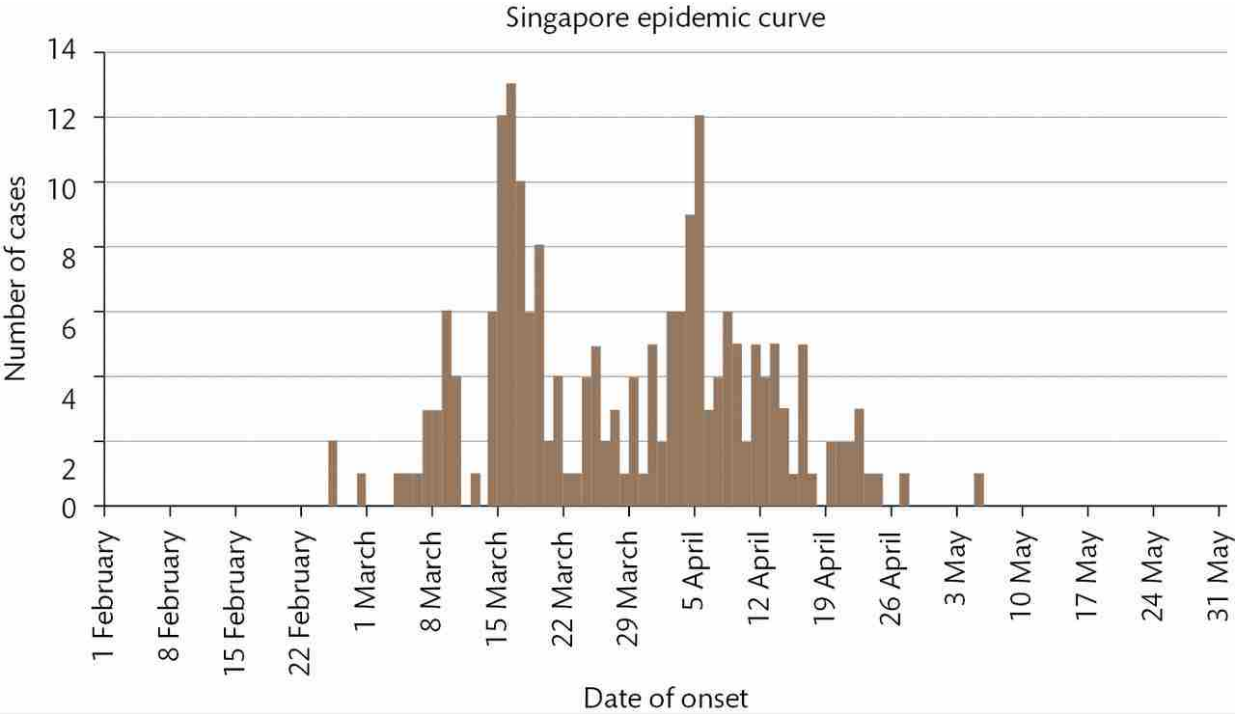


Fig. 5.3.4 SARS cases by date of onset—Singapore, 2003.

Reproduced with permission from World Health Organization Regional Office for the Western Pacific, *SARS: How a Global Epidemic was Stopped*, World Health Organization, Western Pacific Region, Manila, Philippines, Copyright © 2006.

Step 5: testing the hypotheses by an analytic study

In an outbreak of infectious disease, the investigator needs to answer the following questions:

What is the aetiology of the disease?
What is the source of infection?
What is the pattern of spread?
What are the risk factors for an individual to get the disease?
What are the determinants of the outbreak or the factors which, when combined together, result in the outbreak?

The aetiology of the disease should be derived from the laboratory study. The pattern of spread can be identified by a careful descriptive study. Sometimes it is not as simple as expected, and a descriptive study does not give enough clues. The investigator must carefully re-examine the descriptive information, carry out more active case finding, and re-analyse the data. The investigator also needs to directly observe the location, lifestyle, and/or behaviour of the cases. These additional investigations usually help the investigator to generate some reasonable hypotheses.

A hypothesis needs to be tested by an analytical study design. The most common is a case–control study. The investigator needs to define the cases and measure the odds of suspected exposed factors found among cases in comparison to the odds of exposure found among appropriately selected controls. The case definition for the analytical study may be more specific than that used for a descriptive study in order to reduce misclassifying non-cases as cases. Conversely, the controls may also need to be tested to avoid classifying non-apparent cases as controls.

An example is the analytic study for identifying risk factors for haemolytic–uraemic syndrome in a large outbreak of Shiga toxin-producing *Escherichia coli* O104:H4 infections in Germany during May 2011 (Buchholz et al. 2011). A matched case–control study was carried out to obtain data from 26 case subjects with the haemolytic–uraemic syndrome and compared to data from 81 controls. The investigation revealed ingestion of raw sprouts as the vehicle of infection (matched odds ratio, 4.35; 95% confidence interval, 1.05–18.0). A month after, a similar food poisoning outbreak of the same organism was detected in France (King et al. 2012). The investigators conducted a retrospective cohort study and they found an increased risk of developing the disease among adults who attended the buffet at the Children's Community Center Event on June 8, 2011 (relative risk, 5.1; 95% confidence interval, 2.3–11.1) in multivariable analysis (Table 5.3.1).

Table 5.3.1 Multivariable Poisson regression with food-specific relative risks and 95% confidence intervals among adults who attended the buffet at the Children's Community Center Event, France, June 2011 (n = 64)

Food item	RR	95% CI	P-value
Sprouts (any of the 3 kinds)	4.2	1.7–10.0	0.001
Gazpacho (tomato-based cold soup)	2.4	0.9–6.4	0.08
Carrots	2.3	0.8–7.1	0.14
Water (bottled)	2.0	0.8–5.2	0.14
Dip sauce (mayonnaise)	1.7	0.8–3.3	0.15
Lettuce	1.4	0.6–3.6	0.43
Green peppers	0.4	0.1–1.3	0.15

CI, confidence interval; RR, relative risk.

[a] All but 1 of the covariables in the presented final multivariable model have *P* values that are non-significant at the 0.05 level. These variables were left in the final model to ensure a sufficient numeric stability of the model.

Reproduced from King LA et al., Outbreak of Shiga toxin-producing *Escherichia coli* O104:H4 associated with organic fenugreek sprouts, France, June 2011, *Clinical Infectious Diseases*, Volume 54, Issue 11, pp. 1588–94, Copyright © The Author 2012, published by Oxford University Press on behalf of the Infectious Diseases Society of America, by permission of Oxford University Press.

Step 6: environmental or other studies to supplement epidemiological findings

Although an analytical study might be able to confirm a hypothesis, the investigator still needs to find environmental or other evidence to support and explain the epidemiological evidence. In an outbreak of unknown illness in a rural village of Egypt, in which the cases developed severe abdominal pain, persistent vomiting, and generalized weakness, the investigator was able to detect abnormally high blood lead levels among the cases. The analytical study revealed an association between high lead levels in the blood and eating flour from a particular mill factory. The mill implicated in the outbreak was visited, and upon arrival at the mill, the investigators noted a lead smelting pot in the corner of the mill. Lead was used by the miller to attach the crosspiece to the grinding stone. Occasionally the lead would

break off and contaminate the flour. The miller reported using about 2 kg of lead per year. Analysis of grain from the mill showed no lead; however, lead was found in flour from the surface of the mill stone and in samples of flour after grinding was complete (Abdel-Nasser 1996).

Step 7: establishing the causes of the outbreak

Once the hypothesis has been tested and other necessary studies have been carried out, the investigators can draw their conclusions about the causes of the outbreak. This conclusion is very important, because action should follow. Many restaurants or factories have been closed because they were implicated as the source of the outbreak. Outbreak investigations are naturally faced with constraints of time and other uncontrolled conditions that do not favour perfect design and methodology. The findings potentially include both random and systematic errors. Before drawing any conclusions, it is important that the investigator carefully examines the weaknesses or limitations of the investigation. In principle, the investigator must identify the cause of the outbreak based on the agreement of the following four pieces of evidence.

Laboratory evidence

The aetiology of the disease is identified in the patients and in the suspected source of infection. If the investigator cannot identify the aetiology from the suspected source, it is still possible to use certain markers to support their conclusions. For example, although a *Vibrio cholerae* type O1 could not be cultured from the drinking water supply of the affected community, the observation of human faecal coliform bacteria in the water helped to indicate that human faeces had contaminated the water.

Clinical evidence

It is necessary to verify that the incubation period and clinical manifestations of cases are compatible with the aetiological agents reported from the laboratory. In the SARS outbreak, most patients developed respiratory symptoms within 7 days after an exposure, then having atypical pneumonia with rapid progression to respiratory failure, and a fraction of them also had

diarrhoea, which could be explained by the findings of SARS coronavirus both in the respiratory tract and stools of patients.

Environmental evidence

Careful investigation of the environment should reveal clues that the causative agent can pass from the source to the cases or the index case to subsequent cases. The investigation of early SARS cases revealed that the index case passed the virus to several people who stayed on the same floor and used the same elevator (CDC 2003).

For an outbreak of food poisoning, a cooking environment that is dirty or located close to the toilet is convincing evidence of food being contaminated by faeces. Crowding in bedrooms is very convincing evidence for a respiratory tract infection outbreak.

Epidemiological evidence

Epidemiological evidence found in the descriptive and analytical studies should clearly explain the following aspects:

Pattern of spread, as described by an epidemic curve.
Statistical strength of association between exposure and developing the disease.
Dose–response relationship, which demonstrates a higher strength of association when the exposure is increasing.
Exposure should precede an illness.

It is not uncommon to see some disagreement in the evidence from different sources. In this event, the investigator needs to verify the validity of each piece of evidence and discuss the data with people who have expertise in that particular area in order to obtain more information and draw a valid conclusion.

Step 8: on-site reporting to and recommendations for concerned authorities

The most important step leading to a timely response is to report the findings to the responsible individuals both at the local and national levels so that

they can take appropriate action. Keeping this in mind, the investigators need to complete two tasks before leaving the outbreak area:

Complete analysis and data interpretation: leaving the field without completing these tasks will reduce the sense of urgency necessary to finish the work. Data are easier to retrieve when staying in the field rather than instructing field people to send them to the investigation team later.
Present the main findings with recommendations: some findings may be very sensitive because they reflect the weaknesses or mistakes of health personnel or other authorities. The investigator needs to select the appropriate approach, either formal or informal, with those responsible. Leaving the field without providing the information that the team obtains from the investigation will be detrimental to other people who could be prevented from becoming ill if the information were known. In many urgent situations, the findings and recommendations might be regularly given to the people responsible during the outbreak investigation process instead of waiting until the end. The investigators also need to present or communicate the findings and recommendations to the national authorities as soon as possible. The recommendations for action should be based on the findings of the investigation. These might include the following aspects.

What can be done to control the outbreak?

With timely investigation, some interventions can be implemented to stop further spread. In the outbreak of measles in a rural community in Peru, measles vaccines were given to children aged between 6 months and 15 years who were not measles cases, regardless of their previous immunization status. Using knowledge about the complication of measles in previous studies, the investigators estimated that the action prevented 87 cases of diarrhoea and 46 cases of pneumonia, and averted five deaths (Sniadack et al. 1999). Outbreaks can be controlled even if the disease aetiology is not known at the moment, but the mode of transmission is known from the investigation. One good example is the early story of the HIV/AIDS epidemic when sexual transmission was identified as a major mode of transmission among sexually active, young gay men. Therefore, promotion of condom use and safe sex practice was immediately implemented to prevent the transmission.

How to prevent future outbreaks

In many instances, interventions cannot be implemented for the current outbreak, but the findings can be used to set up new practices or policies. These recommendations can help prevent future outbreaks. In Thailand, the recommendation was made to change practices of chicken farming and avoiding direct contact with sick poultry, which was recommended through the results of the investigation of human H5N1 cases. The Egyptian government, after reviewing the results of the investigation of lead contamination in flour mills, agreed to ban the use of lead in privately owned flour mills (Abdel-Nasser 1996).

How to improve the investigation

The investigator should review the performance of the investigation and identify the weaknesses of the methodology or the field operation so that improvements can be made. These might include a more appropriate case definition, a better investigative design, improved laboratory support, a different team composition, lower costs, a shorter time course, etc.

How to improve surveillance

The best time to evaluate the surveillance system is during an outbreak investigation. In most outbreaks, the investigator must review data from existing surveillance systems. With this direct involvement, the investigator will be able to evaluate the timeliness, completeness, validity of diagnosis, sensitivity of the system, and utilization of the surveillance information. The investigation of human cases of avian influenza (H5N1) in Thailand led to an improved definition of a human case in the areas experiencing an H5N1 virus outbreak in poultry.

Step 9: Communicating risk

In addition to on-site reporting, the investigator team should or be supported to do risk communication to educate and mobilize cooperation from the health professional community and the general public. Occasionally, there may be other communities that are also prone to a similar outbreak. The proper information will raise the awareness of public health and government authorities to assess their own situation and implement appropriate measures

to prevent the potential outbreaks. Key messages should promote specific actions that people can realistically take to protect their health. Risk communication is extremely critical and must be delivered in a timely manner with the purpose to build trust and engage community for minimizing consequences of the outbreak. To enhance risk communication capacity of the member states in emergency situation including outbreak setting, WHO has published guidelines for emergency risk communication policy and practice to improve performance of concerned authorities and health workers in the field (WHO 2017).

Dissemination of key messages should occur in a timely manner through weekly or monthly reports. With the advancement of information and communications technology including real-time access to large amount of data, ability to conduct sophisticated analytical techniques with high-performance personal computers, wireless and mobile technology, and availability of various functional applications packages, it is feasible and convenient to disseminate the information on disease situation using online and interactive graphical presentation. The way WHO developed online interactive maps, graphics, and reports to illustrate Ebola outbreak situation in West Africa during 2014–2016 (Fig. 5.3.5) is a wise example of using information technology to visualize epidemic situation in order to inform policy decision and raising situation awareness of the public (WHO 2017).

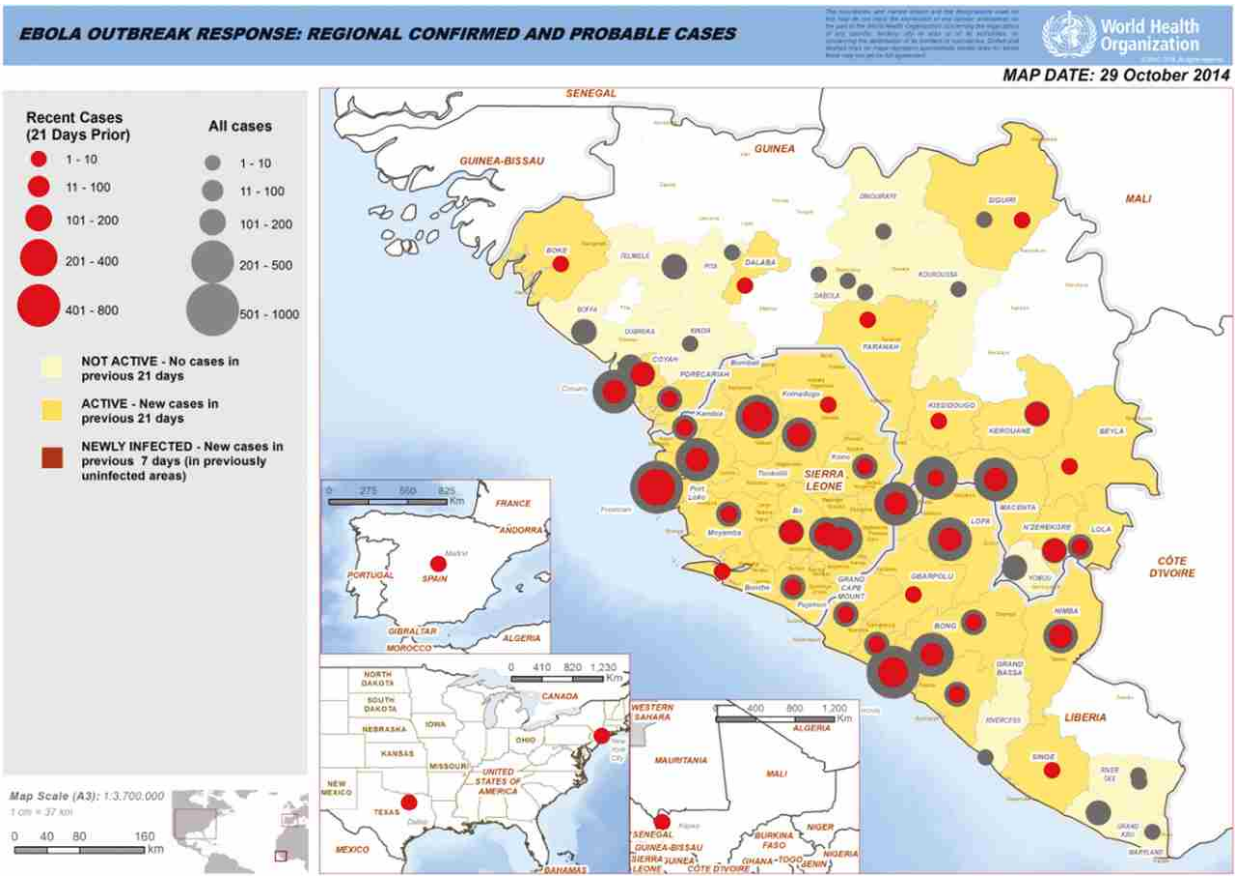


Fig. 5.3.5 Example of maps of Ebola outbreak situation in West Africa, 29 October 2014. Reproduced with permission from Word Health Organization (2014), http://www.who.int/csr/disease/ebola/photos/geographic-map-29-oct-2014.png?ua = 1, accessed on 30th Apr 2018).

Release of important findings through the mass media is also very useful for educating and communicating risk to the general public. Before releasing the investigation results through the media, the investigator must ensure that all the facts are delivered in a constructive manner and do not result in blame of any organization. The investigator should also report the investigation results in an international journal or bulletin such as the *Weekly Epidemiological Record* of the WHO, the *Morbidity and Mortality Weekly Report* of the US CDC, the *Eurosurveillance* of the European Centre for Disease Prevention and Control, or the epidemiological bulletin of own country. This practice is necessary to alert health professionals in other countries and keep them informed of the problem.

In recent years, sharing of information for outbreak detection and control through the internet and email has been increasingly practised for timely

responses. Telecommunications played an important role in the outbreaks of SARS when the disease was spread rapidly by air transportation to countries on the other side of the world. An example is a teleconference on the clinical features and management of SARS that was held by the WHO, which united 80 clinicians from 13 countries worldwide (World Health Organization Regional Office for the Western Pacific 2006).

Step 10: follow-up to ensure implementation of control measures

Finally, the investigator should follow-up the investigation by maintaining close communication with local health authorities. An absence of new cases for at least two incubation periods of the infectious disease under investigation could suggest that the outbreak is subsiding. A good investigator should follow-up on the recommendations. An outbreak investigation is a waste of time if sound recommendations have not been implemented. The investigator should learn the reasons why the recommendations were or were not implemented. If the recommendations were implemented, the investigator can also learn the impact by observing the trend of the disease.

Cooperation for international outbreak investigation and emergency response

The world today is especially prone to outbreaks because of frequent cross-border movement, civil war and migration, rapid transportation, international trading, tourism, etc. An outbreak in one country can spread to other countries very easily. For example, in the multicountry SARS outbreak, international travellers were at risk of disease spreading from a single infectious case in a hotel (CDC 2003). Ultimately, there were over 8,000 cases reported in 26 countries worldwide. The world's experience with the SARS outbreak led to the commitment of countries around the world to improve surveillance and international joint investigation of emerging infectious diseases through International Health Regulations (Connolly 2005), which were legally effective in 2007, requires all countries to improve capacities 'to prevent, protect against, control and provide a public

health response to the international spread of disease in ways that are commensurate with and restricted to public health risks, and which avoid unnecessary interference with international traffic and trade' (WHO 2017). Each country needs to prepare resources and implement strategies to develop, maintain, and monitor its own disease prevention and control capacities according to the IHR (Connolly 2005) in which outbreak investigation and response are among key components.

A massive outbreak of botulism in northern Thailand in March 2006 tested the international capacity to respond to a public health emergency. Botulism poisoning due to contaminated home-canned bamboo shoots caused illness in 209 villagers, of whom 134 persons were hospitalized and 42 required mechanical ventilation. A global search for botulinum antitoxin began, involving international agencies, embassies, national laboratories, airlines, and commercial organizations in seven countries. Sufficient antitoxin was obtained from four sources for treatment of 90 patients, with a delay in treatment of 5–9 days from time of exposure. Rapid local outbreak detection and an effective international response likely prevented mortality and additional morbidity (WHO 2007).

An outbreak anywhere in the world must therefore be treated as a threat to all countries. It is important for each country to build up its capacity for surveillance, outbreak preparedness, and investigation. In addition to the US CDC Epidemic Intelligence Services, over 100 countries have participated in training programmes that is ready to investigate all kinds of outbreaks. These applied epidemiology training programmes are known by different names, such as the Field Epidemiology Training Programme (FETP) or the Field Epidemiology and Laboratory Training Programme (FELTP), and The European Programme for Intervention Epidemiology Training (EPIET). The WHO has supported the establishment of the Global Outbreak Alert and Response Network, a technical collaboration of existing institutions and networks to pool human and technical resources for the rapid identification, confirmation, and response to outbreaks of international importance. It is a wise investment, and a commitment that is needed by the international community for information exchange and collaborative investigation of outbreaks in a timely manner. With continuous growth of emerging disease outbreaks across borders, joint international investigations to control those outbreaks at source continue are increasingly needed.

Conclusion

Outbreak investigation is an essential function of public health professionals who care for the well-being of the community. It is an opportunity to gain new knowledge of diseases and to discover the weaknesses of current public health practices and systems. A good public health professional must always be alert to the possibility of outbreaks. Normal surveillance systems or unofficial sources, such as the mass media, can be a good source for detection of an outbreak. Before starting field investigations, investigators should organize the team, review previous knowledge, prepare the technical and management aspects, and start the investigation as soon as possible. The investigation can be conducted by following the ten steps outlined in this chapter. The investigation usually starts by confirming the existence of the outbreak, verifying the diagnosis, gathering case information, obtaining laboratory support, descriptive epidemiology, formulating and testing the hypothesis when necessary, conducting environmental surveys to supplement epidemiological evidence, providing timely on-site reporting of the findings, with practical recommendations to local and national responsible authorities and communicating risk to health professional community and public. Competent outbreak investigators combine sound scientific knowledge and good management skills. Involvement of other professionals with specific skills and training may be indicated. Direct participation in conducting the investigation is needed to gain the necessary skills. A good investigator should be a field-oriented person with strong levels of perseverance, scepticism, and common sense. The investigator should not end his or her work with the report, but should follow-up on the recommendations and continue vigorous surveillance of the problem. In the future, increasing joint international investigations will be needed. Rich countries and resource-limited countries need to collaborate to detect and stop outbreaks before they get out of control. Through this cooperation, the world will be a safer place amidst the threat of old and emerging disease outbreaks.

REFERENCES

Abdel-Nasser, M.A. (1996). *Outbreak Investigation of an Unknown Illness in a Rural Village, Egypt, 1996.* Cairo, Egypt: Field Epidemiology Training Programme.

Buchholz, U., et al. (2011). German outbreak of *Escherichia coli* O104:H4 associated with sprouts. *The New England Journal of Medicine*, 365, 1763–70.
Centers for Disease Control and Prevention (CDC) (1990). Mass sociogenic illness in a day-care center, Florida. *Morbidity and Mortality Weekly Report*, 39, 301–4.
Centers for Disease Control and Prevention (CDC) (1996). Outbreak of unexplained illness in a middle school-Washington, April 1994. *Morbidity and Mortality Weekly Report*, 45, 6–9.
Centers for Disease Control and Prevention (CDC) (1999a). Food-borne botulism associated with home-canned bamboo shoots—Thailand, 1998. *Morbidity and Mortality Weekly Report*, 48, 437–9.
Centers for Disease Control and Prevention (CDC) (1999b). Outbreak of Hendra-like virus—Malaysia and Singapore, 1998–1999. *Morbidity and Mortality Weekly Report*, 48, 265–9.
Centers for Disease Control and Prevention (CDC) (1999c). Outbreak of Nipah virus—Malaysia and Singapore, 1999. *Morbidity and Mortality Weekly Report* 48, 335–7.
Centers for Disease Control and Prevention (CDC) (1999d). Trichinellosis outbreaks—North Rhine-Westphalia, Germany, 1998–1999. *Morbidity and Mortality Weekly Report* 48, 488–92.
Centers for Disease Control and Prevention (CDC) (2003). Update: outbreak of severe acute respiratory syndrome-worldwide, 2003. *Morbidity and Mortality Weekly Report* 52, 241–8.
Charles-Smith, L.E., et al. (2015). Using social media for actionable disease surveillance and outbreak management: a systematic literature review. *PLoS One*, 10(10), e0139701.
Connolly, M.A. (ed.) (2005). *Communicable Disease Control in Emergencies: A Field Manual*. Geneva: World Health Organization.
Conrad, C., et al. (2015). Community outbreak of HIV infection linked to injection drug use of oxymorphone—Indiana, 2015. *Morbidity and Mortality Weekly Report*, 64, 443–4.
Corley, C.D., Cook, D.J., Mikler, A.R., and Singh, K.P. (2010). Text and structural data mining of influenza mentions in web and social media. *International Journal of Environmental Research and Public Health*, 7, 596–615.
Frost, W.H. (1936). *Introduction to Snow on Cholera*. London: Oxford University Press and Commonwealth Fund.
Gregg, M.B. (1996). Conducting a field investigation. In: Gregg, M.B. (ed.) *Field Epidemiology*. New York: Oxford University Press, pp. 44–59.
Jongcherdchootrakul, K., Henderson, A. K., and Jiraphongsa, C. (2011). Injuries and deaths at a pub fire in Bangkok, Thailand on New Year's Eve 2009. *Burns*, 37, 499–502.
Kamps, B.S. and Hoffmann, C. (2003). Epidemiology. In: Kamps, B.S. and Hoffmann, C. (eds.) *SARS Reference* (3rd ed.). Paris: Flying Publisher, p. 71.
Karnjanapiboonwong, A., et al. (2011). Control of a pandemic influenza A (H1N1) 2009 outbreak in a prison, Saraburi Province, Thailand, August 2009. *Outbreak, Surveillance, Investigation & Response*, 4, 12–16.
King, L.A., et al. (2012). Outbreak of Shiga toxin-producing *Escherichia coli* O104:H4 associated with organic fenugreek sprouts, France, June 2011. *Clinical Infectious Diseases*, 54, 1588–94.
Lee, S.Y., Mak, K.H., and Saw, T.A. (1999). The avian flu (H5N1): one year on. *Public Health and Epidemiology Bulletin*, 8, 1–8.
Lewis, H.C., et al. (2009). Outbreaks of *Shigella sonnei* infections in Denmark and Australia linked to consumption of imported raw baby corn. *Epidemiology and Infection*, 137, 326–34.
Porta, M. (ed.) (2008). *A Dictionary of Epidemiology* (5th ed.). New York: Oxford University Press.
Samitsuwan, P., Sermkaew, T., and Laosiritaworn, Y. (2005). An outbreak of tetrodotoxin poisoning associated with puffer fishes, Bangkok, Thailand, 2002. *Journal of Health Science*, 14, 203–8.
Siraprapasiri, T., et al. (1991). Risk factors for HIV among prostitutes in Chiangmai, Thailand. *AIDS*, 5, 579–82.

Sniadack, D.H., et al. (1999). Measles epidemiology and outbreak response immunization in a rural community in Peru. *Bulletin of the World Health Organization*, 77, 545–52.

Tom-Aba, D., et al. (2015). Innovative technological approach to ebola virus disease outbreak response in Nigeria using the open data kit and form hub technology. *PLoS One*, 10, e0131000.

World Health Organization (WHO) (1978). Report of the International Commission. Ebola hemorrhagic fever in Zaire, 1976. *Bulletin of the World Health Organization*, 56, 271–90.

World Health Organization (WHO) (2007). The need for global planned mobilization of essential medicine: lessons from a massive Thai botulism outbreak. *Bulletin of the World Health Organization*, 85, 238–40.

World Health Organization (WHO) (2012). *Global Alert and Response (GAR): WHO interim case definition—Novel Coronavirus, as of 29 September 2012.* [Online] Available at: http://www.who.int/csr/disease/coronavirus_infections/case_definition_29092012/en/index.html

World Health Organization (WHO) (2017). *Communicating Risk in Public Health Emergencies: A WHO Guideline for Emergency Risk Communication (ERC) Policy and Practice*. Geneva: World Health Organization.

World Health Organization Regional Office for the Western Pacific (2006). *SARS: How a Global Epidemic was Stopped*. Manila: World Health Organization, Western Pacific Region.

5.4

Case–control studies

Noel S. Weiss

Introduction

In 1971, Herbst et al. (1971) * reported that mothers of seven of the eight teenage girls diagnosed with clear cell adenocarcinoma of the vagina in Boston during 1966–1969 claimed to have taken a synthetic hormone, diethylstilboestrol, while pregnant. None of the mothers of the 32 girls without vaginal adenocarcinoma, matched to the cases mothers with regard to hospital and date of birth, had taken diethylstilboestrol during their pregnancy. Within a year, a New York study of five girls with and eight girls without vaginal cancer obtained similar results (Greenwald et al. 1971). The introduction of prenatal diethylstilboestrol use into obstetric practice in the United States during the 1940s and 1950s followed by the appearance of this hitherto unseen form of cancer some 20 years later supported a causal connection between *in utero* exposure to diethylstilboestrol and vaginal adenocarcinoma. The means by which *in utero* diethylstilboestrol exposure might predispose to the occurrence of clear cell vaginal adenocarcinoma was unknown in 1971. (It is now believed that diethylstilboestrol acts by interfering with normal development of the female genital tract, resulting in the persistence into puberty of vaginal adenosis in which adenocarcinoma can arise.) Nonetheless, a causal inference was made at that time by the Food and Drug Administration, which specified pregnancy as a contraindication for diethylstilboestrol use.

The investigation by Herbst et al. was a case–control study: A comparison of prior exposures or characteristics of ill people (cases) with those of people at risk for the illness in the population from which the cases arose. Generally, the prior experience of people at risk is estimated from

observations of a sample of that population (controls). A difference in the frequency or levels of exposure between cases and control—that is, an association—may reflect a causal link.

A case–control, cohort, or any other form of non-randomized study has the potential to identify associations that are not causal, either because of chance, or because of the influence of some other factor associated with both the exposure and outcome. Even so, the evidence that is provided by well-performed case–control studies can carry great weight when evaluating the validity of a causal hypothesis. Indeed, a number of causal inferences have been based largely on the result of case–control studies. These include, in addition to the diethylstilboestrol–vaginal adenocarcinoma relationship, the connection between aspirin use in children and the development of Reye's syndrome, and the use of absorbent tampons and the incidence of toxic shock syndrome.

One of the criteria used to assess the validity of a causal hypothesis is the strength of the association between exposure and disease, usually as measured by the ratio of the incidence rate in exposed and non-exposed people. In most case–control studies, it is not possible to measure incidence rates in either of these groups. Nonetheless, from the frequency of exposure observed in cases and controls, it is usually possible to estimate closely the ratio of the incidence rates.

To understand how this can be done, consider a cohort study in which exposed and non-exposed people are followed for a certain period of time. Table 5.4.1a–c summarizes their experience with regard to a particular disease.

Table 5.4.1a Disease incidence in relation to exposure

	Diseased	Non-diseased	Total
Exposed	*a*	*b*	*a*+*b*
Non-exposed	*c*	*d*	*c*+*d*

The cumulative incidence of the disease in exposed and non-exposed people over a given period of follow-up is $a/(a + b)$ and $c/(c + d)$,

respectively. The relative risk (RR) is the ratio of these:

$$RR = \frac{a/(a+b)}{c/(c+d)}.$$

If the incidence of the disease is relatively low during the follow-up period in both exposed and non-exposed people, then *a* will be small relative to *b* and *c* will be small relative to *d*. Therefore,

$$RR = \frac{a/(a+b)}{c/(c+d)} \approx \frac{a/b}{c/d} = \frac{a/c}{b/d}.$$

Table 5.4.1b Disease incidence in 100,000 people

	Diseased	Non-diseased	Total
Exposed	100	9,900	10,000
Non-exposed	300	89,700	90,000

If a case–control study had been carried out in this population, in which 50% of the cases were included but only 1% of the non-cases, then the following results would have been obtained:

Table 5.4.1c Exposure status in a sample

Cases	Diseased	Non-diseased
Exposed	100 × 0.5 = 50	9 900 × 0.01 = 99
Non-exposed	300 × 0.5 = 150	89 700 × 0.01 = 897

In this expression, the numerator (*a/c*) is the odds of exposure in people who develop the disease and the denominator (*b/d*) is the odds exposure in

people who remain well. Therefore,

$$\text{Odds ratio (OR)} = \frac{a/c}{b/d}.$$

The numerator can be estimated from a sample of cases, and the denominator can be estimated from a sample of non-cases. Neither estimate is influenced by the proportion of cases among the subjects actually chosen for study.

In the following hypothetical example, assume that 100 of 10 000 people exposed to a particular substance or organism developed a disease, compared with 300 of 90 000 non-exposed people:

Therefore,

$$RR = \frac{100/10{,}000}{300/90{,}000} = 3.00$$

$$RR = \text{OR} = \frac{50/150}{99/897} = 3.02.$$

In many studies, controls are chosen as they were in this example; that is, from people who have not developed the disease by the end of the same time period during which other people (the cases) have become ill. In such studies, the less common the disease in both the exposed and non-exposed people during the period, the better the odds ratio will estimate the ratio of cumulative incidence. In the example, only 0.1% and 0.33% of exposed and non-exposed people, respectively, developed the illness: So, the relative cumulative incidence and odds ratio are nearly identical (3 vs. 3.02). However, it is also possible to choose controls from people free of disease only until the corresponding cases have been diagnosed; a person can appear in the study first as a control and later as a case. If this approach is used, the odds ratio will be a valid estimate of the ratio of incidence rates (i.e. cases divided by person–time at risk) irrespective of the disease frequency (Greenland and Thomas 1982; Pearce 1993).

* This chapter is excerpted from *Epidemiologic Methods* (2nd edition) by N.S. Weiss and T.D. Koepsell (2014).

Retrospective ascertainment of exposure status in cases and controls

Epidemiological studies seek to obtain information on exposures present during an aetiologically relevant period of time; that period varies across aetiological relationships. For example, although excess consumption of alcohol predisposes to both motor vehicle injuries and cirrhosis of the liver, it does so during considerably different time intervals before the occurrence of the injury or the onset of the illness, respectively.

Some case–control studies are nested within cohort studies in which specimens (e.g. blood or urine) have been obtained before diagnosis on all cohort members, but have not yet been analysed for the exposure(s) in question. When these analyses are carried out on cohort members who developed a particular illness and on controls selected from the cohort, the results obtained are not influenced by the events occurring after the diagnosis of the illness. (To avoid the possibility of occult illness in cases influencing levels of a suspected aetiological factor, many studies of this type exclude from the analyses specimens obtained in the period before diagnosis that might correspond to the preclinical stage of disease.) Also, among the large majority of case–control studies in which exposure status is not measured until the illness or injury has been diagnosed, some are concerned only with an exposure or characteristic that would have been the same at all times in a person's life. This is true for a genetically determined characteristic such as ABO blood type or the absence of glutathione transferase M_1 activity (an enzyme that metabolizes several potentially carcinogenic constituents of cigarette smoke). Clearly, these studies are no less valid for having had to measure exposure in retrospect.

However, most case–control studies are required to consider explicitly how best to assess in retrospect the subjects' exposure status during one or more possible aetiologically relevant time periods. Possible sources of exposure data include interviews or questionnaires, available records, and physical or laboratory measurements.

Interviews or questionnaires

For many exposures, a subject's memory is an excellent window to the past. A number of important aetiological relationships have been identified through interview-based case–control studies. Generally, study participants will report longer-term and more recent experiences with greater accuracy than shorter-term and more distant ones. Attention to the ways in which questions are asked (White et al. 2008), along with the use of visual aids when appropriate (e.g. pictures of medicines, or containers of household products, and calendars for important life events to enhance recall of the timing of other exposures), will maximize the accuracy of the information received. These efforts, along with the use of the same questions for cases and controls asked in the same way, will also minimize the potential for bias that could result from the subject's or interviewer's awareness of case or control status.

One advantage of exposure ascertainment via interview or questionnaire is that information can be sought for multiple points in the past. It is possible that a given exposure plays an aetiological role only if present at a certain age, for a certain duration, or at a certain time in the past. Because there is often little guidance before a study starts that suggests the most relevant age, length, or recency, key exposures are often elicited throughout much of the subject's lifetime. However, care must be taken not to include exposures that took place after the illness began. An instructive example is provided by (Victora et al. 1989) in a case–control study of infant death from diarrhoea as related to the type of feeding. These investigators asked mothers whether their child was or was not being breastfed immediately before the onset of the fatal illness (mothers of controls were queried about type of feeding prior to a comparable point in time). Mothers were also asked if subsequent to the onset of the illness there had been any changes in the type of feeding; following the development of diarrhoea, many children were switched to formula and cow's milk. Relative to infants who were solely breastfed, those who were supplemented with powdered or cow's milk prior to their illness had about four times the risk of diarrhoeal death. However, the authors showed that if one inappropriately considered the feeding method that was present during the illness, about a 13-fold increase in risk associated with supplementation would have been estimated.

Records

Case–control studies have exploited the presence of vital, registry, employment, medical, and pharmacy records, to name a few, as a means of obtaining information on exposures. However, because the information contained in the records is usually assembled for purposes other than epidemiological research, they may not provide precisely the information desired by the epidemiologist. For example, a death certificate or an occupational record may state an individual's job, but often not his or her actual exposure to the substance(s) of interest to the study. A pharmacy record will indicate a prescription having been filled, but not necessarily whether the patient took the medication on a given day, or took it at all. This sort of imprecision will impair a study's ability to discern a true association between an exposure and a disease—the greater the imprecision, the greater the impairment.

Nonetheless, some very strong associations have been identified through record-based case–control studies. For example, Daling et al. (1982) conducted a tumour-registry-based case–control study to test the hypothesis that homosexual men have a relatively high incidence of anal cancer. Although registry data do not specify a man's sexual preference, they do contain information regarding his marital status. The investigators found that three times more men with anal cancer than controls (men with colon or rectal cancer) had never been married. Being single is far from a perfect predictor of homosexuality, of course. Nonetheless, the presence of such a large case–control difference, given the very poor means of gauging the relevant exposure, was a stimulus to conduct interview-based studies that could elicit information regarding sexual history with greater precision. The latter studies showed an exceedingly strong association between anal cancer in men and a history of sexual intercourse with another man (Daling et al. 1987).

In case–control studies in which medical records are used to characterize exposure status, care must be taken to restrict the information obtained to that which precedes the case's diagnosis (and the presence of symptoms, if any, that led to the diagnosis). The records of controls must be truncated at similar points in time. Without this safeguard, it is possible that bias will arise because there are systematically more records available to review on cases than controls; the case's illness may have stimulated an enquiry by

medical personnel into his or her past, whereas no corresponding enquiry would necessarily have occurred for control subjects.

Physical and laboratory measurements

The recognized limitations of interviews and records in characterizing a variety of potentially relevant exposures have stimulated the conduct of epidemiological studies that use laboratory and other methods of measurement. A woman cannot tell an investigator the level of her reproductive hormones, the concentration of various micronutrients in her blood, or whether her cervix is infected with human papillomavirus, but laboratory tests can. Unfortunately, such tests tell us what these things are only at the time that the specimens have been obtained. For some exposures, there will be a high correlation between the measured level following case and control identification, and that which was present during the aetiologically relevant time period. For example, lead enters and does not leave the dentine of teeth. Therefore, in young school-age children, lead dentine levels are an indicator of cumulative lead exposure, a good portion of which could be relevant to the development of intellectual impairment and other adverse neurological outcomes. In contrast, one would not rely on serum levels of reproductive hormones of postmenopausal cases of breast cancer and controls to indicate what their premenopausal levels were, much less their hormonal status during their very early reproductive years (at which time the hormones may be exerting their greatest impact on future risk of breast cancer).

Case definition

Ideally, the cases in a case–control study would comprise all (or a representative sample of) members of a defined population who develop a given health outcome during a given period of time. For studies of disease aetiology, that outcome is the disease incidence. For studies that seek to determine the efficacy of early disease detection or treatment, the outcome generally is the occurrence of complications of the disease or mortality; such studies have been described in detail elsewhere (Selby 1994; Weiss 1994), and will not be covered any further here.

The population from which cases are to be drawn may be defined geographically or on the basis of other characteristics such as membership in a prepaid healthcare plan or an occupational group. The identification of all newly ill people in a defined population can be facilitated by the presence of a reporting system such as a cancer or malformation registry that seeks to accomplish this identification for other purposes. Occasionally, care for the condition being studied may be centralized, so that it will be necessary to review the records of only one or a few institutions to identify all cases in the population in which those institutions are located. However, in many instances, it is not feasible to identify all cases that occur in a given population; therefore, case–control studies are often based on cases identified from hospital records, or from the records of selected providers from whom patients had sought healthcare. The study by Herbst et al. (1971) on vaginal adenocarcinoma was of this type. Whether or not the cases are derived from a defined population, it is necessary that they be drawn in an unselected manner with regard to exposure status; for example, by including in the study all otherwise eligible cases diagnosed or receiving care during a defined time period.

Although the goal of a case–control study of aetiology is to enrol incident cases, under some circumstances it may be necessary to enrol prevalent cases at a particular point in time, irrespective of when each one's illness had begun. For some conditions, the date of occurrence may simply not be known; for example, in the absence of very close sero-monitoring, one generally cannot determine when a person acquired an HIV infection. Furthermore, for uncommon diseases of long duration, an incidence series may yield too few cases for meaningful analysis. The disadvantages of using prevalent cases in a case–control study relate in part to the added problems of accurate exposure ascertainment. For prevalent conditions whose date of diagnosis is known, pre-illness exposure information on study subjects must be obtained for more distant points in the past, on average, than would be necessary for an incident series. For prevalent conditions whose date of occurrence is unknown (e.g. HIV infection), there will be uncertainty as to the best point in time before which one should elicit exposure information. Also, by studying people remaining alive with a given condition, one is simultaneously studying not only aetiological factors but also those that influence survival from the condition.

Ideally, the criteria used to identify and select individual cases for study should be objective with high sensitivity and specificity for the disease. Specificity is of particular concern, because the inadvertent inclusion of people without disease into the case group will generally obscure any true association with the exposure. With this in mind, in the case–control study of Reye's syndrome and antecedent analgesic use conducted by the Centers for Disease Control (Hurwitz et al. 1985), only cases with a substantial degree of neurological impairment (stage 2 or higher) were included. The use of this criterion minimized the chances that children with a disease other than Reye's syndrome, which generally would have a lesser degree of severity, would be included in the case group. It also was intended to serve as protection against selective misclassification of Reye's syndrome based on knowledge of exposure status, because the hypothesis of the association of aspirin with Reye's syndrome was well-known by the time the study took place. Conceivably, the knowledge that the child had consumed aspirin could have led some doctors to diagnose Reye's syndrome in cases with an atypical illness.

Control definition

Occasionally, the proportion of ill people who have had a specific exposure is so high, unequivocally more than would be expected in the population from which they were derived, that the presence of an association (although not its magnitude) can be surmised from a case series alone (Cummings and Weiss 1998). For example, when it was learned that all cases of a form of pneumonia that was epidemic in Spain in 1981 had ingested adulterated rapeseed oil, a causal inference was drawn, leading to efforts to eliminate further use of this oil. This action was taken before any formal comparison of cases with controls was made (Tabuenca 1981).

However, in the vast majority of instances, an explicit control group is needed to estimate the frequency and degree of exposure that would have taken place among cases in the absence of an exposure–disease association. An ideal control group would be one that comprises individuals:

Selected from a population whose distribution of exposure is that of the population from which the cases arose;
Who are identical to the cases with respect to their distribution of all characteristics (a) that influence the likelihood and/or degree of exposure, and (b) that, independent of their relation to the exposure, are also related to the occurrence of the illness under study or to its recognition;
In whom the presence of the exposure can be measured accurately and in a manner that is identical to that used for the cases.

If these criteria are not met in a particular study, then selection bias, confounding, or information bias, respectively, will be present.

Minimizing selection bias

If the cases identified in the study are all or a sample of those that occurred in a defined population, one can seek to achieve comparability by choosing as controls people sampled from that same population. For geographically defined populations, a number of different sampling methods have been used, including random digit dialling (RDD) of telephone numbers, area sampling, neighbourhood sampling, voters' lists, population registers, motor vehicle licences, and birth certificates, among others. When cases are members of a prepaid healthcare plan who develop an illness or injury, a sample of people who were members of the health plan when the illness or injury occurred can serve as controls. When cases are ill or injured members of an employed population, controls can be selected from this same population.

Enumeration of the defined population from which the cases arose may require an innovative approach. For example, Smith et al. (2001) identified recreational boating fatalities in Maryland and North Carolina among persons 18 years and above during 1990–1998, seeking to determine the presence and magnitude of an association with alcohol consumption. Of 221 fatal cases on whom blood alcohol levels were measured, 55% tested positive. To provide a basis for comparison, the investigators devised a plan to randomly sample recreational boats on the waterways of the two states, and then to obtain a breath sample for alcohol testing from the boat operator and a random sample of any passengers aboard (n = 3,943). Only 17% of

these persons chosen as controls had a positive result, and the disparity between cases and controls grew with increasing alcohol levels.

If cases have not been selected from a definable population at risk for the disease, but from people treated for a particular illness at one or a few hospitals or clinics, then selection bias may be introduced if controls are not chosen from people who, had they developed the illness under study, would have received care at these hospitals or clinics, and if people who do and do not receive care from these sources differ with regard to the frequency or level of exposure. Therefore, when cases are chosen from a narrow range of healthcare providers, controls are often chosen from other ill people treated by these providers. Such ill controls may also be used if, irrespective of the source of the cases, there is no feasible way to sample from the population at large, or if sampling from the general population would result in a substantial level of non-response or information bias. For these reasons, in some studies of fatal illness, exposures in people with a given cause of death are compared with exposures in a sample of people who died from other causes.

However, the choice of ill or deceased controls can itself give rise to selection bias if the illnesses (causes of death) represented in the control group are in some way associated with the exposure of interest. For example, ill or recently deceased people tend to have been smokers of cigarettes more often than other people (McLaughlin et al. 1985), because smoking is associated with a variety of causes of illness and death. Because smoking histories of ill people overstate the use of cigarettes in the population from which the cases arose (even if that population cannot be defined), the odds ratio associated with smoking based on the choice of ill people as controls will be spuriously low.

To minimize selection bias related to having chosen ill or deceased controls, an attempt can be made to omit potential controls with conditions known to be related (positively or negatively) to the exposure. For example, in the analysis of a hospital-based case–control study of bladder cancer and earlier use of artificial sweeteners, the investigators excluded from their control group people who were admitted to hospital for obesity-related diseases (Silverman et al. 1983). They showed that without this restriction, the control group would have a spuriously high proportion of users of artificial sweeteners relative to the population from which their cases had actually come. This approach will succeed to the extent that one judges

correctly which conditions are truly exposure-related, and how accurately the presence of those conditions can be determined. For many exposures, this may pose little problem, and judicious exclusion will yield a control group capable of providing an unbiased result. For others, such as cigarette smoking or alcohol consumption, it has been shown that admitting diagnoses or statements of cause of death are incapable of identifying all people with illnesses related to these exposures (McLaughlin et al. 1985).

Some case–control studies simply compare exposure status among subgroups of patients with the same outcome. For example, Laumon et al. (2005) obtained blood samples of 10,748 drivers involved in fatal road accidents in France during 2001–2003, and had these analysed for a metabolite of cannabis. The investigators contrasted the proportion testing positive between drivers judged to have been at fault for the crash (‘cases’) and drivers judged not to have been at fault (‘controls’). The results of such a study will be valid when it is possible to identify those drivers who were at fault. To the extent that this is not possible, the misclassification that ensues will tend to make the size of the observed association smaller than the true one.

Occasionally, controls are chosen from individuals who are tested for the presence of the disease under study and are found not to have it. For example, in a study of the effectiveness of influenza vaccines, nasal and oropharyngeal swab specimens were obtained in persons with an acute respiratory illness. Vaccination histories were compared between persons who tested positive for the presence of influenza virus and those who tested negative (Jackson et al. 2017). As another example, the prior use of oral contraceptives was compared between women diagnosed with venous thromboembolism and women seen at the same institution for suspected venous thromboembolism who were diagnosed as not having this condition (Bloemenkamp et al. 1999). It may be relatively inexpensive to select controls from people who receive the same diagnostic evaluation as do cases, and it is also possible to achieve case–control comparability with regard to the choice of a healthcare provider (and the correlates of that choice). This approach can have an impact on the study’s validity if the frequency or degree of exposure differs between otherwise comparable members of a population who do and do not receive the test. It will increase the validity if the disease being investigated is generally asymptomatic, and so would not be detected in the absence of testing. Thus, the relation of the

use of oral contraceptives with the incidence of *in situ* cancer of the cervix is best studied in women who have undergone cervical screening, by comparing oral contraceptive use between cases of *in situ* cancer and women with a negative screen. This is because:

In most societies, women who use oral contraceptives are screened more commonly than women who do not.
In situ cancers are asymptomatic and will not be identified in the absence of cervical screening.

Therefore, if controls are chosen from women in general, who may or may not have received cervical screening, an apparent excess of oral contraceptive users would be present among cases of *in situ* cancer even if no true association were present.

However, the choice of test-negative controls can detract from a study's validity if the large majority of the people who develop the disease would soon be diagnosed whether or not the test was administered. There was a controversy in the late 1970s regarding the suitability, in case–control studies, of postmenopausal oestrogen use and endometrial cancer, of a control group restricted to women with no evidence of cancer on endometrial biopsy. Among women without endometrial cancer, oestrogen use differs greatly between those who have and have not undergone biopsy, because oestrogen use predisposes to uterine bleeding of non-malignant causes, which often leads to endometrial biopsy. Those investigators who believed that there was a great prevalence of occult endometrial cancer in the population suggested that the optimal control group ought to be women undergoing endometrial biopsy and found not to have cancer (Horwitz and Feinstein 1978). However, the majority of the investigators believed that no such large pool of prevalent, occult disease existed, and that choosing biopsy-negative controls would lead to a spuriously high estimate of oestrogen use in the population at risk, and thus, a spuriously low odds ratio (Shapiro et al. 1985).

No matter how controls are defined in a case–control study, selection bias may be introduced to the extent that exposure information is not obtained on all who have been selected to participate. The magnitude of this bias will increase in relation to the frequency of missing data and the degree to which exposure frequencies or levels differ between study subjects on whom

exposure status is and is not known. The problem of incomplete ascertainment of exposure on study subjects is particularly common in interview- or questionnaire-based case–control studies. Strategies for minimizing the degree of non-response in case–control studies are discussed in detail elsewhere (White et al. 2008).

Minimizing confounding

Characteristics of confounding variables in case–control studies

Confounding occurs when the estimate of the relationship between an exposure and a disease is distorted by the influence of another factor. In any study design, confounding will occur to the extent that the other factor is associated with both the exposure (although not as a result of the exposure) and the disease or its recognition. In case–control studies alone, a factor may confound even if it is not associated with an altered risk of disease, if the proportions of cases and controls vary across levels or categories of the factor. For example, in a collaborative study of ovarian cancer and use of oral contraceptives (Weiss et al. 1981), an attempt was made to identify and interview all incident cases over several years in two American populations. In one of the populations (western Washington State), several controls per case were interviewed, whereas the control-to-case ratio in the other (Utah) was 1:1. Because oral contraceptive use was more common among women in Washington than in Utah, failure to take into account the state of residence in the analysis (e.g. by adjustment) would have led to a spuriously high estimate of the frequency of oral contraceptive use by controls relative to that by the cases.

Means of controlling for confounding

One straightforward way of preventing confounding is to restrict cases and controls to a single category or level of the potentially confounding variable. For example, in their study on physical activity and primary cardiac arrest, Siscovick et al. (1982) excluded people with conditions, such as clinically recognized heart disease, that could both predispose to cardiac arrest and might be expected to alter level of activity. A second way is to obtain information on exposures or characteristics that may differ between cases and controls, and then make statistical adjustments for those that are

also found to be related to the exposure or characteristic under investigation (Rothman et al. 2008).

Finally, it is possible to match one or more controls to each case's category or level of a potentially confounding factor. It is appropriate to match if:

The variable is expected to be strongly related to both the exposure and disease. Thus, in a case–control study of breast cancer and use of hair dye, it would make sense to match on gender (if the study has not already been restricted to women) because use of dye is more common among women than men in most cultures. Although confounding by gender could be prevented even without matching by adjustment in the analysis, the statistical precision of the unmatched study would be substantially reduced relative to that of a case–control study having a more similar proportion of female cases and controls.

Information on possible matching variables can be obtained inexpensively. There are some means of control selection in which information regarding some confounders can be obtained at no cost. For example, from voters' lists or prepaid health plan membership records, it would generally be possible to choose directly one or more controls who were identical to a given case's age. Conversely, if a population sampling scheme such as RDD were being employed, the age of the respondent would not be known in advance of approaching him or her. Rather than omitting already contacted controls who did not match a particular case's age, the matching can be done much more broadly; additional control for finer categories of age can be accomplished in the data analysis.

Information on exposure status cannot be obtained inexpensively. The higher the cost of exposure ascertainment, the greater the incentive to limit the number of control subjects to the number of cases. Case–control differences regarding confounding factors particularly will reduce the statistical power of a study that does not have a surplus of controls. Enriching the group of controls selected with people more similar to the cases with regard to confounding factors (i.e. matching) can prevent this loss of statistical power.

In case–control studies of genetic characteristics as possible aetiological factors, some investigators have used a matched design in which a specific

type of relative (e.g. parent, sibling, or cousin) is chosen as a control for each case (Yang and Khoury 1997; Witte et al. 1999). This approach has the advantage of minimizing potential confounding by other genetic characteristics with which the one of interest is associated. However, it has the disadvantage of excluding a possibly large fraction of cases for whom there is no relative available of the type needed to provide a sample for genetic analysis.

It should be remembered that matching alone is not sufficient to eliminate a variable's confounding influence; failure to consider a matching variable in the analysis of the study can lead to a biased result (Rothman et al., 2008). Analyses of studies that have matched controls to cases on a given characteristic can adjust for that characteristic as if no matching had taken place. Alternatively, these analyses can explicitly consider cases and controls as matched sets. In the instance of matched case–control pairs and a dichotomous exposure variable, the following table could be constructed. See Table 5.4.2.

Table 5.4.2 Matched case-control study

Cases	Controls	
	Exposed	Non-exposed
Exposed	*a*	*b*
Non-exposed	*c*	*d*

Only the *b* pairs in which the case was exposed but not the matched control, and the *c* pairs in which the reverse was true, would enter the analysis. The odds ratio would be calculated as *b/c*. When there is more than one control per case, the matched odds ratio can be calculated as well (Breslow and Day 1980).

Minimizing information bias

For case–control studies in which information on exposure status is sought via an interview or questionnaire, the chief safeguards against information

bias entail asking questions about events that are salient to the respondent, that are framed in an unambiguous way, and that are presented identically to both cases and controls. Employment of these safeguards, however, will not prevent differential accuracy of reporting between cases and controls in all circumstances. Some past exposures or events will simply be more salient to people with an illness, who might have dwelled on possible reasons for its occurrence, than to people without that illness. Other exposures may be viewed as socially undesirable, and there may be a difference between cases and healthy controls in their willingness to admit to them. If the anticipated difference in the quality of information between cases and otherwise appropriate controls is too great, a control group that is less than ideal in other respects may be selected instead so as to minimize the potential for information bias. For example, some studies of prenatal risk factors for a particular congenital malformation that utilize maternal interviews as the source of exposure data have selected as controls infants with other malformations (Lieff et al. 1999). This control group will provide a more valid result than a control group that consists of infants in general if mothers of malformed and mothers of normal infants report prenatal exposures to a different degree even in the absence of an association, and the exposure in question is not associated with the occurrence of the malformations present in control infants.

Similar reasoning led Daling et al. (1987), when conducting their case–control study on anal cancer and a history of anal intercourse, to eschew the geographic population from which their cases had arisen as a sampling frame for controls. They feared that interviews that sought information about prior anal intercourse might be more complete among men with cancer than men in the population at large. Thus, they chose as controls men with a cancer of a different site (colon), which they believed was unlikely to have been aetiologically related to prior anal intercourse.

When the exposure under consideration is sufficiently imprecise or is open to subjective interpretation, there may not be any control group that will provide information comparable to that provided by cases. An instructive example comes from a case–control study on Down's syndrome (Stott 1958), conducted shortly before the chromosomal basis for the aetiology of this condition had been learned. The study sought to determine whether emotional 'shocks' during pregnancy might be a risk factor. The author interviewed mothers of children with Down's syndrome with regard

to the occurrence of a ‘situation or event [that would be] stress- or shock-producing if this would have been its expected effect on an emotionally stable woman’. Identical interviews were administered to mothers of normal children, and also to mothers of children with mental retardation who did not have Down’s syndrome. Even though it is not possible that an emotional shock in pregnancy could play any aetiological role in a condition already determined at conception, a far higher proportion of mothers of cases of Down’s syndrome reported an emotional shock than did mothers of normal controls (RR estimated from the data, 17.0). The use of other children with mental retardation as controls only partially reduced the spuriously high RR, to a value of 4.3.

When conducting an interview-based study of a rapidly fatal disease, or a disease that impairs a person’s ability to provide valid interview data, it is necessary to obtain information from at least some surrogate respondents; typically, these respondents are close relatives of the cases. In general, for purposes of comparability, similar information ought to be obtained from surrogates of controls, even though the control would be expected to provide more accurate data. Results of case–control studies based on exposure information provided by surrogate respondents need to be interpreted with particular caution. Although by no means present in every instance (Nelson et al. 1990, Campbell et al. 2007), there can be a large difference in the validity of the responses given by case and control surrogates. For example, Greenberg et al. (1985) investigated the basis for an apparent strong association between cancer mortality and ‘nuclear’ work among employees of a naval shipyard, which had been found during a comparison of work histories provided by surrogates of men who died from cancer and of those who died of other conditions. They observed that, regarding work in the nuclear part of the industry, surrogates of the cases generally provided information similar to that contained in employment records of the shipyard; in contrast, the surrogates of controls substantially misclassified the nature of their relatives’ jobs as not involving radiation. Using the data provided by employment records, which included individual radiation dosimetry, little or no association was found between cancer mortality and radiation exposure received at the shipyard (Rinsky et al. 1981).

What was undoubtedly a spuriously negative association was found in a case–control study of lung cancer and passive cigarette smoking that used,

for one analysis, information obtained from surrogate respondents (Janerich et al. 1985). In this analysis, the RR of lung cancer among non-smokers associated with a spouse's having smoked—0.33 (i.e. a 67% reduction in risk)—would seem almost certainly due to a spurious minimization or the denial of smoking by spouses of cases, who may have feared their habit caused their spouse to develop lung cancer.

In most interview-based case–control studies, information is not sought from cases until days to months have elapsed following the diagnosis of their condition. These persons are queried regarding exposures that occurred during what is presumed to be the aetiologically relevant period prior to diagnosis. For controls, for whom there is no 'diagnosis' date, what is the appropriate timeframe in which to focus the questions that are asked of them? If exposures that occurred sometime in the past, or that occurred over an extended period, are most likely to be relevant, controls are generally asked about events during the same calendar time as the cases. However, when studying exposures that could act acutely to lead to illness, having the case recall events immediately prior to illness and the control do the same for the corresponding time in the past may lead to bias: Only the cases will have the onset of their illness to help in recalling events that took place shortly prior to that time. This difference could lead to relatively more complete ascertainment of exposure among cases.

In response to this concern, some case–control studies investigating potentially rapidly acting exposures query controls about the time period prior to the date of interview instead of the date of their case's diagnosis (or, in an unmatched study, the date of a typical case's diagnosis). So, in a study of risk factors for meningococcal disease in adolescents (Tully et al. 2006), cases were asked about events during the two weeks prior to diagnosis (which occurred, on average, 53 days before the interview) and controls about events during the 2 weeks just prior to the interview itself. The investigators felt that the plausibility of case–control comparability on recall of exposures involved in transmission (e.g. 'intimate kissing with multiple partners') would be greater using this approach than one in which the controls were asked about a 2-week period ending 53 days prior to the interview. However, although a sound approach when trying to estimate the short-term influence of many exposures, this strategy can backfire if recall of an exposure diminishes over time to the same degree in cases and controls. For example, in the study of meningococcal disease, a far smaller

proportion of cases than controls reported attending religious observances during their respective two-week intervals (adjusted odds ratio = 0.1, 95% confidence interval = 0.02–0.6). Almost certainly, this difference had more to do with relatively poorer recall of churchgoing among cases (who had to think back some 2 months) than to a genuine salutary effect of church attendance on the incidence of meningococcal disease.

Incomparable assessment of exposure status between cases and controls is not confined to interview- or questionnaire-based studies. Most laboratory-based studies seek to prevent this by testing samples blind to case and control status. If feasible, it is desirable to also carry out this blinding for studies in which exposure is to be determined from medical or other records. However, there are instances in which the nature of the information available in records has already been influenced by whether the subject is a case or a control. For example, it was found that among 100 infertile women who underwent laparoscopy (Strathy et al. 1982), 21 had endometriosis. Only 2% of the 200 women who underwent laparoscopy for another indication, tubal ligation, were noted in the procedural records as having endometriosis. However, the interpretation of this association is unclear, because the identification and/or recording of endometriosis in cases and controls (women undergoing tubal ligation) may well have been incomparable—only in the infertile women was the laparoscopy expressly done as a diagnostic tool to investigate the possible presence of conditions such as endometriosis.

Estimating attributable risk from results of case–control studies

Occasionally, a case–control study identifies a large odds ratio relating an exposure to a disease, and for this and other reasons, a causal influence of the exposure may be suspected. The decision to seek to limit or eliminate that exposure requires weighing its negative and positive consequences. This weighing must be done in absolute, rather than in relative terms, because the same relative increase (or decrease) in risk is of far greater consequence for common than for rare outcomes. The absolute increase in the risk of disease believed to be due to a dichotomous exposure, sometimes referred to as the ‘attributable risk’ (AR), can be estimated directly from

data gathered in cohort studies or randomized trials as the difference between the incidence among exposed (I_e) and non-exposed people (I_n). The formula $I_e - I_n$ can be rewritten as RR(I_n) − I_n, or as I_n(RR − 1). Because the RR can be estimated from the results of a case–control study by means of the odds ratio, the only additional piece of information needed to estimate the attributable risk is an estimate of I_n. For the population in which the study has been conducted, I_n can be estimated if:

The overall incidence (I) of the disease in that population is known or can be approximated.
The frequency of exposure (P_e) in the controls selected for study reasonably reflects that of the population that gave rise to the cases.

Given (1) and (2),

$$\begin{aligned} I &= I_e(P_e) + I_n(1 - P_e) \\ &= I_n \mathrm{RR}(P_e) + I_n(1 - P_e) \\ &= I_n\left(P_e[\mathrm{RR} - 1] + 1\right) \end{aligned}$$

So,

$$In = \frac{I}{P_e(\mathrm{RR} - 1) + 1}$$

Therefore,

$$\mathrm{AR} = \frac{I(\mathrm{RR} - 1)}{P_e + (\mathrm{RR} - 1)} = \frac{I}{P_e + 1/(\mathrm{RR} - 1)}.$$

For example, consider a disease with an incidence rate of 10 per 100,000 per year in a population in which 5% of people have been exposed during a relevant period of time. Table 5.4.3 summarizes data from a case–control study conducted in that population:

Table 5.4.3 Data from a case-control study

Exposed	Cases (%)	Controls (%)	OR
Yes	15	5	3.35
No	85	95	1

The attributable risk that corresponds to the estimated 3.35-fold increase in risk is

$$10 = 0.05 + 1/(3.35 - 1)\ 21.0 \text{ per } 100{,}000 \text{ per year}$$

From the results of case–control studies that suggest a causal relation, it is also possible to estimate the percentage of exposed people with the disease who developed it because of their exposure, rather than through one or more causal pathways not involving the exposure. This measure, often termed the 'attributable risk per cent' (AR%) among exposed people, is defined as

$$\frac{I_e - I_u}{I_e} \times 100.$$

It can be described in terms of the RR alone:

$$\text{AR\%} = \frac{I_e}{I_e} - \frac{I_u}{I_e} = 1 - \frac{1}{\text{RR}} = \frac{\text{RR} - 1}{\text{RR}} \times 100.$$

Therefore, the results of a case–control study that provide a valid estimate of the RR (via the odds ratio) can provide the attributable risk per cent as well, with no additional assumptions or sources of data. It is also possible to estimate the percentage of a disease's occurrence in the population as a whole that resulted from the actions of given exposure. This measure, the 'population attributable risk per cent' (PAR%) or 'aetiological fraction', is simply the attributable risk per cent multiplied by the proportion of cases in that population who were exposed (P_e):

$$\text{PAR\%} = \text{AR\%} \times P_e$$

In the present example,

$$AR\% = \frac{3.35 - 1}{3.35} = 70.1\%$$

and

$$PAR\% = 70.1\% \ (0.15) = 10.5\%$$

The role of case–control studies in understanding disease aetiology

Randomized trials will not be able to answer all questions regarding the reasons for disease occurrence. Many potential disease-causing or disease-preventing exposures cannot be manipulated, either at all—for example, most genetic characteristics—or in any practical way for purposes of a study. For many exposure–disease relationships, either the disease is too uncommon or the induction period is too long to conduct a randomized trial that is not unfeasibly large in size or long in duration. Finally, it generally will not be possible to conduct separate randomized trials to measure the impact of all potential types, amounts, and durations of a class of exposure.

Also, it is not possible to rely solely on cohort studies for answers. Just as with randomized trials, the disease outcome being studied may be too rare to allow a cohort approach to be useful. This explains why the aetiologies of vaginal adenocarcinoma and Reye's syndrome, for example, have been evaluated exclusively by case–control studies—these diseases are simply too uncommon for most cohort studies to generate any cases, even in 'exposed' individuals. Prospective cohort studies are also of limited use when the induction period for the exposure–disease relationship is either very short or very long. If the induction period is very short and the exposure status of an individual varies over time, a cohort study would need to assess exposure status repeatedly among cohort members. For this reason, studies of alcohol consumption and the occurrence of injuries typically are case–control in nature (Holcomb 1938). Similarly, unless information on exposure status can be ascertained retrospectively, it would

not be feasible to initiate a cohort study of a suspected aetiologic relation that requires a very long time (perhaps several decades) to manifest itself.

Although case–control studies may be of particular value in the evaluation of the aetiology of uncommon disease, they may have difficulty in obtaining statistically precise results if the frequency of the exposure in the population under study is either extremely common or extremely uncommon (Crombie 1981). Thus, only an association as strong as the one between cigarette smoking and lung cancer could have emerged reliably from case–control studies of several hundred British men conducted in the late 1940s (Doll and Hill 1950), given that well over 90% of that population were cigarette smokers. For very uncommon exposures—for example, occupational exposure to a specific substance suspected of posing a risk to health, or an infrequently prescribed drug—barring a strong observed association based on a large number of subjects, even the best-designed case–control study will usually offer no more than a suggestion of the presence or absence of a relation with regard to the occurrence of a given illness.

Conclusion

Case–control studies compare ill or injured persons (cases) with those at risk of the illness or injury (controls) in terms of prior exposures or characteristics.

From most case–control studies of a given condition it is possible to obtain an estimate of the RR associated with an exposure by dividing the odds of exposure in cases by the odds of exposure in controls (odds ratio).

The validity of the results of a case–control study can be compromised by differences between cases and controls with regard to exposures or characteristics that are also correlated with the likelihood or degree of the exposure in question.

In most case–control studies, ascertainment of exposure status is undertaken after the cases have sustained their illness or injury. These studies will produce an unbiased estimate of an exposure–disease association to the extent that exposure status prior to disease onset can be ascertained with a high degree of sensitivity and specificity.

In a setting in which (a) randomization of potential study participants is unethical or infeasible and (b) the outcome under study is an uncommon one, data from case–control studies may be the best available by which to judge a potential cause–effect relationship between the exposure and outcome.

REFERENCES

Bloemenkamp, K.W.M., et al. (1999). Risk of venous thrombosis with use of current low-dose oral contraceptives is not explained by diagnostic suspicion and referral bias. *Archives of Internal Medicine*, 159, 65–70.

Breslow, N.E. and Day, N.E. (1980). *Statistical Methods in Cancer Research. Vol. 1: The Analysis of Case-Control Studies. IARC Scientific Publication No. 32*. Lyon, France: IARC Press.

Campbell, P.T., Sloan, M., and Kreiger, N. (2007). Utility of proxy versus index respondent information in a population-based case-control study of rapidly fatal cancers. *Annals of Epidemiology*, 17, 253–7.

Crombie, I.K. (1981). The limitations of case-control studies in the detection of environmental carcinogens. *Journal of Epidemiology and Community Health*, 35, 281–7.

Cummings, P. and Weiss, N.S. (1998). Case series and exposure series: the role of studies without controls in providing information about the etiology of injury or disease. *Injury Prevention*, 4, 34–57.

Daling, J.R., et al. (1982). Correlates of homosexual behavior and the incidence of anal cancer. *The Journal of the American Medical Association*, 247, 1988–90.

Daling, J.R., et al. (1987). Sexual practices, sexually transmitted disease, and the incidence of anal cancer. *New England Journal of Medicine*, 317, 973–7.

Doll, R. and Hill, A.B. (1950). Smoking and carcinoma of the lung. *British Medical Journal*, 2, 739–48.

Greenberg, E.R., et al. (1985). An investigation of bias in a study of nuclear shipyard workers. *American Journal of Epidemiology*, 121, 301–8.
Greenland, S. and Thomas, D.C. On the need for the rare disease assumption in case-control studies. *American Journal of Epidemiology* 1982, *116*, 547–53.
Greenwald, P., et al. (1971). Vaginal cancer after maternal treatment with synthetic estrogens. *New England Journal of Medicine*, 285, 390–93.
Herbst, A.L., Ulfelder, H., and Poskanzer, D.C. (1971). Adenocarcinoma of the vagina: association of maternal stilbestrol therapy with tumor appearance in young women. *New England Journal of Medicine*, 284, 878–81.
Holcomb, R.L. (1938). Alcohol in relation to traffic accidents. *Journal of the American Medical Association*, 111, 1076–85.
Horwitz, R.I. and Feinstein, A.R. (1978). Alternative analytic methods for case-control studies of estrogens and endometrial cancer. *New England Journal of Medicine*, 299, 1089–94.
Hurwitz, E.S., et al. (1985). Public Health Service Study on Reye's syndrome and medications. Report of the pilot phase. *New England Journal of Medicine*, 313, 849–57.
Jackson, M.L., et al. (2017). Influenza vaccine effectiveness in the United States during the 2015–2016 season. *New England Journal of Medicine*, 377, 534–43.
Janerich, D.T., et al. Lung cancer and exposure to tobacco smoke in the household. *New England Journal of Medicine* 1985, 323, 632–6.
Laumon, B., et al. (2005). Cannabis intoxication and fatal road crashes in France: population-based case-control study. *British Medical Journal*, 331, 1371–4.
Lieff, S., et al. (1999). Selection bias and the use of controls with malformations in case-control studies of birth defects. *Epidemiology*, 10, 238–41.
McLaughlin, J.K., et al. (1985). Problems in the use of dead controls in case-control studies. II. Effect of excluding certain causes of death. *American Journal of Epidemiology*, 122, 485–94.
Nelson, L.M., et al. (1990). Proxy respondents in epidemiologic research. *Epidemiologic Reviews*, 12, 71–86.
Pearce, N. (1993). What does the odds ratio estimate in a case-control study? *International Journal of Epidemiology*, 22, 1189–92.
Rinsky, R.A., et al. (1981). Cancer mortality at a naval nuclear shipyard. *Lancet*, 1, 231–5.
Rothman, K.J., Greenland, S., and Lash, T.L. (2008). *Modern Epidemiology* (3rd ed.). Philadelphia, PA: Lippincott Williams & Wilkins.
Selby, J.V. (1994). Case-control evaluations of treatment and program efficacy. *Epidemiologic Reviews*, 46, 91–101.
Shapiro, S., et al. (1985). Risk of localized and widespread endometrial cancer in relation to recent and discontinued use of conjugated estrogens. *New England Journal of Medicine*, 313, 969–72.
Silverman, D.T., Hoover, R.N., and Swanson, G.M. (1983). Artificial sweeteners and lower urinary tract cancer: hospital vs. population controls. *American Journal of Epidemiology*, 117, 326–34.
Siscovick, D.S., et al. (1982). Physical activity and primary cardiac arrest. *Journal of the American Medical Association*, 248, 3113–17.
Smith, G.S., et al. (2001). Drinking and recreational boating fatalities: a population-based case-control study. *The Journal of the American Medical Association*, 286, 2974–80.
Stott, D.H. (1958). Some psychosomatic aspects of casualty in reproduction. *Journal of Psychosomatic Research*, 3, 42–55.
Strathy, J.H., et al. (1982). Endometriosis and infertility: a laparoscopic study of endometriosis among fertile and infertile women. *Fertility and Sterility*, 38, 667–72.
Tabuenca, J.M. (1981). Toxic-allergic syndrome caused by ingestion of rapseed oil denatured with aniline. *Lancet*, 2, 567–68.

Tully, J., et al. (2006). Risks and protective factors for meningococcal disease in adolescents. *British Medical Journal*, 332, 445–50.

Victora, C.G., et al. (1989). Infant feeding and deaths due to diarrhea. *American Journal of Epidemiology*, 129, 1032–41.

Weiss, N.S., et al. (1981). Incidence of ovarian cancer in relation to the use of oral contraceptives. *International Journal of Cancer*, 28, 669–71.

Weiss, N.S. (1994). Application of the case-control method in the evaluation of screening. *Epidemiologic Reviews*, 16, 102–8.

White, E., Armstrong, B.K., and Saracci, R. (2008). *Principles of Exposure Measurement in Epidemiology: Collecting, Evaluating and Improving Measures of Disease Risk Factors* (2nd ed.). Oxford: Oxford University Press.

Witte, J.S., Gauderman, W.J., and Thomas, D.C. (1999). Asymptotic bias and efficiency in case-control studies of candidate genes and gene-environment interactions: basic family designs. *American Journal of Epidemiology*, 149, 693–705.

Yang, Q. and Khoury, M.J. (1997). Evolving methods in genetic epidemiology. III. Gene-environment interaction in epidemiologic research. *Epidemiologic Reviews*, 19, 33–43.

5.5

Cohort studies

Alvaro Muñoz and F. Javier Nieto

Introduction to cohort studies

Cohort studies constitute one of the basic types of designs in epidemiological research. The key element of cohort studies is temporality (i.e. exposure preceding an outcome). Specifically, a cohort (i.e. group of individuals) is enrolled at the beginning of the study (baseline) to examine how characteristics (exposures) predict the occurrence of a discrete event (typically a disease onset) or changes in markers of health/disease status. In the former case, the cohort is typically formed by individuals who are free from the outcome of interest and have sufficiently heterogeneous profiles of putative risk factors (whose simplest form is as exposed or unexposed). In the latter case, the central aim is to characterize the heterogeneity of the change or trajectory of markers of disease progression according to constellations of putative risk factors.

The identification of exposures and risk factors for disease provides a basis for prevention. There are numerous examples of cohort-based exposure–disease associations that have resulted in beneficial prevention strategies: for instance, to prevent lung cancer, cigarette smoking should be avoided; to prevent infection with human immunodeficiency virus (HIV), high-risk sexual practices and injection with unsterile needles should be avoided; and to prevent heart disease, adequate lipid profiles should be achieved (by diet, exercise, and/or lipid-lowering medications). Needless to say, there are instances in which the design of prevention strategies faces the complexity of multifactorial aetiology and of exposures that cluster so that exposures have to be taken into account jointly. When an exposure is not a *sufficient* cause but a *component* of a more complex constellation of causal factors (Rothman et al. 2007; Szklo and Nieto 2018), acting only on that exposure may not result in successful prevention strategies; for example, based on data from cohort studies, high levels of beta carotene markers (purportedly resulting from appropriate diets) had been associated with protection against cancer (Willett et al. 1984), but supplementation with beta carotene only has not resulted in a lower incidence of cancer in clinical trials (Alpha-

Tocopherol, Beta Carotene Cancer Prevention Study Group 1994; Virtamo et al. 2003).

Types of cohort studies

The observational cohort design shares with the experimental design or controlled clinical trial the temporal relationship between exposure and disease. The clinical trial design is a special type of cohort study in which the study subjects are randomly assigned to the different experimental (exposure) groups. Given the inherent limitation of observational designs to control for potential confounding variables unidentified by the investigator, the difference between these study designs lies on the validity of the conclusions on causal inferences. The role of randomization in clinical trials is to make the two groups comparable so that the effect of a therapy can be determined. But randomized controlled trials are not ethical or feasible when exploring the effects of potentially harmful exposures; moreover, the generalizability of results in highly selected trial participants is often limited. This is the reason why observational cohort studies, with all their limitations and caveats, will always be an important study design in epidemiological research.

In some studies, the investigator collects the data *concurrently* with the study being conducted, and in others, the data are available before the study is designed and the study is conducted to obtain that data (i.e. the data for the study precedes its conduct or the study is *non-concurrent*). This type of design is called ‘retrospective study’ by some authors (Vandenbroucke 1991), a term that can cause confusion because it is also used for case–control studies.

The simplest cohort design is to obtain exposure data at baseline and follow-up individuals to only obtain data for when the event of interest occurs. A richer design includes regularly scheduled visits to update exposure data, including repeated physical exams and collection of biological samples, as appropriate. When the outcome of interest is not an event but is change in the profile of a biomarker over time (e.g. change in kidney function measured by filtration rate), cohort studies are often referred to as *longitudinal studies*.

Design issues in cohort studies

The nature of a cohort study provides the temporal structure necessary to associate a particular exposure with a subsequent event of interest. The time at which an exposure first occurs, placing individuals at risk for event development, is referred

to as the *origin* (e.g. birth in population-level life-course epidemiology, infection with a microorganism in infectious disease epidemiology). Ideally, a cohort study will have complete data of the time elapsed between the origin and the event of interest for as many participants as possible. However, in most studies, not all individuals are followed up from the origin: some will have been at risk for some time since the origin when enrolled into the cohort study. Those followed up from the origin are called the *incident* cohort and those entering the study after a certain duration from the origin are called the *prevalent* cohort. Implementation of the appropriate statistical methods to combine incident and prevalent cohorts may not only yield more efficient estimates but also provide the means for appropriately describing the timing of events beyond the study duration (Muñoz et al. 1997).

For some diseases, the origin is well defined and observable, such as employment into an occupation deemed to place workers at risk for health events. For others, the origin is well defined but difficult to observe, such as infection with a virus. Still others do not have a well-defined origin, as in the case of coronary heart disease in which the origin may be the time of initiation of atherosclerotic plaques in the inner layer of the coronary arteries, the time when a certain level of coronary artery stenosis is reached, or the time when an atherosclerotic plaque ruptures and acutely occludes an arterial branch leading to clinical manifestations (angina or myocardial infarction). Cohort studies with data on individuals from origin to the event of interest as well as with baseline information and updated information on predictors at regular intervals between the origin and event provide the most complete data to describe the natural history of a disease (Muñoz et al. 1992; Muñoz and Xu 1996; Muñoz et al. 1997).

Study population

The cohort can be the totality of a geographically and temporally defined population. For example, Hoffmans et al. (Hoffmans et al. 1988) constructed a non-concurrent study of mortality among Dutch males born in 1932. This cohort included the totality of the 84,349 Dutch males who were 18 years old in 1950, when they were examined for eligibility to military service; the investigators used the data collected at baseline (when the individuals were 18 years old) to describe the relationship between obesity at 18 years and subsequent cause-specific mortality (Hoffmans et al. 1989).

Instead of the totality of a population, it is more common to include only a fraction or a sample of a population defined temporally and geographically. For instance, as outlined in Table 5.5.1, the Framingham study was initiated at the beginning of the 1950s in Framingham, Massachusetts, United States, with the goal of studying the incidence of and risk factors for cardiovascular diseases

(Dawber 1980). From a list of residents aged 35–59 years (18,000 persons in total), 6,507 adults were randomly selected, among whom 4,469 agreed to participate in the initial examination. To compensate for the unexpectedly high rate of refusal, 734 volunteers were also added as members of the cohort. Of the resulting 5,203 participants in the baseline examination, 76 persons were excluded for having clinical manifestations of cardiovascular disease, leaving a total of 5,127 participants in the Framingham Heart Study cohort. The follow-up consisted of physical examinations every 2 years with comprehensive data collection on suspected risk factors.

In order to improve generalizability, subsequent cohort studies of cardiovascular disease conducted in the 1980s aimed at recruiting more diverse geographic, ethnic, and socioeconomic groups in the United States. Specifically, the CARDIA study (Friedman et al. 1988) recruited a total of 5,115 persons between 18 and 30 years of age in four states; the Atherosclerosis Risk in Communities (ARIC) study (ARIC Investigators 1989) cohort comprised 15,800 persons aged 45–64 years at the beginning of the study and residents of four states; and the CHS (Fried et al. 1991) followed up 2,955 participants 65 years or older in another four communities of the United States.

In addition to strict temporal and geographic criteria, a cohort can also be defined by a group of individuals sharing common characteristics, which facilitates the follow-up and the accuracy of the data to be collected. For example, the Multicenter AIDS Cohort Study (MACS) (Kaslow et al. 1987; Dudley et al. 1995) comprised homosexual men in four metropolitan areas of the United States with the overall aim of describing the natural history of HIV infection and AIDS (see Table 5.5.1).

From April 1984 to March 1985, a cohort of 4,954 men were recruited into the MACS in four metropolitan areas of the United States (Baltimore, Chicago, Los Angeles, and Pittsburgh). To increase minority enrolment, an additional 668 men, 69.3% of whom were non-Caucasian, were recruited in 1987–1991. To address issues of long-term effectiveness of therapies and their putative adverse effects, 1,397 men were additionally enrolled in 2001–2010. Therefore, the entire MACS cohort as of September 2012 consisted of 7,019 men with three primary inclusion criteria: men having sex with men, older than 18 years, and AIDS-free at enrolment. Of them, 2,905 (41%) were seroprevalent for HIV at entry into the study. All men were followed up every 6 months with repeat interviews, physical examinations, and collection of blood used to characterize the level of HIV-induced immunosuppression. Confirmation of AIDS diagnoses was made by obtaining physician and hospital summaries and by reviewing medical records and deaths were monitored by follow-up and through an ongoing search of death records.

When a specific registry with baseline data is available, it provides the opportunity to conduct a non-concurrent cohort study. For example, Nieto et al. (1992) used a registry of physical examinations carried out on schoolchildren between 1933 and 1945 in Hagerstown (Maryland, United States) to assemble a non-concurrent cohort of about 13,000 children in order to study the relationship between childhood obesity and adult mortality; the baseline information obtained between 1933 and 1945 was linked to the mortality experience of the cohort up to 1985. In other historical cohorts, the study population is defined from the beginning as a function of the exposure factors to be investigated. Thus, participants are selected explicitly according to the presence or absence of the characteristics considered to be possible risk factors, such as a cohort of Norwegian electricians. As summarized in Table 5.5.1, this historical cohort of 37,945 workers was defined from the registry of occupations of the working population in Norway according to the 1960 census (Tynes et al. 1992).

Table 5.5.1 Examples of cohort studies

	Framingham (Dawber 1980)	**MACS (Kaslow et al. 1987; Dudley et al. 1995)**	**Norwegian electricians (Tynes et al. 1992)**
Study aims	Incidence and risk factors for cardiovascular diseases and hypertension	Natural history of HIV infection and AIDS	Relationship between electromagnetic radiation and cancer incidence
Type of study	Concurrent	Concurrent	Non-concurrent
Date	1949– current	1984– current	1960–1991
Study Population	Adults resident in Framingham, Massachusetts, USA	Homosexual men in four metropolitan areas of the USA	Electricians in 1960 census in Norway
Simple size	5,127	7,019	37,945
Sex	Male and female	Male	Male
Age at baseline (years)	30–62	18–70	20–70
Active follow-up	Study visits every 2 years including physical exam, electrocardiogram, and laboratory exams	Study visits every 6 months with extensive interviews including behaviour and medical care, physical exam, laboratory exams, and collection of blood samples for national repository	–
Passive follow-up	• Clinical records (in case of hospitalization) • Death certificates	• Clinical records • Death certificates • National Death Index	• Cancer registries • Death certificates

With the increasing availability of electronic health records, new opportunities to use administrative healthcare databases as a basis for defining and conducting large-scale cohort studies are emerging (Phillips et al. 1999; Sterne et al. 2009; Manolio et al. 2012). These approaches can be particularly powerful when combined with a comprehensive system of national registries, such as those offered by countries like Denmark (Frank 2000). On the other hand, they may be limited by the relative scarcity of ‘negative’ controls (i.e. healthy subjects) and by an under-representation of people who do not have access to (or choose not to take

advantage of) healthcare services. In addition, the quality of the data collected for clinical purposes is variable, not usually subject to the rigours of standardization and quality typical in research settings, especially when studying phenomena that are not directly related to clinical care (e.g. upstream community determinants, psychosocial factors).

Finally, cohort studies can outlive themselves such as when the offspring of the cohort members are recruited with the purpose of conducting family and genetic studies. The Framingham Heart Study was again a pioneer in this methodology, when in the late 1970s the daughters and sons of the original cohort members (and their spouses) were recruited to constitute the Framingham Offspring Study (Kannel et al. 1979; Djoussé et al. 2002).

Baseline visit

The objective of the baseline visit in a concurrent cohort study is generally the classification of the study participants according to the exposure factor or factors under investigation, as well as according to potentially relevant covariates (confounders) and/or effect modifiers. For example, Furth et al. (2006, 2011) designed a cohort of children with chronic kidney disease to determine the risk factors for progression towards end-stage renal disease; for this, they recruited 586 children aged 1–16 years with an estimated glomerular filtration rate of between 30 and 90 mL/min/1.73 m^2, indicating a level of insufficiency in their kidney function that put them at risk of end-stage renal disease. Also important is the collection of information that facilitates the subsequent tracking of study participants during the follow-up including information on contacts.

In its simplest form, the baseline visit can take place over a single continuous time period until the desired sample size is reached. In many cases, however, more participants are recruited: (1) to replace those who have developed the event of interest or who have been lost to follow-up, (2) to include groups of individuals who were under-represented in the original recruitment, and/or (3) to respond to new scientific challenges due to the changing nature of the epidemiology of the disease of interest.

Follow-up

The follow-up can be active, by repeated contacts via successive visits to a health centre, mailed questionnaires, phone calls, or online contacts. Sometimes, the different forms of contacting participants are alternated. For example, in the follow-up of the ARIC study, participants were invited every 3 years to return to the clinic for a new physical examination and complementary assessments (ARIC

Investigators 1989); in addition, each participant was called every year and asked about illnesses and hospitalizations occurring in the interim period since the last clinic visit. This is in contrast to clinic-based cohort studies in which visits take place according to the health needs of the participants.

For certain diseases and events (e.g. death, cancer), external sources (i.e. national death index, registries) can be used to record events occurring among the individuals enrolled in the study. The type and automation of the record linkage system is determined in each case by the availability of data and the characteristics of each particular registry.

The quality of the information about the occurrence of events is fundamental for correct estimation of the incidence and its determinants. Therefore, in many cohort studies, a special effort is made to ensure that the information about disease occurrence or death is accurate. Generally, events recorded in interviews with study participants are verified by requesting documents such as the clinical records, not only to verify the occurrence of the event but also to improve the information regarding diagnosis (e.g. to obtain the histological type in a study of lung cancer; to distinguish between thrombotic and haemorrhagic episodes in a study of cerebrovascular events). A major challenge in cohort studies with cause-specific mortality as the event of interest is the determination of the cause of death, which may require confirmation by inspecting clinical records or by interviewing the physician who attended to the participant before his or her death.

Types and measures of exposures

The classification of study participants according to exposure factors is one of the most complex challenges in a cohort study. *Exposure factor* means any attribute of the subject or any external agent that can influence the subject's health (Armstrong et al. 2008).

Types of exposures

Exposure factors include host factors (such as age, sex, race, genetic factors, metabolic factors, and immune function) and environmental factors (such as viruses and bacteria, environmental pollutants, diet, alcohol, and/or tobacco consumption). Sometimes, an exposure factor represents a set of factors that are difficult to separate. For example, social class encompasses external environmental factors, related to individual behaviour, access to healthcare services, socioeconomic status, and others.

The exposures can be either fixed over time (e.g. race) or variable. The latter could be exposures that change directly with time (e.g. age and calendar) or change at their own pace with time (e.g. cigarettes smoked per day, biological marker levels, diet). When exposures not only change over time but do so heterogeneously (i.e. change differently over time in different individuals), they are referred to as *internal* (i.e. at the individual level) time-dependent exposures. In contrast, exposures that change over time but affect groups or populations who are subject to the exposure in a similar way are referred to as *external* (i.e. at the population level) time-dependent exposures (i.e. they change equally over time for different individuals comprising a given population) such as for example, calendar time, seasonal weather changes, levels of air pollution in a given region, supplements in the food, and introduction of newly approved medical therapies.

Follow-up visits provide the opportunity for collecting data to update internal time-dependent exposures. External time-dependent exposures may be obtained from external sources, but data obtained at the individual level can sometimes be used to define eras (i.e. calendar periods) when there are changes at the population level. For example, the MACS started in 1984 and has since been conducted concurrently with major advances in therapies effective against viral replication in HIV-infected individuals. In 1998, Detels et al. (Detels et al. 1998) determined the effectiveness of the potent antiretroviral therapies introduced in 1995 by showing a doubling of the AIDS-free times under the conditions of the first 2 years (July 1995–July 1997) of the highly active antiretroviral therapy (HAART) era (Detels et al. 1998). In 2005, Schneider et al. (Schneider et al. 2005) determined the population effectiveness of prolonged use of HAART by showing that in the era when HAART was well established (i.e. July 2001–December 2005, 6–10 years after its introduction in July 1995), the median survival time after AIDS diagnosis increased tenfold from the 1.5 years in the era when no therapies were available (Schneider et al. 2005).

In cohort studies, two types of effects of interventions can be measured. The first can be termed ‘individual effectiveness’, which mimics clinical trials by using treatment data at the individual level. These analyses must overcome the lack of randomization and the confounding by indication, whereby those individuals at more advanced disease stages are the ones more likely to receive the therapies (Ahdieh et al. 2000). Measures of individual effectiveness *supplement* the results of clinical trials, but are usually subject to an unknown amount of residual confounding (Phillips et al. 1999). Data analysis methods for individual effectiveness require the use of multivariate methods of survival analysis, including the use of therapy as an internal time-dependent covariate; however, the complexity of the selection of individuals receiving therapies may require the use of more elaborate methods for causal inferences (Robins et al. 2000; Cole et al.

2003). These methods include propensity score matching methods (Stuart 2010), inverse probability weights based on propensity scores (Robins et al. 2000), and marginal structural models (Hernán et al. 2001). The guiding principle of these approaches is to construct a sample in which two or more groups (e.g. exposed/unexposed) have balanced distributions of confounders (Stuart 2010).

There is growing interest in determining the effectiveness of the many therapies participants in cohort studies are observed to receive but there is strong confounding by indication as therapies tend to be given to individuals who fulfil guidelines for who and when to treat. Standard regression adjustments to compare outcomes of treated to those untreated seldom achieve control for the indication for receiving treatment. Alternative approaches include the use of inverse probability weights based on probability of receiving treatment. Inverse probability weights estimators provide a means for estimating the causal effects of treatment among individuals with different covariate values at study baseline (Robins et al. 2000). However, covariates are often dynamic and changing values may alter the impact of therapy on the outcome. Further, marginal structural models are limited in their flexibility, becoming unstable if certain covariate values yield very large or very small probabilities of treatment. Inverse probability weights methods offer powerful tools to deal with complex patterns of confounding, particularly in longitudinal data, where outcomes at one time point may influence both subsequent treatments and outcomes. More elaborate methods such as marginal structural models have been proposed to deal with these types of patterns, whereby the weights in these models are updated longitudinally using the complete past history. In addition, the weights are stabilized, resulting in improved performance. Many scientific initiatives present research questions involving strong confounders for which these approaches may overcome the limitations of classical multivariate regression methods.

The second type of effect that can be measured in a cohort study can be termed ‘population effectiveness’ which quantifies the reduction of disease in the population when (typically) the most ill are treated relative to the occurrence of disease in the population when none or only few are treated. Because the introduction and use of therapies are closely linked to calendar time, the primary comparison may be characterized by time periods. When periods are characterized by differential availability of therapies, they could then play the role of an *instrumental variable* (related to therapy but not to disease progression or markers indicative of therapy need) and thus provide an accurate estimate of the effect of the benefits of therapy at the population level (Greenland 2000; Martens et al. 2006). To control for survival bias and overall disease progression, this approach requires the comparison of groups reaching similar time at risk (e.g. duration of infection or time since disease diagnosis) in different eras of therapy defined by

calendar periods (Muñoz and Hoover 1995; Detels et al. 1998; Schneider et al. 2005). Measures of population effectiveness *complement* the efficacy measured in clinical trials and provide a key public health index: the amount of disease burden that is reduced when only some (typically the most ill) receive the therapy of interest (Muñoz et al. 2000). Comprehensive data collected by cohort studies are essential to eliminate possible ecological fallaciousness (i.e. effectiveness due to changes over time other than changes in the therapies of interest). This includes not only prospectively collected data on therapy use, but also information on access, healthcare utilization and practices, and adherence at the population level.

Measures of exposures

The exposure can be defined as a quantitative (continuous) variable (e.g. blood pressure) or qualitative (categorical) variable (e.g. non-smoker, ex-smoker, current smoker).

When exposures are assessed repeatedly over time, the investigators can take into account not only the amount (dose) but also the duration of the exposure (Armstrong et al. 2008). There are several ways of quantifying dose and duration: (1) cumulative exposure (e.g. number of pack-years of smoking, number of rad-years of exposure to a radioactive compound at work), (2) the average exposure during the latest years (e.g. average number of cigarettes per day, average number of rads in the work environment), or (3) maximum exposure (e.g. maximum number of cigarettes per day, maximum level of radioactivity registered in the industry).

Instruments for the collection of individual data include questionnaires, physical examinations, and registries. Questionnaires allow not only the obtaining of information on demographic data, personal and family history, disease symptoms, behaviours, and attitudes, but also the recording of self-reported physical characteristics such as weight and height. They can be administered through personal, telephone interviews or online, they can be completed by the participants themselves, or they can be administered by a computer-assisted method, which is particularly useful for the collection of sensitive data (e.g. sexual behaviour, drug abuse). Modern technologies are significantly expanding the range of possible modalities for data collection. Web-based instruments, smartphones, and other types of mobile devices are being used to collect information on dietary habits, physical activity, blood pressure and others (Illner et al. 2012). In addition, registries are external sources that can be used to complement the active follow-up and are particularly useful to obtain vital status events or diagnosis of a condition of interest (e.g. cancer registries).

Timescales and types of outcomes

The scientific aims of a cohort study should dictate the type of outcome or measure of disease occurrence that should be collected during follow-up. The most common outcome measure is the time individuals at risk take to develop an end point. To properly describe this time, the following need to be defined: (1) the origin, which marks the time at which individuals are placed at risk for the development of the particular event of interest; (2) the end point, which typically is the onset of a clinical condition or the occurrence of death; and (3) the scale of time.

Time to event

The use of time-to-event outcomes requires that the *time* variable be anchored at the origin, resulting in the characterization of the event incidence across the time since origin. Examples of time-to-event outcomes used in epidemiological research include the following: (1) in clinical epidemiology, the classic time from diagnosis of a clinical condition to death or the time from initiating therapy to response; (2) in infectious disease epidemiology, the time from infection to overt disease (i.e. the incubation period); (3) in cancer epidemiology, the time from remission to relapse; (4) in behavioural epidemiology, the time from enrolment into a smoking cessation programme to the time the individual quits smoking; and (5) in occupational epidemiology, the time from employment to the occurrence of an occupationally related health outcome.

Types of timescales

Traditionally, the timescale used in cohort studies is the time from the beginning of the study (when the cohort is identified and the baseline exam is conducted) until the event of interest (or censoring, as discussed in the next subsection) occurs. For example, in many of the analyses of the Framingham Study data, the origin (time 0) of the follow-up is defined as the date when the baseline interview/exam of each of the participants took place in or at around 1950. The fact that individuals start at different ages is dealt with in the analysis by statistically *controlling* for baseline age.

Whether or not an origin is well defined, the use of *age* as the timescale is an appealing alternative (Korn et al. 1997). The caveat of this approach when applied to a cohort of individuals entering at different ages is that individuals are not under observation during the years from birth to the age at which they are recruited into the cohort. This needs to be handled in the analysis allowing for late

entry (left truncation—see following subsection). This method was used, for example, in a study of the relation between sleep apnoea and cancer mortality in the Wisconsin Sleep Cohort (Nieto et al. 2012) and in a study of life expectancies of HIV-infected individuals in the era of HAART (Wada et al. 2013; Wada et al. 2014).

In most cases, even for groups of individuals with a fixed exposure, the pattern of incidence of disease over the follow-up time is not the same at different durations at risk. An exposure whose deleterious effect monotonically accumulates over time will result in an incidence rate that increases with time (e.g. death after AIDS diagnosis in the absence of treatment). An exposure that first affects susceptible individuals so that those remaining free of the event over time represent resistance (i.e. some degree of immunity to the effects of exposure) will result in an increasing incidence followed by a decreasing incidence rate (e.g. a viral infection to which a subset of individuals are immune). An exposure that initially carries risk that ameliorates over time will result in a decreasing incidence rate (e.g. death after transplantation). A beneficial intervention that decreases the risk of death but whose effect is obliterated by other processes including old age will result in a bathtub hazard (Cox et al. 2013).

Incomplete observation from the origin and to the event

In some cohort studies, not all individuals are observed from the origin, but some enter the study after having been at risk for varying lengths of time. This type of incompleteness in the data is referred to as *late entries* or *staggered entries* and occurs when some individuals have experienced the origin that places them at risk for the event of interest at a date prior to their enrolment into a cohort study. Since these individuals have survived free of the event of interest from the origin to study enrolment, they have peers with similar origins who did not survive event-free and developed the event of interest prior to the cohort study enrolment, precluding their participation in the study. Such unseen individuals are *truncated*, and methods to adjust for these truncated observations need to be implemented to prevent survival bias. The presence of late entries in a study is a case of missing data, where missing data are the number and timing of the events of the truncated peers of those who entered late.

With respect to the event, observations in a cohort study are incomplete because it is usually unfeasible for the investigator to observe the event in all participating individuals. This type of incompleteness in the data is called *censoring* and can be due to any of the following: (1) individuals remaining event-free at the date of analysis, sometimes called *administrative* censoring; (2) individuals being lost to follow-up (prior to the date of analysis) before developing the event (e.g. because

migration or refusal to continue participating); or (3) deaths due to unrelated causes, precluding the observation of the event of interest. In this latter case, it is more appropriate to use methods of analysis based on competing risk models (Fine and Gray 1999; Putter et al. 2007; Checkley et al. 2010; Ng et al. 2016).

For the most part, subjects who are lost to follow-up are *right-censored*, which means that their exit from observation is the last possible time they contributed information to the study (i.e. on the timeline represented by a scale running from left to right, no data are available to the right of the time of censoring). Occasionally, information from sources external to the study (e.g. death certificates) are used to identify the interval of time during which individuals lost to follow-up and free of the event of interest are documented to have had the event because they died of the event of interest at some time point after the censoring. This interval is defined as the time between the date the individual was lost to follow-up and the date of death. Given that the individual died of the event of interest, the event must have occurred during this interval. These observations are called *interval-censored* observations. When using age as the timescale with mortality as the end point, censored observations are all interval-censored as all individuals are expected to die by age 125 (Wada et al. 2013; Wada et al. 2014).

Change of biomarkers

In some cohort studies, the primary event is not an event but whether a marker of disease changes over time differently in individuals exposed to different risk factors. For example, in the cohort study of children with moderate kidney function insufficiency conducted by Furth et al. (Furth et al. 2006), the primary objective was to identify the factors that predicted fast decline of kidney function (Pierce et al. 2011). Another example is the cohort study conducted by Tager et al. (Tager et al. 1983) in East Boston in the 1970s, showing the effect of maternal smoking on slowing the expected increase of pulmonary function in growing children. The study was seminal to the subsequent mounting evidence that second-hand (involuntary) smoking had detrimental health effects and guided the issuing of policies to ban smoking from work and public places as fundamental prevention measures for the health of the public. Furthermore, cohort studies have not only contributed substantially to the elucidation of disease aetiology but have also spurred the development of statistical methods to properly analyse data (Rosner et al. 1985).

Analysis

A fundamental condition for valid epidemiological inferences is that for the disease to be causally linked to a given exposure, the exposed and unexposed groups should be comparable with respect to other factors that are known to explain the heterogeneity of disease occurrence (i.e. absence of confounding). In cohort (observational) studies, comparability needs to be achieved by the use of appropriate analytical methods which often require approaches above and beyond classical stratification and/or regression methods. The goal of causal inference methods is to try to mimic randomization so as to assess the effect of a given exposure in groups of individuals who differ in the exposure of interest, but who are otherwise comparable with respect to all variables known to be associated with disease.

The type of outcome of interest determines the type of analytical technique most appropriate to yield a valid inference. Table 5.5.2 provides a succinct summary of the analytical methods reviewed here.

Table 5.5.2 Main methods for the analysis of cohort studies

Outcome		Summary measure	Comparison		Measure of association
Exposed/unexposed (2-sample)	Multiple (regression)				
I)	Events in person-years	Incidence rate	$(O-E)^2$/variance	Poisson	Relative incidence
II)	Time to event	Kaplan–Meier/ maximum likelihood estimates	Logrank or Mantel–Haenszel/likelihood ratio test	Proportional hazards/parametric	Relative hazard/ relative time
III)	Biomarker repeatedly measured at follow-up visits	Change	t-test for simple measures of change between baseline and last visit	Regression for correlated data; marginal, conditional, random effects	Differences in change over time

When the outcome data are collected as events in person-years, Poisson regression methods are apt for data analysis (outcome I in Table 5.5.2). These methods are particularly useful for the analysis of trends and changes in incidence of disease over calendar time or when the origin is not well defined or not of interest.

When the outcome data are times to event from a well-defined origin (outcome II in Table 5.5.2), the percentage of individuals remaining free of the event of interest at *t* units of time after the origin can be estimated by the Kaplan–Meier (Kaplan and Meier 1958) curve or by maximum likelihood methods for parametric approaches (Cox et al. 2007). Two survival curves can be compared using the Mantel–Haenszel (Mantel and Haenszel 1959) test or the likelihood ratio test (Cox et al. 2007). Regression analysis can be accomplished using semi-parametric methods for hazard functions (Cox 1972) or under richer parametric models (Cox et al. 2007) which lend themselves more directly to estimation of relative times quantifying not only how the hazard is modified by an exposure but also to what extent an exposure shortens (if deleterious) or extends (if beneficial) the event-free times. Software in several statistical packages to incorporate not only censoring but late entries is widely available (Cox et al. 2007).

When the outcome is a biomarker repeatedly measured at follow-up visits (outcome III in Table 5.5.2), regression methods for correlated data need to be used. This is so because the biomarkers measured in the same individual over the follow-up visits are correlated so that the standard methods of regression methods for statistically independent outcomes are inadequate.

Analysis based on person-time

Calculation of person-time and rates

The methods described in this subsection are useful for the analysis of the frequency of events in person-years in groups of individuals with different exposures and when anchoring at the origin is not important. To illustrate the methods, we present data from the Precursors Cohort study. This cohort comprised 1,337 medical students at the Johns Hopkins University, Baltimore, United States, who graduated between 1948 and 1964 (Klag et al. 1993). A total of 1,271 students (95%) completed a questionnaire and were examined during their stay at the school. In order to study the determining factors (‘precursors’) of general mortality and the occurrence of cardiovascular diseases, this cohort was followed up through mailed questionnaires. When the participants reported that they have had a cardiovascular event, the hospital records were requested in order to validate the diagnosis. Table 5.5.3 shows data obtained in the follow-up for each of the groups (smokers and non-smokers), the column named ‘person-years’ showing the totality of time units lived by the members of the cohort between the indicated years of age. Thus, for example, the non-smokers in the cohort lived a total of 862.5 years while being 20–24 years old and 1929.5 years while being 25–29 years of age.

Table 5.5.3 Internal comparison of the rates based on person-years in the Precursors Study. Occurrence of mortality (all-cause) until 1989, as a function of cigarette consumption at the baseline examination

	Non-smokers			Smokers			
Age (years)	Deaths	Person-years	Rate	Deaths	Person-years	Rate	Expected deaths[a]
20–24	1	862.5	0.00116	1	886.5	0.00113	1.0
25–29	2	1,929.5	0.00104	2	2,195.5	0.00091	2.3
30–34	0	2,052	0	1	2,360	0.00042	0
35–39	1	2,064.5	0.00048	4	2,354	0.00170	1.1
40–44	1	2,029.5	0.00049	6	2,280.5	0.00263	1.1
45–49	3	1,940.5	0.00155	2	2,182.5	0.00092	3.4
50–54	5	1,501	0.00333	13	1,819	0.00715	6.1
55–59	5	827	0.00605	11	1,187	0.00927	7.2
60–64	4	370.5	0.01080	13	582	0.02234	6.3
65–69	0	103.5	0	4	106	0.03774	0
70–74	1	17.5	0.05714	0	23	0	1.3
75–79	0	1.5	0	0	0.5	0	0
Total	23			57			29.8

[a] Numbers of deaths that would occur in the group of smokers if they had the age-specific rates observed in the non-smokers

Source: data from Klag MJ et al., Serum cholesterol in young men and subsequent cardiovascular disease, *New England Journal of Medicine*, Volume 328, pp. 313–18, Copyright ©1993 Massachusetts Medical Society. All rights reserved.

Univariate analysis

Internal comparison

The comparison of exposed and unexposed within a cohort can be done through the calculation of the standardized mortality ratio (SMR). Using data from the Precursors Study, the last column of Table 5.5.3 shows the deaths ‘expected’ in the smokers if they had had the age-specific rates of the non-smokers. For

example, the number of deaths expected in smokers aged 45–49 years old would be 0.00155 (deaths/year) × 2,182.5 (person-years) = 3.4 deaths. The SMR is the ratio of total observed deaths to the total expected deaths in smokers (if the smokers had the same age-specific rates as the non-smokers), which in this case is 57/29.8 = 1.9.

External comparison

In other situations, it can be of interest to compare the mortality in the study group with the mortality of the total population that is used as reference. In this case, it is necessary to obtain the expected number of deaths using population-level rates from demographic statistics. Thus, in the Precursors Study, the observed mortality can be compared with the mortality of the population of white males in the United States, as shown in Table 5.5.4. The total number of observed deaths (23 + 57 = 80) divided by the total number of expected deaths based on the demographic statistics provides the SMR of the cohort in comparison with the general population, which in case of data from Table 5.5.4 is 80/148.1 = 0.54. This result implies that the mortality is almost half of what would be expected if the cohort had the same rates as that of the general population.

Table 5.5.4 External comparison of the rates based on person-years in the Precursors Study. All-cause mortality until 1989, observed and expected according to national rates in 1980

Age (years)	National mortality rate[a]	Study cohort		
		Deaths	Person-years	Expected deaths[b]
20–24	0.001909	2	1,749	3.3
25–29	0.001747	4	4,125	7.2
30–34	0.001676	1	4,412	7.4
35–39	0.002087	5	4,418.5	9.2
40–44	0.003161	7	4,310	13.6
45–49	0.005217	5	4,123	21.5
50–54	0.008668	18	3,320	28.8
55–59	0.013756	16	2,014	27.7
60–64	0.021406	17	952.5	20.4
65–69	0.033121	4	209.5	6.9
70–74	0.050338	1	40.5	2.0
75–79	0.074685	0	2	0.1
Total		80	29,676	148.1

[a] Mortality rates in white males, USA, 1980.

[b] According to national rates.

Source: data from Klag MJ et al., Serum cholesterol in young men and subsequent cardiovascular disease, *New England Journal of Medicine*, Volume 328, pp. 313–18, Copyright ©1993 Massachusetts Medical Society. All rights reserved.

Multivariable analysis

The aim of multivariable analysis is to identify the constellation of exposures that explain the variability of incidence of disease, allowing for random variation under a probability distribution model. The Poisson distribution provides a useful and reasonable model to carry out the multivariable analysis of incidence rates observed in cohort studies. Ideally, the final regression model from a multivariable analysis should have an appropriate goodness of fit and be optimally parsimonious (i.e. with most regression coefficients being statistically significant).

Survival analysis

Time-to-event outcomes are used to accomplish the following epidemiological objectives: estimate the percentage of individuals that survive (i.e. remain event-free) for differing lengths of time from the origin, compare the survival of exposed and unexposed groups, and characterize the possibly multifactorial nature explaining observed heterogeneities in survival of different subgroups of individuals using multivariable analysis and regression methods.

Descriptors of the distribution of times to event

If T denotes the variable of time from origin to event, the simplest descriptor of the distribution of T is the *survival function* at time t (where t is a positive number), which is denoted by $S(t)$ and corresponds to the probability that an individual in the baseline cohort survives beyond time t. The survival function starts with a value of one and monotonically decreases towards zero. Its complement from $1–S(t)$ is the cumulative incidence that an individual in the baseline cohort has the event before or at time t. The *pth percentile*, which is denoted by $t(p)$, is the time by which p per cent develop the event of interest; that is, when $1–p = S(t(p))$. A widely used descriptor of survival times is the median, which is simply the 50th percentile, or the time since origin by which half of the individuals develop the event of interest.

The *hazard function* at t, which is denoted by $h(t)$, corresponds to the proportion of individuals who develop the event in the next unit of time among those who survive event-free until time t. That is, $h(t)$ equals the rate of decline of S at t divided by $S(t)$. It is important to note that $h(t)$ is a conditional probability and refers only to the subset of the cohort that survives event-free by time t.

Figure 5.5.1 depicts the 25th percentile and hazard at $t = 12$ *years* of a survival function. The 25th percentile is 6 years (i.e. $t(0.25) = 6$ years), because it corresponds to the time by which 25% of the individuals develop the event (i.e. 75% survive). The hazard at $t = 12$ years is 10% and corresponds to the ratio of 0.04/0.40, which is the value of the rate of decline of the survival function at $t = 12$ (i.e. 0.04) divided by the value of S at $t = 12$ (i.e. 0.40).

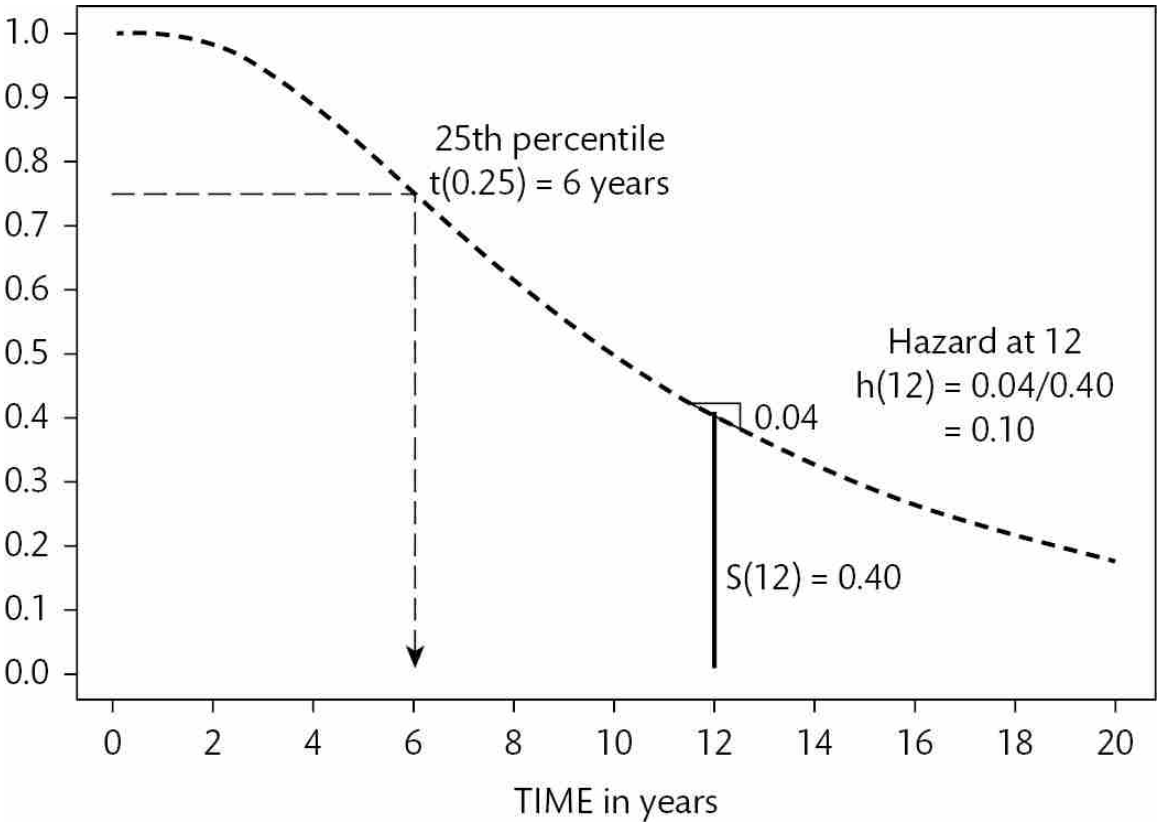


Fig. 5.5.1 Survival function, the 25th percentile ($t(0.25)$) and hazard at $t = 12$ years (h(12)).

Estimation of the survival function

The presence of censored observations is a case of missing data, and the fundamental role of the assumption that censored observations are not a selected subgroup is that, at any time during the study, individuals who remain under observation *represent* those subjects who were censored beforehand. The observation of events beyond the time an individual is censored provides the data to estimate when the censored individual is expected to develop the event had she or he remained in the study. Furthermore, because the individuals who are observed to develop the event are representative of the observations censored before their time of event, estimates of the hazard functions based on the observed individuals are unbiased, and thus they represent the hazards that the censored individuals would have experienced had they remained under observation in the study. Indeed, this forms the basis of the Kaplan–Meier method (Kaplan and Meier 1958), based on the calculation of the survival probability through the product of one minus the observed instantaneous hazards.

Figure 5.5.2 shows five Kaplan–Meier curves according to the number of copies of HIV detected in 500 ml of plasma measured by branched-DNA (bDNA) in HIV-infected individuals followed up in the MACS cohort (Mellors et al. 1997).

It clearly shows the great predictive value that viral load has for the development of AIDS. Indeed, investigators in the MACS have developed methods to quantify the variability of time to AIDS explained by viral load and have shown that it alone explained approximately 50% of the variability of time to AIDS (Mellors et al. 2007).

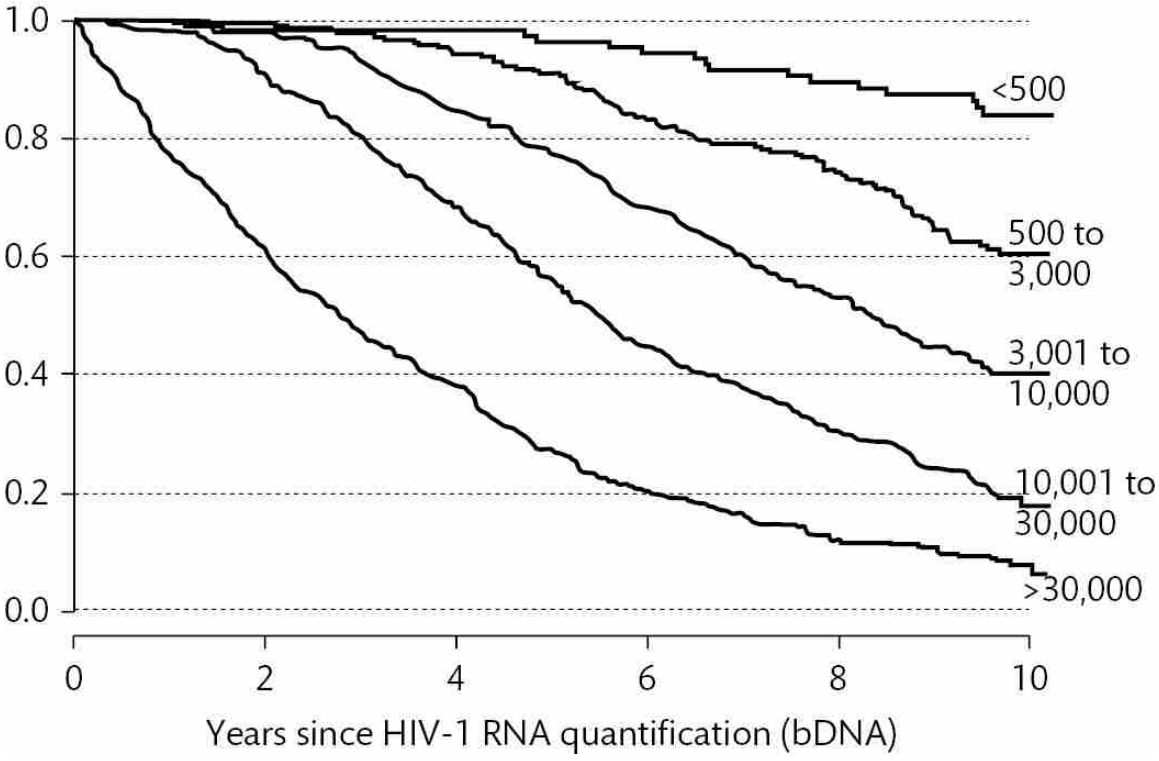


Fig. 5.5.2 Proportion AIDS-free by plasma HIV-RNA (copies/ml).
Source: data from Mellors JW et al., Prognostic value of HIV-RNA, CD4 cell count, and CD4 cell count slope for progression to AIDS and death in untreated HIV-1 infection. Journal of the American Medical Association, 297:2349–50, Copyright 2007 American Medical Association. All rights reserved.

Univariate analysis

Graphical display of the survival curves is generally quite useful to obtain an overall comparison of the experiences of groups subjected to different exposures (Fig. 5.5.2). There are two different ways to measure the association between exposure and survival, which correspond to horizontal (i.e. relative time) and vertical comparisons (i.e. relative hazard) of survival curves. The first one is the relative times defined as $t_1(p)/t_0(p)$, where $t_i(p)$ corresponds to the time taken by p per cent of group i = (*for* i = 0.1) to develop the event. The second one is the relative hazard, defined as $h_1(t)/h_0(t)$, where $h_i(t)$ is the hazard for the *i*th group

(*for* i = 0.1) at time t. If exposure shortens survival, the relative time will be less than one and the relative hazard will be greater than one. If the relative times are independent of p and/or the relative hazards are independent of t, the times, and/or the hazards are said to be proportional. In such cases, the relative times and/or hazards will be constant across all values of p and/or t, respectively. Since the percentiles are relative to the whole cohort but the hazard function is conditional on survivors, proportionality in one does not imply proportionality in the other. Methods based on parametric models (e.g. lognormal regression) are more amenable to estimation and testing of relative times (e.g. to compare, for instance, the median survival time in exposed and unexposed) and they can also provide measures of relative hazards (Cox et al. 2007). Semi-parametric methods based on the proportionality of hazards focused attention to estimation and testing of relative hazards but caution must be exercised because proportionality of hazards is seldom attained.

There are hypothesis-testing procedures to assess whether the differences between two or more survival curves are statistically significant such as for example the logrank test (Peto et al. 1977).

Multivariable analysis

In the context of survival analysis, the most used method of multivariate analysis is the proportional hazards regression (Cox 1972). The semi-parametric proportional hazards regression model corresponds to the hazard of the group with covariates representing exposures and confounders $x_1 = x_2 = \ldots = x_K = 0$ (i.e. the reference group) being completely arbitrary, and the hazard of the group characterized by any other values of x_1 to x_K being a fixed multiple of the hazard of the reference group according to a linear combination of the covariates. Specifically, the hazard of the group with covariates x_1,
$x_2, \ldots, x_K$ denoted by $h_x(t)$ is $h_0(t)exp(\beta_1$
$x_1 + \beta_2$
$x_2 + \ldots + \beta_K$
x_K). The interpretation of the parameters is directly linked to relative hazards. In particular, if we have two groups of individuals described by the constellation of covariates taking values x for one group and x^* for the other, the hazard at time t of the group described by x^* relative to the hazard at the same time of the group defined by x (i.e. the relative hazards) is given by. $RH(t) \equiv \exp\left[\sum_{i=1}^{k} \beta_i \left(x_i^* - x_i\right)\right]$ Because such an expression does not depend on t, the relative hazard is constant (i.e. the hazards are proportional). In particular, if x and x^* only differ in one unit in the k

th covariate (i.e., $x^* = x_1, x_2, \ldots, x_k + 1, \ldots, x_K$) 1then the relative hazard of the x^* group to the x group is simply $RH(t) \equiv exp(\beta_k)$.

When investigators are not only interested in relative hazards but also in describing the underlying hazard with the aim of characterizing mechanisms and the desire to quantify how exposures modify the magnitude of event-free times, parametric models are a more attractive option (Cox et al. 2007). To carry out parametric analyses of time-to-event data, the generalized gamma distribution offers a succinct yet flexible family of distributions. The generalized gamma distribution GG(β, σ, κ) has three parameters whereby the location parameter β governs the value of the median for fixed values of σ and κ; the scale parameter σ determines the value of the interquartile ratio (i.e. IQR = 3rd quartile/1st quartile) for a fixed value of κ and independently of β; and the parameter κ determines the percentiles of the standard $\beta = 0$, $\sigma = 1$ generalized gamma. The generalized gamma contains commonly used parametric models in that the particular cases when κ is zero and one correspond to the lognormal and Weibull models, respectively. More importantly, the parameters κ and σ together determine the *type* of hazard function and yield a graphical taxonomy of the rich variety of hazards covered by the generalized gamma (Cox et al. 2007). Specifically, the taxonomy shows that the generalized gamma family includes all four of the common types of hazard functions: increasing, decreasing, U or bathtub shape, and inverted U or arch shape. More importantly, the pth percentile of the generalized gamma $t(p;\beta,\sigma,\kappa)$ can be expressed in terms of the three parameters by the simple equation $t(p;\beta,\sigma,\kappa) = e^{\beta}\,[t(p;0.1,\kappa)]^{\sigma}$ which dissects the role of each of the three parameters in determining the percentiles of the distribution.

The conventional generalized gamma regression model corresponds to GG($\beta'x$, σ, κ), whereby the distribution of the times for the group of individuals with covariate vector x has different location parameters but same scale and shape. In this case, the relative times of the group with covariates x^* to the group with covariates x are constant (i.e. proportional times) and are equal to $\exp[\beta'(x^* - x)]$. Lack of proportionality of times can be easily incorporated by allowing σ to depend on covariates and full generalization of the model can be accomplished by further allowing κ to depend on covariates. Software is commonly available to implement generalized gamma regression in its full general form (e.g. streg command in STATA).

An example of the richness of parametric regression analyses for survival data is illustrated by Schneider et al. (Schneider et al. 2005), who used data from the MACS cohort of homosexual men and the WIHS cohort of women to characterize the changing pattern of survival after development of clinical AIDS from 1984 to 2004 when different antiretroviral therapies were introduced. Table 5.5.5 presents the relative hazards obtained from a multivariable Cox proportional hazards

regression model and the relative times obtained from a multivariable Weibull regression model. Substantial differences in survival were seen between the last two HAART eras (July 1998–June 2001 and July 2001–December 2003) and the no/monotherapy era (prior to 1989); the time from AIDS to death (adjusted by age, sex, type of AIDS diagnosis, and CD4 cell count at diagnosis) expanded by eightfold and 11-fold in the two eras with stable HAART, respectively. This analysis illustrates the role that cohort studies can play in providing measures of the effectiveness of interventions at the population level.

Table 5.5.5 Descriptive statistics, adjusted relative hazards, and relative times for survival after an initial AIDS diagnosis in five calendar periods from July 1984 to December 2003. Data from the Multicenter AIDS Cohort Study and The Women Interagency HIV Study

Variable	Calendar period				
	July 1984–Dec 1989	Jan 1990–Dec 1994	Jan 1995–June 1998	July 1998–June 2001	July 2001–Dec 2003
Therapy era	**no/monotherapy**	**monotherapy/ combination therapy**	**HAART introduction**	**short-term stable HAART**	**moderate-term stable HAART**
Number seen	633	660	472	496	464
% women	NA[a]	NA[a]	65%	72%	74%
Incident AIDS	633 (100%)	660 (100%)	472 (100%)	143 (29%)	57 (12%)
Median (IQR)[b] Date of AIDS diagnosis	Oct 1987 (April 1986–Dec 1988)	April 1992 (Feb 1991–June 1993)	Jan 1996 (July 1995–Nov 1996)	Nov 1996 (Dec 1995–Sept 1998)	March 1997 (Jan 1996–Oct 1999)
Median (IQR)[b] Age at AIDS diagnosis	36.5 (32.1–41.1)	39.7 (35.3–43.9)	39.7 (34.5–44.5)	39.8 (35.1–44.8)	40.1 (35.6–45.2)
Median (IQR)[b] CD4 cell count at AIDS diagnosis	141 (64–273)	90 (35–194)	196 (78–390)	241 (117–439)	268 (130–456)
Number of person-years	685	912	796	1,156	992
Deaths Number % of person-years	388 57%	445 49%	109 14%	71 6%	44 4%
Relative hazard[c, e] (95% confidence interval)	1	0.65 (0.55, 0.77)	0.21 (0.15, 0.28)	0.08 (0.06, 0.13)	0.06 (0.03, 0.10)
Relative time[d, e] (95% confidence interval)	1	1.42 (1.24, 1.62)	3.57 (2.77, 4.61)	7.82 (5.86, 10.45)	10.65 (7.66, 14.80)

[a] NA, not applicable since WIHS began in October 1994.

[b] IQR, interquartile range.

[c] Results of Cox Proportional Hazards Regression.

[d] Results of Weibull Regression under the assumption of proportional survival times.

[e] Adjusted by age at AIDS diagnosis, type of AIDS diagnosis, gender, and CD4 within 1 year prior to AIDS diagnosis.

Reproduced from Schneider MF et al., Patterns of the hazard of death after AIDS through the evolution of antiretroviral therapy: 1984–2004, *AIDS*, Volume 19, Issue 17, pp 2009–18, Copyright © 2005 Lippincott Williams & Wilkins, Inc, with permission from Wolters Kluwer.

Longitudinal data analysis

In cohort studies, it is often the case that the dependent variable of interest in a cohort study is not the occurrence of an event but the trajectory of a determined biomarker such as blood pressure, weight, leucocyte count, viral load, glomerular filtration rate, pulmonary function, C-reactive protein, or CD4 cell count.

To illustrate the power of longitudinal data in cohort studies collecting data at regularly scheduled visits, we used data collected at semi-annual visits by the MACS from 1984 to 2006. The person-visits of interest were those happening after the first time individuals tested positive for antibodies to HIV and with data on therapy received, CD4 cell count (a marker of immunosuppression), and HIV-RNA (copies of HIV per mL of plasma). The objective was to describe the distribution of HIV-RNA in four categories of CD4 cell count using 200, 350, and 500 cells/mm_3 as cut-off values and in four categories of therapy according to whether the participant at a given visit was not on any therapy, was just receiving antiretroviral therapy but not a triple combination, or was receiving triple combination therapy (HAART) and the visit occurred within 3 years or after 3 years from initiation. Two measures of HIV-RNA were used: (1) the median HIV-RNA and (2) the percentage of person-visits in a given category of CD4 cell count and therapy with undetectable HIV-RNA (i.e. <50 copies/ml). Figure 5.5.3 uses a diamond-shaped equiponderant graphical display of the effects of CD4 cell count and therapy on HIV-RNA whereby the shaded area on each square is proportional to the HIV-RNA (Li et al. 2003). Both panels show the great effectiveness of HAART in suppressing viral replication in all categories of CD4 cell count.

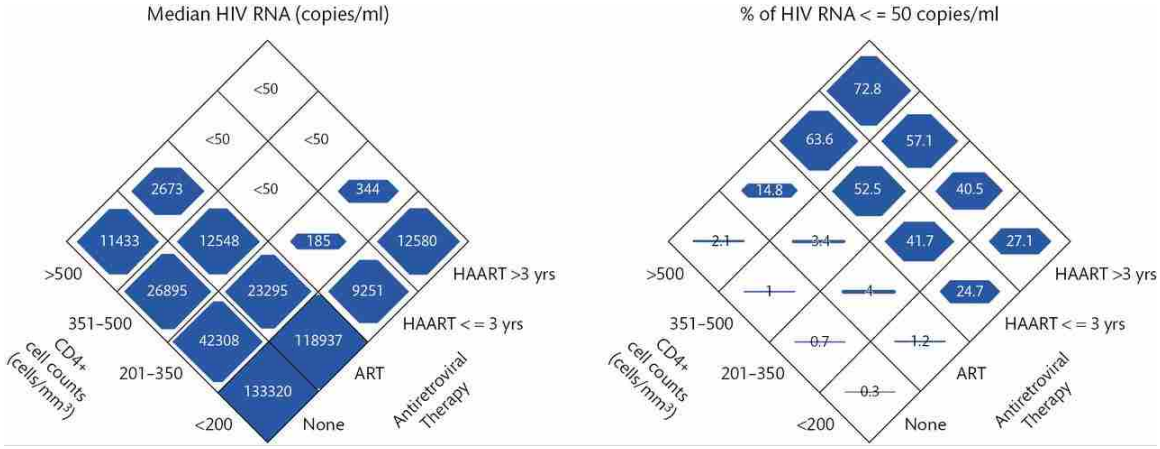


Fig. 5.5.3 Median HIV-RNA (copies/ml) and % with HIV-RNA ≤50 copies/ml in HIV-infected individuals according to CD4 cell count and Antiretroviral Therapy. Multicenter AIDS Cohort Study, 1984–2006.

Methods for the analysis of longitudinal data can be broadly classified in three categories: marginal, conditional, and random effects models. The primary aim of the marginal approach is to combine the multiple cross-sections corresponding to visits so as to provide the most efficient summary of the cross-sectional relationships between exposure and disease. In this approach, the longitudinal component is typically incorporated by including age or time since baseline as a covariate for the regression component of the model. Approaches for the incorporation of the correlation between repeated measurements within individuals include parametric (Jennrich and Schluchter 1986) and non-parametric methods, the latter handling the correlation as a nuisance (Zeger and Liang 1992). The primary aim of conditional models is to regress current outcome on past values of the outcome, and current and previous exposures (Rosner et al. 1985). The primary aim of the random effects models is to allow different individuals to have different intercepts and slopes according to components of variance and to provide direct averages of rates of change across individuals. Figure 5.5.4 depicts the random effects model whereby the population average line has intercept α and slope β and the line for the ith individual has intercept $\alpha + a_i$ and slope $\beta + b_i$ with the departures being random effects (Laird and Ware 1982).

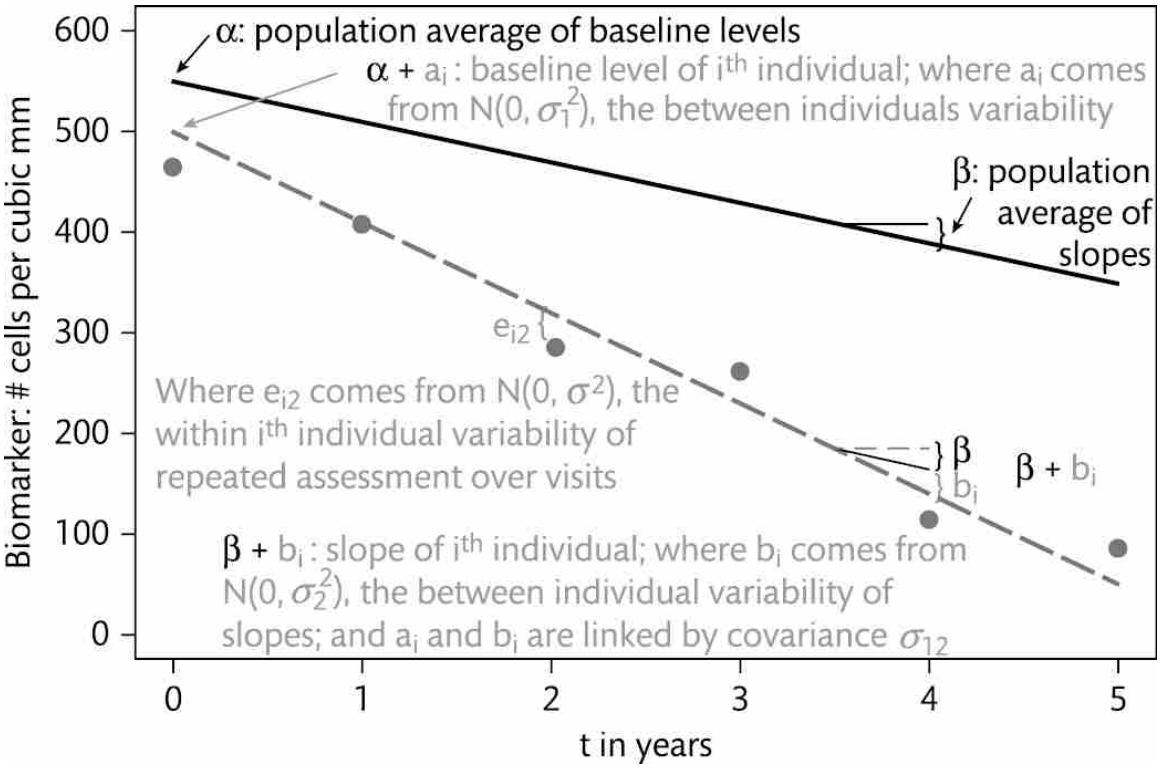


Fig. 5.5.4 Linear model with random intercepts and slopes for the analysis of longitudinal biomarkers data from cohort studies. Population averages in black and random departure of the line of the i^{th} individual in grey.

When data are longitudinal assessments of a biomarker (e.g. pulmonary function, blood pressure, viral load of an infectious agent), linear models for normally distributed biomarkers allowing for both fixed and random effects (i.e. mixed effects) are apt to describe the probabilistic structure of the data. In the simplest case of two groups (i.e. an exposure taking the value zero or one) with linear trends over time, the fixed effects describe the two average lines for each group. That is, for the unexposed group, the average biomarker trajectory is given by $\beta_0 + \beta_1$
t, and for the exposed group the average biomarker trajectory is given by (β0 + β*0) + (β1 + β*1)t. In this case, β*0 represents the fixed effect of exposure on the levels at baseline and β*1 represents the fixed effects of exposure on the rates of change of the biomarker over time. The random effects allow individuals in each of the two groups to vary randomly around the average lines of the fixed effects (i.e. each individual has his or her own line that varies randomly from the average line as depicted in Figure 5.5.4) (Laird and Ware 1982).

Joint analysis of survival and longitudinal data

Since the occurrence of an end point in a cohort study typically marks the end of follow-up in the study, the trajectories of participants who develop end points will typically be shorter than those who remain event-free. When subjects leave the study for reasons related to the effects of exposures and/or the progression of disease, the longitudinal data will be subjected to informative dropout. If the biomarkers of interest have greater changes among those who develop end points, a separate analysis of the longitudinal data may yield biased estimates of the rates of change of the biomarkers as they are unduly weighted towards longer and less changing trajectories. A joint model of the time-to-event data and the longitudinal biomarkers data aims to properly describe the relationships between exposures and the two distinct but related outcome processes. For longitudinal data, a joint model must take into account not only the dependence resulting from repeated measurement of the same outcome, but also the association between the different outcomes, all measured on the same individual. A solution to this problem has been provided by the development of shared-parameter models, in which the necessary dependence structure between the two types of data is provided by a vector of shared random effects that links all the measurements on the same individual. Typically the two outcomes (end point and markers) are assumed to be independent given the random effects, which results in a relatively simple factorization of the joint likelihood. Parametric models for the marginal distributions of each outcome, as well as the standard assumption of a multivariate normal distribution for the random effects, provide a complete description of the joint likelihood. Statistical software to fit these models is available in standard statistical software packages such as SAS and STATA.

Validity and biases

External validity refers to the extent to which the results of a cohort study can be extrapolated or generalized to the reference population. When the cohort is composed of volunteers, representativeness may be limited due to the fact that volunteers tend to be different from the general population. The representativeness or the lack of representativeness of the cohort in relation to the general population is also determined by the inherent characteristics of the sampling framework. Thus, for instance, even in the case that the members of the Framingham cohort were a representative sample of the population of this small town in New England, it is more than dubious that this population was representative of the general population.

Even when the representativeness of the cohort is dubious, it is possible that the internal comparison between cohort subgroups (e.g. exposed and unexposed) may

be valid (*internal validity*). In the Precursors Study, for example, it was observed that the SMR was 54% (see Table 5.5.4), which indicates that the mortality in the cohort of the Precursors Study was almost half that of the general US population of the same sex, race, and age. Obviously, this is due to the fact that this cohort is formed by medical students belonging, in general, to higher socioeconomic strata than the average population, which limits the ability to generalize any estimates of the absolute values of incidence or mortality in this cohort. However, it is reasonable to assume that the *relative* comparison of the mortality in smokers and non-smokers in this population (see Table 5.5.3) is valid, or in other words, that the magnitude of the bias that affects the absolute value of mortality given the lack of representativeness of the cohort is proportionally similar in both groups; under these circumstances, even a non-representative cohort in the strict sense of the word, can yield unbiased estimates of the relative risk (or SMR in this case).

In cohort studies, the internal validity is based not only on the comparability of the selection process of exposed and unexposed when the cohort is established, but also on the comparability between the subjects who are lost to follow-up and those who continue under observation in the study. If the participants who are lost to follow-up have a different survival experience compared to those who remain under observation, the overall estimates of survival are biased (i.e. do not accurately represent the survival experience of the population in the original cohort) and special methods are required to avoid the biases due to informative censoring (Robins and Finkelstein 2000; Howe et al. 2011).

Biases

In the evaluation of the possibility of biases due to the losses to follow-up, it is important to consider which types of losses have occurred. The biases due to differential losses to follow-up in a cohort study are conceptually analogous to the *selection biases* that affect case–control or other types of epidemiological studies.

There are other biases than can affect the internal validity of a cohort study. Among the information biases, one type of bias that could affect these types of studies is *surveillance bias*. This bias is present when the persons exposed to a determined risk factor are more carefully monitored or watched over than the persons without that factor. If the disease has a long subclinical phase, this can exaggerate the apparent magnitude of the association between that factor and the disease outcome. Thus, for example, diabetics are generally subject to periodical check-ups in which blood pressure is taken. Because in the natural history of hypertensive disease there exists a long asymptomatic phase and hypertension is much more likely to be diagnosed in diabetic persons given their periodical clinical assessment, it is possible that part of the strong association that exists

between diabetes and hypertension is due in part to a surveillance bias. Even in cohort studies in which information is collected in a systematic form, if the observer knows the history of the participant, he/she can evaluate the presence of the disease in a biased form (*observer bias*). These biases are removed in large part in cohort studies in which a rigorous protocol for collection of information (preferably blinded with respect to exposure status) and well-trained observers exist.

As in all epidemiological studies, the presence of errors in classification of the participants regarding the presence of the exposure or the occurrence of the disease under study may generate *bias due to misclassification*. When the errors in the determination of disease occurrence are equal in exposed and non-exposed, or when the errors in measurement of the exposure are not differential between diseased and non-diseased, the bias is generally (but not always) conservative; that is, towards the underestimation of the true association (Flegal et al. 1991; Mertens 1993).

Applications and conclusions

Even with the potential limitations regarding internal and external validity discussed in the previous section, the cohort study constitutes the design of reference for all other designs in observational epidemiology. Cohort studies are free from certain types of biases that complicate the interpretation of cross-sectional and case–control studies (Schlesselman 1982), as, for example, recall bias (different recall of the exposure in cases and controls) and cross-sectional bias (when there is a different duration of disease in exposed and unexposed). The cohort design is more powerful for the establishment of the temporal relationship between cause and effect, one of the classical criteria for causality (Hill 1965).

Another advantage of cohort studies is the possibility of studying the occurrence of several events in addition to that initially planned. Thus, for instance, the Framingham cohort has been the basis for the investigation of several hypotheses in addition to those conceived in the initial aims of the study (Kannel 1990) as, for example, the relationship between several factors and cancer (Kreger et al. 1991), senile dementia (Bachman et al. 1993), cognitive function (Elias et al. 1993; Au et al. 2006), osteoporosis (Hannan et al. 1992), hearing loss (Gates et al. 1990), kidney function (Culleton et al. 1999), pulmonary function (Wilk et al. 2009), healthcare preferences, and others. Furthermore, cohort studies are a platform to expand the initial scientific scope of the studies as new knowledge is acquired and new questions arise.

Cohort studies have proven to be extremely valuable in many areas of medical research including cardiovascular disease (Dawber 1980), cancer (Doll and Hill 1952; Doll et al. 2005), and infectious diseases (Samet and Muñoz 1998). Cohort data may also be of great use for the development of public policy, such as treatment guidelines. Follow-up of cohorts before interventions are available provides the most appropriate data for the description of the natural history of disease. Furthermore, if these studies continue to collect data after interventions become available, three uses of these pre-intervention data are apparent: (1) as historical controls for assessing and/or confirming the effectiveness of interventions (Jacobson et al. 2002); (2) as a reference point to establish the effectiveness of interventions at the population level once the interventions have been widely implemented (Detels et al. 1998; Muñoz et al. 2000; Schneider et al. 2005); and (3) as a mechanism to identify the subgroups who should receive treatment once an intervention has been proven efficacious in clinical trials (Cole et al. 2004; Sterne et al. 2009). From the public health perspective, population effectiveness quantifies the reduction of disease achieved by treating the subset of the population that needs therapies the most. In doing so, cohort studies are at the cornerstone of public health and policy (Szklo and Nieto 2018).

REFERENCES

Ahdieh, L., et al. (2000). Selection by indication of potent antiretroviral therapy use in a large cohort of women infected with human immunodeficiency virus. *American Journal of Epidemiology*, 152(10), 923–33.

Alpha-Tocopherol, Beta Carotene Cancer Prevention Study Group (1994). The effect of vitamin E and beta carotene on the incidence of lung cancer and other cancers in male smokers. *The New England Journal of Medicine*, 330(15), 1029–35.

ARIC Investigators (1989). The Atherosclerosis Risk in Communities (ARIC) Study: design and objectives. The ARIC Investigators. *American Journal of Epidemiology*, 129(4), 687–702.

Armstrong, B.K., White, E., and Saracci, R. (2008). *Principles of Exposure Measurement in Epidemiology: Collecting, Evaluating, and Improving Measures of Disease Risk Factors* (2nd ed.). New York: Oxford University Press.

Au, R., et al. (2006). Association of white matter hyperintensity volume with decreased cognitive functioning: the Framingham Heart Study. *Archives of Neurology*, 63(2), 246–50.

Bachman, D.L., et al. (1993). Incidence of dementia and probable Alzheimer's disease in a general population: the Framingham Study. *Neurology*, 43(3 Pt 1), 515–19.

Checkley, W., et al. (2010). Inference for mutually exclusive competing events through a mixture of generalized gamma distributions. *Epidemiology*, 21(4), 557–65.

Cole, S.R., et al. (2003). Effect of highly active antiretroviral therapy on time to acquired immunodeficiency syndrome or death using marginal structural models. *American Journal of Epidemiology*, 158(7), 687–94.

Cole, S.R., et al. (2004). Accounting for leadtime in cohort studies: evaluating when to initiate HIV therapies. *Statistics in Medicine*, 23(21), 3351–63.

Cox, C., et al. (2007). Parametric survival analysis and taxonomy of hazard functions for the generalized gamma distribution. *Statistics in Medicine*, 26(23), 4352–74.

Cox, C., Schneider, M.F., and Muñoz, A. (2013). Quantiles of residual survival. In: Lee, M., et al. (eds.) *Risk Assessment and Evaluation of Predictions*. New York: Springer, pp. 87–106.

Cox, D.R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society*, 34(2), 187–202.

Culleton, B.F., et al. (1999). Prevalence and correlates of elevated serum creatinine levels: the Framingham Heart Study. *Archives of Internal Medicine*, 159(15), 1785–90.

Dawber, T. (1980). *The Framingham Study—The Epidemiology of Atherosclerotic Disease*. Cambridge, MA: Harvard University Press.

Detels, R., et al. (1998). Effectiveness of potent antiretroviral therapy on time to AIDS and death in men with known HIV infection duration. Multicenter AIDS Cohort Study Investigators. *JAMA*, 280(17), 1497–503.

Djoussé, L., et al. (2002). Serum albumin and risk of myocardial infarction and all-cause mortality in the Framingham Offspring Study. *Circulation*, 106(23), 2919–24.

Doll, R., et al. (2005). Mortality from cancer in relation to smoking: 50 years observations on British doctors. *British Journal of Cancer*, 92(3), 426–9.

Doll, R. and Hill, A.B. (1952). A study of the aetiology of carcinoma of the lung. *British Medical Journal*, 2(4797), 1271–86.

Dudley, J., et al. (1995). The Multicenter AIDS Cohort Study: retention after 9 1*/2* years. *American Journal of Epidemiology*, 142(3), 323–30.

Elias, M.F., et al. (1993). Untreated blood pressure level is inversely related to cognitive functioning: the Framingham Study. *American Journal of Epidemiology*, 138(6), 353–64.

Fine, J.P. and Gray, R.J. (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association*, 94(446), 496–509.

Flegal, K.M., Keyl, P.M., and Nieto, F.J. (1991). Differential misclassification arising from nondifferential errors in exposure measurement. *American Journal of Epidemiology*, 134(10), 1233–44.

Frank, L. (2000). Epidemiology. When an entire country is a cohort. *Science (New York, N.Y.)*, 287(5462), 2398–9.

Fried, L.P., et al. (1991). The Cardiovascular Health Study: design and rationale. *Annals of Epidemiology*, 1(3), 263–76.

Friedman, G.D., et al. (1988). CARDIA: study design, recruitment, and some characteristics of the examined subjects. *Journal of Clinical Epidemiology*, 41(11), 1105–16.

Furth, S.L., et al. (2006). Design and methods of the Chronic Kidney Disease in Children (CKiD) prospective cohort study. *Clinical journal of the American Society of Nephrology: CJASN*, 1(5), 1006–15.

Furth, S.L., et al. (2011). Metabolic abnormalities, cardiovascular disease risk factors, and GFR decline in children with chronic kidney disease. *Clinical journal of the American Society of Nephrology: CJASN*, 6(9), 2132–40.

Gates, G.A., et al. (1990). Hearing in the elderly: the Framingham cohort, 1983–1985. Part I. Basic audiometric test results. *Ear and Hearing*, 11(4), 247–56.

Greenland, S. (2000). An introduction to instrumental variables for epidemiologists. *International Journal of Epidemiology*, 29(6), 1102.

Hannan, M.T., Felson, D.T., and Anderson, J.J. (1992). Bone mineral density in elderly men and women: results from the Framingham osteoporosis study. *Journal of Bone and Mineral Research*, 7(5), 547–53.

Hernán, M.A., Brumback, B., and Robins, J.M. (2001). Marginal structural models to estimate the joint causal effect of nonrandomized treatments. *Journal of the American Statistical Association*, 96(454), 440–8.

Hill, A.B. (1965). The environment and disease: association or causation?. *Proceedings of the Royal Society of Medicine*, 58(5), 295–300.

Hoffmans, M.D., Kromhout, D., and Coulander, C.D. (1989). Body mass index at the age of 18 and its effects on 32-year-mortality from coronary heart disease and cancer. A nested case–control study among the entire 1932 Dutch male birth cohort. *Journal of Clinical Epidemiology*, 42(6), 513–20.

Hoffmans, M.D., Kromhout, D., and de Lezenne Coulander, C. (1988). The impact of body mass index of 78,612 18-year old Dutch men on 32-year mortality from all causes. *Journal of Clinical Epidemiology*, 41(8), 749–56.

Howe, C.J., et al. (2011). Limitation of inverse probability-of-censoring weights in estimating survival in the presence of strong selection bias. *American Journal of Epidemiology*, 173(5), 569–77.

Illner, A.K., et al. (2012). Review and evaluation of innovative technologies for measuring diet in nutritional epidemiology. *International Journal of Epidemiology*, 41(4), 1187–203.
Jacobson, L.P., et al. (2002). Evaluation of the effectiveness of highly active antiretroviral therapy in persons with human immunodeficiency virus using biomarker-based equivalence of disease progression. *American Journal of Epidemiology*, 155(8), 760–70.
Jennrich, R.I. and Schluchter, M.D. (1986). Unbalanced repeated-measures models with structured covariance matrices. *Biometrics*, 42(4), 805–20.
Kannel, W.B., et al. (1979). An investigation of coronary heart disease in families. The Framingham offspring study. *American Journal of Epidemiology*, 110(3), 281–90.
Kannel, W.B. (1990). Bishop lecture. Contribution of the Framingham Study to preventive cardiology. *Journal of the American College of Cardiology*, 15(1), 206–11.
Kaplan, E.L. and Meier, P. (1958). Non-parametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53, 457–81.
Kaslow, R.A., et al. (1987). The Multicenter AIDS Cohort Study: rationale, organization, and selected characteristics of the participants. *American Journal of Epidemiology*, 126(2), 310–18.
Klag, M.J., et al. (1993). Serum cholesterol in young men and subsequent cardiovascular disease. *The New England Journal of Medicine*, 328(5), 313–18.
Korn, E.L., Graubard, B.I., and Midthune, D. (1997). Time-to-event analysis of longitudinal follow-up of a survey: choice of the time-scale. *American Journal of Epidemiology*, 145(1), 72–80.
Kreger, B.E., Splansky, G.L., and Schatzkin, A. (1991). The cancer experience in the Framingham Heart Study cohort. *Cancer*, 67(1), 1–6.
Laird, N.M. and Ware, J.H. (1982). Random-effects models for longitudinal data. *Biometrics*, 38(4), 963–74.
Li, X., et al. (2003). A diamond-shaped equiponderant graphical display of the effects of two categorical predictors on continuous outcomes. *The American Statistician*, 57(3), 193–9.
Manolio, T.A., et al. (2012). New models for large prospective studies: is there a better way?. *American Journal of Epidemiology*, 175(9), 859–66.
Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22(4), 719–48.
Martens, E.P., et al. (2006). Instrumental variables: application and limitations. *Epidemiology (Cambridge, Mass.)*, 17(3), 260–7.
Mellors, J.W., et al. (2007). Prognostic value of HIV-1 RNA, CD4 cell count, and CD4 Cell count slope for progression to AIDS and death in untreated HIV-1 infection. *The Journal of the American Medical Association*, 297(21), 2349–50.
Mellors, J.W., et al. (1997). Plasma viral load and CD4+ lymphocytes as prognostic markers of HIV-1 infection. *Annals of Internal Medicine*, 126(12), 946–54.
Mertens, T.E. (1993). Estimating the effects of misclassification. *Lancet (London, England)*, 342(8868), 418–21.
Muñoz, A., et al. (1992). Estimation of time since exposure for a prevalent cohort. *Statistics in Medicine*, 11(7), 939–52.
Muñoz, A., Gange, S.J., and Jacobson, L.P. (2000). Distinguishing efficacy, individual effectiveness and population effectiveness of therapies. *AIDS (London, England)*, 14(6), 754–6.
Muñoz, A. and Hoover, D. (1995). Use of cohort studies for evaluating AIDS therapies. In: Finkelstein, D.M. and Schoenfeld, D. (eds.) *AIDS Clinical Trials*. New York: Wiley, pp. 423–66.
Muñoz, A., Sabin, C.A., and Phillips, A.N. (1997). The incubation period of AIDS. *AIDS (London, England)*, 11(Suppl A), S69–76.
Muñoz, A. and Xu, J. (1996). Models for the incubation of AIDS and variations according to age and period. *Statistics in Medicine*, 15(21–2), 2459–73.
Ng, D.K., et al. (2016). Racial differences in renal replacement therapy initiation among children with a nonglomerular cause of chronic kidney disease. *Annals of Epidemiology*, 26(11), 780–7.e1.
Nieto, F.J., et al. (2012). Sleep-disordered breathing and cancer mortality: results from the Wisconsin Sleep Cohort Study. *American Journal of Respiratory and Critical Care Medicine*, 186(2), 190–4.
Nieto, F.J., Szklo, M., and Comstock, G.W. (1992). Childhood weight and growth rate as predictors of adult mortality. *American Journal of Epidemiology*, 136(2), 201–13.

Peto, R., et al. (1977). Design and analysis of randomized clinical trials requiring prolonged observation of each patient. II. analysis and examples. *British Journal of Cancer*, 35(1), 1–39.
Phillips, A.N., et al. (1999). Use of observational databases to evaluate the effectiveness of antiretroviral therapy for HIV infection: comparison of cohort studies with randomized trials. EuroSIDA, the French Hospital Database on HIV and the Swiss HIV Cohort Study Groups. *AIDS (London, England)*, 13(15), 2075–82.
Pierce, C.B., et al. (2011). Methods for characterizing differences in longitudinal glomerular filtration rate changes between children with glomerular chronic kidney disease and those with nonglomerular chronic kidney disease. *American Journal of Epidemiology*, 174(5), 604–12.
Putter, H., Fiocco, M., and Geskus, R.B. (2007). Tutorial in biostatistics: competing risks and multi-state models. *Statistics in Medicine*, 26(11), 2389–430.
Robins, J.M. and Finkelstein, D.M. (2000). Correcting for noncompliance and dependent censoring in an AIDS Clinical Trial with inverse probability of censoring weighted (IPCW) log-rank tests. *Biometrics*, 56(3), 779–88.
Robins, J.M., Hernán, M.A., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology (Cambridge, Mass.)*, 11(5), 550–60.
Rosner, B., et al. (1985). The use of an autoregressive model for the analysis of longitudinal data in epidemiologic studies. *Statistics in Medicine*, 4(4), 457–67.
Rothman, K., Greenland, S., and Lash, T.L. (2007). *Modern Epidemiology* (3rd ed.). Lippincott Williams & Wilkins.
Samet, J.M. and Muñoz, A. (1998). Evolution of the cohort study. *Epidemiologic Reviews*, 20(1), 1–14.
Schlesselman, J.J. (1982). *Case–Control Studies: Design, Conduct, Analysis*. Oxford, New York: Oxford University Press.
Schneider, M.F., et al. (2005). Patterns of the hazard of death after AIDS through the evolution of antiretroviral therapy: 1984–2004. *AIDS (London, England)*, 19(17), 2009–18.
Sterne, J., et al. (2009). Timing of initiation of antiretroviral therapy in AIDS-free HIV-1-infected patients: a collaborative analysis of 18 HIV cohort studies. *Lancet (London, England)*, 373(9672), 1352–63.
Stuart, E.A. (2010). Matching methods for causal inference: a review and a look forward. *Statistical Science*, 25(1), 1–21.
Szklo, M. and Nieto, F.J. (2018). *Epidemiology, Beyond the Basics* (4th ed.). Sudbury, MA: Jones & Bartlett Publishers.
Tager, I.B., et al. (1983). Longitudinal study of the effects of maternal smoking on pulmonary function in children. *The New England Journal of Medicine*, 309(12), 699–703.
Tynes, T., Andersen, A., and Langmark, F. (1992). Incidence of cancer in Norwegian workers potentially exposed to electromagnetic fields. *American Journal of Epidemiology*, 136(1), 81–8.
Vandenbroucke, J.P. (1991). Prospective or retrospective: what's in a name?. *British Medical Journal*, 302(6771), 249–50.
Virtamo, J., et al. (2003). Incidence of cancer and mortality following alpha-tocopherol and beta-carotene supplementation: a postintervention follow-up. *The Journal of the American Medical Association*, 290(4), 476–85.
Wada, N., et al. (2013). Cause-specific life expectancies after 35 years of age for human immunodeficiency syndrome-infected and human immunodeficiency syndrome-negative individuals followed simultaneously in long-term cohort studies, 1984–2008. *American Journal of Epidemiology*, 177(2), 116–25.
Wada, N., et al. (2014). Cause-specific mortality among HIV-infected individuals, by CD4(+) cell count at HAART initiation, compared with HIV-uninfected individuals. *AIDS (London, England)*, 28(2), 257–65.
Wilk, J.B., et al. (2009). A genome-wide association study of pulmonary function measures in the Framingham Heart Study. *PLoS Genetics*, 5(3), e1000429.
Willett, W.C., et al. (1984). Relation of serum vitamins A and E and carotenoids to the risk of cancer. *The New England Journal of Medicine*, 310(7), 430–4.
Zeger, S.L. and Liang, K.Y. (1992). An overview of methods for the analysis of longitudinal data. *Statistics in Medicine*, 11(14–15), 1825–39.

5.6

Qualitative research imagination

Jennie Popay and Fran Baum

Introduction

In his classic book The Sociological Imagination C. Wright Mills (1970) argued that ‘imagination’ is the creative force that transforms social enquiry from the mundane application of particular techniques into a process that has the potential to generate new ways of thinking about the social world producing new knowledge of both theoretical and practical use. From this perspective, though it may seem counterintuitive, maximizing the contribution to global public health of any type of research requires imagination alongside technical competencies. Imagination enables scientists to envision the potential impact of their research on policy and practice and to ensure it is relevant and high quality. In this chapter we will consider what the qualitative research imagination ‘is’—touching on key questions that should be addressed by all researchers regardless of their approach (see Box 5.6.1). In the first part of the chapter we will consider the first two questions focusing on the type of ‘knowledge’ generated by qualitative research and in the second part on questions about the range of approaches that can be taken to data collection, analysis and interpretation, and the role of theory.

Box 5.6.1 Four key questions for researchers

1. Ontological perspective: What is the nature of the 'social phenomena' I wish to investigate?
2. Epistemological perspective: What might represent knowledge/evidence about these phenomena?
3. Theoretical perspective: What is the intellectual puzzle I wish to 'explain' and what research questions flow from this?
4. Methodological perspective: What approach should I use to collect and analyses data about these phenomena?

What is the qualitative research imagination?

Qualitative research is underpinned by a social constructivist philosophy. it assumes that the 'reality' or phenomena being studied are subjective: the product of people's interpretations informed by their perceptions, beliefs, and experiences. These subjective phenomena are rich, diverse, and fluid. Rather than trying to control them through experiment and measurement, qualitative research therefore studies social phenomena in naturally occurring settings and seeks to develop concepts and theories that help us to see these phenomena from the point of view of the people being studied.

The central question for qualitative research is why individuals, collectives and/or organizations 'act' or 'behave' as they do (what social scientists refer to as *agency*) (Mallinson et al. 2014). It seeks to answer this question by studying the relationship between 'agency' and social context, including in particular *social structures* associated with social class, gender, ethnicity, sexuality, etc. and by illuminating the understandings and meanings people attach to these actions and experiences. Qualitative research can also illuminate power relationships and how these affect the distribution of resources in a society, including the extent of equality in health status between different groups. Concepts are the basis for analysis of qualitative data, rather than numbers, and theory is the basis for explanation and generalization rather than statistical probability.

Given the assumption that the 'reality' or 'phenomena' being studied is socially constructed, a research event, such as an interview, is not viewed as a method for gaining a 'slice of reality' but as a process of contextually bounded 'meaning-making' between the researcher and the researched. The context, perspective and position of the researcher is therefore understood to be an inevitable influence on the findings of qualitative research. So, rather than seeking to eliminate sources of potential bias, the aim is to make

them visible. Processes of *reflexivity* and *transparency* are used by the qualitative researcher to identify relevant aspects of their beliefs/perceptions etc, to consider how these may have influenced the research process and to record this so that readers can assess the strengths, weaknesses, and relevance of a study.

The rest of this chapter is divided into three parts. First, we discuss the type of knowledge produced by qualitative research. We then discuss the potential for community/public involvement as 'partners' in qualitative research and finally consider key methodological issues in the conduct of qualitative research.

What knowledge or evidence can the qualitative research imagination contribute to public health policy and practice?

Armstrong (2000) has written about the separation of two ways of understanding the world: 'mythos' and 'logos' and the limits this imposes on our ability to 'imagine' action for health. improvement. Logos concerns the rational world—how we might understand how the world works, how we solve problems, and grasp causal processes. It is mundane and concerns both logical argument and the accumulation of facts. Mythos, on the other hand, is concerned with meaning: it is what gives life its *telos* or purpose. It is often expressed in different forms, through metaphor, poetry, visual arts: resources commonly used to express the emotional life and our relationships to others (Richards 2011). The point is not that one 'way of knowing' is better than the other but that knowledge itself becomes impoverished by this separation, often unwittingly, because of the narrow disciplinary confines in which people work and their assumption that their expertise marks them out as separate from the everyday. From this perspective, it is vital that global public health is able to access knowledge about both ways of understanding the world: the quantitative and the qualitative (Baum 2016).

The qualitative research imagination can be used to address a wide range of questions (Popay and Williams 2009). These can be groups into four broad domains critical to global public health policy and practice:

Preparing the ground for other forms of enquiry: addressing different kinds of questions and generating hypothesis to be tested by other study designs.
Adding theoretical and/or empirical depth to understandings of pathways to population health and health equity.
Illuminating barriers and enablers to effective policy and practice.
Contributing to greater social justice by addressing inequities between different forms of knowledge and between researcher and researched and by contributing to critical thinking on action for population health and health equity.

Preparing the ground for other enquiries: Imagining public health evidence beyond lifestyles

A large number of qualitative studies relevant to public health focus on the meanings people attach to behaviours/lifestyle. It has been argued that this research, 'tends to produce bland accounts of social life, in which it is difficult to assess the *import* of things for people' (Sayer 2011, p. 6). To avoid this limitation qualitative research has to provide explanations rather than descriptions: exploring the relationship between the social/material context of people's lives and individual behaviours such as smoking and poor diet and revealing the strategies people develop to cope with risks and exposures over which they may have little control. By doing this the qualitative research imagination provides an understanding of 'private troubles' as 'public issues' and reframes the space for action beyond the individual (Wright Mills 1970); see also, Burawoy (2005) on 'public sociology'. For example, qualitative research embedded in a trial aimed at reducing smoking in early teenage years through a 'peer-led' intervention found that some young people continued to smoke, despite having knowledge of the health risks, in order to maintain social relationships with peers, revealing the complex risks assessments involved in individual decision-making processes (Campbell et al. 2008).

Adding depth to understanding pathways to population health and health equity

A recurring finding from qualitative research into the meanings people attach to their actions is the purposeful development and/or maintenance of

valued social identities in situations that threaten to undermine their moral worth. Importantly this body of research often highlights 'fault lines' between lay and professional understanding about health and illness that can impact adversely on the success of policies/practices. For example, on the basis of a synthesis of multiple qualitative studies Graham and McDermott's (2006) argue that teenagers pursued motherhood as a pathway into a valued adult identity—or social inclusion—in the face of limited alternative pathways to adulthood. In contrast, professionals see education and employment as the pathway to adulthood. They fail to see the positive values and identities affirmed through teenage motherhood constructing it as a pathway to social exclusion and lifelong disadvantage. A more recent qualitative study (Wekesa and Coast 2013) highlighted the extent to which stigma and discrimination affected the responses of people living in an informal settlement in Nairobi to a diagnosis of HIV/AIDS. The study showed the need to ask broad questions about the experience of living with HIV/AIDS and the importance of paying attention to structural factors including poverty, gender, and sociocultural norms.

Illuminating barriers and enablers to effective policy and practice

Qualitative research is now a common element of research evaluating the impact of public health policies and practices and factors impinging on successful implementation: either embedded in a mixed method evaluation or stand alone. The qualitative research by Roberts et al. (2004) embedded in a RCT of the installation and use of domestic smoke alarms explored people's perceptions of the risk of fire and barriers and enablers to the installation and maintenance of domestic smoke alarms. The results revealed, for example, that people disabled alarms because they went off when they were cooking or eating. Similarly, using examples from African research Camlin and Seeley (Camlin and Seeley 2018) illustrate how qualitative research conducted within trials of HIV/AIDS treatments can: (i) reveal vital contextual factors that influence implementation and outcomes; (ii) enable an informed adaptation of trials as they are being conducted; and (iii) can lead to the formulation of theory regarding the social and behavioural pathways to outcomes of interventions. This study design can also enable community engagement in trial design and implementation.

Qualitative research conducted independently has also be used to better understanding the results of multiple RCTs. For example, Noyes and Popay (Noyes and Popay 2007) conducted a systematic review of qualitative research on help seeking behaviour in an effort to explain the diverse results of six trials of TB treatment interventions.

Contributing to greater social justice

Public health research has long mapped the processes of exclusion and privilege that create social inequities in the experience of health and illness across and within societies (Dorling 2011; Thomas and Dorling 2011). However, these maps typically do not capture the 'experiences of hardship, subjugation and alienation which cannot easily be … weighed via ratios, metrics and indices' (Trahair 2014, p. 6), but about which people have points of view. These 'knowledgeable narratives' (Williams 2000) people's 'practical wisdom', as Aristotle described it, can be accessed through qualitative research. For the philosopher Martha Nussbaum (1990, p. 55) people's practical wisdom provides a 'profound alternative' to the power of 'scientific' knowledge that pervades 'almost every area of human social life'. Without an understanding of this knowledge, public health policy/practice risks reducing people to unthinking bearers of various risks, deficits or assets and seriously limits the 'evidence' available to improve health and reduce health inequalities.

A particularly important aspect of the experiential knowledge revealed in qualitative research is the emotion it can embody. Conventionally emotion is seen as antithetical to scientific rationality, whereas the Aristotelian view articulated by Nussbaum is that emotions, and the imagination fed by them, are central to a fuller and more grounded understanding of how knowledge of human predicaments is articulated. Rather than bracketing emotions out as residuals in the analysis of health inequalities, it would be more fully rational to include them as components in a 'passionate epistemology' for a critical public health (Krieger 2000). The emotionality embedded in lay knowledge about the experience and consequences of ill-health points to the significance of these experiences for individuals, but can also highlight causal pathways within which different kinds of intervention or action might be developed (Popay et al. 2003a; Eades et al. 2008).

Finally, qualitative research can contribute to the development of the theory and practice of transformative action by illuminating both the need for, and barriers to, collective action as strategies to improve population health and reduce health inequalities (Elliot et al. 2015). Currently, prominent theories in the public health field focused, for example, on social capital, social cohesion, resilience, community assets, etc, describe the way things *are*, and suggest interventions that help people cope with current conditions but offer limited insight into how to overcome the threats themselves. For example, in their literature review Elliot et al. (2011) found that while most contemporary community-based public health interventions were potentially transformative for individuals they failed to support communities to acquire the collective social/political power they need to change the determinants of health and health inequities.

Recent work by the New Economics Foundation (Coote and Angel 2014) points to the need for further critical qualitative research in this field, arguing that while concepts such as social capital are about 'the glue' that holds people together, 'solidarity' concerns 'momentum' and the conditions for change. This conceptual reframing locates the sources of transformative action in civil society, but also highlights the need for better understanding of barriers and enablers to such action. One important contribution qualitative research has made to this understanding is to reveal how the strategies people develop to manage their lives can undermine solidarity. For example, Popay et al. (2003b) and Ruston (2009) in the United Kingdom and Bustos Castro and Lindbladh (2004) in Sweden have described how residents of low-income neighbourhoods frequently withdraw from the public sphere and adopt strategies of avoidance, surveillance, and fortification in order to distance themselves for the stigma attaching to living in these neighbourhoods.

Public/community involvement in research can also contribute to the development of capabilities for transformative action in public health. Health researchers have argued for decades that, whenever possible, research should be participative (Wadsworth 1984; Feuerstein 1986; Baum 1988; Bruce et al. 1995; de Koning and Martin 1996; Minkler and Wallerstein 2003; Popay 2010). Participative approaches to research have tended to be associated with qualitative research methods. In Australia the National Lowitja Institute—the National Institute for Aboriginal & Torres Strait Islander Health Research—has developed supports to guide

participatory and ethical research with Indigenous peoples (Laycock et al. 2011). Similar participatory approaches have been encouraged in Canada (Canadian Institutes of Health Research 2014).

Minkler and Wallerstein (2003) have developed a practice of community-based participatory research and described its growth and increasing acceptability in the past years. However, as Oakley (1989) has argued, quantitative methodologies, such as randomized control trials, can be emancipatory in practice and could be designed to involve research participants more than at present. Today, though still underdeveloped in the public health field, the active involvement of 'lay people' (e.g. residents of low-income neighbourhoods, users of services, or people with particular health conditions) as partners in research is a requirement of major funding bodies in many countries. In the United Kingdom, for example, the National Institute for Health Research has published National Standards for Public Involvement in research and expects this to be adequately funded in proposals.[1] The Australian National Health and Medical Research Council (NHMRC) requires researchers to demonstrate the degree to which Aboriginal and Torres Strait Islander people have been involved in the planning and will be involved in the execution the research (National Health & Medical Research Council 2003).

The movement for more participative research approaches partly reflects the increasing recognition that lay people can offer unique expertise. In addition, it reflects a growing commitment to democratize the research process. While participants in all types of research have the right to receive information about the research, formally agree to their participation and withdraw at any time, traditional research models involve what Wadsworth (1984) called 'data raids'. In this approach, researchers swooped down from their ivory towers, collected data, returned to their towers, and never communicated the results of the raid to the subjects. In contrast, the new public health is calling for a more meaningful form of involvement, including involving the public/communities in defining research questions, determining methods, collecting and analysing data, interpreting findings, and deciding how they could be reported back in a readily comprehensible way. In particular, there is an increasing emphasis on approaches, such as action research, that enable traditionally marginalized groups and communities to become full partners in gaining and creating knowledge and mobilizing for change. AS we have noted a participative research approach

is essential when conducting Aboriginal research given the history of the use of research as part of the processes of colonization and inhumane treatment of Indigenous peoples. The People's Health Movement conducted participative research to examine its own form of collective action to improved health and has suggested this approach presents particular methodological challenges (Bodini et al. 2019). These concerned the following issues: making meaning; aligning research and action; managing power relations; valuing experiential knowledges, chaos, and contingency; challenging preconceptions. Specific approaches to participative research are discussed in more detail in the next section of the chapter (see also Minkler and Wallerstein 2003, p. 2).

Doing qualitative research

In this section we move on to consider the range of study designs that can be used in qualitative research, the specific methods, or tools, available for data collection, and approaches to analysis. Patton (2014) suggests that qualitative data can be grouped into three categories:

in-depth, open-ended narrative data collected from individuals or groups;
direct observation and description of people's activities, behaviours, actions, and interactions, including audio- and video-taped material. Kellehear (1993) extends the definition to include the study of 'material culture' (graffiti, garbage, cemeteries);
written data, usually excerpts, quotations, or entire passages from organizational, clinical, or programme records, personal diaries, official records or publications, and open-ended written responses to questionnaires.

These data can be collected in the context of different study designs the main ones being:

ethnographies (based on observations);
interview-based studies (individual or group based);
case studies (which may involve both qualitative and quantitative methods and most commonly use in-depth and/or focus group interviews to collect qualitative data);
document-based research and content analyses;
participative action research (using a variety of methods).

Next we discuss the first four of these study designs, the methods available to collect qualitative data, approaches to analysing these data and how 'quality' in qualitative research can be assessed. Approaches to participative action research and then discussed and described in the final section of the chapter.

Study designs and data collection methods

Ethnographic approaches and observation

Ethnographers seek to immerse themselves in a study context to study 'everyday' activities and practices in their natural setting in order to develop a holistic, in-depth, understanding of the 'way of life' of a group, institution, and/or community. Denzin (Denzin 1989, p. 83) described this as producing a 'thick description', which:

> goes beyond mere fact and surface appearances. It presents details, context, emotions, and the webs of social relationships that join persons to one another. Thick description evokes emotionality and self-feeling. It inserts history into experience. It establishes the significance of an experience, or the sequence of events, for the person or persons in question. In thick descriptions, the voices, feelings, actions and meanings of interacting individuals are heard.

Two 'classic' ethnographies in the health field are Asylums (Goffman's (1961) study of a single mental hospital for which he posed as a member of staff for over a year) and Boys in White (Becker et al.'s 1961 study of medical education). Reading these studies, which have had profound impacts on health policy and practice, is a good way of becoming familiar with the methodology.

Ethnographers used a range of methods but participant observation, where the observers are fully active members of the activities being studied is the central method of data collection. Observational approaches to data

collection are, however, used in other study designs, often less intensively when the observer may be non-participant (maintaining distance by not having a formal role in the activities there). Observation typically involves the use of field diaries to record data, though video and audio recording may be used. Field notes are written up as soon as possible after the observation and include description of events that occurred (who, what, when, where, how), theoretical notes that try to derive meaning from the descriptive observation and notes on methodological issues that may impinge on the credibility of the research. As in all qualitative research observers are typically reflecting continually on the meaning of their observations during fieldwork and beginning to analyse by generating categories and ideas.

The ethical challenges of being an observer have been a source of debate as have the methodological issues: researcher presence may change the context under investigation. Building trust between researcher and researched, important to the quality of all empirical research, is vital. However, a balance has to be struck between building trust, understanding issues through the eyes of the researched and maintaining a critical observer stance.

Interview-based studies

Interviews are an almost ubiquitous element of qualitative research but they can take various forms: individuals or groups; short informal conversation or long sessions over many hours; one-off or longitudinal. They are a way of gaining a more ‘in-depth’ perspective on experiences than that provided by more structured forms of research (Olesen 1994). They are typically used to explore people’s understanding or experience of a phenomenon (e.g. poverty or housing quality), the meanings they attach to these experiences and the impact on their lives. An example of this type of research is Williams and Elliot’s Health Impact Analysis (Elliott and Williams 2008) of plans to demolish substandard housing in a South Wales community, which included a qualitative study involving both individual in-depth interviews and focus groups. The qualitative findings revealed that despite recognizing housing problems residents and local professionals were ambivalent and uncertain about the developments because of the potential disruption to social and family networks. Payment of expenses (travel and childcare) may often be necessary, especially for people on low incomes. However,

payment of a fee as an inducement to participate in individual or group interviews, which is becoming more common, raises ethical questions. Some ethics committees are uncomfortable with the practice. In terms of validity, payment may encourage people to take part primarily to earn the money, possibly even claiming membership of a group in order to qualify.

In-depth interviews are either unstructured (e.g. oral history or life history) or guided by a set of topics to be used flexibly with each interviewee. Interviewees are encouraged to express ideas and experiences in their own words, generating. 'textual' or 'narrative' data. Interviews are usually tape recorded and transcribed for analysis. Respondents are often selected purposively or 'theoretically' to reflect a range of experience of the phenomenon of interest (e.g. a health condition, an intervention, a place). Sample sizes vary but typically would be around 20–30 interviewees. Decisions about the style of interview to use will be influenced by the:

Topic of the research. More sensitive research topics (e.g. fertility) demand longer interviews than others.
Resources available. Long, in-depth interviews are costly to conduct, transcribe, and analyse. Tape recording is the preferred option (assuming respondents agree): responses can be handwritten but this is hard work as the interviewer also has to concentrate on maintaining eye contact and responding appropriately.
Tolerance of the interview group. In-depth interviews can be quite intrusive. People also have to give up time, possibly 2 or 3 hours. This may be easier for some people. Health professionals may be less willing, but older people are often more willing to spend time being interviewed.

In-depth interviewing is a skilled process but all members of a team doing research involving in-depth interviews should do at least some of the interviewing. Textbooks dealing with in-depth interviewing contain a considerable amount of advice for researchers (Glesne and Peshkin 1992; Seidman 2005; Minichiello et al. 2008). It is important to consider issues of gender, culture, language, and social class when planning interviews and in addition to establishing rapport, researchers need to: be able to explain the purpose of the research clearly; to keep an open mind and see 'the respondent as teacher' (Glesne and Peshkin 1992, p. 81); to learn and adapt as the interviews proceed; and to adopt an analytical approach from the

beginning of the research. In-depth interviewing has also been used to good effect to study policy processes such as Townsend et al. (2019) study of how paid parental leave was able to get in the policy agenda.

Focus groups are interviews with groups of people on a particular focused issue typically for up to 2 hours. They are often used to develop questionnaires for quantitative research or to obtain participants interpretation of results from the analysis of other data/studies. Increasingly, however, they have been used as a self-contained means of qualitative data collection. For example, Davidson, Kitzinger, and Hunt (2006) used 14 focus group discussions in Scotland and the north of England to explore views on health inequalities among people from different socioeconomic background. Focus groups are also commonly used in health promotion needs assessment. Similarly, Cortie et al. (1996) used four focus groups (stratified according to low and high socioeconomic status and extent of physical activity) to determine the factors in a local community that determine people's decision to exercise or not while Jacobs and Roberts (2004) used focus groups to collect data from health professionals on approaches to addressing malnutrition in Cambodia.

As Brown et al. (1989, p. 40) comment focus group 'are not just a convenient way to accumulate the individual knowledge of their members. They give rise synergistically to insights and solutions that would not come about without them'. They also enable participants to control the agenda and formulate opinions in a way that would not be possible in individual interviews (Morgan 1988, p. 28) and to explore topics in an open format at the beginning of research. Researchers also need to be aware of potential problems including:

'Group think' which may discourage individuals from expressing opinions at odds with the majority: 'devil's advocates' may overcome this tendency (MacDougall and Baum 1997).
Dominant individuals who can influence results: skilful moderation to minimize this risk is vital.
Less control by the researchers over the data that are generated.
Challenges associated with recording discussions.

The number and size of focus groups will be determined by research goals and available resources. Krueger and Casey (2009) recommends using

groups of between 5 and 10 but say size can be as few as 4 to as many as 12: the key point is that the group must be small enough for everyone to have an opportunity to provide insights and yet large enough to give a range of perceptions. As with individual interviews, the number of group interviews required to answer the research question cannot always be predicted in advance. It will also be necessary to decide how to divide the 'population'. For instance, if gender is likely to be a significant variable, it may be appropriate to have separate groups for men and women. In an evaluation, more 'honest' discussion may ensure if for example the providers of services are in a separate group from users. The research topic will usually determine how easy it is to recruit people for focus groups. Generally, the more sensitive the topic, the more challenging is recruitment.

As with all interviews, the quality of data obtained from focus groups is dependent on how they are run. Skills in facilitation, knowledge of groups dynamics and an understanding of the research area (so that interesting 'leads' are followed up in the discussion) are essential. For Morgan (1988) the role of the facilitator is to set up the group with instructions and then move to the side, intervening only if it goes badly off track. The venue will also be important: participants need to feel comfortable and relaxed and it needs to be suitable for achieving high-quality recordings. The data from a focus group are most easily recorded on audio tapes or digital media. It is usual to have one researcher facilitating the group and another taking notes. Depending on the research it may be sufficient to have a co-researcher take notes during the focus group, merely supplementing them with a recording but if more detailed content analysis is required then the recording would be transcribed verbatim. Focus group information, like other textual data, may be analysed according to its content and themes drawn out as discussed in the next section.

Case studies

Case studies involve the systematic study of an individual, group or community, an institution, or an event with a view to understanding why something happens in a particular context. Typically case studies place less emphasis on members' tacit knowledge than in ethnography, involve less intensive fieldwork, and are more likely to combine quantitative and qualitative methods. Data is usually collected from a variety of sources, including in-depth individual or group interviews, self-completion surveys,

and documents. Single or multiple cases can be studied, they can be descriptive or explanatory and may involve testing hypotheses.

Case studies were once seen as an exploratory phase of research, but are now widely accepted as evaluation tools in their own right (Stake 1995; Yin 2018). They are used extensively by development agencies such as the World Bank and US AID because 'it is more desirable to have a few carefully done case studies with results one can trust than to aim for large, probabilistic, and generalizable samples with results that are dubious because of the multitude of technical, logistic and management problems in third world settings' (Patton 1990, p. 100) Policymakers and politicians may relate to a case study and its story better than to a set of statistics so potentially increasing the likelihood that the research will be impactful. Combining statistics with case studies may be a particularly effective way of politicians. The Healthy Cities project is an international initiative based on the evaluation of international case studies. A summary evaluation of The Noarlunga Healthy Cities initiative (Baum et al. 2006a), for example, involving a narrative review of findings from a number of reported qualitative case studies of initiative activities and other project documents identified nine factors that lead to the initiative's sustainability over 18 years. Case studies have also been used to study the extent to which primary healthcare services are comprehensive and highlight the value of comparative case studies (Baum et al. 2013).

According to Yin (2018, pp. 242–7) criteria for exemplary case studies include: significance (i.e. focused on issues that are unusual and of general public interest, or important in theoretical, policy, or practical terms); *c*ompleteness (i.e. having clear boundaries and giving full attention to all relevant evidence); consider alternative perspectives on how the data should be interpreted; provide a *s*ufficiency of evidence to allow readers to judge the quality and robustness of conclusions; and use *e*ngaging approaches to dissemination, *such as videos.*

Document-based research

The collation and analysis of documents (such as minutes of meetings, annual reports, newspapers, etc.) is a valuable method in qualitative research. They can provide a 'particularly rich source of information about many organizations and programs' (Patton 2014, p. 176), and are often used in combination with other methods (Bowen 2009). Document analysis has

been used to assess uptake of research evidence in health policies (Flitcroft et al. 2011; Rosella et al. 2013) and the extent to which health policies address equity issues (Keleher 2013), and is emerging as a method to assess uptake of evidence on social determinants of health and health inequities (Pinto et al. 2012; Borrell et al. 2013; Fisher et al. 2014). Document analysis is also highly relevant to the commercial determinants of health, Anaf et al. (2017; see also Anaf et al. 2019) were able to conduct corporate health impact assessments by referring to company annual reports, other corporate documents, and media analysis.

Diaries/autobiographies are a particular form of documents used in qualitative research. Some studies make use of existing diaries other require participants to keep a diary for a set period focusing on key events they judge to be memorable. The method is particularly good for longitudinal data collection where recall may be a challenge, and where repeated interviews are not feasible. Diaries may be more or less structured and may be paper-based, computer based, or online blogs. The data are likely to be analysed in the same way as an interview transcript.

Sampling in qualitative research

The rationale for sampling in qualitative studies is purposive rather than random as in quantitative research. The aim is to select cases that will provide the most relevant and richest data. For this reason, as Kuzel (1992, p. 41) notes:

the sample can be adjusted as the theory develops, the aim being to thoroughly explore the topic of interest and to consider as many different angles and perspectives as possible. Glaser and Strauss (1967) refer to this as ‘theoretical sampling’;
people/cases are selected serially and sampling continues until little new information is gained (referred to as saturation);
‘negative or deviant cases’ are searched for to test ideas emerging from the analysis.

‘Snowball sampling’ is common. It involves contacting a few members of the group that is the focus of the research, and asking them to help find others. This technique is particularly useful for reaching groups that are not

readily identifiable, such as illicit drug users or voluntarily childless people. Researchers may also have to advertise. As already noted, sample size depends on what you want to know, what will be useful, what will have credibility, and what can be done with available time and resources (Patton 2014, p. 311). However, as a rule of thumb, six to eight data sources or sampling units will often be sufficient where the group is homogeneous, while up to 20 might be needed when looking for disconfirming evidence or trying to achieve maximum variation (Lincoln and Guba 1985; Patton 1990; Kuzel 1992). Qualitative researchers may be asked to specify their sample size in advance because of the requirements of a research funding body. The crucial feature is the logic underlying the sampling strategy, as this is the main criterion used to judge the strength of the sampling method. A proposal should be able to describe the sampling strategy clearly, state the selection criteria and provide an approximation of the sample size (National Health & Medical Research Council 2007).

Analysing qualitative textual data

In qualitative research analysis is the point at which data and theory are brought together to try and generate new understandings and explanations of social phenomena. Two elements in the analysis process can be distinguished: the purpose (what is being sought) and the practice (how it is done).

The purpose of data analysis

Although visual or audio data may be collected most qualitative research produces textual data, which typically comprise a mix of field or observational notes and transcripts from interviews or focus groups. Some examples of different approaches to analysis of these data are summarized next, but the key point is that the choice of approach to data analysis should be driven by the purpose of the research. For example, a narrative, life history, discourse, or conversation analysis can be conducted on the same extract from an interview, hence the importance of surfacing researcher standpoint and the theory driving a particular piece of research.

Content analysis: looks at the way themes and issues arise across texts (including interview transcriptions). Analysis may focus on the context,

frequency, and/or how themes are patterned by e.g. gender or ethnicity. This is a descriptive level of analysis.

Narrative analysis: looks at the way a person constructs a 'story' in the light of the audience and their purpose for giving the account. There is a focus on language, imagery, metaphor, and rhetorical purpose in the story being told.

Conversation analysis (CA): focuses on the structure of communication and conversation management such as turn taking or pausing. The aim is to reveal how meaning is constructed in interaction. CA is very specialized and marks both words spoken and how conversation proceeds (for example intonation).

Discourse analysis: explores the way knowledge is produced in particular contexts through the use of specialized language or theories and through performances, interaction, and rhetorical devices used to persuade. A range of texts can be subjected to discourse analysis (e.g. interview transcripts, video, letters, policy documents).

The practice of data analysis

In the broadest terms qualitative data analysis (QDA) involves the imposition of meaning and interpretation on mainly textual, usually unstructured, and unwieldy data. There are several schools of thought about methods of QDA but three theoretical traditions are particularly significant: interpretivism, social anthropology and collaborative research (Miles and Huberman 1994, p. 8). Interpretivism emphasizes the meaning of the data, and works from the premise that researchers' perceptions and understanding are affected by, and affect, the process of research. Included within this general tradition are ethnomethodologists, phenomenologists, deconstructionists, and discourse analysts. Social anthropology, including ethnography, focuses more on the accuracy of description and less on the conceptual or theoretical meaning of the observations. There is an interest in 'discovering' the underlying patterns or rules in a situation, which often begins with a theory that is tested during analysis. Within this broad tradition come researchers in life history, grounded theory, and much applied qualitative research and evaluation in education and health. Finally, there is collaborative research, in which analysis is either through reflective (questioning) or dialectic (opposing) enquiry with the research participants interspersed with further rounds of data collection.

Not-with-standing these different traditions, QDA always involves two distinct processes: description and interpretation, with the former comprising the identification of themes, concepts and categories in order to develop ideas or 'theories' about the data and relationships within it. There are many textbooks describing specific approaches to qualitative data but the Framework Analysis developed in the 1980s by applied qualitative researchers Ritchie and Spencer (1994) in 1994 and updated in 2011 is now extensively used. This involves the five stages briefly described next.

Familiarization with the range and diversity of data collected. A selection of the data should be read and (if recorded) listened to. Key ideas and themes should be listed.

Identifying a thematic framework within which textual data can be sifted and sorted. The framework should draw upon the original research ideas, theories informing the research, issues raised by respondents and themes identified during familiarization.

Indexing is the process whereby the thematic framework is applied to the data. Each transcript or set of field notes is coded by reference to the themes, usually using a computer package. A commonly used package is NVIVO (Bazeley and Jackson 2013). The system of annotating the text means the process is visible and accessible to others, and can be checked by two or more researchers.

Charting refers to the process of taking the data from its original context and rearranging it according to the appropriate thematic reference. Data are usually analysed by considering each theme across all respondents/cases. Charts can be produced manually but computer packages can also be used to compile all the data relating to a particular theme, which is then available for further analysis.

Mapping and interpretation involve reviewing the themes, comparing accounts and experiences, searching for patterns and connections, and seeking explanations for them within the data: processes that require 'leaps of intuition and imagination' (Ritchie and Spencer 1994, p. 186).

The use of computer packages is now common in QDA. Importantly, however, the intellectual work of description and interpretation is the responsibility of the researcher. Software cannot conduct analysis: it is simply a tool that can help make the analytical process more systematic by helping to:

Manage and keep track of all data including raw data files from interviews, survey and focus groups, or field observations, images, diagrams, audio, video, web pages, policy documents, ideas recorded as memos.
Organize and provide rapid access to conceptual and theoretical knowledge you generate.
Retrieving all relevant information from your data base in response to questions you pose. Queries can be saves and so becomes part of an ongoing enquiry process.
Visualize data to show the content of cases, ideas, concepts timelines, etc. and to represent these relationships visually.
Report from the data: including information about the original data source, the ideas and knowledge developed from them and the processes by which outcomes were reached (Bazeley and Jackson 2013, p. 3).

Assessing the quality of qualitative research

Validity, reliability, and generalisability are concepts that enable the value of positivist research to be judged. There is uncertainty in the literature about the extent to which these concepts can be applied to qualitative research. Kirk and Miller (1986) argue that they can be adapted to qualitative research methods, but others suggest alternative terms. Lincoln and Guba (1985) suggest using the terms ‘credibility’, ‘transferability’, ‘dependability’, and ‘confirmability’.

Credibility: Patton (2014) suggests this can be determined by reference to:

the rigour of the techniques and methods used
the credibility of the researcher (e.g. training, experience, track record, and status)
the philosophical orientation and assumptions that underpinned the study.

Triangulation can be a useful way of increasing the credibility of qualitative research. Denzin (1978) originally defined four types of triangulation:

data source triangulation—the use of a variety of data sources
researcher triangulation—the use of several different researchers or evaluators
theory triangulation—the use of multiple theoretical perspectives to interpret the same data
methodological triangulation—the use of multiple methods to study a particular problem.

To these Janesick (1994) adds a fifth type, which is relevant to public health:

interdisciplinary triangulation—having researchers from a variety of professions. This can be extended to include non-researchers who have an interest in the findings of research, such as residents of neighbourhoods receiving interventions that are being evaluated.

The analysis of negative cases that do not fit broader patterns identified in the data is also an important way of increasing credibility. Patton (2014, p. 656) comments: 'Dealing openly with the complexities and dilemmas posed by negative cases is both intellectually honest and politically strategic.'

Transferability refers to the ability of qualitative researchers to extend their findings to other settings. Lincoln and Guba (1985) suggest that research reports should provide sufficient detail (about methods, parameter, setting) for other researchers and users to make judgements about transferability. The use of theory in qualitative research is central to its transferability: Hilary Graham's (1994) research re-conceptualizing smoking among poor white women in England as a coping mechanism rather than a health damaging behaviour is an example of how the findings from qualitative research can be generalized on the basis of theory.

Dependability and reliability refer to whether the research is likely to be consistent over time and across researchers and methods and should be considered at all stages of the research. The reliability of data depends on skills in interviewing, observing, and recording. Analysis procedures, such as having two people involving in coding of data to themes, are likely to increase the reliability of interpretation.

Confirmability refers to the need to confirm research results with a source outside the research team and overlaps with interdisciplinary triangulation

discussed earlier. De Laine (1997, p. 279) suggests that an 'audit trail' of sampling decisions, methods of recording, analysis, coding, construction of themes, and personal notes would enable an outsider to follow the logic of each stage and understand how and why the researchers had drawn the conclusions they had.

There are, therefore, distinct ways of assessing qualitative research, but it is still common for the validity and reliability of public health qualitative research to be called into question by those who consider it to be subjective, and so invalid and unreliable. The issue of subjectivity and objectivity in science has been extensively debated by science philosophers, and today there are few who still maintain that positivist science is as objective as it was once thought to be, and there is broad acceptance that qualitative data can make a crucial contribution to aspects of public health. Patton (2014, p. 58) takes a pragmatic position that appears also to make sense for public health when he suggests we 'stay out of futile debate about subjectivity versus objectivity'.

Participatory action research

As we discussed earlier, participatory approaches are embedded in a broader movement to democratize the research process and to contribute to greater social justice. These approaches, which are not confined to qualitative research, challenge the privileged position of the professional researcher. They have drawn on the work of the adult educator Paulo Freire (1972) and place as much emphasis on the process as on the product of research.

The term 'action research' is often used interchangeably with participatory research; but sometimes the two are combined into 'participatory action research' (PAR). In medical and public health circles action research has received more attention in recent years but it had its origins in the work of educationists during the 1940s, particularly the work of the North American Kurt Lewin and it is now an accepted and respected research tradition in educational and management research (McTaggart 1991).

Drawing on the work of Minkler and Wallerstein (2003) and Grbich (1999) Baum et al. (2006b) define PAR as an approach that seeks to

understand and improve the world by changing it. At its heart it is collective, self-reflective enquiry that researchers and researched undertake together so they can understand and improve upon the practices in which they participate and/or the situations in which they find themselves. The reflective process is directly linked to action, influenced by understanding of history, culture, and local context and embedded in social relationships. The process of PAR should be empowering and lead to people having increased control over their lives. PAR pay careful attention to power relationships, advocating for power to be deliberately shared between researcher and researched: blurring the line between them until the researched becomes the researcher. The researched cease to be objects and become partners in the whole research process: including selecting the research topic, data collection, and analysis and deciding what action should happen as a result of the research findings. PAR may use a range of qualitative and quantitative approaches to data collection.

PAR is therefore a process that includes the generation of knowledge alongside 'education and development of consciousness and of mobilisation' (Gaventa 1988, p. 19). Some of the most exciting experimentation with participatory research has been in low-income countries where there is a potential to raise people's consciousness and encourage them to become involved in actions that could liberate them from poverty and despair. Bloem and colleagues (1996), for example, describe the development of the People's Participatory Planning (PPP) process by a non-government organization in Bangladesh in 1990. These authors stress that the PPP is not a universal panacea but overall they conclude that 'the continual process of reflection and action can enable people to change and transform from one level of functioning to another' (1996, p. 150). Similar conclusions were reached by Howard-Grabman (1996) in her analysis of a PAR designed to address maternal and neonatal health problems in rural Bolivia. The involvement of men and women in the project led to the development of new reproductive health practices that have seen maternal, perinatal, and neonatal mortality decline in the communities that were part of the project. Similarly, the PAR study by Hague and colleagues (2011), which found that the practice of paying for wives in rural Uganda was associated with increased domestic violence and male power over women, developed skills and capacity at the local level and made recommendations for strategies for action at both the national and the local level.

PAR does not require the researcher to downplay their specialist knowledge but to use that knowledge in dialogue with the expert knowledge of participants in the research process. As Reason (1994, p. 328) argued 'it is through dialogue that the subject–object relationship of traditional science gives way to a subject–subject one, in which the academic knowledge of formally educated people works in a dialectic tension with the popular knowledge of the people to produce a more profound understanding of the situation'. Similarly, PAR must avoid eulogizing the potential contribution of research participants. Just as professional researcher perspectives must be subject to critical reflection, so too must those of participants through a rigorous process of 'critical subjectivity' to arrive at new forms of interpretation and knowledge (Reason 1994, p. 333).

The practice of PAR presents dilemmas in the public health domain as it does in other fields. Boutilier, Mason, and Rootman (1997, p. 76) describe a number of these and suggest that key issues to be resolved in the practice of what they call 'community reflective action research' relate to: whose knowledge is valued; who owns the research; why the research is being done; and recognition of the complementary skills of lay people, researchers, and practitioners. Community participation and control of research processes may also lead to tensions with the demands of scientific rigour (Allison and Rootman 1996), as participants are generally not experts in research or primarily interested in the validity of research. These tensions should be addressed so that evaluations are scientifically acceptable and ethically participatory but they are not easy to reconcile. Inevitably, then, the processes of negotiation in participatory action research can be long and complicated and may in turn conflict with the needs of funding bodies. The key point is that PAR creates new dialogical spaces in which experiential knowledge is taken seriously: contributing to the creation of a democratic public health science that builds knowledge and understanding about population health and health equity through 'narratives of living and being'.

New directions in qualitative research

Qualitative public health researchers have recently begun to think of more creative ways in which to engage with and represent the experiences of people and communities by, for example, working more closely with the

arts and new methods including photovoice, digital storytelling and poetry which have been used to engage homeless women in research (Walsh et al. 2010). These kinds of initiatives are asking us to 'imagine' what cultural representations of population health and health inequity might look like, beyond the graph and the quote and they point to ways in which public health research might take more seriously the knowledge and concerns of local publics (Burawoy 2005). Working inside the fabric of everyday life can reveal the complex relationships between private troubles and local, national, and global public forces more clearly, illuminating the limits of behaviour change and demonstrating the value of experiential knowledge. This surely offers the hope of public health research with a stronger connection to broader social and political action in today's increasingly unequal, unjust social systems.

Conclusion

Qualitative research is vital to the study of global public health. Although primarily emanating from social epidemiology, public health research is increasingly utilizing more integrative approaches to explanation bringing together insights from different disciplines. Qualitative research has a key role to play in this context by opening up new theoretically informed analyses of the relationships between individual lives and social structures, including 'materially-informed' ways of thinking about the contribution of 'individual lifestyle' factors to health inequalities (Bartley 2004). Qualitative data can add depth to our understanding of the nature and impacts of the relationships between the social, commercial, and political determinants of population health and health inequalities described in quantitative research. Qualitative research can open up understanding of what things *mean* to people, what drives them to act (or not) in the situations they are in (Sayer 2011). Much public health research neglects the *particulars* of people's situation and the knowledge they develop and utilize to deal with these situations. Importantly, good quality qualitative research makes theoretical connections from these 'particulars' to more general social structures and processes. This is illustrated by our earlier example of qualitative research in which the concept of smoking as a coping mechanism was developed to demonstrate how white working-class

women caring for children while living in poverty relied on a cigarette to help them manage very stressful situations (Graham 1994). By making these connections (in this case between smoking and coping with poverty) and involving people as partners in the research process qualitative research can support public health practitioners and policymakers and the public more widely to develop new forms of solidarity: to 'imagine' and implement mechanisms for personal and collective change that will contribute to improving health and reducing inequities (Elliot et al. 2015)

REFERENCES

Allison, K.R. and Rootman, I. (1996). Scientific research and community participation in health promotion research: are they compatible? *Health Promotion International*, 11(4), 333–40.

Anaf, J., et al. (2017). Assessing the health impact of transnational corporations: a case study on McDonald's Australia. *Globalization and Health*, 13(7), 1–7.

Anaf, J., et al. (2019). The health impacts of extractive industry transnational corporations: a study of Rio Tinto in Australia and Southern Africa. *Globalization and Health*, 15(13), 1–13.

Armstrong, K. (2000). *The Battle for God: Fundamentalism in Judaism, Christianity and Islam*. London: Knopf/HarperCollins.

Bartley, M. (2004). *Health Inequality: An Introduction to Concepts, Theory and Methods*. Cambridge: Polity Press.

Baum, F. (1988). Community-based research for promoting the new public health. *Health Promotion International*, 3, 259–68.

Baum, F. (2016). Qualitative research methods. In: Baum, F. (ed.) *The New Public Health* (4th ed.). Melbourne: Oxford University Press.

Baum, F., et al. (2013). Health promotion in Australian multi-disciplinary primary health care services: case studies from South Australia and the Northern Territory. *Health Promotion International*, 29, 705–19.

Baum, F., et al. (2006a). What makes for sustainable Healthy Cities initiatives?—A review of the evidence from Noarlunga, Australia after 18 years. *Health Promotion International*, 21, 259–65.

Baum, F., MacDougall, C., and Smith, D. (2006b). Participatory action research [Glossary]. *Journal of Epidemiology and Community Health*, 60, 854–7.

Bazeley, P. and Jackson, K. (2013). *Qualitative Data Analysis with NVivo*. London: SAGE Publications.

Becker, H.S., et al. (1961). *Boys in White: Student Culture in Medical School*. Chicago, IL: University of Chicago Press.

Bloem, M., Biswas, D., and Adhikari, S. (1996). Towards a sustainable and participatory rural development: recent experiences of an NGO in Bangladesh. In: De Koning, K. and Martin, M. (eds.) *Participatory Research in Health: Issues and Experiences*. London: Zed Books.

Bodini, C., et al. (2019). Methodological challenges in researching activism in action: civil society engagement towards health for all. *Critical Public Health*, 30(4), 386–97.

Borrell, C., et al. (2013). Comparison of health policy documents of European cities: are they oriented to reduce inequalities in health? *Journal of Public Health Policy*, 34, 100–20.

Boutilier, M., Mason, R., and Rootman, I. (1997). Community action and reflective practice in health promotion research. *Health Promotion International*, 12, 69–78.

Bowen, G.A. (2009). Document analysis as a qualitative research method. *Qualitative Research Journal*, 9, 27–40.
Brown, J.S., Collins, A., and Duguid, P. (1989). Situated cognition and the culture of learning. *Educational Researcher*, 18, 32–42.
Bruce, N., et al. (1995). *Research and Change in Urban Community Health*. Aldershot: Avebury Press.
Burawoy, M. (2005). For public sociology. *American Sociological Review*, 70, 4–28.
Bustos Castro, P. and Lindbladh, E. (2004). Place, discourse and vulnerability—a qualitative study of young adults living in a Swedish urban poverty zone. *Health & Place*, 10, 259–72.
Camlin, C.S. and Seeley, J. (2018). Qualitative research on community experiences in large HIV research trials: what have we learned? *Journal of the International AIDS Society*, 21, 55–9.
Campbell, R., et al. (2008). An informal school-based peer-led intervention for smoking prevention in adolescence (ASSIST): a cluster randomised trial. *The Lancet*, 371, 1595–602.
Canadian Institutes of Health Research (2014). *Health Research in Canada and You*. Ottawa: Canadian Institutes of Health Research, p. 20.
Coote, A. and Angel, J. (2014). *Solidarity: Why It Matters for A New Social Settlement*. London, UK: New Economics Foundation, pp. 1–34. [Online] Available at: https://b.3cdn.net/nefoundation/207c255d8a0c04cba0_sum6b1yy7.pdf
Cortie, B., Donovan, R.J., and Holman, C.D. (1996). Factors influencing the use of physical activity facilities: results from qualitative research. *Health Promotion Journal of Australia*, 6, 16–21.
Davidson, R., Kitzinger, J., and Hunt, K. (2006). The wealthy get healthy, the poor get poorly? Lay perceptions of health inequalities. *Social Science and Medicine*, 62, 2171–82.
de Koning, K. and Martin, M. (1996). *Participatory Research in Health: Issues and Experiences*. London: Zed Books.
de Laine, M.D. (1997). *Ethnography: Theory and Application in Health Research*. Sydney: MacLennan & Petty.
Denzin, N.K. (1978). *The Research Act: A Theoretical Introduction to Sociological Methods*. New York: McGraw Hill.
Denzin, N.K. (1989). *Interpretive Interactionism.* Newbury Park, CA: SAGE Publications.
Dorling, D. (2011). *So You Think You Know About Britain?* London: Hatchette UK.
Eades, S., et al. (2008). Bibbulung Gnarneep (‘solid kid’): causal pathways to poor birth outcomes in an urban Aboriginal birth cohort. *Journal of Paediatrics and Child Health*, 44, 342–6.
Elliot, E., Popay, J., and Williams, G. (2015). Knowledge of the everyday: confronting the causes of health inequalities. In: Smith, K., Bambra, C., and Hill, S. (eds.) *Health Inequalities: Critical Perspectives.* Oxford: Oxford University Press.
Elliott, E., et al. (2011). *Connected Communities: A Review of Theories, Concepts and Interventions Relating to Community-level Strengths and Their Impact on Health and Well-being*. [Online] Available at: https://ahrc.ukri.org/documents/project-reports-and-reviews/connected-communities/a-review-of-theories-concepts-and-interventions-relating-to-community-level-strengths-and-their-impact-on-health-and-well-being/
Elliott, E. and Williams, G. (2008). Developing public sociology through health impact assessment. *Sociology of Health & Illness*, 30, 1101–16.
Feuerstein, M.T. (1986). *Partners in Evaluation: Evaluating Development and Community Programmes with Participants*. London: Macmillan.
Fisher, M., et al. (2014). A qualitative methodological framework to assess uptake of evidence on SDH in health policy. *Evidence & Policy*, 10.1332/174426414X14170264741073.
Flitcroft, K., et al. (2011). Getting evidence into policy: the need for deliberative strategies? *Social Science and Medicine*, 72, 1039–46.
Freire, P. (1972). *Pedagogy of the Oppressed*. Harmondsworth: Penguin.

Gaventa, J. (1988). Participatory research in North America. *Convergence*, 21, 17–46.
Glaser, B.G. and Strauss, A.L. (1967). *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Chicago, IL: Aldine.
Glesne, C. and Peshkin, A. (1992). *Becoming Qualitative Researchers: An Introduction*, White Plains, NY: Longman.
Goffman, E. (1961). *Asylums: Essays on the Social Situation of Mental Patients and Other Inmates*. Harmondsworth: Penguin.
Graham, H. (1994). Gender and class as dimensions of smoking behaviour in Britain: insights from a survey of mothers. *Social Science and Medicine*, 38, 691–8.
Graham, H. and McDermott, E. (2006). Qualitative research and the evidence base of policy: insights from studies of teenage mothers in the UK. *Journal of Social Policy*, 35, 21–37.
Grbich, C. (1999). *Qualitative Research in Health. An Introduction*. St Leonards: Allen & Unwin.
Hague, G., Thiara, R.K., and Turner, A. (2011). Bride-price and its links to domestic violence and poverty in Uganda: a participatory action research study. *Women's Studies International Forum*, 34(6), 550–61.
Howard-Grabman, L. (1996). Planning together: developing community plans to address priority maternal and neonatal health problems in rural Bolivia. In: Koning, K.D. and Martin, M. (eds.) *Participatory Research in Health: Issues and Experiences*. London: Zed Books, pp. 153–63.
Jacobs, B. and Roberts, E. (2004). Baseline assessment for addressing acute malnutrition by public-health staff in Cambodia. *Journal of Health, Population and Nutrition*, 22(2), 212–19.
Janesick, V. (1994). The dance of qualitative research design: metaphor, methodolatry, and meaning. In: Denzin, N.K. and Lincoln, Y.S. (eds.) *Handbook of Qualitative Research.* Thousand Oaks, CA: Sage, pp. 209–19.
Keleher, H. (2013). Policy scorecard for gender mainstreaming: gender equity in health policy. *Australian and New Zealand Journal of Public Health*, 37, 111–17.
Kellehear, A. (1993). *The Unobtrusive Researcher: A Guide to Method*. St Leonards: Allen & Unwin.
Kirk, J. and Miller, M. (1986). *Reliability and Validity in Qualitative Research.* Newbury Park, CA: Sage.
Krieger, N. (2000). Passionate epistemology, critical advocacy and public health: doing our profession proud. *Critical Public Health*, 10, 287–94.
Krueger, R.A. and Casey, M.A. (2009). *Focus Groups: A Practical Guide for Applied Research*. Thousand Oaks, CA: Sage Publications.
Kuzel, A.J. (1992). Sampling in qualitative inquiry. In: Miller, B. (ed.) *Doing Qualitative Research*. Newbury Park, CA: Sage, pp. 31–44.
Laycock, A., et al. (2011). *Researching Indigenous Health: A Practical Guide for Researchers.* Carlton South: Lowitja Publishing.
Lincoln, Y.S. and Guba, E.G. (1985). *Naturalistic Inquiry.* Beverly Hills, CA: Sage.
MacDougall, C. and Baum, F. (1997). The devil's advocate: a strategy to avoid groupthink and stimulate discussion in focus groups. *Qualitative Health Research*, 7, 532–41.
Mallinson, S., Popay, J., and Williams, G. (2014). Qualitative research for public health practitioners. In: Guest, C., et al. (eds.) *Oxford Handbook of Public Health Practice*. Oxford: Oxford University Press.
McTaggart, R. (1991). *Action Research: A Short Modern History*. Melbourne: Deakin University.
Miles, M.B. and Huberman, A.M. (1994). *Qualitative Data Analysis*, Thousand Oaks, CA: Sage.
Minichiello, V., Aroni, R., and Hays, T. (2008). *In-depth Interviewing*. Frenchs Forest: Pearson Australia.
Minkler, M. and Wallerstein, N. (2003). *Community-based Participatory Research for Health*. San Francisco, CA: Jossey-Bass.
Morgan, D.L. (1988). *Focus Groups as Qualitative Research*. Thousand Oaks, CA: Sage.

National Health & Medical Research Council (2003). *Values and Ethics: Guidelines for the Ethical Conduct in Aboriginal and Torres Strait Islander Research*. Canberra: Commonwealth of Australia. [Online] Available at: https://www.nhmrc.gov.au/about-us/publications/values-and-ethics-guidelines-ethical-conduct-aboriginal-and-torres-strait-islander-health-research

National Health & Medical Research Council (2007). *National Statement on Ethical Conduct in Human Research.* Canberra: National Statement on Ethical Conduct in Human Research. Available at: https://www.nhmrc.gov.au/about-us/publications/national-statement-ethical-conduct-human-research-2007-updated-2018#block-views-block-file-attachments-content-block-1

Noyes, J. and Popay, J. (2007). Directly observed therapy and tuberculosis: how can a systematic review of qualitative research contribute to improving services? A qualitative meta-synthesis. *Journal of Advanced Nursing*, 57, 227–43.

Nussbaum, M. (1990). *Love's Knowledge: Essays on Philosophy and Literature*. New York: Oxford University Press.

Oakley, A. (1989). Who's afraid of the randomised controlled trial? Some dilemmas of the scientific method and 'good' research practice. *Women's Health*, 15, 25–9.

Olesen, V. (1994). Feminisms and models of qualitative research. In: Denzin, N.K. and Lincoln, Y.S. (eds.) *Handbook of Qualitative Research.* Thousand Oaks, CA: SAGE Publications, pp. 158–174.

Patton, M.Q. (1990). *Qualitative Evaluation and Research Methods.* Newbury Park, CA: SAGE Publications.

Patton MQ. (2014). *Qualitative Research and Evaluation Methods*, London: SAGE Publications.

Pinto, A.D., et al. (2012). Equity in public health standards: a qualitative document analysis of policies from two Canadian provinces. *International Journal for Equity in Health*, 11, 1–10.

Popay, J. (2010). Community Engagement for Health Improvement: questions of definition, outcomes and evaluation. In: Morgan, A., et al. (eds) *Health Assets in a Global Context: Theory, Methods, Action*. New York: Springer, pp. 183–97.

Popay, J., et al. (2003a). Beyond 'beer, fags, egg and chips'? Exploring lay understanding of social inequalities in health. *Sociology of Health & Illness*, 25, 1–23.

Popay, J., et al. (2003b). A proper place to live: health inequalities, agency and the normative dimensions of space. *Social Science & Medicine*, 57, 55–69.

Popay, J. and Williams, G. (2009). Equalizing the people's health: a sociological perspective In: Gabe, J. and Calnan, M. (eds.) *The New Sociology of the Health Service*. London: Routledge, pp. 222–43.

Reason, P. (1994). Three approaches to participative inquiry. In: Denzin, N.K. and Lincoln, Y.S. (eds.) *Handbook of Qualitative Research*. Thousand Oaks, CA: SAGE Publications, pp. 324–39.

Richards, G. (2011). *Psychology, Religion, and the Nature of the Soul*. Kent: Springer Science.

Ritchie, J. and Spencer, L. (1994). Qualitative data analysis for applied policy research. In: Bryman, A. and Burgess, R.G. (eds.) *Analyzing Qualitative Data*. London: Routledge, pp. 173–94.

Roberts, H., et al. (2004). Putting public health evidence into practice: increasing the prevalence of working smoke alarms in disadvantaged inner city housing. *Journal of Epidemiology & Community Health*, 58, 280–5.

Rosella, L.C., et al. (2013). Pandemic H1N1 in Canada and the use of evidence in developing public health policies: a policy analysis. *Social Science and Medicine*, 83, 1–9.

Ruston, A. (2009). Isolation: a threat and means of spatial control. Living with risk in a deprived neighbourhood. *Health, Risk & Society*, 11, 257–68.

Sayer, A. (2011). *Why Things Matter to People: Social Science, Values and Ethical Life.* Cambridge: Cambridge University Press.

Seidman, I. (2005). *Interviewing as Qualitative Research: A Guide for Researchers in Education and the Social Sciences*. New York: Teachers College Press.
Stake, R.E. (1995). *The Art of Case Study Research*. Thousand Oaks, CA: SAGE Publications.
Thomas, B. and Dorling, D. (2011). *Bankrupt Britain: A Post-recession Atlas*. Bristol: Policy Press.
Townsend, B., et al. (2019). How can multiple frames enable action on social determinants? Lessons from Australia's paid parental leave. *Health Promotion International*, Sep 16; daz086. doi: 10.1093/heapro/daz086.
Trahair, E. (2014). Editorial. *Planet: The Welsh Internationalist*, 213, 6–9.
Wadsworth, Y. (1984). *Do It Yourself Social Research*. Melbourne: Victorian Council of Social Services.
Walsh, C.A., Rutherford, G., and Kuzmak, N. (2010). Engaging women who are homeless in community-based research using emerging qualitative data collection techniques. *International Journal of Multiple Research Approaches* 4, 192–205.
Wekesa, E. and Coast, E. (2013). Living with HIV postdiagnosis: a qualitative study of the experiences of Nairobi slum residents. *BMJ Open*, 3, e002399.
Williams, G. (2000). Knowledgeable narratives. *Anthropology and Medicine*, 7, 135–40.
Wright Mills, C. (1970). *The Sociological Imagination*. Harmondsworth: Penguin Books.
Yin, R.K. (2018). *Case Study Research and Applications*. Thousand Oaks, CA: SAGE Publications.

---

[1] https://www.invo.org.uk/posttypepublication/national-standards-for-public-involvement/

5.7

Methodological issues in the design and analysis of cluster randomized trials

Kathy J. Baisley, Richard J. Hayes, and Lawrence H. Moulton

Introduction

Randomized controlled trials are the accepted gold standard for evaluating the effects of interventions to improve health. In the majority of such trials, individuals are randomly allocated to the experimental conditions under study, for example, to treatment and control arms. However, in some situations it is more appropriate to randomly allocate groups of individuals to the treatment arms. These groups are referred to as clusters, and trials of this kind are known as cluster randomized trials (CRTs). Examples of clusters include schools, villages, workplaces, or health facilities, but there are many other possible choices. In some CRTs, all individuals within the selected clusters are automatically included. In others, there may be additional eligibility criteria. Similarly, the impact of the intervention may be measured in all individuals in the cluster, or in a random subsample. This chapter aims to discuss methodological issues that arise in the design and analysis of CRTs. Some of the content has been condensed from the textbook by Hayes and Moulton (2017).

There are a number of reasons for using a cluster randomized design. Some interventions by their nature have to be implemented at the cluster level, such as an intervention to improve water supplies in villages. Others could in principle be implemented at the individual level, but cluster randomization may offer greater logistical convenience, or may be more acceptable to the community. Cluster randomization may help avoid contamination, where individuals in one treatment arm are exposed to or receive the treatment assigned to the other arm. CRTs are also the best design to capture the population-level effects of an intervention applied to a

large proportion of a population. For example, a CRT in Uganda evaluated the effect of intermittent preventive treatment (IPT) for malaria in school children (Staedke et al. 2018). Each cluster included one primary school and 100 surrounding households. The main trial outcome was malaria parasite prevalence in the community, measured in a random sample of households in each cluster.

When considering a cluster randomized design, the advantages need to be weighed against the limitations. The power and precision of a CRT is lower than for an individually randomized trial of the same size, and the logistical aspects of working in several different clusters may make a CRT expensive to implement. Other issues to consider are selection bias, imbalance between study arms, and generalisability, which will be discussed in this chapter.

CRTs have become increasingly more common in health research. In the early years, the special design and analysis requirements of these trials were not always understood, and there were many trials that were designed, analysed, or reported inappropriately. However, over the past two decades, much work has been done on methods for CRTs, and the quality of these trials has improved.

Methodological implications of cluster randomization

The design and analysis of individually randomized trials rely on individual observations being independent of one another. However, when observations are naturally grouped within clusters, this assumption of independence is violated, since observations on individuals in the same cluster are usually correlated. That is, responses among individuals in one cluster will tend to resemble each other more than the responses of individuals in different clusters. This phenomenon is termed within-cluster correlation. It can be equivalently viewed as between-cluster variation, which reflects the natural variation among clusters that exists even in the absence of an intervention effect. Therefore, within-cluster correlation and between-cluster variation can be thought of as corresponding concepts that provide two different perspectives on the same underlying phenomenon.

Within-cluster correlation, or equivalently between-cluster variation, can occur in CRTs for three main reasons. The first reason is the most obvious,

which is that health conditions tend to vary between populations, owing to differences in the individuals that make up each cluster or differences in geographical or environmental characteristics of the cluster. A perhaps less obvious point is that different clusters may respond differently to the intervention, so that there may be no between-cluster variation in the outcome of interest before the intervention, but significant between-cluster variation after the intervention is introduced. The final reason that health outcomes may be clustered is due to interactions between members of the cluster. For example, in interventions against infectious diseases, the effect of an intervention in one individual may affect the risk of contracting the disease in other individuals in the same cluster.

The correlated nature of the data from CRTs requires careful thought at the design stage to ensure sufficient information is collected. As a result of this correlation, CRTs require larger sample sizes than conventional individually randomized trials to be adequately powered to detect treatment effects, and to obtain sufficiently precise estimates of the treatment effect.

Since standard statistical techniques rely on the independence of individual observations, we cannot use these methods to analyse data from a CRT. Using standard statistical techniques will result in p-values that are too small and confidence intervals for the treatment effect that are too narrow. Instead, analysis techniques that account for between-cluster variability are required to analyse CRTs.

Measures of between-cluster variability

There are two main measures of between-cluster variability, corresponding to the two ways of viewing this phenomenon: the coefficient of variation between clusters (k) and the intracluster correlation coefficient (ICC).

Consider a CRT in which a quantitative variable, for example systolic blood pressure, is measured on each individual. The parameter of interest is the mean, μ. Suppose there are c clusters, and that the true value of the mean in the jth cluster is given by μ_j. We assume that the μ_j are randomly sampled from a probability distribution with mean $E(\mu_j) = \mu$ and variance $Var(\mu_j) = \sigma_B^2$, where σ_B^2 represents the between-cluster variance. Then the coefficient of variation is defined as:

$$k = \frac{\sigma_B}{\mu}.$$

If the parameter of interest is a proportion or rate, μ is replaced by π or λ, the true value of the proportion or rate, respectively. It is important to recognize that the μ_j represent the true value of the mean in a notional wider population in each of the c clusters. The μ_j are not directly observable, since we collect data only on a finite sample of individuals in each cluster. However, it is possible to use the observed values of the mean in each cluster to obtain valid estimates of σ_B, μ, and hence *k*, as we describe in a later section.

The ICC, usually denoted by the Greek letter *ρ*, represents how much more similar outcomes are in individuals from the same cluster compared with individuals from different cluster. Another way of describing the ICC is to say that it measures the proportion of the total variance in the outcome that may be attributed to between-cluster variation. The ICC is calculated differently depending on the outcome of interest. The methods for quantitative and binary outcomes are described next. The ICC is not defined for a rate; therefore, *k* is used to estimate between-cluster variability when outcomes are event rates.

For quantitative outcomes, the ICC is defined as:

$$\rho = \frac{\sigma_B^2}{\sigma_B^2 + \sigma_W^2},$$

where σ_B^2 is the between-cluster variance and σ_W^2 is the within-cluster variance.

For binary outcomes, where an individual either does or does not satisfy a certain condition (for example, has hypertension or does not), the objective is usually to compare the proportion of individuals with the outcome. The ICC for binary outcomes is defined as:

$$\rho = \frac{\sigma_B^2}{\pi(1-\pi)},$$

where π is the true proportion for the entire population, and σ_B^2 is the variance of the cluster-specific true proportions.

In general the ICC will fall between zero and one. An ICC of zero implies that there is no clustering, so that individuals within the same cluster are no more similar to each other than to individuals from different clusters, in other words, there is no between-cluster variability. An ICC of one implies that individuals in the same cluster are perfectly correlated, in other words they all have the same outcome, and there is no within-cluster variability.

Design effect and effective sample size

A CRT can be considered equivalent to collecting data from a cluster sample of individuals, rather than a simple random sample as in a conventional individually randomized trial. Cluster sampling provides less precise estimates of parameters of interest than simple random sampling. The increase in variance can be measured by the design effect (DEff) (Kish 1965):

$$\mathrm{DEff} = \frac{\text{variance for cluster sampling}}{\text{variance for simple random sampling}}.$$

Since the variance for a simple random sample is inversely proportional to the sample size, the design effect also gives the effective sample size for the cluster sampling design:

$$\mathit{Effective\,sample\,size} = \frac{\mathit{actual\,sample\,size}}{\mathit{DEff}}.$$

For example, if DEff = 2, this means that the cluster sample will provide the same precision as a simple random sample of half the size.
DEff can be calculated from the ICC as follows:

$$\mathrm{DEff} = 1 + (m-1)\rho.$$

where *m* is the number of individuals per cluster.

From this, we can see that the impact of clustering in a trial depends not only on the value of the ICC, but also on the size of the clusters. DEff may be substantial when we have large clusters, even if the ICC is small. A further discussion of the relationship between DEff, the ICC, and *k*, can be found in Hayes and Moulton (2017, pp. 19–22). A detailed discussion of how to account for the impact of clustering in sample size calculations and analyses is presented in later sections of this chapter.

Experimental and observational units

The key difference between a CRT and an individually randomized trial is that the intervention is randomly allocated to clusters of individuals rather than to individual participants. The cluster therefore constitutes the experimental unit in a CRT. While the cluster is always the experimental unit, observations may be made at different levels so there may be several types of observational unit. In the simplest case, observations may be made only on the individual participants in each cluster. More often, data will also be collected on cluster-level variables and, in some CRTs, there may also be intermediate levels of observation. For example, in a CRT of a school-based health education programme, in which the intervention is randomly allocated to schools, units of observation might include the individual student (e.g. age, sex), class (e.g. grade, size, number of students, number of sessions delivered), and school (e.g. urban/rural, type of school, size of school). Data from all these levels may be used when making inferences about the main parameters of interest.

Design considerations in CRTs

Choice of cluster type and size

One of the first decisions when designing a CRT relates to the choice and definition of the clusters that are to be randomized. In practice, trials have used a wide variety of cluster types, ranging from CRTs of families or households, with only a few individuals in each cluster, to CRTs of large geographical regions in which each cluster may include millions of

individuals. While the general principles of CRT design may be similar for these widely varying situations, the practical aspects of implementing such trials are clearly very different.

Having decided on the type of cluster that will be randomized, the investigator may still have some latitude in choosing the cluster size. This is particularly the case with geographical clusters, where arbitrary boundaries may be drawn. Even when we have institutional clusters, however, there may be some flexibility—for example to group neighbouring schools or health facilities into one cluster.

A number of factors may be considered when deciding on cluster size. The formula for DEff shows that, for a given total sample size, precision is maximized with a cluster size of m = 1, corresponding to individual randomization. Assuming the ICC remains constant, DEff increases with cluster size. This implies that a large number of small clusters is statistically more efficient than a small number of larger clusters. By statistical efficiency, we mean that our estimate of the intervention effect will be more precise. However, in practice the ICC may not remain constant, and generally tends to decrease with increasing cluster size. For example, suppose the choice is between randomizing towns or neighbourhoods. There may be substantial variation in health outcomes between neighbourhoods, owing in part to socioeconomic differences. However, since each town is likely to have a mix of neighbourhoods of higher and lower socioeconomic status, the differences may average out so that there is little variation in the outcome between towns. As a result, the ICC may be larger for neighbourhoods than for towns. Thus it is important to have estimates of the ICC or *k* which relate to the types and sizes of cluster under consideration, when estimating the sample size of a CRT.

Statistical issues are not the only consideration when deciding on cluster size. Although designs using a large number of small clusters are statistically more efficient, it may be logistically and financially more feasible to carry out field work in a smaller number of large clusters. The relative cost and feasibility of designs based on different cluster sizes will depend critically on the type of cluster and geographical setting.

Contamination is another consideration in the choice of cluster size. Contamination can occur through contacts between intervention and control clusters, or through contacts between intervention clusters or control clusters and the wider population. In general, contamination is likely to

decrease as cluster size increases. Rates of travel and migration across area boundaries generally decrease as the size of the area increases. Edge effects are also likely to diminish with the size of the cluster, since the proportion of the population living close to the boundary of an area decreases as the area is increased.

One of the main strategies for reducing contamination in a CRT is to select clusters that are well-separated. This is particularly useful for reducing the level of contact between intervention and control clusters. The separation needed will depend on the intervention under study, for example, much greater geographical separation may be needed for a mass media intervention. Investigators have sometimes specified that the trial clusters should be separated by buffer zones, so that there is little or no common boundary between any two clusters. Additional strategies may be needed to reduce contact between trial clusters and the wider population.

For example, a CRT in rural Burkina Faso evaluated the effect of a comprehensive radio campaign on key family behaviours for improving child survival (Sarrassat et al. 2018). Clusters were 14 distinct geographical areas, each centred around a community radio station, and comprising sufficient villages to provide a total population of around 40,000 per cluster. Clusters were identified using digital terrain maps, on-the-ground mapping of radio signal strength, and surveys to assess radio listenership. Clusters were selected based on their high radio listenership (>60% of women listening to the radio in the past week), and to have little or no overlap of radio signal between areas to minimize contamination.

Baseline imbalance between trial arms

In individually randomized trials with a large number of participants, we can rely on the randomization process to provide study arms that are similar with respect to all potential confounding factors. In CRTs, owing to practical and financial constraints, the number of clusters randomized is often quite small, for example, 20 or 30. In these circumstances, randomization still has the advantages of impartiality, transparency, and justification for statistical inference as it does in individually randomized trials. However, with such a small number of randomization units, it is quite likely that there will be some imbalance on one or more potential confounding factors. While these factors can be adjusted for in the analysis,

the credibility of the trial results may be weakened in the presence of such imbalances; therefore investigators may choose to employ matching, stratification, or restricted (covariate-constrained) randomization to improve baseline balance.

Matching

Matching involves grouping clusters that are similar with respect to certain pre-defined characteristics, then randomizing one cluster in each group to each of the treatment arms. If there are two treatment arms in the CRT, clusters are grouped into matched pairs and one cluster in each pair is randomly allocated to each treatment arm. Matched trials with three treatment arms involve identifying matched triplets and so on. Matching can help to minimize differences between treatment arms with respect to baseline characteristics, and can improve the precision and power of a CRT, since it can reduce between-cluster variability in the outcome of interest.

For example, an HIV prevention trial reported by Grosskurth et al. (1995) was carried out in 12 communities in Mwanza, Tanzania that included rural villages, settlements along a major road, and villages along the shore of Lake Victoria. Baseline HIV prevalence in these communities varied from 1.6 to 8.6%, and the primary outcome (HIV incidence) was expected to be similarly variable. When such high levels of between-cluster variability exist, the coefficient of variation in the outcome is large, which reduces the power of the trial for a given sample size. Therefore, the trial communities were grouped into matched pairs, based on baseline HIV prevalence and geographical characteristics. This helped to ensure that the two arms were similar with respect to these factors at baseline, and it was presumed that they would therefore be more similar with respect to the primary outcome. A matched analysis was employed by comparing the outcome between the intervention and control communities within matched pairs. Since the variance in the outcome within matched pairs may be considerably smaller than the overall variance, replacing the overall coefficient of variation k with the coefficient of variation within matched pairs, k_m, may increase the power of a trial to detect a treatment effect.

However, there are several important disadvantages to matching. Matching is most effective if the matching variables are highly correlated with the outcome variable. In the Mwanza trial, the matching correlation

coefficient was 0.94. However, this level of success in pair-matching is unusual, and indeed in some trials, efficiency may even be lost relative to that of an unmatched design. This is because of the loss of degrees of freedom associated with a matched analysis. For example, to test for an effect of the intervention, a CRT with c matched pairs of clusters is typically analysed using a paired t-test with c − 1 degrees of freedom. This is half the 2(c − 1) degrees of freedom available for an unpaired t-test. The power to show an intervention effect will depend on both the value of the t-statistic and the critical value of the *t* distribution, which in turn depends on the degrees of freedom. It has been shown that below a particular break-even value of the matching correlation coefficient, an unmatched design will have higher power than a matched design, because the loss in degrees of freedom will outweigh the reduction in variance (Martin et al. 1993; Hayes and Moulton 2017, pp. 80–1).

Matching may also present problems if an entire cluster drops out of the trial. Such drop-out also reduces the power and precision of an unmatched CRT and raises questions over selection bias. However, in a matched CRT, the entire matched pair is lost, since the remaining cluster no longer has a matched cluster for comparison.

Another disadvantage of the matched design is that it is not possible to assess whether the intervention effect varies between matched sets. This is because the difference in outcomes between clusters in a matched set depends on both the intervention effect and the intrinsic between-cluster variability within that matched set. It is impossible to separate out these two effects without making additional assumptions. This difficulty results from the lack of replication within each matched set, meaning that there is only one cluster per treatment arm in each matched set. Therefore, it is not possible to calculate the ICC or *k* based on data from a matched trial (unless additional assumptions are made), because the between-cluster variability is confounded with variation in the intervention effect. Although matched trials can still be analysed without an estimate of the ICC or k, the lack of a good estimate of between-cluster variation may be a limitation when investigators are calculating sample sizes for future trials.

Stratification

Stratification is a less rigid method than matching for reducing baseline imbalance between treatment arms and reducing between-cluster variability. Stratification involves grouping clusters with similar characteristics into strata, then randomizing more than one cluster in each stratum to each treatment arm. Thus, unlike matching, stratification achieves replication within strata, which overcomes the main disadvantages of matching. For example, the HIV Treatment as Prevention (TasP) trial in rural South Africa used a stratified design to allocate 34 communities to either the intervention or control arms (Iwuji et al. 2013). The investigators created six strata on the basis of antenatal HIV prevalence. Within each stratum, communities were then randomly allocated to the two trial arms.

Restricted randomization

Although matching and stratification may help reduce imbalances between the treatment arms, they are less useful if balance is required on a large number of variables. In these circumstances, restricted (covariate-constrained) randomization can be used to ensure an acceptable level of overall balance on a number of variables. This method uses baseline or pre-existing data from each cluster to generate a subset of possible allocations that satisfy certain predetermined balance criteria (Raab and Butcher 2001; Moulton 2004). One allocation is then randomly selected from this restricted subset. For example, the ZAMSTAR trial of a tuberculosis control intervention in Zambia and South Africa used both restricted randomization and stratification to randomly allocate 24 clusters to four intervention groups (Sismanidis et al. 2008). Stratification was by country and tuberculous infection prevalence and restricted randomization was by tuberculous infection prevalence, HIV prevalence, urban/rural, social context, and geographical location. A further overview of matching, stratification, and restricted randomization for achieving baseline balance in CRTs is provided by Ivers et al. (2012).

Choice of design strategy

Matching and stratification are designed to achieve baseline balance and reduce between-cluster variability, so thus improve power and precision. In

contrast, restricted randomization is mainly intended to achieve baseline balance.

Matching has been most commonly used for CRTs with a small number of clusters per arm (fewer than 10–15). Unfortunately, while matching is likely to improve balance, it may reduce the power and precision of the trial if the matching is ineffective. In contrast, stratification loses fewer degrees of freedom, so power and precision are more likely to be maximized. In trials with a small number of clusters, care should be taken to limit the number of strata: as a general rule, no more than two strata should be used in trials with ≤6 clusters per arm. Although stratification loses fewer degrees of freedom compared with matching, it is less likely to achieve close balance. In particular, if balance is sought on a large number of variables, the number of strata required would result in a large loss of degrees of freedom. In this situation, restricted randomization would be more effective in achieving balance.

Baseline balance is more likely in trials with larger numbers of clusters. However matching and stratification may still be useful to improve power and precision. In trials with many clusters, the loss of degrees of freedom has very little effect, so either method can be an effective design. In most practical situations, it is likely that stratification (with, say, three or four strata) will capture most of the additional power provided by matching unless there are matching factors that are very closely related to the endpoints of interest. Box 5.7.1 provides a summary of important points for CRT design.

Box 5.7.1 Summary of important points for CRT design

- In CRTs, the intervention is randomly allocated to groups of individuals, rather than individual participants
- Observations in the same cluster are correlated; therefore, CRTs require larger overall sample sizes than individually randomized trials
- The degree to which observations are correlated can be measured by k or the ICC
- The design effect (DEff) measures the impact of clustering on sample size
- CRTs often randomize a small number of clusters, so are susceptible to baseline imbalance
- Matching, stratification, and restricted randomization can be used to help ensure balance
- Matching and stratification may also improve power and precision
- However, if the matching is ineffective, power and precision may be reduced in a matched CRT

Alternative trial designs

Most CRTs involve the randomization of clusters to two parallel treatment arms. A number of alternative CRT designs exist, and may be advantageous in certain situations. However, the analysis and interpretation of trials with these alternative designs are not always straightforward. Investigators would generally be advised to use simpler deigns unless there are convincing reasons for using these more complex designs.

Factorial designs allow the study of the independent effects of two interventions in the same trial. This has the advantage of being cost effective, and conserving sample size. The design most commonly used in CRTs takes a 2 × 2 layout resulting in four treatment arms: one arm receiving the first intervention, another receiving the second intervention, an arm receiving both interventions, and a control arm. For example, the ZAMSTAR trial evaluated enhanced tuberculosis (TB) case-finding (ECF) and a household intervention (HH) including TB screening and preventive treatment (Sismanidis et al. 2008). Clusters were randomly allocated to one of four arms: ECF, HH, both ECF and HH (ECF + HH), or control. The effect of ECF was assessed by combining the ECF and ECF + HH arms and comparing them with the combined HH and control arms. The effect of HH was assessed in a similar manner. This approach is only valid if there is no interaction between the interventions. Where interactions are expected, factorial designs can be used to identify the joint effect of two interventions; however, larger sample sizes will be required.

The crossover design is commonly used in individually randomized trials, and has been adopted for CRTs. Each cluster receives two interventions, one after the other, in an order determined at random (Parienti and Kuss 2007). Delivery of the two interventions is separated by a washout period, to avoid any ‘carry-over’ effects. Since each cluster serves as its own control, the design has potentially greater efficiency than a parallel group CRT. However, its validity depends on the assumption that carry-over effects are absent, that is, that the estimated effects of intervention are independent of the order in which they are assigned. Even with the use of

washout periods, crossover trials are still often subject to carry-over effects, particularly for interventions that aim to effect behaviour change.

Crossover trials are also subject to ‘period’ effects, where participants respond differently in one treatment period for reasons unrelated to the intervention (e.g. because their health status has changed). A review of crossover designs is provided in Arnup et al. (2017).

Stepped wedge trials have received increasing attention in the literature. This design is a modification of the crossover design, and involves all of the clusters beginning the trial in the control arm (Hussey and Hughes 2007). After each successive time period of the trial, one or more clusters commences the intervention and remains in the intervention arm for the rest of the trial, so that at the end of the trial all of the clusters are in the intervention arm. The order in which the clusters commence the intervention is decided randomly to avoid confounding by variables related to the order of introduction. For example, a stepped wedge trial of an intervention to reduce the risk of stillbirth was conducted in 33 maternity hospitals in the United Kingdom and Ireland (Norman et al. 2018). Hospitals were randomly allocated to one of nine intervention starting dates, approximately 3 months apart. At baseline, none of the hospitals received the intervention. Every 3 months, the intervention was implemented in the hospitals allocated to that starting date, until all of the hospitals had received the intervention. Although the stepped wedge design is becoming more widespread and there have been recent advances in the methodological literature, it is still viewed as a relatively new design with gaps in knowledge remaining around their uses, ethics, conduct and analysis (e.g. Hargreaves et al. 2015; Thompson et al. 2017).

An important disadvantage of the stepped wedge design is that there is a systematic difference in calendar time of observation between the two treatment arms, so the design is susceptible to bias introduced through secular trends in the outcome. Stepped wedge trials often take place over a period of years, increasing the opportunity for a gradual overall improvement or deterioration in outcomes. Furthermore, it may be difficult to ascribe a particular outcome to the effect of the control or intervention, if the outcome takes a long time to occur. Similarly, there may be difficulties in interpretation when the intervention effect waxes or wanes. Further discussion of the use of stepped wedge designs in CRTs can be found in Rhoda et al. (2011), Copas et al. (2015), and Hemming et al. (2015).

Approaches for impact evaluation

The impact of an intervention in a CRT can be evaluated either through repeated cross-sectional surveys or by following up a cohort of individuals over time. With a cross-sectional survey design, surveys are conducted at different points throughout the trial; either the entire cluster is surveyed or a random sample may be selected. Individuals who enter the cluster during the trial may be included in surveys at later time points. For example, in the CRT of a radio campaign in Burkina Faso mentioned earlier (Sarrassat et al. 2018), trial outcomes were assessed in cross-sectional surveys at 20 months and at 32 months after the trial start. In each survey, a random sample of around 360 mothers per cluster were interviewed regarding their demographic characteristics, radio listenership, and family behaviours of relevance to child survival. Cross-sectional surveys are useful for assessing the effect of an intervention on the prevalence of a binary outcome, such as hypertension prevalence, or on a quantitative outcome, such as serum cholesterol. However they are less useful for measuring incidence rates of an event of interest.

An alternative approach is to enrol a cohort of individuals who are followed over time. Either the entire cluster, or a random sample of each cluster, may be enrolled. If the entire cluster is enrolled, the investigator must specify whether new people entering the cluster during the trial will be included (referred to as an open or ‘dynamic’ cohort) or if the cohort will be limited to those seen at baseline (a closed cohort). Cohort designs are particularly useful for measuring event rates. They also allow direct measurement of change over time, such as a change in serum cholesterol or a change in disease incidence over time. For example, a CRT in rural India of a mobile phone-based system and nurse-managed care for hypertension and diabetes randomly allocated 40 community health clinics to the intervention or control arms (Jha et al. 2017). Individuals with hypertension and/or type 2 diabetes were invited to participate in the trial. A total of 90 individuals were recruited at each clinic and followed for 12 months. The primary outcome was change in systolic blood pressure and glycated haemoglobin (HbA1c) over 1 year. However, cohort designs are subject to selection bias owing to loss to follow-up, are generally more logistically complex, and may not be as generalizable since they represent the effect of the intervention in the subgroup of the population who remain in the cohort

throughout follow-up. In a closed cohort, the disease incidence may decrease over time as the susceptible members acquire the disease early on and only the less susceptible members remain at risk. The choice of design will also depend in part on the primary questions of scientific interest. If interest focuses on change at the broader community level, cross-sectional designs may be the more natural choice, while cohort designs may be more appropriate if focus is on change at the individual level. Further discussion may be found in Feldman and McKinlay (1994).

Sample size determination for CRTs

Statistical approaches to estimating sample size

The formal assessment of statistical power is a fundamental step in the design of any intervention trial. The consequences of trials that are too small for their purpose have been widely discussed in the literature. CRTs are often more costly and logistically complex than individually randomized trials, placing even more responsibility on researchers to ensure that trials are designed with sufficient power to achieve their main objectives. A fundamental principle of randomized trials is that sufficient replication is required to draw reliable conclusions. The correlated nature of the data from CRTs means that the information gained is less than that in an individually randomized trial of the same size. Even if each cluster contains a large number of individuals, the trial may be insufficiently powered if there are too few clusters. We need several clusters in each treatment arm to draw any useful conclusions about the effect of an intervention. For example, imagine an extreme case in which just two clusters are randomly allocated, one to each treatment arm. This is equivalent to a clinical trial in which just two patients are randomly allocated to the two treatment arms. Because there is no replication within each treatment arm, we have no way of measuring and taking account of between-cluster variation, and so no statistical inference is possible. Only by randomizing sufficient clusters to the different treatment arms can we obtain an accurate estimate of average response in each arm, and of differences between arms.

Sample size formulae for CRTs are easily available in journals (e.g. Hemming et al. 2011; Rutterford et al. 2015) and textbooks (e.g. Donner and Klar 2000; Eldridge and Kerry 2012; Hayes and Moulton 2017). In

their simplest form, if we have an estimate of DEff, sample sizes for CRTs can be estimated by calculating a sample size for an individually randomized trial then multiplying this by the DEff (Donner et al. 1981).

Next we give a brief overview of formulae for a standard unmatched CRT design. The formulae assume that the cluster size is fixed, and that we need to calculate the number of clusters required per arm. This is the most useful formulation in CRTs where the sizes of the clusters are defined in advance and the outcome is measured in all individuals in each cluster. If instead the number of clusters is fixed, for example a health service intervention in a fixed number of clinics, the formulae can be rearranged to calculate the required cluster size.

Binary outcomes are usually analysed by comparing the proportions with the outcome in the two arms of the trial. The number of clusters, c, required in each arm can be calculated as:

$$c = 1 + (z_{\alpha/2} + z_{\beta})^2 \frac{\pi_0(1-\pi_0)/m + \pi_1(1-\pi_1)/m + (k_0^2\pi_0^2 + k_1^2\pi_1^2)}{(\pi_0 - \pi_1)^2},$$

where $z_{\alpha/2}$ and z_{β} are standard normal distribution values corresponding to the upper tail probabilities of $\alpha/2$ and β, respectively, k_0 and k_1 are the coefficients of variation in the control and intervention arms, π_0 and π_1 are the true proportions with the outcome in the two arms, and m is the number of individuals per cluster.

If we can assume that the intervention has a similar proportional effect across clusters, then the coefficient of variation for the two arms will be similar and the equation can be simplified to:

$$c = 1 + (z_{\alpha/2} + z_{\beta})^2 \frac{\pi_0(1-\pi_0)/m + \pi_1(1-\pi_1)/m + k^2(\pi_0^2 + \pi_1^2)}{(\pi_0 - \pi_1)^2},$$

where *k* is the common value of the coefficient of variation in the two arms. An important practical point is that, if the expected proportion with the outcome is above 0.50, it is usually better to work in terms of $1-\pi$, rather than π, when calculating the sample size.

We illustrate the calculation of sample size for a CRT using the HIV TasP trial described earlier as an example (Iwuji et al. 2013). The trial's primary outcome was the impact of the intervention on the cumulative incidence of HIV over 24 months in adults aged ≥16 years who were HIV negative at baseline. Clusters comprised contiguous neighbourhoods in a demographic surveillance area (DSA). The study area was mapped before the trial; average cluster size was 1,000 adults. Based on previous HIV surveys in the DSA, HIV prevalence was estimated to be 20%. Allowing for 20% non-participation and loss to follow-up, the estimated mean number of HIV negative adults to be followed over 2 years in each cluster was m = 1,000 × 0.80 × 0.80 = 640.

HIV cumulative incidence over two years in the control communities was estimated to be 5% (π_0), based on surveys in the DSA. On the basis of mathematical modelling, the intervention was expected to reduce cumulative incidence by 30% (i.e. from 5% to 3.5% (π_1)). A value of k = 0.25 was used, based on other studies in the region with HIV incidence as an outcome. The proportional effect of the intervention was assumed to be roughly constant over clusters so that $k_0 \approx k_1 \approx 0.25$. To have 80% power ($z_\beta = 0.84$) to detect a significant difference (2-sided $p < 0.05$, $z_{\alpha/2} = 1.96$), the number of clusters that are needed per arm is:

$$c = 1 + (1.96 + 0.84)^2 \frac{[0.05(1-0.05)]/640 + [0.035(1-0.035)]/640 + 0.25^2(0.05^2 + 0.035^2)}{(0.05-0.035)^2} = 13.5.$$

We would therefore select 28 clusters and randomly allocate 14 to each arm. The investigators decided to randomize 34 clusters (17 per arm), which would give them 88% power to detect an intervention effect.

For trials in which outcomes are expressed as rates per person-year (e.g. incidence rates or mortality rates), the required number of clusters per arm can be calculated as:

$$c = 1 + (z_{\alpha/2} + z_\beta)^2 \frac{(\lambda_0 + \lambda_1)/y + (k_0^2\lambda_0^2 + k_1^2\lambda_1^2)}{(\lambda_0 - \lambda_1)^2},$$

where y is the number of person-years in each cluster, and λ_0 and λ_0 are the true rates in the control and intervention arms, respectively. If the intervention effect is assumed to be similar across all clusters, the equation can be simplified by using a common value of k as above.

For CRTs with quantitative outcomes (e.g. cholesterol levels), the required sample size can be calculated as:

$$c = 1 + (z_{\alpha/2} + z_{\beta})^2 \frac{(\sigma_0^2 + \sigma_1^2)/m + (k_0^2\mu_0^2 + k_1^2\mu_1^2)}{(\mu_0 - \mu_1)^2},$$

where μ_0 and μ_1 are the true means in the control and intervention arms, and σ_0 and σ_1 are the within-cluster standard deviations. Again, the equation can be simplified assuming a common value of k.

Simple extensions of these formulae can be used for matched trials, using the coefficient of variation of the outcome between clusters within matched pairs, k_m, and adding an extra cluster to each treatment arm, to make up for the loss of degrees of freedom when using the paired t-test (Hayes and Moulton 2017, pp.141–2). Sample sizes for stratified trials can be calculated using the same equations as for matched trials. The value of k_m for stratified trials will generally lie between k_m for matched trials and the k for the unmatched design.

Alternative formulae are available for binary and quantitative outcomes based on the ICC (Donner et al. 1981; Donner and Klar 2000). Application of these formulae generally provides very similar results to those based on the coefficient of variation.

Our discussion so far has assumed that the number of individuals or person-years in each cluster is the same. However, in many CRTs, the cluster size may be variable, for example, because villages vary in size, or individuals in some clusters may be lost to follow-up. If the variability in cluster size is large, this should be taken into account in the sample size calculations to ensure that the power is adequate. Methods to account for variable cluster size are reviewed in Rutterford et al. (2015). Sample size calculations for more complex designs are discussed in Hayes and Moulton (2017), Campbell et al. (2007), and Rutterford et al. (2015). Sample size

determination for stepped wedge trials is discussed in Baio et al. (2015), Hemming and Taljaard (2016), and Thompson et al. (2017).

In practice sample size calculations are rarely performed by hand. Statistical packages such as R and Stata have commands to perform sample size calculations that take account of between-cluster variability. Sample size estimation by simulation is an alternative to using an analytical formula, particularly for more complex designs (Reich et al. 2012).

Obtaining estimates of *k* or ICC

Sample size calculations for CRTs require an estimate of between-cluster variation in the outcome of interest. Recall that the coefficient of variation in the i^{th} treatment arm, k_i, is defined as σ_{Bi} $/\lambda_i$, σ_{Bi} $/\pi_i$, or σ_{Bi} $/\mu_i$ for outcomes measured as event rates, proportions, and means, respectively. The values of λ_i, π_i, or μ_i need to be specified in any case, as these are intrinsic to any sample size calculation. In addition, we need an estimate of the between-cluster standard deviation σ_{Bi} in order to obtain a value for k_i.

In many cases, empirical data from which σ_{Bi} can be estimated are not available at the time when a CRT is being designed. In these cases, sample sizes can be calculated using different plausible assumptions regarding *k*. As a rough guideline, experience from field trials suggests that *k* is often ≤0.25, and seldom exceeds 0.5 for most health outcomes. However, values of *k* in excess of 0.5 can occur, especially if the distribution of cluster values is skewed.

Sometimes researchers may have data on between-cluster variation either from previous studies or a baseline survey. A more accurate estimate of σ_B^2 can then be obtained from the empirical variance s^2 of cluster-specific results. The parameter σ_B^2 relates to the true cluster-level values, whereas s^2 relates to the observed cluster-level values and therefore incorporates random sampling error. An estimate of this random sampling error needs to

be subtracted from s^2 to get an estimate of σ_B^2. A detailed explanation of the approach is provided in Hayes and Moulton (2017, pp. 145–7).

As in unmatched trials, there are often no data available on k_m in matched and stratified trials and so a range of plausible values must be used to estimate the required sample size. Occasionally, empirical data are available, and data on the endpoint of interest from clusters grouped according to the matching or stratification variables can be used to estimate k_m.

One advantage of calculating sample sizes based on ρ is that estimates of the ICC have been published for a range of outcome variables and settings, and it is now fairly common for investigators to publish estimates of the ICC when reporting trial results. However, as with k, the ICC will depend on both the study outcome and the setting, so investigators should use caution when using published values to plan studies in different populations or with different outcomes. A range of likely values of the ICC (or k) should be explored, to ascertain the sensitivity of the sample size to these values.

Other considerations for power

CRTs may be particularly vulnerable to problems of low power given the nature of the interventions evaluated and the heterogeneity of the populations in which they are applied. For example, many interventions are delivered to large communities with relatively little attention given to individual study participants. Thus the risk of non-compliance and loss to follow-up may be greater than in individually randomized trials.

Since the power of a given CRT will depend on the degree of between-cluster variation, one strategy for increasing precision is to use relatively narrow cluster-level eligibility criteria. For example, in the Mwanza HIV prevention trial described previously (Grosskurth et al. 1995), clusters were deliberately selected to be similar within pairs in order to increase statistical power.

For trials where there is some latitude in how the primary outcome variable is defined, quantitative endpoints (e.g. cholesterol level) usually lead to more statistical power than their binary counterparts (e.g. high

cholesterol or not), although issues of measurement error and ease of interpretability must also be considered.

As discussed, the statistical efficiency of a CRT is increased if we use a larger number of small clusters rather than a few large clusters. As the value of *k* increases, the number of clusters required increases substantially. The effect of increasing the sample size per cluster is subject to diminishing returns, particularly for large values of *k*. A useful rule of thumb in terms of the ICC is that very little increase in power will be obtained for a fixed number of clusters after the cluster size exceeds the value 1/ρ (Donner and Klar 2004). For example if ρ = 0.01, then the increased power obtained by enrolling more than 100 participants per cluster will be small. However, researchers will often have practical reasons for enrolling large numbers of participants per cluster, even if not needed from a statistical perspective.

Plans to adjust for baseline covariates can lead to reduced sample size requirements. Teerenstra et al. (2012) present a simple formula for quantitative outcomes, that provides a reduction in the sample size depending on the expected correlation between the baseline and follow-up values of the outcome. However, reducing the sample size may not always be advisable, since the degree of correlation may change between the start and end of the trial.

Once the appropriate sample size has been decided, there remains the practical problem of ensuring that recruitment targets can actually be met. Feasibility studies are essential, particularly for CRTs that involve complex intervention components. Box 5.7.2 provides a summary of important points for sample size considerations.

Box 5.7.2 Summary of important points for CRT sample size considerations

- Sample size calculations for CRTs require an estimate of the correlation between observations
- Data from previous studies may be used to estimate k or ICC; however, the value may vary in different settings so a range of likely values should be explored
- In general, power and precision of a CRT will be higher with a large number of small clusters, than with a small number of large clusters
- CRTs tend to have diminishing returns in terms of power and precision as cluster size increases

Analysis strategies in CRTs

A wide range of analytical methods have been proposed for the analysis of CRTs. The aim of this section is not to provide a comprehensive review of all such methods but instead to focus on a limited number of methods that are simple to understand, can be easily applied, and have proven to be robust and efficient. A detailed description of these methods, including examples of all procedures discussed in this section, is given by Hayes and Moulton (2017).

Cluster-level analysis

There are two main approaches to the analysis of data from CRTs: analysis of cluster-level summaries, and analysis of individual-level observations. The cluster-level approach is conceptually very simple, and performs well under a wide range of conditions. It is essentially a two-stage process. In the first stage, a summary measure is obtained for each cluster, based on data collected on the outcome of interest among individuals in that cluster. Thus, the individual observations are reduced to 2c numbers, if there are c clusters in each of two study arms. At the second stage, we compare these two sets of cluster-specific measures using an appropriate statistical method. The most common approach is to use a simple two-sample t-test. Sometimes we start by applying a logarithmic transformation to the cluster-level summaries, especially if these have a skewed distribution. The *t* distribution can be used both to carry out a significance test of the null hypothesis of no intervention effect, and to obtain a confidence interval for the intervention effect. The methods based on the t-test have been shown to be highly robust to departures from the underlying assumptions.

A common objection to cluster-level analysis is that information is wasted when the data are reduced to a set of summary measures. Investigators note that, in a trial with 10 clusters per arm but following many thousands of individuals, the final analysis is based on two columns of 10 numbers. This objection, however, is not valid. The power of the study and the precision of the effect estimates will depend on the observed variability in the outcome between clusters. The observed between-cluster

variance will be reduced if a large number of individuals are studied in each cluster.

As an example, we illustrate the cluster-level approach by considering the comparison of a binary outcome in an unmatched CRT with two arms, although similar equations can be used for event rates or quantitative outcomes. The first step is to calculate the proportion of individuals with the outcome in each cluster (the cluster-level summaries). Suppose that p_{ij} is the observed proportion in the jth cluster of the ith treatment arm (i = 1: intervention; i = 0: control). Then we can estimate the mean proportion in each treatment arm as

$$\bar{P}_i = \frac{\sum p_{ij}}{c_i},$$

where c_i is the number of clusters in the i^{th} treatment arm.

When working with proportions, we are generally interested in the risk difference or risk ratio as a measure of the intervention effect. We can use the observed mean proportions to obtain a *point estimate* of the effect:

Risk difference: $\bar{P}_1 - \bar{P}_0$

Risk ratio: $\frac{\bar{P}_1}{\bar{P}_0}$

We can then use the unpaired t-test to test the null hypothesis of no intervention effect:

$$t = \frac{\bar{P}_1 - \bar{P}_0}{s\sqrt{\frac{1}{c_1} + \frac{1}{c_0}}},$$

where s^2 is the *pooled estimate* of the variance of observed cluster-specific risks (i.e. proportions) within treatment arms, computed as

$$s^2 = \frac{\sum_{ij} (p_{ij} - \bar{P}_i)^2}{c_1 + c_0 - 2},$$

The resulting test statistic is compared with the *t* distribution with ($c_1 + c_0 - 2$) degrees of freedom.

The validity of the t-test depends on the observed cluster proportions being normally distributed. However, simulations have shown that the t-test is remarkably robust to deviations from normality, even for quite small numbers of clusters (e.g. Gail et al. 1996; Bennett et al. 2002). For CRTs with a very small number of clusters, a more conservative option is to use non-parametric methods such as Wilcoxon's rank sum test or a permutation test (Eldridge and Kerry 2012; Hayes and Moulton 2017).

Application of the t-test to cluster-level summaries is robust, in the sense that it provides reliable p-values and confidence intervals with approximately correct coverage. However, it may not be fully efficient in terms of providing optimal power and precision. The t-test described just now gives equal weight to each cluster, even though some clusters may be much larger than others and thus provide more information on the outcome of interest. It is theoretically possible to use a weighted t-test which weights clusters according to the information they provide. However, calculation of the weights depends on having an accurate estimate of the between-cluster variance. If the weights are not estimated accurately, the precision of the weighted analysis may be worse than for an unweighted analysis. Therefore, we would not generally recommend use of the weighted t-test.

In addition to testing the evidence against the null hypothesis, it is important to compute a confidence interval which represents the precision in our estimate of the intervention effect. We can obtain a 95% confidence interval for the true risk difference as:

$$(\bar{p}_1 - \bar{p}_0) \pm t_{v,0.025} \times s\sqrt{\frac{1}{c_1} + \frac{1}{c_0}},$$

where $t_{v,0.025}$ is the upper 2.5% point of the *t* distribution with $v = c_1 + c_0 - 2$ degrees of freedom.

Note that this formula assumes equal variances for the risks in the intervention and control arms. While this assumption is reasonable under the null hypothesis of no intervention effect, it may not be when there is a non-null effect. A robust alternative is to compute the variance of the risk

difference based on separate estimates of the variances in the two treatment arms.

We can also obtain a 95% confidence interval for the risk ratio computed based on cluster-level summaries. Since ratio estimates of this kind tend to have skewed distributions, it is usual to work instead in terms of the logarithm of the risk ratio. The variance of the natural logarithm ln(risk ratio) can be estimated by:

$$V = \frac{s_1^2}{c_1 \bar{p}_1^2} + \frac{s_0^2}{c_0 \bar{p}_0^2}.$$

Then a 95% confidence interval for the logarithm of the risk ratio can be obtained as

$$\log(\text{risk ratio}) \pm t_{v,0.025} \times \sqrt{V}.$$

The confidence interval for the risk ratio can be obtained by dividing and multiplying the estimated risk ratio by the error factor (EF):

$$EF = \exp\left(t_{v,0.025} \times \sqrt{V} \right).$$

We illustrate the cluster-level approach with data from the MEMA kwa Vijana (MkV) trial in Tanzania. In this CRT, 20 rural communities were randomly allocated to an intervention arm receiving an integrated adolescent sexual health programme or a control arm (Ross et al. 2007). The final impact of the intervention was evaluated in a cross-sectional survey of young people aged 15–30 years who had attended the trial schools (Doyle et al. 2010). A range of outcomes were examined, but we shall focus on the effect of the intervention on sexual attitudes. Participants were classed as having a desirable attitude based on their answers to three questions on sexual behaviour and gender norms. The trial used a stratified design to allocate communities to each arm; however, for the purpose of this example, we will ignore the stratification in the analysis.

Table 5.7.1 shows the number of adolescents surveyed, and the number and proportion with a desirable attitude in each cluster. The mean

proportion in the control arm $\bar{p}_0$ is 0.161, and in the intervention arm $\bar{p}_1$ is 0.201. Thus the risk difference is 0.040, and risk ratio is 1.246.

Table 5.7.1 Proportion of participants with desirable sexual attitudes at the final survey in the 20 clusters of the MEMA kwa Vijana trial

Intervention clusters			Control clusters		
N with desirable attitude	Total N	Proportion	N with desirable attitude	Total N	Proportion
102	557	0.183	88	568	0.155
119	659	0.181	128	640	0.200
153	519	0.295	49	495	0.099
147	705	0.209	71	569	0.125
123	567	0.217	116	518	0.224
216	786	0.275	165	646	0.255
123	544	0.226	77	756	0.102
95	827	0.115	144	626	0.230
92	603	0.153	69	614	0.112
109	706	0.154	69	639	0.108
Mean		0.201			0.161
SD		0.0556			0.0606

To test the null hypothesis of no intervention effect, we carry out an unpaired t-test. First we compute the pooled variance, using the aforementioned equation. Alternatively, we can take the weighted average of the variance in each treatment arm shown in Table 5.7.1:

$$s^2 = \frac{(c_1 - 1)s_1^2 + (c_2 - 1)s_0^2}{c_1 + c_0 - 2} = \frac{(9 \times 0.0556^2) + (9 \times 0.0606^2)}{10 + 10 - 2} = 0.00338,$$

giving an estimated pooled standard deviation of s = 0.0582. Then we compute the t-test as:

$$t = \frac{0.201 - 0.161}{0.0582\sqrt{\frac{1}{10} + \frac{1}{10}}} = \frac{0.040}{0.0260} = 1.524.$$

Referring this value to tables of the t distribution with ν = 18 degrees of freedom, we obtain a two-sided p-value of 0.14. Thus, while there is a trend towards more desirable attitudes in the intervention arm, this difference may have occurred by chance. A 95% confidence interval for the risk difference is obtained as 0.040 ± 2.101 × 0.0260, or (− 0.0147, 0.0943), where 2.101 is the upper 2.5% point of the t distribution with 18 degrees of freedom.

To obtain the 95% confidence interval for the risk ratio, we first estimate the variance of the log(risk ratio) as

$$V = \frac{0.0556^2}{10 \times 0.201^2} + \frac{0.0606^2}{10 \times 0.161^2} = 0.0218.$$

We can then calculate the error factor as $EF = \exp(2.101 \times \sqrt{0.0218}) = 1.364$ and divide and multiply the risk ratio by the EF to obtain the 95% confidence interval as 1.246/1.364 and 1.246 × 1.364, or (0.914, 1.700).

Using the logarithmic transformation

When using cluster-level summaries to obtain estimates of the intervention effect, it is advisable to inspect the distribution of the summary measures within each treatment arm. These distributions are sometimes markedly skewed, particularly for proportions and rates. Since the statistical inference methods used assume the cluster-level summaries are normally distributed, a logarithmic transformation may be useful before analysis, if the cluster-level summaries are positively skewed. For proportions, this would involve computing the logarithm of the risk for each cluster as $l_{ij} = \log(p_{ij})$, then working with the l_{ij} which are likely to be more symmetrically distributed. It is common in statistical practice to use the natural logarithm (ln),

although the choice of logarithmic base has no effect on the estimates of the intervention effect.

We can then calculate the difference in the mean of the cluster-specific log risks ($\bar{l}_1 - \bar{l}_0$), and apply the t-test to the log-transformed values. The log transformation may help to ensure that the variances of the cluster summaries in the two treatment arms are similar as well as providing a more symmetric distribution of cluster summaries, thus improving the validity of the t-test. We can also calculate a 95% confidence interval for the difference in log risks ($\bar{l}_1 - \bar{l}_0$), and exponentiate to get these estimates on the ratio scale.

Adjusting for covariates

In an individually randomized trial, we can expect that randomization will achieve good baseline balance between the trial arms, and the primary analysis is generally an unadjusted analysis. However, as we have seen, CRTs often randomize a relatively small number of clusters (≤30), so substantial baseline imbalance between arms is not unusual. Thus adjustment is much more important in this context and often the primary analysis would be adjusted. Furthermore, adjustment may also reduce k, by accounting for some of the variation between clusters, so may increase power and precision.

There is limited guidance for choosing which covariates to adjust for in the analysis of CRTs. Ideally they should be covariates considered a priori to be important risk factors for the outcome of interest. These covariates must be specified in advance in a statistical analysis plan, to protect against allegations of data dredging or selective reporting. Crucially, the decision to adjust for baseline variables should not be based on the statistical significance of the imbalance between arms. Instead, the decision should depend on how strong a risk factor a covariate is thought to be, the magnitude of any imbalance, how many variables show an imbalance and the number of clusters randomized. For example, in the HIV TasP trial in South Africa described earlier (Iwuji et al. 2013), the analysis was adjusted for temporal changes in national ART guidelines, and cluster-level measures of sex (proportion female), age (proportion <30 years and >=60 years) estimated HIV prevalence, and estimated population ART coverage

at baseline, all characteristics known to be associated with HIV transmission risk in the population (Iwuji et al. 2018).

Adjustment for covariates in a cluster-level analysis can be done using a two-stage procedure. In the first stage, we carry out a standard individual-level regression analysis of the outcome, including all the covariates of interest, *except for the intervention effect*, in the regression model. At this stage, we ignore the effects of clustering, and use standard logistic regression for binary outcomes, Poisson regression for rates, and linear regression for quantitative outcomes. The regression model is used to obtain the expected number of events (for binary outcomes or rates) or the expected mean in each cluster, ignoring the intervention effect. After fitting the regression model, the expected and observed values are compared and used to calculate a residual for each cluster. For example, if we are interested in estimating the risk difference, we calculate the difference-residual as:

$$R_{Dij} = \frac{d_{ij} - e_{ij}}{m_{ij}},$$

Where d_{ij} is the observed number of events in the *jth* cluster of the *i*th treatment arm, e_{ij} is the expected number of events from the regression model, and m_{ij} is the number of individuals.

If the rate ratio is our parameter of interest, then the ratio-residual is calculated as

$$R_{Rij} = \frac{d_{ij}}{e_{ij}},$$

In the second stage of the analysis, the residuals take the place of the cluster-level summaries when calculating point estimates and performing statistical tests, using the methods described in the previous section. If the intervention has no effect, then the average of the residuals will be zero (for difference-residuals) or 1 (for ratio-residuals), whereas an intervention effect will result in systematic differences in the residuals between the two arms.

Individual-level analysis

As we have discussed, the unpaired t-test is based on the assumption that the cluster summary measures are normally distributed and have equal variance. If cluster sizes vary, this latter assumption is clearly invalid, and greater power and precision would be achieved if the analysis were weighted to take into account the different variances. Such weighting is incorporated automatically when we carry out regression analyses based on individual-level data. It is essential that the regression method takes proper account of between-cluster variation. A wide range of regression methods have been suggested for this purpose, but we will focus on two approaches that have received the most attention and which have advantages over other methods: random effects regression models and generalized estimating equations (GEE).

An advantage of individual-level regression methods over the simpler cluster-level methods is that, in an adjusted analysis, the effects of the covariates are estimated and presented simultaneously with the intervention effects. This is not the case in a cluster-level analysis, since we are comparing cluster-level residuals from which the effects of other covariates have already been removed.

An important disadvantage of individual-level regression methods is that they do not appear to perform robustly when there are a relatively small number of clusters. Formally, both the individual-level and cluster-level methods are valid asymptotically, that is when there are a very large number of clusters per treatment arm. However, the cluster-level methods seem to perform much more robustly for small numbers of clusters.

Random effects models

In random effects models, we take account of between-cluster variation by assuming that there are cluster-level effects (over and above the random sampling variation within each cluster) which follow a specified probability distribution. Random effect models take the general form:

$$\theta_{ijk} = \alpha + \beta_i + \sum_j \gamma_j z_{ijkl} + u_{ij}$$

In this equation, θ_{ijk} is a parameter related to the outcome of interest and refers to the k^{th} individual in the j^{th} cluster in the i^{th} treatment arm. For event rate data, for example, θ might be the logarithm of the rate in that individual. The parameter β_i represents the intervention effect, the γ_l parameters represent the effects of a set of covariates, z_1, z_2, … z_L, each of which may be measured at individual or cluster level, and α is an intercept term representing the expected outcome in individuals in the control arm (I = 0) and with zero values for all the covariates.

The term u_{ij} is a random effect relating to the j^{th} cluster in the i^{th} treatment arm, and takes account of between-cluster variation. Therefore, models of this form are known as random effects models.

The random effects are assumed to be sampled from a probability distribution. To fit a random effects model, we need to make an assumption about the distribution of these effects, which will depend on the outcome of interest. Random effects models represent full probability models for the data and so likelihood inference is used to estimate the parameters associated with the random effects as well as the effects of the intervention and the covariates. These models can be readily fitted using standard statistical software such as R, Stata, or SAS.

We can use random effects Poisson regression for event rates, such as disease incidence, and random effects logistic regression for binary outcomes. For quantitative outcomes, for example systolic blood pressure, these models are known as mixed effect linear regression models, because they have both fixed effects, indicated by the parameters α, β_i and γ_l, as well as the random effects represented by u_{ij}.

Random effect models perform well for event rate data and quantitative (normally distributed) outcomes, since the assumptions that are made about the distribution of the random effects and the variation in the outcome within clusters mean that the likelihood can be written down in analytical form. However, the likelihood for the random effects logistic regression model assumes a mixture of normal and binomial distributions, and the combination of these distributions does not provide a simple composite distribution. Therefore, numerical approximations have to be used to maximize the likelihood, and to obtain parameter estimates, confidence intervals, and significance tests. The numerical procedure used for this purpose is known as quadrature and involves approximating the normal

distribution. With older versions of statistical software, the parameter estimates were not always reliable and were dependent on the number of quadrature points. However, recent advances in statistical software have resulted in significant improvements in the procedure of quadrature, so quadrature problems are rarely encountered.

Generalized estimating equations

GEE provide an alternative method of individual-level regression modelling and make allowance for clustering without incorporating additional terms in the model (Zeger and Liang 1986). They do this by assuming that observations in the same cluster are correlated and using the data to estimate the strength of the correlation.

While GEE models do make allowance for within-cluster correlation, they do not correspond to a full probability model for the data. For this reason, we cannot define the likelihood of the data and so we cannot obtain maximum likelihood estimators of the parameters or use likelihood ratio tests for hypothesis testing. Instead, GEE uses an iterative generalized least squares method to estimate the parameters together with robust standard errors. The Wald test can be used for hypothesis testing. As with random effects models, GEE only works reliably when there are a sufficiently large number of clusters in each treatment arm.

For the analysis of CRTs, the usual assumption in a GEE model is that observations on individuals in different clusters are uncorrelated, while observations on individuals in the same cluster all have the same correlation coefficient ρ. In this situation an exchangeable correlation matrix is specified in the statistical model, although the method is asymptotically robust to this specification. The observed data are used to estimate the value of ρ, and this is used to obtain estimates of the regression coefficients and their standard errors that are adjusted for the correlations in the data. If ρ is large, less weight will be given to each individual in large clusters, since the outcomes within a cluster will tend to be similar. Conversely, if ρ is close to zero, this means that there is little clustering and all individuals will be given similar weight in the analysis. An excellent non-technical discussion of GEE is given in Hanley et al. (2003).

GEE models may be fitted for event rates, binary outcomes, and quantitative outcomes. GEE models are more robust than random effects

models, since they do not rely on assumptions about the distribution of the between-cluster variation. Although GEE and random effects models estimate the same parameter for the intervention effect when the outcome variable is normally distributed or an event rate, this equivalence disappears in the case of a binary outcome, and thus their interpretation is different. In a GEE logistic regression model, we are estimating a population-averaged odds ratio of the outcome, whereas with random effects models we are estimating a cluster-specific odds ratio. The cluster-specific odds ratio represents the effect on the odds of the outcome if the intervention is applied to any given cluster, whereas the population odds ratio represents the effect on the odds in the combined population of all the clusters. In many circumstances, there will be little difference between the cluster-specific and population-averaged estimates. The difference will only be large if there is substantial between-cluster variation and, in at least a proportion of clusters, the outcome is common.

Choosing an appropriate analysis

Individual-level regression methods are popular for analysing CRTs but require a sufficient number of clusters to ensure unbiased estimates. Based on a review of simulation studies, McNeish and Stapleton (2016a) recommended that there are at least 15 clusters (8 per arm) for quantitative outcomes and 30 (15 per arm) for binary outcomes, in order to ensure unbiased estimates of the treatment effect and its standard error with random effects models. The distributional assumptions of random effects models can also be difficult to verify without large numbers of clusters.

GEE models have been shown to have inflated type I error when there are few clusters per arm (Bellamy et al. 2000). That is, we are more likely to reject the null hypothesis, when there is no true underlying effect of the intervention. In contrast, analyses of cluster-level summaries based on the t-test provide reliable p-values, and confidence intervals with approximately correct coverage under a wide range of conditions.

Therefore, for CRTs with a small number of clusters, less than around 15 per arm, we recommend the use of cluster-level analyses. For trials with larger numbers of clusters, either approach can be used but the individual-level regression methods are likely to be statistically more efficient than the cluster-level methods and provide greater convenience when analysing the

effects of covariates. If the assumptions of the random effects models are roughly correct, we advise using these models since they are likely to be the most efficient. Another advantage of random effects models is that they provide direct estimates of the between-cluster variance, and hence the ICC, quantities which are treated as nuisance parameters using the GEE approach. If however, the assumptions of the random effects models are violated, GEE with an exchangeable correlation matrix and robust standard errors should be the method of choice. Box 5.7.3 provides a summary of important points for analysis of CRTs.

New literature on the performance of methods for analysing CRTs is continually emerging, including methods based on restricted maximum likelihood estimation and Bayesian techniques (e.g. McNeish and Stapleton 2016b). Adjustments to GEEs for small number of clusters and variation in cluster sizes are also being investigated (e.g. Huang et al 2016). Therefore, particularly where there are few clusters or very small clusters it may be advisable to seek the advice of a statistician with experience of analysing CRTs.

Box 5.7.3 Summary of important points for CRT analysis

- The analysis of data from CRTs must take clustering into account
- Analyses based on cluster-level summaries are robust and perform well under a wide range of conditions
- Individual-level analyses have greater power and precision, but are generally less reliable when there are relatively few clusters (<15 per arm)
- Matched designs should take account of the matching in the analysis

The matched-pair design

Matched-pair CRTs are generally analysed using cluster-level summaries. Individual-level regression methods are usually not suitable for matched-pair designs, for reasons discussed at the end of this section.

Some investigators have suggested that, in matched-pair CRTs with a small number of clusters, an unmatched analysis may be used, particularly when the matching correlation is not strong. This is known as breaking the

matching, and helps to achieve balance while avoiding loss of degrees of freedom in the analysis. If the matching is broken, we can use either cluster-level or individual-level analytical methods (providing there is a sufficient number of clusters). However, the investigator needs to specify in advance in an analysis plan that an unmatched analysis will be conducted, resulting in a loss of power if it turns out that the matching was highly effective. If instead the investigator makes the decision based on the observed matching correlation after the trial is completed, this increases the type I error (Diehr et al. 1995). Therefore, we do not recommend this approach.

We will begin by considering how to obtain point estimates, confidence intervals, and significance tests for pair-matched CRTs where no adjustment is made for covariates. We use the comparison of proportions as an example, but similar equations can be derived for event rates and quantitative outcomes. The simplest method of obtaining an estimate of the risk difference is to calculate the mean of the pair-wise risk differences, as follows:

$$\frac{1}{c}\sum_{j}(p_{1j}-p_{0j}),$$

where p_{1j} and p_{0j} are the observed proportion with the outcome in the j^{th} matched pair in the intervention and control arms, respectively, and c is the number of matched pairs of clusters.

This can also be written as:

$$\bar{p}_1-\bar{p}_0,$$

so that the point estimate for the risk difference is identical to that obtained from an unmatched analysis.

We can write the observed difference in the risks in the j^{th} matched pair as:

$$p_{1j}-p_{0j}=h_j,$$

A 95% confidence interval for the risk difference can then be obtained as follows:

$$\bar{h} \pm t_{\upsilon,0.025} \times \frac{s_m}{\sqrt{c}},$$

where $\bar{h}$ is the mean of the differences in the risks across matched pairs, $t_{\upsilon,0.025}$ is the t distribution value with one-sided p-value of 0.025 with $\upsilon = (c-1)$ degrees of freedom and $s_m^{\,2}$ is the empirical variance of these differences:

$$s_m^{\,2} = \frac{1}{c-1}\sum_j (h_j - \bar{h})^2.$$

We can perform a paired t-test on the pair-wise differences to test the null hypothesis of no intervention effect:

$$t_m = \frac{\bar{h}}{\sqrt{\frac{s_m^{\,2}}{c}}},$$

The computed value of t_m is referred to tables of the t distribution with $\upsilon = (c-1)$ degrees of freedom. For small numbers of pairs, it may be advisable to additionally carry out a non-parametric test, for example the signed rank test. Note, however, that at least six pairs are needed to achieve a significant result (two-sided $p < 0.05$) when using such a non-parametric test.

If our parameter of interest is the risk ratio, since ratios often have skewed distributions, we may work in terms of the log (risk ratio), and calculate the mean of the log of the pair-wise risk ratios, obtain the 95% confidence interval and apply the paired t-test in a similar manner. As with an unmatched analysis, we can exponentiate our results to obtain an estimate of the risk ratio and its corresponding 95% confidence interval.

If close matching is achieved, the treatment arms will be similar with respect to the variables matched on. However there may be other variables

that are not balanced between arms and we may want to adjust for these in the analysis. For example, in the Mwanza HIV prevention trial reported by Grosskurth et al. (1995), the arms were similar at baseline with respect to HIV prevalence and other factors that were matched on. However, there was some imbalance in reported history of sexually transmitted infections (STIs), travel away from the village, and male circumcision, and so these variables were adjusted for in the analysis. As with analysis of unmatched CRTs, adjustment for covariates is a two-stage process. The first step is to fit a regression model excluding the intervention effect, but including terms for the matched pairs and the covariates we are adjusting for. We then compare the observed and expected numbers of events in the two clusters in each matched pair. If the intervention has a protective effect, then the observed number of events in the intervention arm will tend to be less than the expected number of events, and the observed number of events in the control arm will tend to exceed the expected number.

Individual-level regression methods are generally not suitable for the analysis of pair-matched CRTs when the trial endpoint is an event rate or binary outcome. Both random effects logistic and Poisson regression models are fitted using likelihood inference, which relies on large sample approximations for its validity. More specifically, it assumes that the number of parameters to be estimated is much smaller than the number of data points used to estimate them. However in the pair-matched model, we must estimate a large number of parameters, including a separate parameter for each matched pair, one for the intervention effect, and a parameter for each covariate in our model. Therefore in a pair-matched CRT, the number of parameters to be estimated approaches the number of clusters. Under these conditions the regression analysis cannot be relied upon to give robust results. Other regression methods, such as GEE, are similarly unreliable for the pair-matched design. Cluster-level methods are therefore recommended when analysing proportions or event rates in pair-matched trials.

An exception are matched CRTs with quantitative outcomes, because the analysis does not rely on the approximate methods of likelihood inference. Providing the random effects can be assumed to be normally distributed, valid estimates, confidence intervals, and significance tests can be obtained that do not depend on large sample assumptions. However, if the number of clusters is small, the dependence on assumptions of normality becomes critical, and so individual-level methods are unlikely to be as robust as the

cluster-level methods, even for quantitative outcomes. A more detailed discussion of the analysis of pair-matched CRTs is provided in Hayes and Moulton (2017, pp. 281–93).

Stratified CRTs

Stratified CRTs can be analysed using cluster-level summaries, using extensions of the methods described for matched CRTs (Hayes and Moulton 2017, pp. 293–303). Unlike with matched trials, stratified CRTs can also be analysed using individual-level regression methods with an additional fixed effect parameter for the strata. Stratified designs fall between the extremes of pair-matched and unmatched designs, because there is replication within the strata, and only one parameter is fitted for each stratum rather than one for each matched pair. This replication also allows us to obtain direct estimates of between-cluster variation, and to assess any differences in the intervention effect between strata.

Less research has been conducted on the performance of small stratified CRTs, but as a rough guideline, individual-level regression is not recommended for stratified trials with less than 20 clusters per arm, due to the extra parameters that are fitted. It is also recommended to carry out cluster-level analysis, to check that the conclusions are robust.

Ethical issues in the design of CRTs

The ethical principles underpinning research on human subjects are well recognized, and there are internationally accepted guidelines setting out the minimum requirements for such studies. However, cluster randomization introduces a number of additional ethical issues to which investigators need to give careful consideration when planning a CRT.

Obtaining informed consent is a key ethical requirement in all intervention trials. In individually randomized trials, individuals are asked for consent both to be randomly allocated to one of the treatment arms and to participate in the trial. In CRTs, consent for randomization and for participation generally need to be considered separately. Since randomization is at the cluster level, obtaining individual consent for randomization is rarely possible. Instead, approval is given at the group

level on behalf of the entire cluster, by a person who genuinely represents the interest of cluster members. Investigators need to carefully consider who should provide this approval. It should be noted that while a person may provide approval for randomization on behalf of the individual members of the cluster, they cannot provide formal consent (as defined by good clinical practice, GCP) on behalf of these individuals.

Although cluster randomization may preclude seeking consent from individuals before randomization, it is often possible to obtain individual consent at other stages. In many CRTs, individuals are able to consent to receive the intervention. Even if individuals cannot be asked for consent to receive the intervention, for example, a mass media intervention, most CRTs collect data from individuals in order to evaluate the intervention. Consent to participate must be obtained from these individuals.

For example, in the CRT in Uganda of IPT for malaria described earlier (Staedke et al. 2018), clusters were each a primary school and the surrounding households. Consent for randomization was given by national and district level ministries of health and education. Head teachers were asked for consent to participate in the trial. In the intervention clusters, children and their parents/guardians were asked for consent for the child to receive IPT. The outcome was measured in household surveys, in which all individuals were asked for consent to participate.

Potential benefit and harm also need to be considered at the cluster level. Many CRTs evaluate public health interventions such as educational programmes where the potential for harm may not be obvious, or may be poorly understood. For example, a social marketing intervention aimed at promoting condom use may have an unintended effect of encouraging casual sexual relationships. Another consideration is that the risk/benefit ratio may not be the same for all cluster members. For example, in the CRT of malaria IPT (Staedke et al. 2018), adults in the intervention clusters were expected to benefit from the reduction in malaria transmission, without taking IPT.

Ethical guidelines for CRTs are less readily available than for individually randomized trials. However, the Ottawa Statement on the Ethical Design and Conduct of Cluster Randomized Trials (Weijer et al. 2012) sets out 15 recommendations on the ethical design and conduct of CRTs. Further discussion of ethical issues in the context of CRTs can be found in Hayes and Moulton (2017).

Reporting of CRTs

CRTs are often large studies involving substantial financial, logistical, and human resources. There is perhaps a particular onus on CRT investigators to ensure that the reporting of such trials is carried out to the highest standards, to maximize the chances that the findings are interpreted correctly and the appropriate lessons drawn for future health policy. The Consolidation of Standards for Reporting of Trials (CONSORT) guidelines were developed over 20 years ago as an evidence-based checklist for reporting individually randomized trials (Begg et al. 1996). Subsequently, an extended statement was published that provided guidelines for use with CRTs (Campbell et al. 2004; Campbell et al. 2012).

The results of a CRT refer to the specific set of clusters that are studied. Therefore, when reporting a CRT, the characteristics of the clusters should be carefully described, including a clear definition of a cluster and an explanation of how the specific study clusters were selected. Eligibility criteria and sampling methods should be described at both cluster and individual levels.

A key requirement of the CONSORT guidelines is the provision of a flow diagram showing, by treatment arm, the number of clusters, and the number of individuals, that were eligible, randomized, and lost to follow-up at each stage of the trial, and numbers analysed for the main trial endpoints. This provides useful information for the assessment of selection bias and the generalisability of trial findings.

A clear description of the intervention is also important if the results of the trial are to be put into practice in other settings. This is particularly true for CRTs since they often evaluate complex interventions, comprising a package of measures that may be applied at either the cluster level or individual level or both.

Assumptions made in the sample size calculations should also be reported. This includes the number of individuals per cluster and the values of k or the ICC that were used. In addition, assumptions about the baseline rate, proportion, or mean outcome, as well as the expected intervention effect, should be stated. If matching, stratification, or restricted randomization were used, these aspects of the design need to be described carefully. Since the analysis methods for CRTs are less standardized than for individually randomized trials, it is essential to provide an adequate

description of the analytical methods and reasons for their choice. Any covariates used for adjustment should be described.

The first table in a report of the results of a randomized trial is usually a description of baseline characteristics of each trial arm, which shows whether the randomization has been effective in achieving comparable treatment groups. For a CRT, it is important to present baseline characteristics of both the clusters and individuals. Clearly, the most important table in any trial report is that which shows the estimated effects of the intervention on the main outcome. For a CRT, tables should show the estimated intervention effects and associated confidence intervals, obtained using appropriate statistical methods that take account of between-cluster variability. If there are fewer than 15–20 clusters per arm, we recommend that the cluster-level summary outcomes also be presented, at least as supplementary material, for transparency. It is also good practice to report the level of between-cluster variability in the form of the ICC or k. This provides information on appropriate values of these coefficients for investigators planning new studies.

As with individually randomized trials, the discussion section of a report of a CRT should focus on the analyses of primary endpoints, paying particular attention to the confidence interval associated with the effect estimate. Investigators should also consider the generalisability of trial results, since both the implementation of an intervention and the intervention effects may differ between populations or over time.

Conclusions

CRTs have become increasingly common in the literature in recent years. In this chapter, we have attempted to provide investigators with a practical overview of the special methods of design and analysis that are needed for these trials. Cluster randomization may be preferred over individual randomization in a number of situations; however, the advantages of this design need to be weighed against its limitations. If used appropriately, CRTs have the potential to make an important contribution to the development of evidence-based interventions to improve health.

REFERENCES

Arnup, S.J., McKenzie, J.E., Hemming, K., Pilcher, D., and Forbes, A.B. (2017). Understanding the cluster randomised crossover design: a graphical illustration of the components of variation and a sample size tutorial. *Trials*, 18, 381.

Baio, G., Copas, A.J., Ambler, G., Hargreaves, J., Beard, E., and Omar, R.Z. (2015). Sample size calculation for a stepped wedge trial. *Trials*, 16, 354.

Begg C., et al. (1996). Improving the quality of reporting of randomized controlled trials. The CONSORT statement. *Journal of the American Medical Association*, 276(8), 637–9.

Bellamy, S.L., et al. (2000). Analysis of dichotomous outcome data from community intervention studies. *Statistical Methods in Medical Research*, 9(2), 135–59.

Bennett, S., Parpia, T., Hayes, R., and Cousens, S. (2002). Methods for the analysis of incidence rates in cluster randomized trials. *International Journal of Epidemiology*, 31, 839–46.

Campbell, M.J., Donner, A., and Klar, N. (2007). Developments in cluster randomized trials and statistics in medicine. *Statistics in Medicine*, 358, 2–19.

Campbell, M.K., Elbourne, D.R., and Altman, D.G., for the CONSORT Group (2004). CONSORT statement: extension to cluster randomised trials. *British Medical Journal*, 328, 702–8.

Campbell, M.K., Piaggio, G., Elbourne, D.R., and Altman, D.G. (2012). CONSORT 2010 statement: extension to cluster randomised trials. *British Medical Journal*, 345, e5661.

Copas, A.J., Lewis, J.J., Thompson, J.A., Davey, C., Baio, G., and Hargreaves, J.R. (2015). Designing a stepped wedge trial: three main designs, carry-over effects and randomisation approaches. *Trials*, 16, 352.

Diehr, P., Martin, D.C., Koepsell, T., and Cheadle, A. (1995). Breaking the matches in a paired *t*-test for community interventions when the number of pairs is small. *Statistics in Medicine*, 14, 1491–504.

Donner, A., Birkett, N. and Buck, C. (1981). Randomization by cluster-sample size requirements and analysis. *American Journal Epidemiology*, 114, 906–14.

Donner, A. and Klar, N. (2000). *Design and Analysis of Cluster Randomization Trials in Health Research*. New York: Oxford University Press.

Donner, A. and Klar, N. (2004). Pitfalls of and controversies in cluster randomization trials. *American Journal of Public Health*, 94, 416–21.

Doyle, A.M., et al. (2010). Long-term biological and behavioural impact of an adolescent sexual health intervention in Tanzania: follow-up survey of the community-based MEMA kwa Vijana Trial. *PLoS Medicine*, 7(6), e1000287.

Eldridge, S. and Kerry, S. (2012). *A Practical Guide to Cluster Randomised Trials in Health Services Research*. Chichester, UK: John Wiley.

Feldman H.A. and McKinlay, S.M. (1994). Cohort versus cross-sectional design in large field trials: precision, sample size, and a unifying model. *Statistics in Medicine*, 13, 61–78

Gail M.H., Mark, S.D., Carroll, R.J., Green, S.B., and Pee, D. (1996). On design considerations and randomization-based inference for community intervention trials. *Statistics in Medicine*, 15, 1069–92.

Grosskurth, H., et al. (1995). Impact of improved treatment of sexually transmitted diseases on HIV infection in rural Tanzania: randomised controlled trial. *Lancet*, 346, 530–6.

Hanley, J.A., Negassa, A., Edwardes, M.D., and Forrester, J.E. (2003). Statistical analysis of correlated data using generalized estimating equations: an orientation. *American Journal of Epidemiology*, 157, 364–75.

Hargreaves, J.R., et al. (2015). Five questions to consider before conducting a stepped wedge trial. *Trials*, 16(1), 1.

Hayes, R.J. and Moulton, L.H. (2017). *Cluster Randomised Trials* (2nd ed.). London: Chapman & Hall/CRC Biostatistics.

Hemming, K., Girling, A.J., Sitch, A.J., Marsh, J., and Lilford, R.J. (2011). Sample size calculations for cluster randomised controlled trials with a fixed number of clusters. *BMC Medical Research Methodology*, 11, 102.

Hemming, K., Haines, T.P., Chilton, P.J., Girling, A.J., and Lilford, R.J. (2015). The stepped wedge cluster randomised trial: rationale, design, analysis, and reporting. *British Medical Journal*, 350, h391.

Hemming, K. and Taljaard, M. (2016). Sample size calculations for stepped wedge and cluster randomised trials: a unified approach. *Journal of Clinical Epidemiology*, 69, 137–46.

Huang, S., Fiero, M.H., and Bell, M.L. (2016). Generalized estimating equations in cluster randomized trials with a small number of clusters: review of practice and simulation study. *Clinical Trials*, 13(4), 445–9.

Hussey, M.A. and Hughes, J.P. 2007. Design and analysis of stepped wedge cluster randomized trials. *Contemporary Clinical Trials*, 28(2), 182–91.

Ivers, N.M., et al. (2012). Allocation techniques for balance at baseline in cluster randomized trials: a methodological review. *Trials*, 13, 120.

Iwuji, C.C., et al. (2013). Evaluation of the impact of immediate versus WHO recommendations-guided antiretroviral therapy initiation on HIV incidence: the ANRS 12249 TasP (Treatment as Prevention) trial in Hlabisa sub-district, KwaZulu-Natal, South Africa: study protocol for a cluster randomised controlled trial. *Trials*, 14, 230.

Iwuji, C.C., et al. (2018). Universal test and treat and the HIV epidemic in rural South Africa: a phase 4, open-label, community cluster randomised trial. *Lancet HIV*, 5(3), e116–25.

Jha, D., et al. (2017). Protocol for the mWellcare trial: a multicentre, cluster randomised, 12-month, controlled trial to compare the effectiveness of mWellcare, an mHealth system for an integrated management of patients with hypertension and diabetes, versus enhanced usual care in India. *BMJ Open*, 7(8), e014851.

Kish, L. (1965). *Survey Sampling*. New York: John Wiley & Sons.

Martin, D.C., Diehr, P., Perrin, E.B., and Koepsell, T.D. (1993). The effect of matching on the power of randomized community intervention studies. *Statistics in Medicine*, 12, 329–38.

McNeish, D.M. and Stapleton, L.M. (2016a). The effect of small sample size on two-level model estimates: a review and illustration. *Educational Psychology Review*, 28, 295–314.

McNeish, D.M. and Stapleton, L.M. (2016b). Modeling clustered data with very few clusters. *Multivariate Behavioral Research*, 51(4), 495–518.

Moulton, L.H. (2004). Covariate-based constrained randomization of group-randomized trials. *Clinical Trials*, 1, 297–305.

Norman, J.E., et al. (2018). Awareness of fetal movements and care package to reduce fetal mortality (AFFIRM): a stepped wedge, cluster-randomised trial. *Lancet*, *392*(10158), 1629–38.

Parienti, J.J. and Kuss, O. (2007). Cluster-crossover design: a method for limiting cluster level effects in community-intervention studies. *Contemporary Clinical Trials*, 28, 316–23.

Raab, G.M. and Butcher, I. (2001). Balance in cluster randomized trials. *Statistics in Medicine*, 20, 351–65.

Reich, N.G., Myers, J.A., Obeng, D., Milstone, A.M., and Perl, T.M. (2012). Empirical power and sample size calculations for cluster randomized and cluster-randomized crossover studies. *PLoS One*, 7(4), e35564.

Rhoda D.A., Murray, D.M., Andridge, R.R., Pennel, M.L., and Hade, E.M. (2011). Studies with staggered starts: multiple baseline designs and group-randomized trials. *American Journal of Public Health*, 101, 2164–9.

Ross D.A., et al. (2007). Biological and behavioural impact of an adolescent sexual health intervention in Tanzania: a community randomized trial. *AIDS*, 21, 1943–55.

Rutterford, C., Copas, A.J., and Eldridge, S. (2015). Methods for sample size determination in cluster randomized trials. *International Journal of Epidemiology*, 44(3), 1051–67.

Sarrassat, S., et al. (2018). Effect of a mass radio campaign on family behaviours and child survival in Burkina Faso: a repeated cross-sectional, cluster-randomised trial. *Lancet Global Health*, 6(3), e330–41.

Sismanidis, C., et al. (2008). Restricted randomization of ZAMSTAR: a 2x2 factorial cluster randomized trial. *Clinical Trials*, 5(4), 316–27.

Staedke, S.G., et al. (2018). Assessment of community-level effects of intermittent preventive treatment for malaria in schoolchildren in Jinja, Uganda (START-IPT trial): a cluster-randomised trial. *Lancet Global Health*, 6(6), e668–79.

Teerenstra, S., Eldridge, S., Graff, M., de Hoop, E., and Borm, G.F. (2012). A simple sample size formula for analysis of covariance in cluster randomized trials. *Statistics in Medicine*, 31, 2169–78.

Thompson, J.A., Fielding, K.L., Davey, C., Aiken, A.M., Hargreaves, J.R., and Hayes, R.J. (2017). Bias and inference from misspecified mixed-effect models in stepped wedge trial analysis. *Statistics in Medicine*, 36(23), 3670–82.

Weijer, C., et al. (2012). The Ottawa Statement on the ethical design and conduct of cluster randomized trials. *PLoS Medicine*, 9(11), e1001346.

Zeger, S. and Liang, K.Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, 42, 121–30.

5.8

Community intervention trials in high-income countries

John W. Farquhar and Lawrence W. Green

Introduction to community intervention trials in high-income countries

As background on theories and methods underlying successful multilevel, multimethod, *community-based*, *health promotion intervention trials*, this chapter first reviews community organizational and mass media intervention developments in the first half of the twentieth century and the origin of randomized trials outside laboratory settings early in that century. It then reviews specific health examples from tuberculosis mobile X-ray and immunization campaigns in the 1950s and 1960s, family planning programmes in the 1960s and 1970s, and HIV/AIDS in the 1980s. We then turn to the most prominent and influential community health intervention trials for the remainder of the twentieth century as touchstones for development of twenty-first-century public health challenges, such as obesity and diabetes control, substance abuse prevention, and cancer control. These touchstones were the cardiovascular disease prevention and risk reduction trials (including most notably hypertension, tobacco, dyslipidaemia, and physical inactivity) mounted in California with the Stanford Three-Community Study and in North Karelia, Finland in 1972. Parallel national and state-wide successes in hypertension control in the 1970s, and tobacco control in the 1980s and beyond, have drawn significantly on these and subsequent exemplary trials from North America, Europe, Australia, and elsewhere.

These later trials exemplify the scope and innovation of *experimental* epidemiology in advancing disease prevention and health promotion science, involving total populations of communities, initiated often following insights from *observational* epidemiology, or from *randomized clinical* trials in

smaller populations. Some were derived from ‘natural experiments’ that studied the impact of policies and programmes not initiated by academic investigators.

The earliest of mid-century trials were often in developing countries, designed to reduce social development problems associated with communicable diseases and high birth rates. The communicable disease studies in Western countries gave opportunities to test new polio vaccines and their adoption, and later by emergence of new infectious diseases, particularly HIV/AIDS. Other studies increasingly addressed risk factors for cardiovascular disease (CVD), cancer, and other chronic diseases as the rising mortality curve of chronic diseases crossed the declining curve of communicable diseases in the mid-twentieth century. All relied on locally available mass communication channels, community organizing, and other policy advocacy and educational methods, using relatively low-cost approaches with potentials to change lifestyles or specific behaviours.

One study examined alcohol-involved trauma using only community organizing to promote adherence to existing laws, rather than public education for policy-change advocacy (Holder et al. 1997). Another study, designed to reduce consumption of dietary trans fats, used primarily new laws and regulations (Angell et al. 2009). By community organizing we mean collaborating and seeking support from community leaders and other channels of communication. The World Health Organization characterized this as community participation or participatory planning (Green 1983), and as ‘maximum feasible participation’ (Moynihan 1969) for community projects under President Kennedy’s ‘New Frontier’, and Lyndon Johnson’s ‘New Society’ legislation. But this spin on community organization assumed lay communities set the goals (Green 1989). This orientation was recently termed ‘community-based participatory research’ when organizing was undertaken collaboratively at the initiative of the community or by academic or external health agencies (George et al. 2006; Minkler and Wallerstein 2008). Following further definition of community and community organizing for health, we review this history to identify some prominent precedents in the public health and community health promotion literature.

Definitions of community and community organizing for health

A community-based programme was defined historically as one organized locally, and promoted through the community's institutions and communication channels. '*Community-based*' has become more broadly labelled, including both community-initiated and community-administered following initiation from external agencies. Programmes were termed '*community-placed*' when conceived, sponsored by, or driven from organizations outside the community and applied with or without community participation.

The traditional definition of a 'community' is a residential area with legally defined geographic boundaries, where a *local* governmental system regulates many aspects of schools, businesses, law enforcement, transportation, and recreational activities. A community is usually the last designated jurisdiction in a nation's official or governmental regulatory chain, where education and implementation must ultimately occur. In rural areas the county or district becomes the agent for education and other official governance. For programmes funded by non-governmental organizations, local agencies or branch organizations are the implementing agents within communities.

The following summarizes changing precedents in effectiveness trials of community health promotion, reviewed by twentieth-century decades. It illustrates that technologies for solving emerging public health problems led to research in high-income countries, developing theories, and evidence needed for community interventions for large segments of the population at risk.

Some converging developments for community health trials

Table 5.8.1 summarizes a chronology of community trials that dominated eras of action research across seven decades of the twentieth century, and a few citations of works illustrating trial types and theoretical and epidemiological problems they tested, either as proof-of-concept trials, or as health promotion model evaluations.

The origin of randomized controlled trials (RCTs) is usually attributed to Sir Ronald Fisher (1925) for its application to field trials comparing effectiveness of alternative cultivation methods (Fisher 1926). Some rural sociologists began in the 1930s to study the spread among farmers of new hybrid seeds, providing technical assistance to farmers by agricultural

extension agents. This led to the first publication of the diffusion model describing the normal-curve pattern of adoption of new ideas, behaviour, or innovations in populations (Ryan and Gross 1943). The diffusion model and its theoretical framing of social and behavioural change in populations has been chronicled by the rural sociologist Everett Rogers (2003), through five editions of his book, *Diffusion of Innovations*, and specifically for public health applications (Green et al. 2009). Rogers converged at Stanford with others who brought their various sociobehavioural and health disciplines to bear on development of the first large-scale, interdisciplinary US trial in community-based cardiovascular risk factor reduction. Table 5.8.1 recapitulates this history of community trials leading to and including the cardiovascular trials and beyond to tobacco control and now obesity control.

Tobacco control successes, from the 1960s through to the 1990s

Of all chronic disease prevention and health promotion efforts, tobacco control is the most cited as a great success of health behaviour change on a population scale. Ironically, at least in North America, Europe, Australia, and New Zealand, this success, initially, was not the direct application of controlled trials, either clinical or community. It followed many planned media and educational events, popular media, policy initiatives and the environmental changes associated with them, and their evaluation through yearly measures of tobacco consumption. The mass media ‘softened’ public opinion, denormalizing smoking in public places, and assisting smoking cessation policy changes (Fig. 5.8.1).

Table 5.8.1 Decades in which community trials of particular types dominated in development of a body of evidence, methods, and theories

Decade	Influential community trials	Examples and reviews
1930s	Agricultural extension trials and diffusion studies	Fisher 1926; Ryan and Gross 1943; Rogers 2003; Green et al. 2009
1940s	Mass media and war-years' propaganda studies; advertising, radio, and film studies	Hovland 1948; Maccoby 1951; Michael and Maccoby 1953
1950s	Community mobile X-ray and various immunization studies	Hochbaum 1958; Rosenstock et al. 1959; WHO 1954; Green 1970
1960s	Community family planning and polio immunization campaign trials	Stycos 1962, pp. 481–482; Berelson 1966; Bogue 1962; Northcutt et al. 1964; D'Onofrio 1966; Gray et al. 1966; Green et al. 1972
1970s	The first cardiovascular community risk factor prevention trials	Farquhar et al. 1977; Puska et al. 1981
1980s	The HIV/AIDS and substance abuse prevention trials	Miller et al. 1990; Ruiz et al. 2000; Cain et al. 2001; Darrow et al. 2009; Kagan et al. 2012
1990s	Culmination of four decades of tobacco control	Centers for Disease Control and Prevention 1999; Rogers 2010; Levy et al. 2012
2000s	Obesity control	Institute of Medicine 2010a, 2010b, 2012a, 2012b

These associations offer 'lessons' from tobacco control for chronic disease control at the population level, demonstrating that changing risk *conditions* in the social and physical environment is needed for behaviour change.

In addition to denormalization, research shows other policies and programmes have additive or multiplicative effects on reducing tobacco consumption and its premature mortality. The greatest impacts include tobacco price increases, high-intensity media campaigns, comprehensive cessation treatments, strong health warnings, stricter smoke-free air regulations, advertising bans, and youth access laws (e.g. Levy et al. 2012). Price increases have an independent effect; but the last three depend on both laws and their enforcement. These, in turn, have depended on building an informed and activated electorate through community-level organizing and communications, because community policies controlling tobacco

promotions and smoking in public places could be passed locally when they could not overcome industry lobbying at state and national levels.

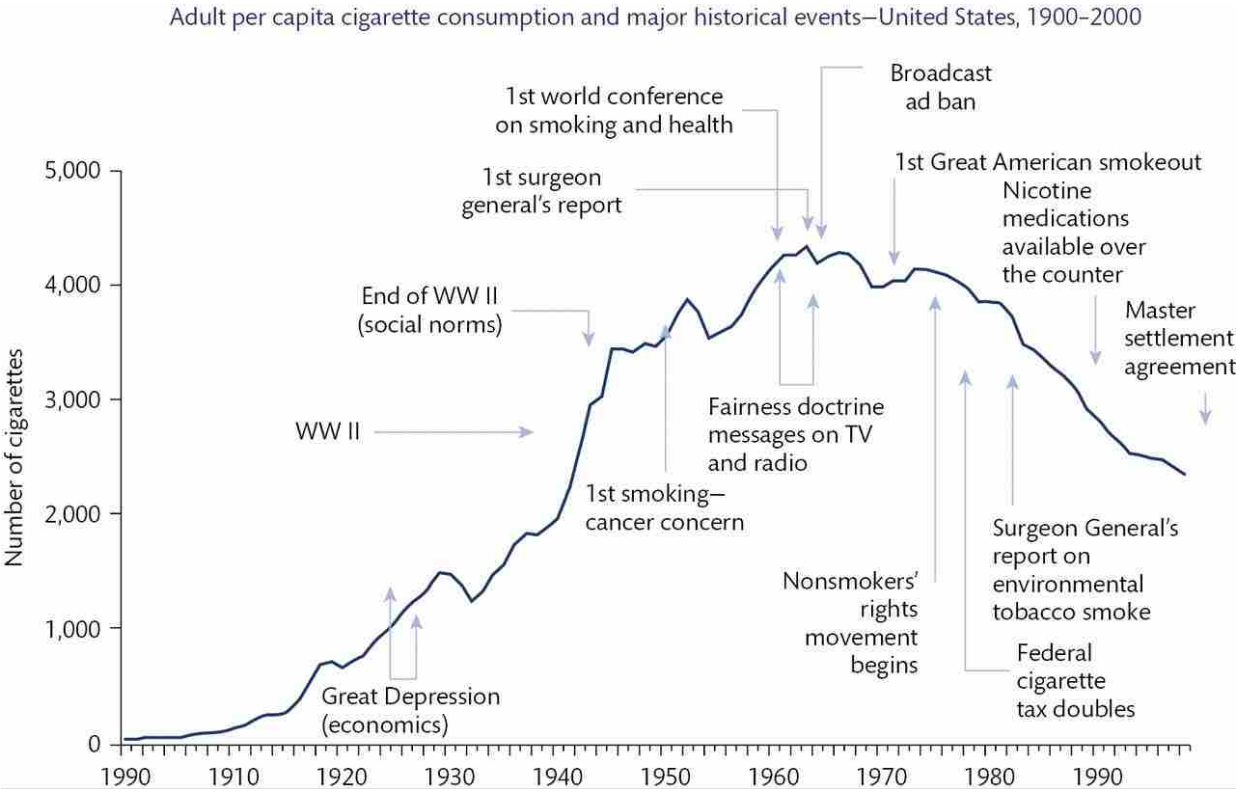


Fig. 5.8.1 Linkage in tobacco consumption trends with historical events in the social, economic, and policy environments.

Adapted from US Department of Health and Human Services, *Reducing Tobacco Use: A Report of the Surgeon General*, US Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health, Atlanta, Georgia, USA, 2000, available from http://www.cdc.gov/tobacco/data_statistics/sgr/2000/complete_report/pdfs/fullreport.pdf

Lessons drawn from the last seven decades of the twentieth century

We draw 'lessons' or generalizations for community prevention and health promotion interventions from seven decades of trials and natural experiments just reviewed, drawing especially on CVD trials and adoption and cessation trends of tobacco seen during the twentieth century (Fig.

5.8.1). Many ‘lessons’, derived from the tobacco consumption trends, were more readily adoptable within the CVD community trials, which began in the 1970s, and included smoking prevention and cessation. The tobacco experiences were also then seen as applicable to other risk behaviours at the community level. Finally lessons from these CVD trials lead to a description of the mix and sequence of the intervention steps needed.

The economic imperative

Cigarettes became increasingly popular and inexpensive in the century’s first three decades after manufacturing ‘advances’ reduced mass production costs, accounting for a steep rise in consumption. Consumption plummeted in the late 1920s during the Great Depression. This strong association in usage was early evidence for the wisdom of suppressing consumption through tobacco excise taxation. It demonstrated that even an addictive product is sensitive to price relative to disposable income. The argument for excise taxes gained momentum as evidence accumulated. A 10% increase in cigarette price, on average, reduces consumption by about 4% (Chaloupka et al. 2002), and more for young, lower-income and pregnant smokers (Hyland et al. 2006).

Two other advantages spurred more cigarette tax initiatives. One is that revenues can finance state-wide *and* community tobacco control programmes, making cigarette taxes even more palatable to the public, thus launching and sustaining some very successful programmes. The other is growing evidence that tobacco consumption increases healthcare expenditures, and since taxes decrease tobacco use, healthcare costs may eventually fall by billions of dollars in a state the size of California (Max et al. 2013).

The economic imperative also relates to obesity control and diabetes prevention, partly due to decades of increasing agricultural efficiency, causing progressive food price decreases. Food manufacturers and vendors have increased portion sizes, thus many have become accustomed to consuming larger portions and more frequent indulgences of cheaper, unhealthful food; much of it processed rather than home-cooked, and often with empty calories. These trends have led a contemporary movement advocating taxation of sugar-sweetened beverages (e.g. Novak and Brownell 2011).

The social normative or denormalization imperative

Two early notable changes in the trajectory of tobacco consumption were acceleration during the Second World War, and a less dramatic but significant rise during the First World War (Fig. 5.8.1). These are partly attributable to official condoning, facilitating, and even encouraging of smoking among soldiers to reduce boredom, and cope with stress, and smoking became integrated with their camaraderie. A modest drop in smoking occurs following the Second World War as soldiers' social networks fragment. Later recognition of these social norm changes led to using mass media to portray smoking as less glamorous, less fashionable, less smart, and portrayed the tobacco industry as deceiving and exploiting the public. Experiments within the community CVD trials began to highlight use of social influence strategies and media use to provide role models for changing behaviours, such as smoking (Ramirez et al. 1995).

The public health education and informed electorate imperative

Before mass media were used to create denormalization, most public health education was limited to smoking's health hazards, which had effects as noted. In 1954 a single mass media message created a precipitous drop in tobacco use. That year, *The Reader's Digest*, a popular and credible magazine with the American public, featured an article showing that tar of tobacco residues induced cancerous tumours when painted on mice. Although dramatic in its drop, consumption soon bounced back and continued upwards with continued tobacco industry advertising (Fig. 5.8.1).

As the public became more aware and cautious about tobacco, America was ready for the first Surgeon General's Report on Tobacco and Health in 1965, the peak of tobacco consumption growth. This and subsequent events led to legislation restraining the tobacco industry (e.g. the Fairness Doctrine requiring equal broadcast media time for antitobacco and pro-tobacco advertising). This indicated that public health education's most powerful method, certainly in relation to tobacco, was creation of an informed electorate—more supportive of regulations restraining industry advertising, promotion, and sales of a harmful product, and raising tobacco products' taxes.

The surveillance imperative

The tracking of tobacco consumption illustrates how helpful a consistently measured outcome or risk factor metric related to particular chronic diseases can be. Ideally, such measures should be collected on whole populations and also on smaller subpopulations requiring particular attention. Regrettably, such surveillance—nearly universal for communicable diseases—remains sparse for most chronic disease risk factors. They are usually absent at the community level, so special surveys are needed for community intervention effects on behavioural and environmental issues. This expense alone makes trials prohibitively costly and dependent on a national government's funding.

The comprehensiveness imperative

The major tobacco consumption declines, following the first Surgeon General's Report on Smoking and Health in 1965, reflect the historical creation of national policies and programmes supportive of state and local efforts to create comprehensive programmes (Centers for Disease Control and Prevention 1999, 2007). Two states in particular (California and Massachusetts) in the late 1980s and early 1990s increased tobacco taxes, with a portion of the revenue earmarked for state-wide and community tobacco control, including use of the social normative approach through schools, worksites, and mass media. These programmes achieved tobacco consumption reductions that doubled and then tripled decline rates of the other 48 states within a few years. Given results of these and later state-wide and community trials, and success in measuring and tracking their effects, the Centers for Disease Control and Prevention's Office on Smoking and Health published in 1999 and updated in 2007 the 'Best Practices' document cited earlier, guiding other states' comprehensive state-wide programme development.

The formative evaluation and tailoring imperative

Formative evaluation must begin before and be continued throughout implementation of any planned community intervention campaign. Success requires programmes well designed in mix and sequence, delivered through varied channels, adjusted for audience responses. This integration, with goals set in advance and adjustments from early results, is analogous to a

commercial marketing campaign, hence the term ‘social marketing’. (Ironically, *commercial* marketing has adopted many novel methods from academic research on *social* marketing.) It includes recognition of audience segmentation, or tailoring of messages and appeals to the varied motives, needs, and preferences of subpopulations (Slater and Flora 1991). Needs analysis, message design, and pre-testing are formative evaluation phases. Message sequences should first increase awareness, then knowledge, then motivation, and last provide training in the skills and resources to enable adoption and maintenance of new behaviours (Bandura 1986). Electronic media can furnish the first two parts of this sequence and stimulate use of more information-dense print media of newspapers, booklets, CDs, and websites, which are inherently more effective in *skills training* than are electronic media (Flora et al. 1997).

The ecological imperative

Health behaviours and health status are influenced by economic, political, cultural, and socioeconomic factors, all non-biological (Schooler et al. 1997; Best et al. 2003; Trickett et al. 2011). Healthcare’s reliance on clinical services can neither prevent most chronic disease nor reach large population segments. Therefore, changing behaviour and improving health need to include family, cultural, and social contexts of these populations. A primary tenet of community prevention programmes is the opportunity to intervene in multiple settings and domains (Poland et al. 2000), and at multiple levels from individuals to families, organizations, and whole populations. This ecological approach honours the ‘comprehensiveness imperative’, but also urges communities to encompass interdependencies among environmental, social, and individual factors and the potential for public health strategies to interact with individually tailored therapeutic and medical strategies (Shea and Basch 1990; Schooler et al. 1997; Green et al. 2012; Farquhar in press). This principle is known as the ecological approach to community or population programmes (Green et al. 1996; Kok et al. 2008; Golden and Earp 2012), and has been a central tenet of the Ottawa Charter for Health Promotion (World Health Organization et al. 1986).

Community-based programmes offer opportunities to approximate the ideal multicomponent and multilevel approach by coordinating population-based policy and mass health communication and education strategies with

individual-based clinical intervention. Social norms and community attributes greatly impact health. Community governmental and private institutions influence health behaviours by controlling, for example, health, recreational (e.g. Kokko et al. 2008), and transportation services (Mosquera et al. 2012) and youth access to tobacco. The health and safety policies and worksite health promotion practices of businesses can also affect health. Stores and restaurants also shape health according to the food they sell (Schooler et al. 1997) and whether menus provide information on calories and ingredients.

The ideal programme design, then, would include programmes developed within or channelled through multiple and coordinated worksites, schools, recreation sites, libraries, churches, local business organizations, and other sites to supplement and expand the effects of policy and health education approaches and appropriate delivery of healthcare (Farquhar et al. 1985; Kokko et al. 2006; Dooris 2009). This broad reach into the community is critically appropriate to chronic disease prevention and injury control because most people are at some level of risk for these and will benefit from interventions to predispose, enable, and reinforce healthful behaviour, leading to creating a *population* impact on prevalent risk factors. Indeed, most CVD does not occur among the relatively few adults at highest risk, but among many at moderate risk (Blackburn 1979; Puska et al. 1985; Schooler et al. 1997). Therefore, population health promotion must reach a community's entire population, including those who are neither engaged with the health system nor motivated to reduce a possible distant development of chronic disease.

While some behaviour change reflects individuals recognizing and reducing a risk, a population-wide impact always requires diffusion of behaviour change from a social process—individuals learning about new behaviours from neighbours, or individuals being encouraged to adopt new behaviours by those in their social networks. Such social diffusion results best from increasing opportunities to learn about and implement new behaviours, plus increased social expectations for doing so. The tobacco control literature calls this the social norms approach, 'denormalizing', by making both tobacco smoking and the tobacco industry's practices socially unacceptable, an approach central to California's successful Tobacco Control Program (Rogers et al. 2010).

This social normative and total community approach requires building an environment that coherently and consistently favours behaviour change. In addition to policy changes and mass communications campaigns, this total environmental change may be aided by entrepreneurs recognizing a need and developing products to accelerate behaviour change, through workplaces providing incentives for adherence to wellness programmes (Volpp et al. 2009), through schools that modify food available to children and require physical activity (Lynagh et al. 1997; Michaud 2003; Louriero 2004), through sports clubs (Kokko et al. 2006), and through programmes in other community settings. The goal is to create a behaviour change cascade, arising from synergistic interactive effects when many components are embedded in total community-based programmes that encourage institutional change, leaves space for entrepreneurs, and includes mass media and environmental change (Schooler et al. 1997).

Several synergies arise from combining multicomponent, multilevel (ecological) community intervention approaches with clinical services of physicians and other health professionals (Green et al. 2012). Adherence to dietary and exercise advice is more readily followed when family members, friends, and colleagues are modelling, enabling, and reinforcing these behaviours. Medication adherence improves when others in the patient's social network have learned, for example, about the need to control blood pressure and blood lipids. Health professionals, and their practice sites, potentially contribute to the 'total push' for better health, for example, by supplying print materials on health and alerting their patients to forthcoming community health events.

Comprehensive community-based efforts may modify the environment to support healthful or inhibit unhealthful individual actions, create organizational and institutional support for programmes, and influence the knowledge, attitudes, and behaviours of individuals (Schooler et al. 1997). However, a pragmatic view on implementing community-based programmes includes healthy scepticism about depending on institutions if reducing the CVD burden is not central to their primary missions. For workplaces, schools, and other community organizations the advantages may be made more evident when improved health increases performance or reduces absenteeism, for example, of students or employees.

Committing limited resources from community-level organizations to programmes may be unsustainable if incentives are weak or if interventions

are too complex to implement and maintain, leading to dissipation despite early enthusiasm (Thompson et al. 2000). For example, businesses that have insurance incentives to reduce smoking and obesity might invest in CVD prevention programmes, but few precedents exist for long-term sustainability and effectiveness even in high-income countries—and the potential incentives for businesses in low- and middle-income countries are less certain. However, recent findings suggest that financial incentives increase effectiveness of many worksite programmes, with evidence limited so far to their use in large worksites (Volpp et al. 2009). Similarly, physical activity is greatly needed in schools, but may not be implementable, given competition for limited educational resources and their primary academic mission. The infrastructure and capacity may be even more fragile in less formalized community organizations. More knowledge is needed on methods to create programmes and disseminate or scale them up in low- and middle-income countries.

'Best processes' to complement 'best practices'

Community intervention trials are typically funded to determine *efficacy*; multicommunity trials test *effectiveness*, that is, what might result when applied to different populations with different community structures. The generalizability from these to many varieties of communities and populations involves external validity (Green et al. 2009). Adaptation is an assessment and planning *process* which considers the history, culture, demographic and political structure, attitudes, community needs, and diversity of a particular community.

Features of this planning, implementation, and evaluation process approach are shown procedurally in models such as PRECEDE-PROCEED (Green and Kreuter 2004), which was informed by community trials following on Stanford and North Karelia, and the PATCH programme model of the Centers for Disease Control and Prevention (CDC) (Green and Kreuter 1992).

Message characteristics

Messages must arouse interest in health improvement and disease prevention and appeal to individuals with varied backgrounds and knowledge. They

must surmount indifference caused by the public's exposure to an *extreme* advertising overload. Since the 1990s, the average adult's annual exposure in the United States was 35,000 television advertisements, equalling 292 hours of personal exposure (Fortmann et al. 1995). They must arouse interest and awareness and have production qualities able to capture the public's attention. Formative evaluation must continue to match message content to the population's readiness. Early campaign messages must raise awareness and increase knowledge; later messages should increase motivation, provide direction to access resources, and develop needed skills—with resultant behaviour change (Schooler et al. 1997).

For adolescents, for example, the following enhanced success: parents and teachers whose personal habits were role models; and amount and quality of school-based education on tobacco, fitness, and nutrition. Also, certain personal characteristics were influential—for example, the person's baseline *self-efficacy* towards one's behaviour-change abilities (Fortmann et al. 1995). Finally, as Perry, of the Minnesota Heart Health Project (Perry 2000) explained, if a peer-led smoking prevention curriculum reduced smoking onset among youth, effect preservation required other support: 'The careful links of behavioural curricula, parental involvement, community support, and peer participation and leadership, with messages that are consistent and coordinated, have demonstrated efficacy beyond the implementation of a single strategy' (p. 122). This lesson is very germane for developing countries, given their proportionately high adolescent population.

The multiplicity and multilevel dimensions of community supports are shown in Table 5.9.2 as channels and mediators, referring to the individuals through whom the interventions can be personally or organizationally adopted and converted into reasons to act themselves, and then to influence others to change behaviour. They connect in the third column to the receiver's social cognitive change processes (Bandura 1986).

Prominent theories applied in community-based interventions

Community-based health education carried out by SPRC and some analogous projects has been guided by two major theories (community organizing and health communication–behaviour change), with success depending on their judicious blend (Flora et al. 1989; Farquhar et al. 1991).

Community organizing theory

This describes how to identify the community's health problem (and resources needed or available), and how to mobilize the community's opinion leaders and organizations. It describes the need to form coalitions, begin and maintain educational programmes, achieve regulatory changes, and *empower communities* to reach and maintain their goals. See 'Functions (for the sender)' in Table 5.8.2. Community organizing has elements central to concepts of 'community self-development' (Green and Kreuter 2005; Minkler 2005) and includes elements of diffusion theory (Rogers 2003)—which has shown how innovations are adopted through natural social networks, aided by a community's opinion leaders. Rogers strongly advised to ensure that these leaders play an important role, given their perceived trustworthiness (Rogers 2003).

Health communication–behaviour change theory

Health communication–behaviour change theory describe the basis for educating the population and its subgroups considering their health needs, cultural attributes, social networks, media habits, attitudes, motivation, knowledge, and self-management skills (see Table 5.8.1). It describes principles underlying education's success, including social cognitive theory (Bandura 1986), which emphasizes an individual's capacity for self-directed change. Bandura's research builds on the long-standing educational principle that 'learning by doing' is more effective than 'learning from observing' (modelling a behaviour), and both are more effective than an 'information-only' approach. Bandura described these principles as *social cognitive theory*: proposing that guided practice in a new behaviour can increase self-efficacy and behaviour change beyond the usual (Bandura 1986). The health communication–behaviour change theory includes efficient ways to reach total populations using methods described earlier in this chapter as 'social marketing' (Kotler 1975; Lefebvre and Flora 1988). The goal of a comprehensive, multicomponent, community-based intervention is to achieve 'community efficacy', just as 'self-efficacy' occurs for an individual after a successful behaviour change programme (Bandura 1986).

Specific methods for CVD community-based interventions

Problem identification

The first task is to identify the problem, initiated from within or outside the community. Local (community) interventions ideally interact with national programmes and link to scientific support from national or international levels, as noted with the family planning and HIV/AIDS histories. CVD prevention (WHO 1986) and alcohol sales control (Holder and Wallack 1986) are problems requiring local action, and any chronic disease requiring widespread education for changing 'lifestyle' behaviours (such as exercise, diet, and cigarette use) will require analogous methods.

Table 5.8.2 The components of health communication–behaviour change

Communication inputs	Functions (for the sender)	Behaviour objectives (for the receiver)
Face-to-face messages	Determine receiver's needs	Become aware
Mediated messages	Gain attention (set the agenda)	Increase knowledge
Community events	Provide information	Increase motivation and interest
Environmental cues	Provide incentives	Take action, assess outcomes
Training for community sources of influence (parents, employers, etc.)	Provide training to enable and reinforce others	Maintain action, practise self-management skills
Diffusion and influence beyond the primary receivers	Provide cues to action, including environmental change	Become an opinion leader (exert peer-group influence)

Adapted from Farquhar, J.W. et al., 'Methods of communication to influence behavior' in Holland, W.W., Detels, R., and Knox, G. (eds) *Oxford Textbook of Public Health*, Second Edition, Volume 2, Oxford University Press, Oxford, UK, Copyright © 1991, by permission of Oxford University Press.

Formation of partnerships

Key opinion leaders as well as relevant organizations usually need to form a coalition or at least a network of strategic partnerships (Green et al. 2001) to assure coordinated agreement on needs assessments, planning and

implementation, development of a community resource inventory, and mobilization of populace support for policy changes and enforcement. Relevant organizations, particularly in developed countries, include the following: the County Medical Society, or its equivalent, the local public health department, municipal hospital community affairs departments, city parks and recreation departments and voluntary health agencies (e.g. the local branches of national heart, lung, cancer, and diabetes organizations). Any organization that will act as a conduit for the public's education should also be represented (e.g. television and radio stations, local newspapers, schools, churches, libraries, pharmacies, clinics, and physicians' and dentists' offices). As the coalition-partner numbers proliferate, a 'steering committee', preferably limited to six members, is needed for more detailed planning, including formation of expert groups ('task forces') in education and evaluation and other topics (Flora et al. 1989; Farquhar et al. 1991).

Matching for comparison

Assuming the high likelihood of a non-randomized or cluster-randomized small number of communities, and therefore a 'quasi-experimental' design (Flora et al. 1989; Mercer et al. 2007), successful implementation requires close matching of intervention communities (treatment) with comparison (control) communities on relevant demographic and socioenvironmental factors, while also minimizing shared media channels (Farquhar et al. 1985). Many community trials have not followed these important principles.

Intervention components

The goals include providing behaviour-change skills training, supplementing goals of increasing knowledge. They should enlist multiple sectors beyond health organizations in health improvement campaigns, including changes in social and built environments, and regulations that promote access to facilities and resources needed for healthful physical activity. Multiple communication channels can facilitate reaching subgroups with differing needs. Comprehensive interventions should involve schools, worksites, senior citizens' centres, churches, and facilities for sport, recreation, and health. Voluntary health agencies (such as the local branches of any national organization that deal with cardiovascular disease, diabetes, and cancer) should be involved. These can serve as education conduits, with the

community's electronic and print mass media organizations assisting in message design, content, and delivery. The internet's community education role (digital health) is now a major force (Baker et al. 2003) and can reach and benefit more than traditional channels—an emerging variant of mass media. Interactive computer learning, including hand-held computers (King et al. 2008) ('mobile health') now provides new points of entry.

The intervention 'dose'

For future community intervention planning, an important question is how much amount and duration of intervention (the 'dose') is needed? Although outcome evaluations measure change achieved in health or disease risk, this 'dose' is rarely measured, thus cost-effectiveness remains a mystery. Dose measurement is difficult, the usual report gives only the intervention's duration and a cursory description of the educational inputs (Farquhar 2014), with little attention to reach or intensity (exposures). One can, at least, measure print and electronic media distributed, all converted to time, as was done uniquely in the Stanford Five-City Project (Farquhar et al. 1990). Public health policy would be served if this were done more commonly.

The community transformative phase

This begins early in the community organizing period, preceding the intervention, but accelerates throughout, particularly in the terminal period when external support diminishes or ceases (Farquhar et al. 1985). Enhancing a community's programme adoption for intervention sustenance is a major goal. This begins earlier, but becomes more definitive towards cessation of educational forces extrinsic to the community. A concurrent goal is to facilitate policy and regulatory changes that foster long-term goals and provide supportive physical infrastructures. The preferred outcome sees a continuing role for the original external academic advisors for community assistance, including the county's health agencies, to create new disease prevention programmes, as new technologies emerge. The ideal long-term outcome is that the 'transformed' community will advocate for local, regional, and national governmental regulations and laws increasing local intervention effects, such as smoke-free places, and extend them beyond the community.

The case of comprehensive community health promotion for CVD

The history of seven decades of twentieth-century community trials described has encompassed a mid-twentieth-century shift from public health and medical strategies appropriate to communicable diseases that accounted for the majority of deaths to those appropriate to a growing chronic disease burden. It outlined a series of formal research projects designed to affect health of communities, and illustrative CVD risk factors. The CVD trials have involved entire populations of at least one *education* community, compared to at least one *control* community. Three projects are described in greater detail here as case studies to provide examples of their methods and results. Dissemination worldwide into practical applications of community organizing and mass communication technologies, derived in part from these projects, occurred throughout these decades, with a relatively small number using research methods analogous to those described here.

Two examples: the first decade of CVD community-based projects

The first example is the Stanford Three-Community Study (3CS) in three small agricultural marketing towns in California (total population 45,000). The second is the North Karelia Study (NKS) in two adjoining predominately rural Finnish counties (North Karelia population about 180,000). Both projects began in 1972.

The first of two Stanford projects, 3CS, was carried out from 1972 to 1975 in both English and Spanish, comparing effects of 2 years of mass media plus ten-session risk reduction classes for 1 year for some high-risk adults in one community with the effects in a comparison community having 2 years of mass media alone, and a third as a control having neither (Farquhar et al. 1977; Schooler et al. 1997). Groups exposed to varied education amounts showed a dose–response change in levels of smoking, blood pressure, and blood cholesterol. Spanish-speaking residents had a proportionately larger effect than did the Anglo majority. A composite measure of CVD risk reduction at 2 years of about 23% and 30% occurred in the mass media-only and mass media-plus classroom conditions, respectively. The risk score (a probability) was derived from each adult's 'before–after' risk levels. These risk parameters predicted the 12-year future probability of a coronary heart disease event, using the scoring system

derived from the long-term Framingham Heart Study (Truett et al. 1967). Thus, a relatively modest amount of mass media alone over 2 years (about 30 television and radio ‘spots’, weekly newspaper columns on heart health, and four separate mass mailings of booklets) was sufficient to change the population’s coronary disease risk by 23%. The education followed the principles outlined in the earlier section ‘Prominent theories applied in community-based interventions’ for prominent theories of community organizing and health communication–behaviour change.

In Finland, NKS evaluated two unmatched rural counties that contained many villages with farming and lumbering as their main occupations. North Karelia (which has a population of about 180,000, with the highest CVD rates in Finland) received an education campaign that began in 1972, continuing to the present and expanded to the entire country. After 10 years, CVD risk factor changes comparable to the 3CS occurred, and significant net reductions in CVD events occurred (Puska et al. 1985; Schooler et al. 1997). Extensive community organizing, resulting in strong partnerships with residents and their organizations marked this study. The NKS influence on Finland’s national policies was unequalled among the CVD projects, giving its most important lesson—that a well-executed project led by respected scientists can change an entire country. Examples included Finland’s food and agricultural industries making large changes: fertilizers were supplemented with selenium (a substance needed for health, but low in Finland’s soil), milk pricing previously based on fat content converted to protein-based pricing, programmes were created to replace dairy farms with berry farms, a new canola industry replaced jobs lost in dairying, and increased production of low-saturated fat foods occurred (Puska et al. 1985, 1995). In 1972, population-wide nutrition change and smoking cessation interventions were an innovation internationally, which may partly explain the success of both 3CS and NKS (Green and Richard 1993).

Early international diffusion, 1977–1983

Building on the 3CS and NKS were studies carried out in various countries:

Italy (The Martignacco Project, 1977–1983, one treatment, mass media and screening, one control—coronary heart disease (CHD) risk fell in men only).
Australia (The North Coast Project, 1978–1980, similar to 3CS, one mass media, one mass media plus classes—effects on smoking, greater in the mass media 'plus' community).
Switzerland (two treatment, two control pairs, German and French speaking, with mass media, classes, and environmental changes—small effects on smoking, blood pressure, and obesity).
South Africa (similar to 3CS, one mass media, one mass media plus classes plus community events—decreased CHD risk, blood pressure, smoking; both treatments were equivalent).

All of these studies, reviewed in Schooler et al. (1997), reported significant risk factor changes, adding evidence both for the effectiveness of the 3CS model, and for the external validity of the evidence. In all, a predominately community-wide mass media campaign was effective, but a supplemental face-to-face instruction usually adds effects (Farquhar et al. 1991). A noteworthy qualification is that all these studies were done in rather small communities (population sizes of about 12,000 to 15,000 residents), thus achieving these effects may be more difficult in larger and more complex cities.

The second and third decades—projects begun in the 1980s and 1990s

The third case study, the Stanford Five-City Project (5CP) (California, United States), had an intervention phase of slightly more than 5 years, 1980–1986, and extended 3CS methods to larger populations (total population of about 360,000) with multifactor CVD prevention directed at two northern California cities, no intervention in three control cities (Farquhar et al. 1990). It differed from 3CS in larger community sizes, in greater use of community organizing, in greater collaboration with the communities' health, media and educational organizations in planning and implementing programmes, and in offering extensive continuing collaboration with Stanford. It was similar in its generous use of mass media, both print and electronic. The *proportion* of message sources received by the average adult over 5 years of education included: television and radio 67%, newspaper 28%, other print (such as booklets and 'tip' sheets) 4%, and face-to-face 1%. This dose measurement method, more fully described elsewhere

(Farquhar 2014), was unique among all reported CVD community intervention programmes in measuring total education dose. In the intervention communities this amounted to about 5 hours/year and about 100 exposure episodes/year to all forms of media and classroom education (Flora et al. 1989; Farquhar et al. 1990). This method's use increased evaluation costs, almost two-thirds of total expenses. This expense may explain why others have not used it to similar degrees. Public health education would be served were more campaigns analysed in this way (Brown et al. 2012).

The first year of 5CP's television messages (characterized as 'high reach/low involvement') led to the public's later use of print media, which contained more specific instructions than from television in skills needed for changing smoking, diet, exercise, or weight control (Flora et al. 1997). Television, however, was also quite effective in skills training. For example, the 5CP presented an 8-week 'live' smoking cessation TV programme, demonstrating smoking cessation skills (Altman et al. 1987). The results of an antismoking campaign of a 'quit smoking' contest, the television and face-to-face classes, and a widely distributed print 'Quit Kit' were analysed separately. The highest cost-effectiveness was achieved by the low-cost print component (Altman et al. 1987). Results were comparable to the 3CS (about a 15% fall in Framingham composite risk of CVD) (Truett et al. 1967), with a major impact on blood pressure (4–5%) and smoking (13%) (Farquhar et al. 1990). The usually timid health policy experts might gain courage from the 5CP's evidence that only 3 hours/year of high-quality television health education can counteract the public's relatively intimidating exposure to about 100 hours/year of television advertising devoted to unhealthful nutrition. The US population's exposure is estimated as 292 hours of TV advertisement as measured by the Nielsen monitoring system (Farquhar 2014). Michael Jacobson, from the Center for Science in the Public Interest, identified that one-third of these promoted unhealthful nutrition (personal communication).

Compared to similar projects, 5CP's greater effects at relatively low cost, support that communication–behaviour change theory's predicted effects may be explained by mass media presentation components, including adaptation of skills training aspects of Bandura's social cognitive theory. An important lesson for policymakers is that adult quitters saved over three times more money from their cessation of cigarette purchases (2,860 quitters @ $274/quitter/year) than the cost of the entire population-wide campaign

($4/adult/year), which would justify cigarette taxes as the campaign's payment. Savings are retained life-long by the community's smoker-quitters, not counting long-term health costs' savings, and absenteeism from the individual's employment decreases, all accruing to the whole community. The intervention's relatively low cost is again emphasized, since the total yearly cost of cigarettes to the 22,000 smokers (over $6 million) was 23 times greater than the yearly campaign costs (about $260,000/year). Additionally, the county's Health Department adopted 5CP's technologies, applying them to seatbelt promotion, changing restaurants' saturated fat use, violence prevention, and adolescent pregnancy prevention.

More variably sized CVD projects occurred in the 1980s and 1990s in the United States, Sweden, Denmark, Canada, Germany, the Czech Republic, and China, including World Health Organization-sponsored projects (Schooler et al. 1997). The Minnesota Heart Health Project (MHHP) (Luepker et al. 1994) emerged as an important innovation source, pioneering introduction of antismoking contests (Lando et al. 1990), and demonstrating synergism in school health intervention effects when multiple schools in a community are involved within a comprehensive community intervention (Perry 2000). Many projects borrowed heavily from and often received training from either the Stanford or the North Karelia groups, although changes seen were often less than in those two projects. Reasons for lower success may defy explanations (e.g. the magnitude of the interventions was incompletely described). For many, the cause was likely due to inadequate use of inherently cost-effective mass media.

To summarize lessons learned, a five-step approach emerges from community trials, much derived from the Stanford 5CP, and some from various worldwide trials. Generalizability, of course, has limitations inherent in the nature of such trials, including designs, locales, and methods:

Define the problem. Choose behaviour and environment change goals following community consultation.
Implement a 2-year campaign. Provide 3–5 hours of multichannel education/year. Year 1: 70% electronic media, 15% print, 15% community classes, contests, health fairs. Year 2: change to 40%, 35%, and 25%.
Evaluate. Use formative, process, and summative evaluations, tracking exposure, and effects of each intervention type. Use repeated cohort surveys to best measure effects, avoiding greater variance and in-migration's dilution effect. Repeated independent sample designs are feasible if in-migration dilution is less than 5%/year.
Institutionalize. Continue intervention. Create a community demonstration project as a regional resource centre.
Leverage. Advocate for local, regional, and national governmental regulations to magnify intervention effects.

Conclusions

Six decades of applying the 'total community' health promotion approach in developed countries show unequivocally that considerable change is achievable at reasonable cost. Well-designed projects have transferred educational technologies and social-behavioural intervention skills to a community's infrastructure (public health, media, schools, etc.), resulting in significant improvements in health risk factors and conditions of populations (Schooler et al. 1997). Although many studies were done in small communities, recent successes in New York City (Angell et al. 2009), suggest that success occurs in large urban populations. Adherence to theory is key: gains in self-efficacy of many individuals through education results in *community efficacy*, enhancing changed institutional policy and practice, thus maintaining and extending community change. Changing communities through organizing and education requires *advocacy*, *activism*, *partnership building*, and *leadership*; success is enhanced by *regulatory change*. When needed laws and regulations come from state, provincial, or national governments, local changes are enhanced. Both local and broad governmental approaches are needed to succeed in preventing chronic disease, which is particularly important for developing countries (Institute of Medicine 2010a). Adding well-planned, well-integrated digital health

programmes, as yet untried, can likely increase effectiveness at decreased cost.

REFERENCES

Altman, D.G., Flora, J.A., Fortmann, S.P., and Farquhar, J.W. (1987). The cost-effectiveness of three smoking cessation programs. *American Journal of Public Health*, 77, 162–5.

Angell, S.Y., et al. (2009). Cholesterol control beyond the clinic: New York City's trans fat restriction. *Annals of Internal Medicine*, 151, 129–34.

Baker, L., Wagner, T.H, Singer, S., and Bundorf, M.K. (2003). Use of the internet and e-mail for health care information: results from a national survey. *Journal of the American Medical Association*, 289, 2400–6.

Bandura, A. (1986). *Social Foundations of Thought and Action: A Social Cognitive Theory*. Englewood Cliffs, NJ: Prentice-Hall.

Berelson, B. (ed.) (1966). *Family Planning and Population Policies: A Review of World Developments*. Chicago, IL: University of Chicago Press.

Best, A., et al. (2003). An integrative framework for community partnering to translate theory into effective health promotion strategy. *American Journal of Health Promotion*, 18, 168–76.

Blackburn, H. (1979). Diet and mass hyperlipidemia: public health considerations—a point of view. In: Levy, R.R.B., Dennis, B., and Ernst, N. (eds.) *Nutrition, Lipids, and Coronary Heart Disease*, pp. 309–47. New York: Raven Press.

Brown, D.R., et al. (2012) Community Preventive Services Task Force. Stand-alone mass media campaigns to increase physical activity. A community guide updated review. *American Journal of Preventive Medicine*, 43, 551–61.

Cain, R.E., Schyulze, R.W., and Preston, D.B. (2001). Developing a partnership for HIV primary prevention for men at high risk for HIV infection in rural communities. *International Journal of Health Promotion and Education*, 8(2), 75–8.

Centers for Disease Control and Prevention (1999). *Best Practices for Comprehensive Tobacco Control Programs* (1st ed.). Atlanta, GA: US Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health.

Centers for Disease Control and Prevention (2007). *Best Practices for Comprehensive Tobacco Control Programs* (2nd ed.). Atlanta, GA: US Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health.

Chaloupka, F.J., Cummings, K.M., Morley, C., and Horan, J. (2002). Tax, price and cigarette smoking: evidence from the tobacco documents and implications for tobacco company marketing strategies. *Tobacco Control*, 11, i62–72.

Darrow, W.W., Montanea, J.E., and Sánchez-Braña, E. (2009). Coalition contract management as a systems change strategy for HIV prevention. *Health Promotion Practice*, 11, 867–75.

D'Onofrio, C.N. (1966). *Reaching our Hard to Reach—The Unvaccinated*. Berkeley, CA: California State Department of Public Health.

Dooris, M. (2009). Holistic and sustainable health improvement: the contribution of the settings-based approach to health promotion. *Perspectives in Public Health*, 129, 29–36.

Farquhar, J.W. (2014). Community-based health promotion. In: Ahrens, W. and Peugot, I. (eds.) *Handbook of Epidemiology* (2nd ed.). Berlin: Springer, pp. 419–38.

Farquhar, J.W., et al. (1990). The Stanford Five-City Project: effects of community-wide education on cardiovascular disease risk factors. *Journal of the American Medical Association*, 264, 359–

65.
Farquhar, J.W., et al. (1991). Methods of communication to influence behaviour. In: Holland, W.W., Detels, R., and Knox, G. (eds.) *Oxford Textbook of Public Health* (1st ed.). Oxford: Oxford University Press, pp. 331–44.
Farquhar, J.W., et al. (1977). Community education for cardiovascular health. *Lancet*, 1, 1192–5.
Farquhar, J.W., Maccoby, N., and Wood, P.D. (1985). Education and communication studies. In: Holland, W.W., Detels, R., and Knox, G. (eds.) *Oxford Textbook of Public Health* (1st ed.), pp. 207–21. Oxford: Oxford University Press.
Fisher, R.A. (1925). *Statistical Methods for Research Workers*. Edinburgh: Oliver & Boyd.
Fisher, R.A. (1926). The arrangement of field experiments. *Journal of the Ministry of Agriculture of Great Britain*, 33, 505–13.
Flora, J.A., Maccoby, N., and Farquhar, J.W. (1989). Communication campaigns to prevent cardiovascular disease: the Stanford studies. In: Rice, R. and Atkin, C. (eds.) *Public Communication Campaigns*. Beverly Hills, CA: SAGE Publications, pp. 233–52.
Flora, J.A., Saphir, M.N., Schooler, C., and Rimal, R.N. (1997). Toward a framework for intervention channels: reach, involvement and impact. *Annals of Epidemiology*, S7, S104–12.
Fortmann, S.P., et al. (1995). Community intervention trials: reflections on the Stanford Five-City Project. *American Journal of Epidemiology*, 142, 576–86.
George, M.A., Daniel, M. and Green, L.W. (2006). Appraising and funding participatory research in health promotion. *International Quarterly of Community Health Education*, 18(2), 181–97.
Golden, S.D. and Earp, J.A. (2012). Social ecological approaches to individuals and their contexts: twenty years of health promotion interventions. *Health Education & Behavior*, 39, 364–72.
Gray, R.M., Kesler, J.P., and Moody, P.M. (1966). The effects of social class and friends' expectations on oral polio vaccination participation. *American Journal of Public Health*, 56, 2028–32.
Green, L.W. (1970). *Status Identity and Preventive Health Behavior*. Pacific Health Education Reports No. 1. Berkeley, CA: University of California School of Public Health.
Green, L.W. (1983). New policies in education for health. *World Health*, April–May, 13–17.
Green, L.W. (1989). The theory of participation: a qualitative analysis of its expression in national and international health policies. In: Patton, R.D. and Cissell, W.B. (eds.) *Community Organization: Traditional Principles and Modern Applications*, pp. 48–62. Johnson City, TN: Latchpins Press.
Green, L.W., Brancati, F.L., Albright, A., Primary Prevention of Diabetes Working Group (2012). Primary prevention of diabetes: integrative public health and primary care opportunities, challenges and strategies. *Family Practice*, 29(Suppl. 1), i13–23.
Green, L.W., Daniel, M., and Novick, L. (2001). Partnerships and coalitions for community based research. *Public Health Reports,* 116, 15–26.
Green, L.W., Gustafson, H.C., Griffiths, W., and Yaukey, D. (1972). *The Dacca Family Planning Experiment*. Berkeley, CA: School of Public Health, University of California.
Green, L.W. and Kreuter, M.W. (1992). CDC's planned approach to community health as an application of PRECEDE and an inspiration for PROCEED. *Journal of Health Education*, 23(3), 140–7.
Green, L.W. and Kreuter, M.W. (2005). *Health Program Planning: An Educational and Ecological Approach* (4th ed.). St. Louis, MO: McGraw-Hill Publishing Co.
Green, L.W., Ottoson, J.M., Garcia, C., and Hiatt, R. (2009). Diffusion theory and knowledge dissemination, utilization and integration. *Annual Review of Public Health*, 30, 151–74.
Green, L.W., Potvin, L., and Richard, L. (1996). Ecological foundations of health promotion. American *Journal of Health Promotion*, 10(4), 270–81.
Green, L.W. and Richard L. (1993). The need to combine health education and health promotion: the case of cardiovascular disease prevention. *Promotion and Education: International Journal of Health Promotion and Education*, 1, 11–17.

Hochbaum, G. (1958). *Public Participation in Medical Screening Programs.* Public Health Service Publication No. 572. Washington, DC: US Government Printing Office.
Holder, H.D., et al. (1997). Summing up: lessons from a comprehensive community prevention trial. *Addiction*, 92(S2), S293–S301.
Holder, H.D. and Wallack, L. (1986). Contemporary perspectives for prevention of alcohol problems: an empirically-derived model. *Journal of Public Health Policy* 7, 324–30.
Hovland, C.I. (1948). Social communication. *Proceedings of the American Philosophical Society*, 92, 371–5.
Hyland, A., et al. (2006). Cigarette purchase patterns in four countries and the relationship with cessation: findings from the International Tobacco Control (ITC) Four Country Survey. *Tobacco Control*, 15(Suppl. 3), iii59–64.
Institute of Medicine (2010a). *Promoting Cardiovascular Health in the Developing World: A Critical Challenge to Achieve Global Health*. Washington, DC: National Academies Press.
Institute of Medicine (2010b). *Bridging the Evidence Gap in Obesity Prevention: A Framework to Inform Decision Making*. Washington, DC: National Academies Press.
Institute of Medicine (2012a). *An Integrated Framework for Assessing the Value of Community-Based Prevention*. Washington, DC: National Academies Press.
Institute of Medicine (2012b). *Measuring Progress in National and Community Efforts in Obesity Control*. Washington, DC: National Academies Press.
Kagan, J.M., et al. (2012). Community–researcher partnerships at NIAID HIV/AIDS clinical trials sites: insights for evaluation and enhancement. *Progress in Community Health Partnerships*, 6, 311–29.
King, A.C., et al. (2008). Promoting physical activity through hand-held computer technology. *American Journal of Preventive Medicine*, 34, 138–42.
Kok, G., Gottlieb, N.H., Commers, M., and Smerecnik, C. (2008). The ecological approach in health promotion programs: a decade later. *American Journal of Health Promotion*, 22, 437–42.
Kokko, S., Kannas, L., and Villberg, J. (2006). The health promoting sports club in Finland—a challenge for the settings approach. *Health Promotion International*, 21, 219–29.
Kotler, P. (1975). *Marketing for Nonprofit Organizations*. Englewood Cliffs, NJ: Prentice-Hall.
Lando, H.A., et al. (1990). Community impact of a localized smoking cessation contest. *American Journal of Public Health*, 80, 601–3.
Lefebvre, R.C. and Flora, J.A. (1988). Social marketing and public health interventions. *Health Education Quarterly*, 15, 299–315.
Levy, D., et al. (2012). Italy SimSmoke: the effect of tobacco control policies on smoking prevalence and smoking attributable deaths in Italy. *BMC Public Health*, 12, 709.
Louriero, I. (2004). A study about effectiveness of health promoting schools network in Portugal. *Promotion and Education* 11(2), 85–92.
Luepker, R.V., et al. (1994). Community education for cardiovascular disease prevention: risk factor changes in the Minnesota Heart Health Program. *American Journal of Public Health*, 84, 1383–93.
Lynagh, M., Schofield, M.J., and Sanson-Fisher, R.W. (1997). School health promotion programs over the past decade: a review of the smoking, alcohol and solar protection literature. *Health Promotion International*, 12, 43–59.
Maccoby, E.E. (1951). Television: its impact on school children. *Public Opinion Quarterly*, 15, 425.
Max, W., Sung, H.-Y., and Lightwood, J. (2013). The impact of changes in tobacco control funding on healthcare expenditures in California, 2012–2016. *Tobacco Control*, 22(e1), e10–15.
Mercer, S.M, et al. (2007). Study designs for effectiveness and translation research: identifying trade-offs. *American Journal of Preventive Medicine*, 33(2), 139–54.
Michael, D.N. and Maccoby, N. (1953). Factors influencing verbal learning from films under varying conditions of audience participation. *Journal of Experimental Psychology*, 46, 411–18.

Michaud, P.-A. (2003). Prevention and health promotion in school and community settings: a commentary on the international perspective. *Journal of Adolescent Health*, 33, 219–25.
Miller, H.G., Turner, C.F., and Moses, L.E. (eds.) (1990). *AIDS: The Second Decade*. Washington, DC: National Academies Press.
Minkler, M. (ed.) (2005). *Community Organizing and Community Building for Health* (2nd ed). New Brunswick, NJ: Rutgers University Press.
Minkler, M. and Wallerstein, N. (2008). *Community-Based Participatory Research for Health* (2nd ed.). San Francisco, CA: Jossey-Bass.
Mosquera, J., et al. (2012). An inside look at active transportation in Bogotá: a qualitative study. *Journal of Physical Activity and Health*, 9, 776–85.
Moynihan, D.P. (1969). *Maximum Feasible Misunderstanding: Community Action in the War on Poverty*. New York: Free Press.
Northcutt, T.J., Jenkins, C.D., and Johnson, A.L. (1964). Factors influencing vaccine acceptance. In: Neill, J.S. and Bond, J.O. (eds.) *Hillsborough County Oral Polio Vaccine Program*. Monograph No. 6. Jacksonville, FL: Florida State Board of Health.
Novak, N.L. and Brownell, K.D. (2011). Taxation as prevention and as a treatment for obesity: the case of sugar-sweetened beverages. *Current Pharmaceutical Design*, 17(12), 1218–22.
Perry, C. (2000). Commentary: the school as a setting for health promotion. In: Poland, B.D., Green, L.W., and Rootman, I. (eds.) *Settings for Health Promotion: Linking Theory to Practice*. Thousand Oaks, CA: SAGE Publications, pp. 120–6.
Poland, B.D., Green, L.W., and Rootman, I. (2000). Reflections on settings for health promotion. In: Poland, B., Green, L.W., and Rootman, I. (eds.) *Settings for Health Promotion: Linking Theory and Practice*. Newbury Park, CA: SAGE Publications, pp. 341–51.
Puska, P. (1995). Experience with major subprogrammes and examples of innovative interventions. In: Puska, P., Nissinen, A., and Vartianinen, E. (eds.) *The North Karelia Project*. Helsinki: Helsinki University Printing House, pp. 159–67.
Puska, P., et al. (1985). The community-based strategy to prevent coronary heart disease: conclusions from the ten years of the North Karelia Project. *Annual Review of Public Health*, 6, 147–93.
Puska, P., et al. (1981). *The North Karelia Project: Evaluation of a Comprehensive Community Programme for Control of Cardiovascular Diseases in 1972–1977 in North Karelia, Finland*. Copenhagen: World Health Organization.
Ramirez, A.G., McAlister, A., Gallion, K.J., and Villarreal, R. (1995). Targeting Hispanic populations: future research and prevention strategies. *Environmental Health Perspectives*, 103(Suppl. 8), 287–90.
Rogers, E. (2003). *Diffusion of Innovations* (5th ed.). New York: Free Press.
Rogers, T. (2010). The California Tobacco Control Program: introduction to the 20-year retrospective. *The Quarter that Changed the World: Celebrating 20 Years of California Tobacco Control Program, special issue of Tobacco Control*, 19 (Suppl. 1), i1–i2.
Rosenstock, I.M., Derryberry, M., and Carriger, B.A. (1959). Why people fail to seek poliomyelitis vaccination. *Public Health Reports*, 74, 98–103.
Ruiz, M.S., et al. (2000). *No Time to Lose: Getting More from HIV Prevention*. Washington, DC: National Academies Press.
Ryan, B. and Gross, N.C. (1943). The diffusion of hybrid seed corn in two Iowa communities. *Rural Sociology*, 8, 15–24.
Schooler, C., Farquhar, J.W., Fortmann, S.P., and Flora, J.A. (1997). Synthesis of findings and issues from community prevention trials. *Annals of Epidemiology*, S7, S54–68.
Shea, S. and Basch, C.E. (1990). A review of five major community-based cardiovascular disease prevention programs. Part I. Rationale, design, and theoretical framework. *American Journal of Health Promotion*, 4, 203–13.

Slater, M.D. and Flora, J.A. (1991). Health lifestyles: audience segmentation analysis for public health interventions. *Health Education Quarterly*, 18, 221–33.

Stycos, J.M. (1962). A critique of the traditional planned parenthood approach in underdeveloped areas. In: Kiser, C.V. (ed.) *Research in Family Planning*, pp. 477–501. Princeton, NJ: Princeton University Press.

Thompson, B., et al. (2000). Durability of tobacco control efforts in the 22 Community Intervention Trial for Smoking Cessation (COMMIT) communities 2 years after the end of intervention. *Health Education Research*, 15(3), 353–66.

Trickett, E.J., et al. (2011). Advancing the science of community-level interventions. *American Journal of Public Health*, 101, 1410–19.

Truett, J., Cornfield, J., and Kannel, E. (1967). Multivariate analysis of the risk of coronary heart disease in Framingham. *Journal of Chronic Disease*, 20, 511–24.

US Department of Health and Human Services (2000). *Reducing Tobacco Use: A Report of the Surgeon General*. Atlanta, GA: US Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health. [Online] Available at: http://www.cdc.gov/tobacco/data_statistics/sgr/2000/complete_report/index.htm

Volpp, K.G., et al. (2009). A randomized, controlled trial of financial incentives for smoking cessation. *The New England Journal of Medicine*, 360, 699–709.

World Health Organization (WHO) (1954). *Expert Committee on Health Education of the Public.* WHO Technical Report Series No. 89. Geneva: World Health Organization.

World Health Organization (WHO) (1986). *Report of a WHO Expert Committee. Community Prevention and Control of Cardiovascular Disease.* Technical Report Series No. 732. Geneva: World Health Organization.

World Health Organization, Health and Welfare Canada, and Canadian Public Health Association (1986). *Ottawa Charter for Health Promotion: An International Conference on Health Promotion—The Move Towards a New Public Health, Nov. 17–21,* Ottawa. Geneva: World Health Organization.

5.9

Natural and quasi-experiments

Peter Craig

Introduction

Natural experiments, broadly defined as events or processes outwith the control of a researcher, which divide a population into exposed and unexposed groups (Craig et al. 2012), have long been used in epidemiology and public health. John Snow's studies of cholera in mid-nineteenth century London (Snow 1855), the most celebrated example of early epidemiology (Davey Smith 2002), exploited a natural experiment brought about by a change in domestic water supply between two cholera outbreaks. At the time, cholera was widely believed to be an airborne disease. Snow used information about cholera deaths and water supply at neighbourhood and household level to show that deaths were more likely in households whose water came from a contaminated source. Although he did not use the term, Snow clearly understood his approach as a natural experiment:

> No fewer than 300 000 people of both sexes, of every age and occupation and every rank and station, from gentle folks down to the very poor, were divided into two groups without their choice, and in most cases without their knowledge; one group being supplied with water containing the sewage of London … , the other group having water quite free from such impurity. To turn this grand experiment into account all that was required was to learn the supply of water to each individual house, where a fatal attack of cholera may occur. (Snow 1855)

What makes this 'grand experiment' useful is that the change in exposure was well-characterized and independent of other factors that may influence the health outcome. In another important early example of turning natural experiments into account, GCM M'Gonigle compared changes in death rates among residents of an 'Unhealthy Area' of a town in the North East of England who were rehoused following a survey of housing conditions, with

changes among residents who remained in their original houses (M'Gonigle 1933). The division of the area was, according to M'Gonigle, purely administrative, and death rates prior to rehousing in 1927 were similar among the two groups of households. In the five years following the rehousing, the death rate was markedly higher among the rehoused residents, but declined slightly in those who remained. Further investigation suggested that increased rents leading to inadequate diets among adults in the rehoused families was the probable cause.

Despite these early demonstrations of the potential of natural experiments for addressing important public health questions, and a substantial body of epidemiological work from the 1940s onwards, including influential studies of famines and mass migrations, the methodology of evaluating natural experiments has developed piecemeal across a range of subject areas and disciplines, including psychology, economics and political science (Bärnighausen et al. 2017b). Developments in these other disciplines have been slow to permeate public health. It is only in the last ten years or so that attempts have been made to pull together the range of methods available to study the different kinds of natural experiments that are of interest to public health, and to identify the circumstances in which each method works best (Craig et al., 2012; Bärnighausen et al. 2017b; Basu et al., 2017; Craig et al., 2017). This chapter introduces the main approaches that researchers can use to evaluate natural experiments, and provides examples of their application in a wide range of contexts. It begins by setting out the advantages, and also the limitations, of a natural experimental approach. The term *quasi-experiment* is often used interchangeably with *natural experiment*, but is also used to refer to true experiments (in which there is manipulation of exposure by the researcher) which are non-randomized. To avoid confusion, the preferred term in this chapter is *natural experiment*.

Trials, observational studies, and natural experiments

Randomized controlled trials and syntheses of evidence from trials occupy the top steps of the conventional hierarchy of evidence (Concato et al. 2000). The hierarchy is organized according to risk of bias, with observational studies regarded as being at higher risk because of the difficulty of isolating the effect of the exposure of interest from other

factors that might affect outcomes. In a trial, random assignment of participants to treatment and control groups, plus a number of other methodological safeguards such as blinding of participants and assessors, and concealment of the allocation sequence, generates exposed and unexposed groups that are, in expectation, balanced in terms of all the factors that may influence outcomes other than exposure to the treatment. Differences in outcomes between the groups can therefore be interpreted as effects of the treatment. Outside an experimental setting, such balance is rare: individuals may be selected for treatment, or may self-select, according to characteristics that influence outcomes, leading to confounding of the treatment effect. In an observational study, statistical adjustment may be used to address such confounding, but only works to the extent that the confounders are known and accurately measured. Estimates of the treatment effect will be biased if there is 'residual confounding' due to unknown or imperfectly measured confounders (see Chapter 5.13). Natural experimental studies are a subset of observational studies that use the process by which exposure is determined to identify treatment effects. Understanding this process is therefore key to the design of a good natural experimental study. Because such processes rarely give rise to exposed and unexposed groups that are perfectly balanced on all of the characteristics that may affect outcomes, natural experiments also rely on statistical models that take into account both the allocation process and any observed differences between the groups in characteristics that may affect outcomes.

Given that randomized controlled trials are such an effective way of eliminating bias associated with selective exposure to treatment, why should natural experiments be considered a useful alternative? One reason is that for many interventions a trial may not be feasible, for ethical, political, or practical reasons (Craig et al. 2012; Bärnighausen et al. 2017b; Derzon 2018). It may be regarded as unethical to withhold a treatment in order to identify its effect on health, if it clearly provides other benefits. This particularly applies to non-healthcare interventions whose main purpose is to influence some other outcome such as participation in education or in the labour market. Recipients of a conditional cash transfer benefit directly from receiving the money. If the purpose of the scheme is to reduce poverty, withholding payments from some potential recipients to study its effect on health, once its effectiveness in alleviating poverty has been established, is unlikely to be seen as fair by the losers, and may also be

unacceptable to other stakeholders. Similar arguments apply to evaluating the wider, non-health-related, or long-term impacts of health interventions. The design of the intervention may make randomization impractical, at least at an individual level—as in the case of international treaties, national legislation, or taxation policies—or stakeholders may be unwilling to selectively withhold an intervention once it has been implemented, even if its effectiveness in delivering its primary outcome is uncertain.

Natural experimental studies, which involve no manipulation of exposure by the researcher and which can be conducted retrospectively, may therefore be feasible when a prospective randomized controlled trial would not. Natural experiments may also avoid threats to validity that are specific to trials. One is the confounding of trial procedures with treatment. This may occur when methods for retaining participants in the trial also improve retention in care among patients in the control arm, leading to an underestimate of the effectiveness of treatment (Bor et al. 2017). Another is that effects on individuals observed in a trial may differ from the effect of exposing the whole population to the intervention. This may occur when interactions between recipients contribute to the overall treatment effect. To use the conditional cash transfer example again, treating a whole population may change social norms regarding the behavioural conditions, strengthening the individual incentive of the cash transfer.

Another reason for considering alternatives to randomized trials is that minimizing bias is not the only aim of good research design. The strength of randomized controlled trials is that they permit strong causal inference with a minimum of assumptions. But additional assumptions are always needed to apply the evidence to a particular context, if it differs in any way from the settings in which the trials were conducted (Deaton and Cartwright 2018). Researchers evaluate public health interventions to provide information that is useful for decision-makers, as well as to contribute to a scientific evidence base. Precise estimates with a small amount of bias from a large natural experimental study, conducted in a relevant context, may be more useful than unbiased but imprecise estimates from an experimental study conducted in a highly artificial setting. Whether this, or any of the other reasons for considering natural experimental approach, is persuasive will depend on the details of the case. Rather than seeing trials and natural experiments as competing for position at the top of a hierarchy of evidence,

researchers and evidence users should see them as belonging to a common methodological toolkit for answering important public health questions.

Natural experimental study designs

The key methodological problem that all natural experimental studies must solve is that of selective exposure to the intervention, and the designs summarized in this section can be seen as alternative solutions to this challenge. They include some that would be ruled out by definitions that restrict natural experiments to cases where a process of 'as-if randomization' determines exposure (Dunning 2012), or to methods that deal with unobserved confounding as well as with confounding by observables (Tugwell et al. 2017). 'As-if randomization' is difficult to define, and rules out cases where the assignment process is clearly non-random, but amenable to modelling. Non-random assignment can provide useful opportunities for natural experimental studies, so long as the assignment process is well understood. None of the available methods automatically solves the selection problem. How far they do so depends on how well their underlying assumptions are met (Bärnighausen et al. 2017a). As this is unlikely to be perfect, or completely demonstrable from available data, a hard and fast distinction between methods that control for both unobserved and observed confounders, and those that only deal with observables, is unhelpful. A more inclusive list is also in keeping with the idea that the value of natural experimental approaches is that they widen the range of public health interventions that can usefully be evaluated. Because the methods originated in a number of different disciplines, terms that are readily understood by some researchers may be unfamiliar or puzzling to others. Definitions of some key terms, according to their usage in this chapter, are provided in Table 5.9.1.

'Real-world' randomization

Sometimes, lotteries or ballots are used in the delivery of an intervention or implementation of a policy. A well-known example is the Vietnam-era draft lottery which has been used to identify the effects of military service on

earnings (Angrist 1990). Another is the practice of reserving council seats for women in a randomly selected proportion of village elections in the Indian states of Rajasthan and West Bengal. Researchers have used the system to study the effects of women's involvement in local government on the provision of public goods (Chattopadhyay and Duflo 2004). An example in the health field is the Oregon Health Plan Lottery in which randomly selected entrants to a public lottery were eligible to apply for health insurance. The lottery broke the link between income, prior health status and health insurance, enabling researchers to identify the effect of health insurance on use of health services and on health outcomes (Baicker et al. 2013).

Randomization may also come about inadvertently, as a consequence of errors in policy implementation. Database errors led to some families being randomly excluded from Brazil's *Bolsa Alimentação*, a conditional cash transfer programme that sought to improve child nutrition. Researchers compared the growth of children in households that were accidentally excluded from the programme and in matched households which received the payments (Morris et al. 2004). Resource constraints may sometimes limit programme reach in arbitrary ways that approximate random assignment. In Scotland's Family Nurse Partnership, which provides teenage first-time mothers with additional support pre-natally and in the first few years of the child's life, access to the programme is determined by the availability of specially trained nurses. When all the trained nurses have a full caseload, recruitment to the programme stops. Assignment should therefore be unrelated to characteristics of the mothers that might affect outcomes (Wimbush et al. 2015). Real-world randomization is rare, but may offer valuable opportunities for natural experiments. In studies that seek to exploit such opportunities, it is important to understand in detail how the assignment process works, and to check for self-selection by potential recipients, or manipulation of assignment by those delivering the programme.

Table 5.9.1 Glossary

Term	Usage
Assignment, allocation	The process that determines whether units (which may be individual people, or aggregates of some kind) are exposed to the intervention that is being studied. Also referred to as the data generating process.
Confounding, confounder (observed, unobserved)	Confounding refers to the mixing up of the effects of the intervention with the effects of characteristics associated with exposure. Confounding by observed characteristics (i.e. those on which data is available) can be addressed using a statistical model. Some natural experimental study designs go a step further, and use a combination of assumptions about the assignment process and statistical modelling to address confounding by unobservables.
Control, counterfactual	Controls are unexposed units. Outcomes among the controls constitute the counterfactual, in the sense that they represent the outcomes that would occur in the absence of exposure.
Exposure	A general term for receipt of an intervention; effectively synonymous with treatment in natural experimental studies.
Identification, identifying assumption	Identification refers to the method used to obtain an estimate of the effect of the intervention. All such methods rely on identifying assumptions, such as the assumption that pre-intervention trends in outcomes will continue in the absence of intervention. Testing how well the assumptions are met is an important element of good natural experimental study design.
Intervention	A general term for any policy, programme, service, or treatment that is being evaluated.
Natural experiment	An event or process that divides a population into exposed and unexposed subpopulations. The division may be spatial, temporal, or based on the characteristics of individual units, such as test scores.
Observational study	A study that does not involve any manipulation of exposure for research purposes. Natural experimental studies are a subset of observational studies.
Outcome, potential outcome	Outcome is used as a general term for the effect of treatment or exposure to the intervention being studied. Potential outcomes are the outcomes that would occur in the absence of treatment. Comparison of actual with potential outcomes is the basis for identifying treatment effects in a natural experimental study, with potential outcomes measured in a group that is not exposed to the intervention but is otherwise similar to the exposed group.
Quasi-experiment	Natural experimental studies are often referred to as quasi-experiments, but quasi-experiment is also sometimes used to refer to non-randomized experiments. To avoid confusion, the term natural experiment is preferred in this chapter.

Term	Usage
Selection	A process that leads exposed and unexposed units to differ in ways other than exposure to the intervention that are associated with differences in outcome. Selection is the source of confounding in natural experimental studies.
Treatment	Treatment is often used as a general term for exposure to an intervention, rather than to denote a medical treatment.

Controlling for observables: regression adjustment and propensity scores

When it is plausible to expect that there are no important differences between exposed and unexposed groups a standard regression model may be sufficient to identify the effect of the intervention. The real-world randomization examples discussed in the previous section are one kind of situation in which this assumption is reasonable. Another is when randomization is built into the timing of data-gathering rather than the timing of exposure. Goodman and colleagues used this approach to evaluate a school-based cycle training scheme (Goodman et al. 2016). Parents of some children in participating schools were interviewed as part of a long-running birth cohort study before the training was implemented in their school; parents of other children in the same schools were interviewed after the training had been delivered. The researchers used Poisson regression to estimate the effects of training on cycling behaviour, with adjustment for a wide range of potential confounders. Unlike previous evaluations which compared children in participating and non-participating schools, the study found no effect of the training on cycling behaviour.

As with real-world randomization, such opportunities are scarce but potentially highly valuable. In the example, data were gathered only from participating schools so selective participation cannot bias the estimate of effect of the programme (though may limit its generalisability). The key identifying assumption is that the timing of interviews was unrelated to delivery of training, a reasonable assumption given that the survey and the training were planned and conducted independently of one another.

When random allocation cannot be assumed, or is known not to occur, as in the case of many social interventions that are targeted at poorer households or poorer neighbourhoods, an alternative is to model the

allocation mechanism using a propensity score, defined as the conditional probability of exposure, given a number of covariates. The propensity score is usually estimated using a logistic regression model. It can be used to match exposed with unexposed units, to stratify analyses of outcomes, or as a covariate in an outcomes model. Propensity score-based methods have been extensively used to evaluate the impacts of targeted social programmes. In their evaluation of the Sure Start programme, an area-based intervention to improve the health and wellbeing of young children in England, Melhuish and colleagues calculated a propensity score from 85 area-level variables, which they used to stratify their analyses of the effects on individuals of exposure to the programme (Melhuish et al. 2008). The key identifying assumption is that the propensity score fully accounts for exposure to the programme—a strong assumption when data are sparse. Although propensity score-based methods will produce effect estimates similar to those of conventional regressions using the same covariates, they encourage a clearer conceptualization of the assignment process. A disadvantage is that units that cannot be matched on the propensity score must be omitted from the analyses.

Difference in differences and fixed effects models

When data are available on exposed and unexposed units before and after the introduction of an intervention, the difference in the change in outcomes can be used to identify the effect of the intervention. This difference in differences (DiD) approach is one of the most widely used natural experimental study designs. Its key assumption, known as the common trends assumption, is that any group specific confounders are time invariant, and any time-varying confounders are common to all groups. Effect estimates can be obtained from a regression model with dummy variables for group (exposed, unexposed) and period (pre/post-intervention), and a group-by-period interaction term that serves as the DiD estimator. The model takes account of any differences in unobserved fixed characteristics of the groups. Additional covariates can be added to take account of observed characteristics of the groups that vary over time. Researchers have developed a range of ways in which the common trends assumption can be tested (Wing et al. 2018). With data on covariates, the

standard model can be fitted with the outcome replaced by each covariate in turn to test for changes in the composition of the groups. With data on multiple time points, the outcome model can be fitted to the pre-intervention data with time as a continuous variable, to identify differences between the groups in the pre-intervention trends. The common trends assumption may also be violated by other interventions introduced at the same time as the intervention under study that affect one or other of the groups. This cannot be dealt with statistically. Instead, researchers should use specialist knowledge of the subject area to identify such threats.

DiD methods are widely applied to the evaluation of national programmes or policies, or interventions introduced at state level within federal jurisdictions. Schram and colleagues used a DiD approach to estimate the effect of joining the World Trade Organization, and subsequent removal of controls on foreign direct investment, on sugar-sweetened beverage (SSB) consumption in Vietnam (Schram et al. 2015). Models comparing SSB consumption in Vietnam and the Philippines in the pre- and post-accession periods indicated a substantially greater increase in Vietnam. The authors included gross domestic product (GDP) as a covariate in the models and fitted further models for foreign company sales growth and consumption of unprocessed foods. They found that the increase in soft drink consumption was not explained by economic growth, and was specific to products manufactured and marketed by foreign companies. The additional analyses support the conclusion that increased consumption of SSBs was the result of trade and investment liberalization.

The two-group, two-period DiD described here can be extended to situations where there are data on multiple groups and multiple time periods. The standard approach is to estimate a regression model with a variable denoting exposure and a complete set of group and time fixed effects to take account of unobserved time invariant characteristics of the groups, and temporal trends that apply to all groups. Nandi and colleagues used this approach to evaluate the effects of paid maternity leave on infant mortality in low- and middle-income countries (Nandi et al. 2016). Their models included fixed effects of country and year, a term indicating whether countries had changed their maternity leave policies in the previous year, plus additional covariates to take account of characteristics of countries that could vary over time, such as the level of wage replacement

and spending in healthcare. The research found that each month of paid maternity leave was associated with a 13% fall in infant mortality.

Webster and colleagues used a similar approach to evaluate the impact on murder rates of withdrawing a restriction on the purchase of handguns in the US State of Missouri (Webster et al. 2014). The researchers fitted models comparing Missouri with all other US states, with fixed effects for state and year, and a range of covariates including unemployment, burglary, and incarceration rates, levels of policing, and other firearm-related legislation. They found a 16–23% increase in gun-related murders associated with withdrawal, but no change in murders by other means. As with the study of trade liberalization and soft drink consumption, this use of outcomes that are not expected to respond to the intervention in order to test the specificity of the effects helps to strengthen confidence in the conclusions.

Synthetic controls

DiD and fixed effects approaches allow the effects of an intervention to be identified under the assumption that any differences between the exposed and unexposed groups other than the intervention are stable over time, or can be modelled using information on covariates. Choosing a comparator which is similar to the exposed group, such as a neighbouring state, or excluding groups that are known to have been affected by other interventions, can increase confidence in this assumption. An alternative approach is to create a synthetic control from a pool of potential comparators (Abadie et al. 2015). A synthetic control is a weighted composite of control units that matches the pre-intervention trend in the outcome in the exposed unit. The weights are constructed to minimize the pre-intervention differences in the outcome and observed predictors of the outcome in the exposed unit and the synthetic control. Matching on the pre-intervention outcome should also minimize differences in unobserved time-varying characteristics associated with the outcome. The difference in the post-intervention trend between the exposed unit and its synthetic control provides the effect estimate. Although standard statistical inference is not possible with this method, the estimated effect can be compared with a distribution of ‘placebo’ effect estimates obtained by swapping the

intervention unit with each control unit in turn and constructing a synthetic control for each.

The first application of synthetic control methods in public health was in a study of a tobacco control programme introduced in California (Abadie et al. 2010). Abadie and colleagues derived their synthetic control from a donor pool of other US states. After excluding those that had implemented similar programmes, the synthetic control was constructed from a weighted combination of five states. The difference in the post-intervention trend in cigarette sales between California and its synthetic control was greater than differences between the other states and their synthetic counterparts, suggesting that the programme was effective. Since that study was published, the use of synthetic control methods, often applied in conjunction with other methods, has grown rapidly (Bouttell et al. 2018). Comparing results of approaches that rely on differing assumptions is a useful way both of strengthening confidence in findings when no single method is unambiguously better, and of learning about the strengths and weaknesses of the methods.

Interrupted time series

Alongside DiD, interrupted time series (ITS) methods are among the most widely used approaches to evaluating natural experiments. They can be used to evaluate the effect of interventions that take place at a well-defined timepoint, and affect a population for whom outcome data are available at regularly spaced intervals. Whereas in DiD studies, the pre-intervention data are used to test whether the common trends assumption is reasonable, ITS studies incorporate the data directly into the outcome model. The simplest approach is to fit a segmented regression model to the sequence of outcome data, with terms for time since start of the observation period, time period (pre/post-intervention) and time post-intervention (Bernal et al. 2017). The coefficients of these terms estimate the pre-intervention trend in the outcome, the change in level of the outcome at the point at which the intervention is introduced, and the change in trend post-intervention.

Inclusion of the pre-intervention trend in the time series model controls for fixed characteristics of the exposed population. Additional covariates can be added and interacted with the time variables to account for observed

characteristics that change more quickly, or for other interventions that might affect the outcome of interest. An unexposed group can be added to control for unobserved time-varying confounders. Outcomes that are not expected to change as a result of the intervention, but may be affected by unobserved underlying trends, can be modelled in the same way as the outcome of interest to test for specificity of effects. Similar changes seen across all the outcomes are likely to represent unobserved confounding rather than an intervention effect. Other modifications can be made to account for seasonal patterns and other forms of serial correlation which are common in time series data. Whatever approach is taken, choosing an appropriate outcome model, based on reasonable assumptions about how the intervention is likely to impact on the outcome, and what other factors need to be taken into account, is key to obtaining useful results. For example, if the effect is expected to take place after a lag, observations in the lag period can be omitted or modelled as a separate segment of the series.

Humphreys and colleagues used a controlled ITS approach to evaluate the impact on homicide rates of a ‘stand your ground’ law enacted in the US State of Florida in 2005 (Humphreys et al. 2017a, b). Such laws, which permit the use of lethal force in self-defence, have been enacted in a number of US states, but previous evaluations had produced conflicting findings. The study used segmented regression models to analyse trends in homicide and suicide rates in Florida and four other US states for which comparable data were available, but which did not have stand your ground laws. The models assumed an immediate and permanent impact of the new law, and allowed for seasonal effects and serial correlation in the data. An interaction term was included to identify changes in homicide and suicide rates in Florida, relative to the control states. Results suggested a 24.4% increase in the monthly homicide rate but no increase in suicide rates in Florida, and no increase in either homicide or suicide rates in the control states. The specificity of the increase strongly suggests a causal interpretation.

Regression discontinuity

DiD, fixed effects, synthetic controls, and ITS methods can be used when the introduction or, more rarely, withdrawal (Craig et al. 2018) of an intervention divides a population into unexposed (i.e. pre-intervention or post-withdrawal) and exposed subpopulations. Other approaches can be applied when the assignment process operates at an individual level, the most widely used being regression discontinuity (RD) designs and instrumental variables (IVs). RD designs can be applied when a cut-off in a continuous variable, such as age or income, is used to determine exposure. Such methods of assignment are commonly used in public policy. Examples include social security benefits awarded on the basis of a means test, school places allocated using a test of cognitive ability, and laws that define a minimum age at which people can hold a driving licence or buy alcohol. The identifying assumption is that individuals with values of the assignment variable close to the cut-off will be similar in terms of other characteristics, both observed and unobserved, that may affect outcomes, especially if there is random error in the measurement of the assignment variable. Differences in outcomes for individuals either side of the cut-off can therefore be used to estimate a treatment effect (Bor et al. 2014).

When probability of exposure changes abruptly at the cut-off from close to zero to close to one, treatment effects can be estimated from a regression of outcomes on the assignment variable (known as the ‘forcing variable’ in RD studies), with a dummy variable denoting exposure. The coefficient of the dummy variable is the RD estimator. Additional terms may be included to allow the slopes to vary above and below the cut-off, and to control for confounders on which information is available. This is known as a sharp RD design. In many cases, known as fuzzy RD designs, probability of exposure takes intermediate values within a range of values of the forcing variable either side of the cut-off, and a two-stage approach to estimation is needed. In the first stage, probability of exposure is estimated from a regression of actual treatment on the cut-off and the forcing variable. In the second stage, the estimated probabilities of exposure replace the dummy exposure variable in the outcomes model.

Care is needed in the interpretation of effect estimates from RD studies. In both sharp and fuzzy RD designs, the effect of exposure to the intervention is identified at the cut-off. In a sharp RD design, the effect estimate can be interpreted as the average effect of treatment across the whole population, so long as it is independent of the forcing variable. In

practice, RD estimates are obtained from local regressions, using observations drawn from a narrow range of values of the forcing variable around the cut-off where the assumption of near-random assignment is most plausible. The effect estimated in a fuzzy RD design is known as a complier average causal effect because it represents the effect of exposure on those who comply with their assignment.

A strength of the RD design is that the key identifying assumption can be tested empirically. A histogram of the assignment variable and scatterplots of the forcing variable against other covariates can be used to identify any bunching either side of the cut-off that would indicate manipulation of treatment assignment (for example, by under-reporting of income or misrepresentation of age). A scatter plot of the outcome against the forcing variable can also be used to inform decisions about the modelling approach, such as what range of data should be used, whether additional terms are needed to take account of non-linearities or changes in slopes at the cut-off. Drawbacks are that additional assumptions are needed to extrapolate effect estimates to populations far from the cut-off, and the need to restrict analyses to a subset of observations close to the cut-off, which reduces statistical power. The latter problem is less serious in studies where the effect close to the cut-off is most relevant to the research questions, such as those where the focus is on a small change in the assignment rule (Moscoe et al. 2015).

Hainmueller and colleagues used a sharp RD design to identify the impact on child mental health of protection against deportation under the US Deferred Action on Childhood Arrivals (DACA) programme (Hainmueller et al. 2017). DACA provided renewable temporary work permits and protection against deportation for unauthorized immigrants who arrived in the US as children, and who were aged under 31 in 2012, the year in which the programme was introduced. The age cut-off determining eligibility for DACA provided the basis for a sharp RD design. Using a sample of unauthorized immigrant mothers with birthdays within 150 days of the age cut-off, drawn from Emergency Medicaid records in Oregon, the researchers found that following the introduction of the programme, children of DACA-eligible mothers were less likely than children of ineligible mothers to be diagnosed with adjustment or anxiety disorders. There was no such difference prior to the introduction of the programme. Robustness checks indicated no discontinuities in measured background

characteristics, and no bunching of mothers' birthdates around the age cut-off which would suggest manipulation of the forcing variable.

Andalón used a fuzzy RD design to investigate the impact of Mexico's *Opportunidades* programme on obesity and overweight among adolescents (Andalón 2011). *Opportunidades* provided cash transfers to households in rural communities below a poverty threshold, conditional on school and health attendance, plus a number of other eligibility criteria, leading to a fuzzy rather than a sharp discontinuity at the poverty threshold. Andalón used two-stage least squares regression to estimate the probability of receipt of the cash transfer as a function of poverty scores and the cut-off, and then to predict obesity and overweight as a function of the estimated probability of receipt. She found that obesity and overweight decreased among female but not among male participants. Note the difference in the interpretation of the two studies: the first estimates the effect of the DACA programme on the children of eligible mothers; the second estimates the impact on programme recipients, rather than all households below the poverty threshold.

Instrumental variables

Fuzzy RD designs are closely related to instrumental variable (IV) methods. Both methods use the same two-stage approach to estimating treatment effects. In IV studies a potentially confounded direct measure of exposure is replaced by an unconfounded proxy measure. To solve the problem of confounding, the instrumental variable must be associated with exposure to the intervention, have no association with other factors associated with exposure, and no association with outcomes other than through its association with exposure to the intervention. Effects are estimated by first obtaining a prediction of the probability of treatment from a regression of treatment status on the instrument, plus any observed confounders. The predicted probabilities of treatment are then fitted in an outcomes model, which also includes any confounders used in the first stage model. As in the fuzzy RD design, the effect estimated is the complier average causal effect, rather than the average treatment effect across the whole exposed population.

Reinhard and colleagues used this approach to evaluate the impact of a policy of providing free bus travel on loneliness and mental health among older people in England (Reinhard et al. 2018). Use of public transport is socially patterned but under the policy, eligibility for free travel depended solely on age. The researchers estimated the impact of the policy by regressing use of public transport on the age cut-off determining eligibility for free travel plus a range of covariates, such as employment status and pension receipt, then regressed the mental health and social outcomes on the predicted probabilities of public transport use plus the same range of covariates. They found that use of public transport was associated with fewer depression symptoms and feelings of loneliness, and higher levels of social engagement.

The key strength of IV methods is that if the three assumptions outlined are met, the instrument controls for unobserved as well as observed confounders. The drawback is that although the first assumption can be tested empirically, it is difficult to show that the instrument has no association with any other factors that may be associated with the outcome of interest, or that it has no direct relation with outcomes other than through its association with exposure to the intervention. Critics of the use of IVs point to the lack of either an empirical or a good theoretical justification for the identifying assumptions in some applications of the method (Deaton 2010).

Conclusion—making good use of natural experiments

Until recently, natural experiments were regarded primarily as a historical curiosity. That view no longer prevails, and natural experimental approaches are now widely, but not universally, seen as central to efforts to improve the evidence base for large-scale public health interventions. Leading national and international public health agencies have invested in initiatives to promote the use of natural experiments (Emmons et al. 2018; National Institutes of Health Office of Disease Prevention 2018; Snell et al. 2018).

This chapter has set out the case for seeing trials and natural experimental methods as belonging to a common toolkit of approaches available to researchers, rather than as occupying separate levels of an

evidence hierarchy that is stratified only by risk of bias. Exactly the same reasoning applies to the different natural experimental approaches described in this chapter. There is no single approach that is best in all circumstances. Instead, there is a range of methods that provide researchers with tools to tackle a wide array of evaluation problems. Choice of methods should depend on the specific details of the evaluation problem, rather than on more general assumptions about the superiority of some methods over others.

Good natural experimental studies require a clear understanding of how the intervention works, and in particular of the processes that determine exposure. An explicit logic model or programme theory is a useful way of capturing this. Whatever method is chosen for identifying the effects of the intervention, careful testing of assumptions and, where possible, the comparison of results from alternative methods helps to build confidence in conclusions. Prior publication or registration of study protocols, so that it is clear which questions were theoretically motivated, rather than driven by the data, is also useful—and increasingly a requirement of journal publication.

A key advantage of natural experimental approaches is that they can be applied retrospectively. A corollary is that they often rely on routinely collected data, from Censuses, long-running surveys, vital events registration, or monitoring carried out in the course of programme implementation. Investment in infrastructure for linking exposure and outcomes data at an individual level is opening up new opportunities for both trials and natural experiments. To enable researchers to make the most of these opportunities, information governance processes need to be rigorous, but also efficient and proportionate to the real risk of disclosure. For their part, researchers need to be alert to the opportunities provided by the withdrawal or scaling back of existing interventions, the fine details of assignment rules, or by imperfections in programme delivery, as well as by the implementation of large-scale policies or programmes.

REFERENCES

Abadie, A., Diamond, A., and Hainmueller, J. (2010). Synthetic control methods for comparative case studies: estimating the effect of California's tobacco control program. *Journal of the American Statistical Association*, 105, 493–505.

Abadie, A., Diamond, A., and Hainmueller, J. (2015). Comparative politics and the synthetic control method. *American Journal of Political Science*, 59, 495–510.

Andalón, M. (2011). Oportunidades to reduce overweight and obesity in Mexico? *Health Economics*, 20, 1–18.

Angrist, J. (1990). Lifetime earnings and the Vietnam-era draft lottery. Evidence from social security administrative records. *American Economic Review*, 80, 313–35.

Baicker, K., et al. (2013). The Oregon experiment—effects of Medicaid on clinical outcomes. *New England Journal of Medicine*, 368, 1713–22.

Bärnighausen, T., et al. (2017a). Quasi-experimental study designs series—paper 7: assessing the assumptions. *Journal of Clinical Epidemiology*, 89, 53–66.

Bärnighausen, T., et al. (2017b). Quasi-experimental study designs series—paper 4: uses and value. *Journal of Clinical Epidemiology*, 89, 21–9.

Basu, S., Meghani, A., and Siddiqi, A. (2017). Evaluating the health impact of large-scale public policy changes: classical and novel approaches. *Annual Review of Public Health*, 38, 351–70.

Bernal, J. L., Cummins, S., and Gasparrini, A. (2017). Interrupted time series regression for the evaluation of public health interventions: a tutorial. *International Journal of Epidemiology*, 46, 348–55.

Bor, J., et al. (2017). Treatment eligibility and retention in clinical HIV care: a regression discontinuity study in South Africa. *PLoS Medicine*, 14, e1002463.

Bor, J., Moscoe, E., Mutevedzi, P., Newell, M.-L., and Bärnighausen, T. (2014). Regression discontinuity designs in epidemiology: causal inference without randomized trials. *Epidemiology*, 25, 729–37.

Bouttell, J., Craig, P., Lewsey, J., Robinson, M., and Popham, F. (2018). Synthetic control methodology as a tool for evaluating population-level health interventions. *Journal of Epidemiology and Community Health*, 72, 673–8.

Chattopadhyay, R., and Duflo, E. (2004). Women as policy-makers: evidence from a randomized policy experiment in India. *Econometrica*, 72, 1409–43.

Concato, J., Shah, N., and Horwitz, R.I. (2000). Randomized, controlled trials, observational studies, and the hierarchy of research designs. *New England Journal of Medicine*, 342, 1887–92.

Craig, P., et al. (2012). Using natural experiments to evaluate population health interventions: new Medical Research Council guidance. *Journal of Epidemiology and Community Health*, 66(12), 1182–6.

Craig, P., Gibson, M., Campbell, M., Popham, F., and Katikireddi, S.V. (2018). Making the most of natural experiments: what can studies of the withdrawal of public health interventions offer? *Preventive Medicine*, 108, 17–22.

Craig, P., Katikireddi, S. V., Leyland, A., and Popham, F. (2017). Natural experiments: an overview of methods, approaches, and contributions to public health intervention research. *Annual Review of Public Health*, 38, 39–56.

Davey Smith, G. (2002). Commentary: behind the Broad Street pump: aetiology, epidemiology and prevention of cholera in mid-19th century Britain. *International Journal of Epidemiology*, 31, 920–32.

Deaton, A. (2010). Instruments, randomization, and learning about development. *Journal of Economic Literature*, 48, 424–55.

Deaton, A., and Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine*, 210, 2–21.

Derzon, J.H. (2018). Challenges, opportunities, and methods for large-scale evaluations. *Evaluation and the Health Professions*, 41, 321–45.

Dunning, T. (2012). *Natural Experiments in the Social Sciences: A Design-Based Approach*. Cambridge: Cambridge University Press.

Emmons, K.M., Doubeni, C.A., Fernandez, M.E., Miglioretti, D.L., and Samet, J.M. (2018). National Institutes of Health pathways to prevention workshop: methods for evaluating natural experiments in obesity. *Annals of Internal Medicine*, 168, 809–14.

Goodman, A., Van Sluijs, E.M., and Ogilvie, D. (2016). Impact of offering cycle training in schools upon cycling behaviour: a natural experimental study. *International Journal of Behavioral Nutrition and Physical Activity*, 13, 34.

Hainmueller, J., et al. (2017). Protecting unauthorized immigrant mothers improves their children's mental health. *Science*, 357, 1041–4.

Humphreys, D.K., Gasparrini, A., and Wiebe, D.J. (2017a). Association between enactment of a 'stand your ground' self-defense law and unlawful homicides in Florida. *JAMA Internal Medicine*, 177, 1523–4.

Humphreys, D.K., Gasparrini, A., and Wiebe, D.J. (2017b). Evaluating the impact of Florida's 'stand your ground' self-defense law on homicide and suicide by firearm: an interrupted time series study. *JAMA Internal Medicine*, 177, 44–50.

M'Gonigle, G. (1933). Poverty, nutrition and the public health: an investigation into some of the results of moving a slum population to modern dwellings. *Proceedings of the Royal Society of Medicine*, 26, 677–87.

Melhuish, E., Belsky, J., Leyland, A. H., Barnes, J., and National Evaluation of Sure Start Research Team (2008). Effects of fully-established Sure Start Local Programmes on 3-year-old children and their families living in England: a quasi-experimental observational study. *Lancet*, 372, 1641–7.

Morris, S.S., Olinto, P., Flores, R., Nilson, E.A., and Figueiro, A.C. (2004). Conditional cash transfers are associated with a small reduction in the rate of weight gain of preschool children in northeast Brazil. *The Journal of Nutrition*, 134, 2336–41.

Moscoe, E., Bor, J., and Bärnighausen, T. (2015). Regression discontinuity designs are underutilized in medicine, epidemiology, and public health: a review of current and best practice. *Journal of Clinical Epidemiology*, 68, 132–43.

Nandi, A., Hajizadeh, M., Harper, S., Koski, A., Strumpf, E.C., and Heymann, J. (2016). Increased duration of paid maternity leave lowers infant mortality in low-and middle-income countries: a quasi-experimental study. *PLoS Medicine*, 13, e1001985.

National Institutes of Health Office of Disease Prevention (2018). *A Report from the Federal Partners Meeting of the National Institutes of Health Pathways to Prevention Workshop: Methods for Evaluating Natural Experiments in Obesity*. Bethesda, MD: National Institutes of Health.

Reinhard, E., Courtin, E., Van Lenthe, F.J., and Avendano, M. (2018). Public transport policy, social engagement and mental health in older age: a quasi-experimental evaluation of free bus passes in England. *Journal of Epidemiology and Community Health*, 72, 361–8.

Schram, A., Labonte, R., Baker, P., Friel, S., Reeves, A., and Stuckler, D. (2015). The role of trade and investment liberalization in the sugar-sweetened carbonated beverages market: a natural experiment contrasting Vietnam and the Philippines. *Globalization and Health*, 11, 41.

Snell, A., et al. (2018). WHO Regional Office for Europe's Natural Experiment Studies Project: an introduction to the series. *European Journal of Public Health*, 28, Supplement 2, 1–3.

Snow, J. (1855). *On the Mode of Communication of Cholera*. London: Churchill.

Tugwell, P., Knottnerus, J. A., Mcgowan, J., and Tricco, A. (2017). Big-5 quasi-experimental designs. *Journal of Clinical Epidemiology*, 89, 1–3.

Webster, D., Crifasi, C.K., and Vernick, J.S. (2014). Effects of the repeal of Missouri's handgun purchaser licensing law on homicides. *Journal of Urban Health*, 91, 293–302.

Wimbush, E., Geddes, R., Woodman, K., Craig, P., and Jepson, R. (2015). Evaluability Assessment of the Family Nurse Partnership in Scotland. Edinburgh: NHS Heath Scotland.

Wing, C., Simon, K., and Bello-Gomez, R.A. (2018). Designing difference in difference studies: best practices for public health policy research. *Annual Review of Public Health*, 39, 29.1–29.17.

5.10

Clinical epidemiology

Fiona F. Stanaway, Naomi Noguchi, Clement Loy, Sharon Reid, and Jonathan C. Craig

Introduction

Clinical epidemiology is a science that extends the principles and methods of epidemiology to clinical practice and clinical research. It incorporates the location, evaluation, and application of the best evidence to patient care by clinicians as well as the generation of high-quality research evidence by clinical researchers. Rather than being focused on health at the population level, clinical epidemiology focuses on individual patients and groups of patients. The application of the science of clinical epidemiology to patient care and its' integration with clinical expertise and patient values is also often referred to as evidence-based medicine or evidence-based practice.

Clinical epidemiology can make an important contribution to global public health through the application of relevant methods to public health decision-making. The field of clinical epidemiology has driven changes in the way in which evidence is generated and synthesized, and has led to more efficient approaches to the identification, retrieval, and appraisal of evidence (Sackett 2002). These methods make an important contribution to evidence-based public health, the term used to describe public health decision-making that is based on the best available evidence, availability of resources, and population needs, values, and preferences (Brownson et al. 2009). Clinical epidemiology has also contributed to the development of methods for research translation, also known as implementation science, to facilitate the rapid uptake of scientific evidence into clinical decision-making (Woolf 2008). These same approaches are being applied to the field of public health, ensuring that public health research evidence is translated into effective interventions and policies.

Clinical epidemiology benefits global public health by its impact on improving healthcare. By providing better diagnostic and therapeutic care to individual patients and groups of patients, clinical epidemiology can contribute to health improvement at the population level. These improvements can come about through better evidence-based decision-making on the part of individual clinicians as well as health system wide approaches.

In this chapter, we provide an overview of clinical epidemiological methods and how these approaches can be used to improve global public health. We have focused primarily on using evidence in decision-making in this chapter, rather than study design and conduct elements, such as randomized controlled trials, which are covered in detail elsewhere. Consequently, we have provided a framework for critical appraisal and reporting of relevant study designs (how to use and report), rather than a detailed discussion about how such studies should be designed and conducted (how to do).

Clinical epidemiological methods

Efficient location of research evidence

Locating a relevant piece of evidence is the first step behind evidence-based decision-making. An important principle behind making a search for evidence efficient is starting the search with higher levels of evidence (primarily systematic reviews) and resources that have already been appraised for quality (Alper and Haynes 2016). In these 'pre-appraised' sources of evidence, the information has been critically appraised and filtered by experts, meaning that it has been appraised for risk of bias. Table 5.10.1 shows the hierarchy in terms of locating pre-appraised evidence. The highest level referred to as 'systems' is currently not applicable to public health decision-making as it involves integration of an individual's electronic health records with computerized decision support systems. However, all of the other levels of pre-appraised evidence can be useful for answering public health questions. The second level of synthesized summaries for clinical reference, also referred to as point-of-care tools, although specifically aimed at clinical decision-making at the point of care, can contain useful summaries of public health-related evidence. However, an important barrier to the use of these resources is their cost as they are all subscription-based tools.

Within the lower three levels of pre-appraised evidence, available content can include original reports that have been pre-appraised by experts (filtered content) and appraisal and extraction of key content of original reports (synopses) (Alper and Haynes 2016). Table 5.10.1 shows resources that can provide these various forms of pre-appraised evidence to aid public health decision-making. In addition to sites that provide one type of pre-appraised resource such as the Cochrane Library for systematic reviews, there are a number of search engines that search simultaneously across multiple sources and organize results according to level or quality of evidence (e.g. TRIP database and Health Evidence).

An efficient search would start at the highest level (point-of-care tools if available) and then work down based on the availability of sources and the success of the search. Within each level, the more pre-appraised the material the better. For example, if searching for evidence from a guideline, a guideline that has been rated as high quality by experts would be a better start than a guideline that has not been appraised. Results of single primary studies such as a single randomized controlled trial, cohort, or case–control study should only be used to inform decision-making when no other higher-level sources are available.

Table 5.10.1 Resources by levels of pre-appraised evidence

Levels of pre-appraised evidence	Publicly available resources	Subscription-based resources
1. Systems	Not applicable for public health	Not applicable for public health
2. Synthesized summaries for clinical reference		DynaMed Plus: http://www.dynamed.com/
		UpToDate: https://www.uptodate.com/home
		IBM Micromedex: https://www.micromedexsolutions.com/home/dispatch/ssl/
		BMJ Best Practice: https://bestpractice.bmj.com/
		Essential Evidence Plus: https://www.essentialevidenceplus.com/
3. Guidelines	UK National Institute for Health and Clinical Excellence (NICE) Public Health Guidelines: https://www.nice.org.uk/	
	US Task Force on Community Preventive Services Guide to Community Preventive Services: https://www.thecommunityguide.org/	
	Centers for Disease Control and Prevention (CDCP) Morbidity and Mortality Weekly Report (MMWR) Recommendations and Reports: https://www.cdc.gov/mmwr/index.html	
	US Preventive Services Task Force (USPSTF) Recommendations: https://www.uspreventiveservicestaskforce.org/Page/Name/recommendations	
	WHO guidelines: https://www.who.int/publications/guidelines/year/en/	
4. Systematic reviews	The Cochrane Database of Systematic Reviews (freely available in some jurisdictions): https://www.cochranelibrary.com/	
	Campbell Collaboration Library of Systematic Reviews: https://campbellcollaboration.org/	
	PubMed Clinical Queries: https://www.ncbi.nlm.nih.gov/pubmed/	
	US Office of Disease Prevention and Health Promotion Healthy People 2020 Evidence-based Resources: https://www.healthypeople.gov/2020/tools-resources/Evidence-Based-Resources	
	University of London Evidence for Practice and Policy Information (EPPI) Centre: https://eppi.ioe.ac.uk/cms/Default.aspx?tabid=53	
	Agency for Healthcare Research and Quality (AHRQ) Evidence Reports: https://www.ahrq.gov/research/findings/evidence-based-reports/search.html	
	Centre for Reviews and Dissemination: https://www.crd.york.ac.uk/CRDWeb/	
5. Studies	PubMed: https://www.ncbi.nlm.nih.gov/pubmed/	
6. Meta-search services across evidence levels	TRIP database: https://www.tripdatabase.com/	ACP Journal Club: http://annals.org/aim/journal-club *
	ACCESSSS: https://www.accessss.org/	Evidence Alerts: https://plus.mcmaster.ca/EvidenceAlerts/QuickSearch.asp
	PDQ-Evidence: https://www.pdq-evidence.org/	

* ACP Journal Club and Evidence Alert list synopses of systematic reviews and primary studies.

Evidence synthesis

An important part of being able to locate pre-appraised evidence for health decision-making is the creation of pre-appraised resources. A key component to this is the methodological approach to synthesizing evidence that has been developed through clinical epidemiology. ‘Evidence synthesis’ refers to the creation of an unbiased summary

of all the available evidence relevant to a specific question. It involves locating all the relevant evidence for a particular question or topic, and using rigorous methods to appraise, synthesize, and present summarized information in a concise, transparent, and accessible way to inform decision-making. Synthesized summaries for clinical reference, guidelines and systematic reviews are important forms of evidence synthesis used in clinical epidemiology and the practice of evidence-based medicine. These are also important sources of pre-appraised information for evidence-based public health

Synthesized summaries for clinical reference

These types of resources are also referred to as point-of-care tools (as they facilitate evidence-based decision-making at the point of care) and online, evidence-driven, clinical textbooks. A large number of these resources are available and examples are listed in Table 5.10.1. While these resources are aimed at clinical decision-making, they also provide summaries of evidence that can be relevant for public health decision-making such as evidence around smoking cessation, diabetes prevention, and vaccine effectiveness.

Ideally, these resources should provide clear descriptions of the processes that are used to locate and appraise the evidence that forms the basis for the summarized information. In addition, there needs to be a clear process of regular updates to the resources as well as the provision of information to users about the recency of the evidence. A recent review of these resources found variation in terms of editorial quality (including items such as disclosure of conflicts of interest and regularity of updates to summaries), evidence-based methodology, and breadth of diseases covered (Kwag et al. 2016). Only three resources received scores in the top quartile across all three domains and these were BMJ Best Practice, DynaMed, and UpToDate.

Guidelines

Clinical practice guidelines (CPGs) are evidence-based statements including recommendations that aim to optimize patient care and guide healthcare decision-making for specific clinical circumstances (National Health and Medical Research Council (NHMRC)). CPGs typically integrate evidence from a number of different systematic reviews to assess the benefits and harms of different management options (AGREE 2018). The CPG development process involves specific methodology including a systematic method of searching for and selecting evidence, involvement of target populations so that their views and preferences can be incorporated, appraisal of the quality of evidence collected, declaration of any conflicts of interest, and clear recommendations for management based on the evidence located.

Although developed specifically for clinical practice, CPGs can also have a public health focus such as guidelines regarding alcohol consumption and dietary recommendations. In addition to CPGs, an increasing number of guidelines are developed specifically for public health and to this end organizations such as the World Health Organization (WHO 2014) and the National Institute for Health and Care Excellence (NICE 2012) have developed manuals outlining the methodological approaches to guideline development for public health decision-making. Both WHO and NICE use the Grading of Recommendations Assessment, Development and Evaluation (GRADE) approach for rating evidence quality.

An important tool for assessing the quality of guidelines is the Appraisal of Guidelines for Research and Evaluation (AGREE) II instrument. AGREE II provides a framework to assess the quality, development and reporting of CPGs and emphasizes the important contribution of systematic reviews in the rigour of CPG development (AGREE 2018). Although developed for appraisal of CPGs, AGREE II is a useful tool for appraising public health guidelines.

Systematic reviews

A systematic review is an evidence synthesis of primary research studies using rigorous and clearly defined methodology. A systematic review aims to collate all empirical evidence that meets pre-specified eligibility criteria, to answer a specific research question. Systematic reviews use explicit, systematic methods to minimize the risk of bias, and aim to provide reliable findings from which conclusions can be drawn and decisions made. There are two organizations, the Cochrane Collaboration (Cochrane) and the Campbell Collaboration that produce high-quality systematic reviews that are useful for evidence-based decision-making.

Cochrane provides an invaluable contribution to the methodological standards and dissemination of systematic reviews. The original focus of Cochrane was on systematic reviews of randomized controlled trials but over recent years the Cochrane Database of Systematic Reviews (CDSR) has expanded to include: intervention reviews

(randomized and non-randomized studies) to assess the benefits and harms of healthcare and health policy interventions; diagnostic test accuracy reviews to assess how well a test diagnoses and detects a particular disease; methodology reviews which assess issues relevant to the conduct and reporting of systematic reviews and clinical trials; qualitative reviews which consider aspects of interventions other than effectiveness; and prognosis reviews which assess the course and future outcomes for people with a particular health outcome. The Cochrane public health and health systems network includes the Cochrane public health review group which focusses on topics that address health and other outcomes at the population level and aim to help government and non-government agencies achieve population-level health, well-being, learning, and social outcomes.

The Campbell Collaboration is a 'sister' organization to Cochrane for the purpose of producing systematic reviews of research evidence on the effectiveness of social interventions. Campbell systematic reviews address the effectiveness of programmes, policies, and practices in the areas of crime and justice, education, international development, and social welfare. In addition to the systematic reviews series, the Campbell Collaboration also includes: the policies and guidelines series; the methods series; and a collection of resources and tools for researchers.

GRADE and assessment of evidence quality

Both systematic reviews and CPGs use the GRADE approach to 'grading' the confidence (or certainty) of evidence. GRADE classifies the quality of evidence for each outcome at one of four levels: high, moderate, low, and very low. In the GRADE approach, the certainty of evidence refers to more than the risk of bias domain (internal validity) of a study. GRADE quality assessment includes consideration of the following five components.:

Study limitations or risk of bias;
Inconsistency of results;
Indirectness of evidence;
Imprecision; and
Other considerations such as publication bias (Guyatt et al. 2008).

Risk of bias

Bias is a kind of systematic error that leads to the results being different from the truth. Risk of bias assessment refers to the process where the risk of bias of a study is assessed in order to judge if the study's results are trustworthy. If there is a systematic difference between the study groups not due to the intervention being studied, the truth may be distorted towards a particular direction. The domains and specific items that are important in terms of risk of bias vary between study types and the type of research question. For example, a cohort study may be used to examine questions of causality (e.g. the link between high blood pressure and stroke) and prognosis (e.g. survival following stroke), but the risks of bias domains are study design/question specific. Epidemiological study types, clinical question types, and examples of tools that can be used to appraise them are provided in Table 5.10.2.

In general, when appraising the risk of bias of primary studies, we need to assess if: (1) the study groups remain comparable to each other throughout the research study (selection bias); (2) the participants are treated equally and their outcomes are measured in the same manner between groups (performance bias and measurement bias); and (3) if there are systematic differences in participant characteristics that can impact on the outcomes, these are taken into account (adjustment for potential confounders). Taking randomized controlled trials (RCTs) as an example, for the study groups to remain equal throughout a RCT, the study should begin with comparable groups by using an appropriate randomization method and ensuring there is allocation concealment to prevent researchers interfering with the randomization process. The groups can then be kept balanced by ensuring that loss to follow-up is minimized throughout the study and by conducting analyses by intention-to-treat, or as randomized, to preserve the effect of the randomization. Blinding is an essential component of RCTs that enables equal treatment of participants during the study and equal outcome measurement regardless of study group allocation. Confounders are factors associated with the outcome that are distributed unevenly between study groups. As randomization should ensure equal distribution of characteristics between study groups, the management of confounding is usually not needed in a RCT.

The approach to risk of bias assessment of systematic reviews, which would be an essential part of guidelines development, is somewhat different to that of primary studies. Some of the key points in assessing risk of bias in systematic reviews are: the comprehensiveness of the search for studies including both published and unpublished research; risk of bias assessment of included primary studies; and appropriate numerical synthesis of included studies (meta-analysis).

An important and common barrier to critical appraisal is that sufficient information may not be reported in the article to allow assessment of the risk of bias. As a result, a number of reporting guidelines have been developed that specify a set of items that authors need to report for each study type to enable risk of bias assessment. These reporting guidelines have also been listed in Table 5.10.2. It is important to note that these guidelines are reporting checklists for authors to use when preparing a manuscript and are frequently mandated by journals. They are not suitable to use as critical appraisal checklists to assess risk of bias.

Table 5.10.2 Checklists for risk of bias assessment and guidance for reporting research studies by study designs and question types

Study design/ Question type		Checklist for risk of bias assessment	Guidance fo
Systematic reviews of intervention studies		ROBIS (Risk of Bias in Systematic Reviews): http://www.bristol.ac.uk/media-library/sites/social-community-medicine/robis/robisguidancedocument.pdf	PRISMA (Pr Systematic R http://www.e guidelines/pri
		AMSTAR: http://www.amstar.ca/	
		DSU NMA methodology checklist: http://scharr.dept.shef.ac.uk/nicedsu/wp-content/uploads/sites/7/2016/03/TSD7-reviewer-checklist.final_.08.05.12.pdf	
		CASP systematic review checklist: http://docs.wixstatic.com/ugd/dded87_a02ff2e3445f4952992d5a96ca562576.pdf	
Systematic reviews of diagnostic test accuracy and prognostic studies		ROBIS: http://www.bristol.ac.uk/media-library/sites/social-community-medicine/robis/robisguidancedocument.pdf	PRISMA-DT http://www.e guidelines/pri
Randomized controlled trials		Cochrane RoB tool: https://sites.google.com/site/riskofbiastool/welcome/rob-2-0-tool	CONSORT (Reporting Tri network.org/
		Effective Practice and Organization of Care (EPOC) RoB Tool: http://epoc.cochrane.org/resources/epoc-resources-review-authors	
		CASP RCT checklist: http://docs.wixstatic.com/ugd/dded87_40b9ff0bf53840478331915a8ed8b2fb.pdf	
Non-randomized controlled trials		Cochrane ROBINS-I: https://sites.google.com/site/riskofbiastool/welcome/home	TREND (Tra with Non-ran https://www.c
		EPOC (Effective Practice and Organization of Care) RoB Tool: http://epoc.cochrane.org/resources/epoc-resources-review-authors	
All observational studies			STROBE (St Observationa http://www.e guidelines/str
	Cohort studies	Cochrane ROBINS-Ihttps://sites.google.com/site/riskofbiastool/welcome/home:	
		EPOC RoB Tool: http://epoc.cochrane.org/resources/epoc-resources-review-authors	
		CASP cohort study checklist: http://docs.wixstatic.com/ugd/dded87_e37a4ab637fe46a0869f9f977dacf134.pdf	
		The Newcastle-Ottawa Scale (NOS): http://www.ohri.ca/programs/clinical_epidemiology/nosgen.pdf	
	Case–control studies	CASP case–control checklist: http://docs.wixstatic.com/ugd/dded87_63fb65dd4e0548e2bfd0a982295f839e.pdf	
		EPOC RoB Tool: http://epoc.cochrane.org/resources/epoc-resources-review-authors	
		The Newcastle-Ottawa Scale (NOS): http://www.ohri.ca/programs/clinical_epidemiology/nosgen.pdf	
	Cross sectional studies	JBI checklist for cross sectional study: https://joannabriggs.org/assets/docs/critical-appraisal-tools/JBI_Critical_Appraisal-Checklist_for_Analytical_Cross_Sectional_Studies.pdf	
		AXIS: http://bmjopen.bmj.com/content/6/12/e011458	
	Diagnostic test accuracy studies**	QUADAS-2: http://www.bristol.ac.uk/population-health-sciences/projects/quadas/quadas-2/	STARD (Star Accuracy Stu network.org/
		CASP diagnostic test accuracy checklist: http://docs.wixstatic.com/ugd/dded87_3815f02af1b34c21b8c3b2b5020024c3.pdf	

Study design type	Question	Checklist for risk of bias assessment	Guidance fo
	Prognostic studies**	QUIPS checklist: http://methods.cochrane.org/sites/methods.cochrane.org.prognosis/files/public/uploads/QUIPS%20tool.pdf	
	Prevalence/ incidence studies**	JBI checklist for prevalence studies: https://joannabriggs.org/assets/docs/critical-appraisal-tools/JBI_Critical_Appraisal-Checklist_for_Prevalence_Studies.pdf	

* Bold type indicates critical appraisal checklists recommended by the UK NICE.

** Critical appraisal checklists for specific types of clinical questions.

Inconsistency of results

Within GRADE, confidence is also downgraded in the presence of unexplained inconsistency or heterogeneity in relative effects between studies (Guyatt et al. 2011c). The quality of evidence is rated down if: there is great variation in point estimates between studies; the confidence intervals around these point estimates demonstrate minimal or no overlap; and there is evidence of statistical heterogeneity in the meta-analysis through a significant statistical test for heterogeneity and a large I^2 value. Inconsistency is particularly important when the variability in results would lead to different decisions. For example, it would be important to rate down evidence quality for inconsistency if some studies demonstrated large beneficial effects while other studies demonstrated large harmful effects, and there was no apparent reason for such discordance. In contrast, inconsistency would have less impact on decision-making if the inconsistency observed was quantitative only, when all studies showed benefit but varied regarding the magnitude of the benefit.

Indirectness of evidence

Indirectness refers to the situation where the research that is being summarized in the systematic review or guideline does not directly answer the question we are interested in. This might be because the population or patients included in the study is not the population of interest. For example, studies may have all been conducted on young or middle-aged people whereas the population of interest is older people. Or, for example, in the development of the WHO guideline regarding avian influenza, evidence about the effectiveness of antivirals for avian influenza was rated down as it was based on evidence of the effectiveness of antivirals for seasonal influenza and it is possible that the avian influenza virus may not respond in the same way to antivirals (Schunemann et al. 2007; Guyatt et al. 2011b).

Another potential reason for indirectness concerns the intervention itself. This could include consideration of situations where the same intervention may be implemented differently due to variation in the availability of technical expertise or other resources. The latter is a particularly important consideration for low- and middle-income countries.

A third cause of indirectness is a lack of outcomes that are considered most important to patients, which may include the timing (long term versus short term) or domain (surrogate versus patient reported outcome). For example, in assessing evidence around a health condition like cardiovascular disease, outcomes such as the incidence of myocardial infarction, stroke, and overall mortality would be considered more important and relevant to patients than cholesterol levels.

Imprecision

Assessment of imprecision in GRADE is based on confidence intervals (Guyatt et al. 2011a). However, rather than considering the width of the confidence interval alone in determining whether or not precision is sufficient, the confidence intervals relative to decision-making thresholds are regarded as more important. When the entire confidence interval lies to either side of a threshold, the imprecision has no impact on decision-making as at either extreme of the confidence interval the clinical decision would be the same—to either use or not use the intervention. However, if a confidence interval crosses the treatment threshold it would mean that decisions may be different at either extreme of the confidence interval and as a result, the quality of the evidence would be rated down for imprecision. The impact of confidence intervals on clinical decision-making in relationship to thresholds is shown in Figure 5.10.1. In study 1, the entire confidence interval is to the left of the threshold of clinical importance and therefore is sufficiently precise for decision-making. In study 2, as the confidence interval crosses

the threshold of clinical importance there is a possibility that while the intervention is likely beneficial that the benefit may not be large enough to be clinically important. In this situation, therefore, we would rate down for imprecision. In study 4, even though the confidence interval is quite wide and crosses the line of no effect indicating lack of statistical significance, the entire confidence interval is to the right of the threshold of clinical importance. This means that we are confident that there is no clinically important effect and would not recommend the intervention.

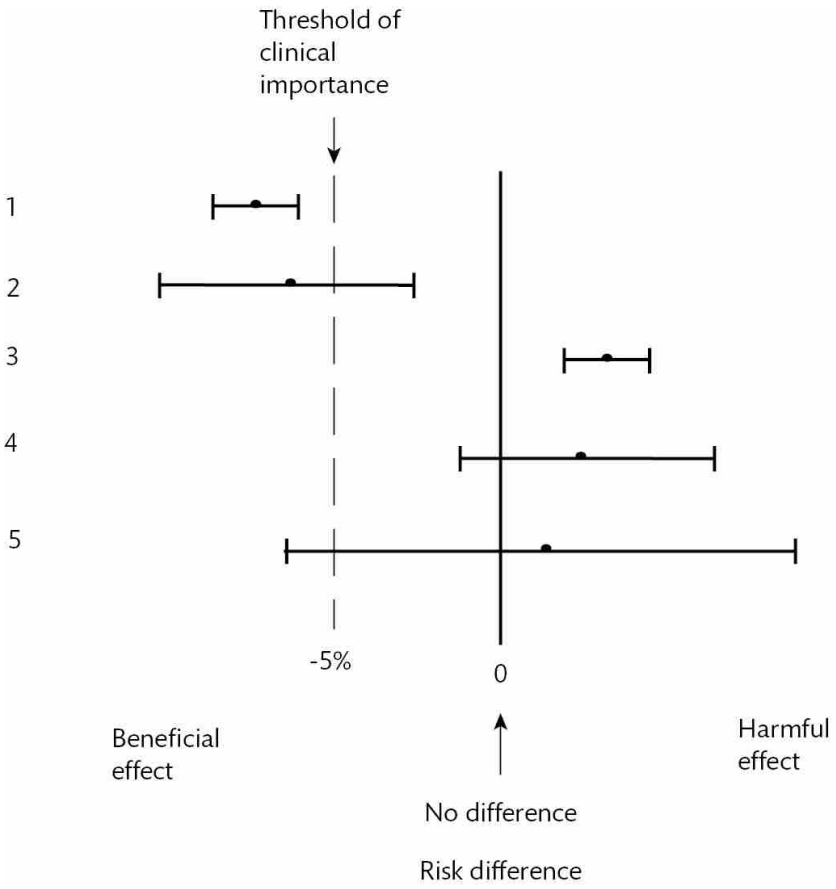


Fig. 5.10.1 Judging precision of confidence intervals in relation to clinically important thresholds
Studies 1 and 3 are sufficiently precise as the entire confidence intervals are to one side of the threshold of clinical importance.
Study 4 is also sufficiently precise as despite lacking statistical significance, the entire confidence interval lies to one side of the threshold of clinical importance.
Studies 2 and 5 are insufficiently precise as the confidence intervals cross the threshold of clinical importance.

Publication bias

Studies that have statistically significant results are more likely to be published than those with non-significant results. Therefore, relying solely on published studies in a systematic review may be misleading (in favour of the intervention) due to such publication bias. In addition, studies with non-significant results that are published are more likely to experience delayed publication which means that an early systematic review of a particular topic may have biased results as the negative results of these later publications will be excluded—a phenomenon described as lag bias (Hopewell et al. 2007). Publication bias can be associated with industry funding as industry sponsored trials may be less likely to publish negative or null results. A high likelihood of publication bias means that the confidence in a body of evidence about a particular topic would be rated down.

Strength of recommendations

Using the GRADE approach, recommendations made in guidelines collapse down into two categories, ‘strong’ or ‘weak’ (also known as conditional). In addition to the confidence in the evidence determined as described earlier, the strength of recommendations also involves consideration of: uncertainty about the balance between benefits and harms associated with an intervention; uncertainty about values and preferences of target populations; and uncertainty about whether the intervention is a wise use of resources (Guyatt et al. 2008; Andrews et al. 2013). The GRADE strength of recommendations are presented in the following four categories:

Strong recommendation against using the intervention
Weak recommendations against using the intervention
Weak recommendation for using the intervention
Strong recommendation for using the intervention.

GRADE and public health guidelines

The GRADE approach is readily applicable to assessment of the quality of evidence in systematic reviews and guidelines developed for public health questions and is being increasingly used in public health guidelines. WHO has adopted the use of GRADE for guidelines and NICE also uses GRADE for determining evidence certainty. However, despite this, there are still some concerns in the public health community about using GRADE (Burford et al. 2012; Rehfuess and Akl 2013). Many of the concerns are to do with the frequent need to use lower levels of evidence in public health decision-making due to the lack of availability and/or lack of feasibility of RCTs. However, it should be noted that although observational studies default to a low-confidence rating, they can be rated up if the size of the effect is large, if there is evidence of a dose-response gradient, and if all the impact of possible confounding is in the direction of reducing the observed effect (Guyatt et al. 2011d). One such example would be rating up the body of evidence from observational studies supporting smoking as a risk factor for lung cancer due to the large size of effect (RR of 5.5 overall and 16.9 for squamous cell carcinoma) and a clear dose-response gradient with greater cancer risk with the amount smoked, duration of smoking, tar level and fraction smoked (Lee et al. 2012). Moreover, many public health questions can be addressed by randomized studies, particularly with the increased use of pragmatic trials and methods such as cluster randomization or stepped wedge design.

Other concerns about the use of GRADE in public health include the importance of taking into consideration other aspects such as cost and the legal implications of particular public health approaches. NICE, although adopting GRADE for assessment of evidence quality, also incorporates the quality of cost-effectiveness studies in the review and uses a different approach for the wording of recommendations (Thornton et al. 2013).

The GRADE working group has taken many of these considerations into account in the development of the GRADE Evidence to Decision (EtD) framework for health system and public health decisions. In this framework, consideration is also given to the priority given to a particular problem, the impact of the proposed option on equity, and the acceptability and feasibility of the proposed public health approach (Moberg et al. 2018).

Research translation

It has been increasingly recognized that the results of medical research are often not translated into patient care, or when they are, it is delayed. For example, as many as 8 years after a large trial demonstrated a 40% reduction in mortality from using beta blockers post myocardial infarction, this treatment was prescribed in only one-third of relevant patients in the United States. It took 20 years from the publication of the study until the rate of prescribing reached high levels and this was only after this was included as a recommendation in guidelines and became part of accreditation standards (Glasziou et al. 2017).

The role of clinical epidemiology methods in evidence synthesis in the form of systematic reviews, meta-analyses, and clinical practice guideline development has been outlined, but alone is not sufficient to change practice. Translating research into individual clinical practice needs to be an active process. Passive uptake of guideline recommendations can be low, even in well-resourced health systems.

As well as failing to use interventions with strong evidence of net benefit in routine clinical care (underuse), overuse of inappropriate healthcare is also common. Overuse refers to the provision of healthcare where the harms to patients outweigh the benefits. For example, in Thailand and China, an estimated 55% of children with acute diarrhoea are reported to receive inappropriate antibiotic prescriptions (Brownlee et al. 2017). As well as non-

evidence-based use of treatments, inappropriate use of diagnostic tests is also very common. For example, in the United States an estimated 36% of patients had a repeat colonoscopy that was not recommended based on guidelines and 40% had inappropriate imaging for low back pain (Brownlee et al. 2017). Research translation is concerned with both the lack of use of effective interventions and tests as well as the use of interventions and tests that are ineffective or unnecessary.

Models of change have been developed to assist the translation of evidence into practice. These models include steps that help integrate theory with interventions to encourage change. The following steps are from a modified version of the model developed by Grol (Grol and Wensing 2004):

Identify and measure an evidence practice gap—it is useful to choose an area of practice that is important clinically, has a good evidence base to support a particular management approach, and where current practice can be easily measured
Conduct a cause analysis to determine reasons for the evidence practice gap—this can be done using tools such as fishbone diagrams or a theoretical domain framework
Choose an intervention aimed at changing practice
Evaluate the effectiveness of the intervention using an appropriate indicator of quality improvement
Sustainability—a plan for sustainability should be incorporated into project planning from the outset as well as how sustainability will be assessed

Optimizing use of health resources

Another area where clinical epidemiology can contribute to the functioning of the health system, is through the interpretation of evidence and, in partnership with health economists, optimization of resource use. Critical evaluation of trial data and meta-analytical techniques form the backbone of decision-making for health technology agencies. Some additional techniques used in this context include network meta-analysis (Lumley 2002) (to estimate the relative efficacy of more than two treatments to determine which intervention is likely to be the best), and a ‘linked evidence approach’ to assessing diagnostic tests (Merlin et al. 2013).

Evaluation reports from health technology agencies not only provide local guidance, but also form a rich resource for other health systems considering adoption of new medicines. These include reports from the Food and Drug Administration in the United States (McDonagh et al. 2013), the European Medicines Agency (European Medicines Agency 2018), the National Institute for Health and Care Excellence in the United Kingdom (National Institute for Health and Care Excellence UK), and the Medicare Benefits Schedule and Pharmaceutical Benefits Scheme in Australia (Pharmaceutical Benefits Scheme 2018).

Applying evidence to individual patients

As well as contributing to system wide approaches to optimizing health resource use, clinical epidemiology methods may optimize resource use through providing clinicians with skills to tailor management decisions according to the circumstances of individual patients. This process has been variously called applicability and transferability of evidence, individualizing therapy, precision medicine, and stratified care. The importance of this approach comes from an understanding that even when there is strong evidence from a trial or meta-analysis of trials that a treatment on average produces a net benefit, there may not be a net beneficial effect from the treatment in every patient in the trial(s) or in every patient in clinical practice (Glasziou and Irwig 1995). Traditionally, it has been common to use the inclusion and exclusion criteria of a trial to decide whether or not an intervention will have a net benefit in a patient. However, these inclusion and exclusion criteria for trials are primarily based upon separate considerations such as improving the power of a study to detect an effect and to maximize safety. There will be patients who meet inclusion criteria for the trial who would not receive a net benefit from the intervention and there will be those who do not meet inclusion criteria who would receive a net benefit.

A more appropriate approach to determining which patients should receive particular interventions is a risk-benefit model approach. This process has several steps which are:

Listing all of the benefits and harms of the intervention
Determining if there is any heterogeneity or effect modification of the relative treatment effect and/or harms
Considering the impact of baseline risk on absolute benefits and harms
Weighing up benefits and harms taking into consideration patient values

In the first step, a list of all patient-relevant outcomes that the intervention will have an impact on should be generated. Evidence about efficacy should come from a systematic review of RCTs. However, it may be necessary to use a cohort study or systematic review of cohort studies to obtain sufficient data about harms.

In the second step, it is important to consider if there are subgroups of patients or particular settings in which the treatment might work differently to the estimate of effect given in the systematic review or RCT. This difference in effect can occur not only for the benefits of an intervention but also for the harms. One important cause of effect modification is socio-economic status which can be a source of variability in observed treatment effects due to resource constraints acting at the individual and/or community level. This variation in treatment effects and harms by patient subgroups will impact on the trade-off between benefits and harms.

The next consideration is the impact of baseline risk on the benefit to harm trade-off. An important part of this step is understanding the difference between relative and absolute measures of effect. Relative risks tend to be less dependent upon baseline risk and so are commonly the measure of effect reported in clinical trials. However, the risk difference which refers to the absolute difference in incidence between groups is more useful for determining benefit to harm trade-off as illustrated in Figure 5.10.2. This figure demonstrates how the baseline risk of disease determines the size of the absolute difference given a stable relative risk. This means that if harms are constant, in those patients with a high baseline risk of disease the benefit to harm trade-off will be greater. Therefore, focussing therapy on this high-risk group would result in both better patient care and more efficient use of resources, at an individual and group level.

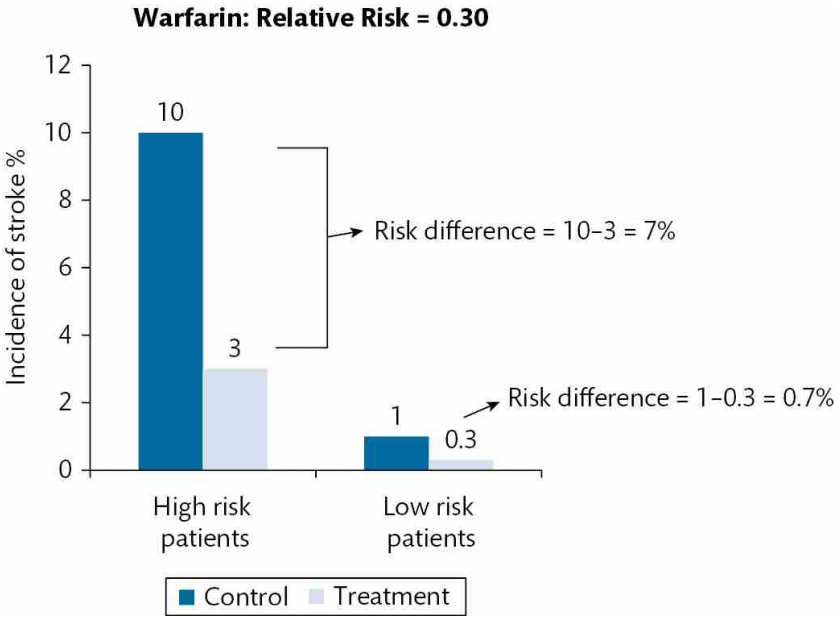


Fig. 5.10.2 Change in absolute risk difference by baseline risk.

This approach of estimating total or absolute risk of disease and tailoring management strategies accordingly is incorporated in many guidelines, such as the current WHO guidelines for cardiovascular disease (WHO 2007). It has been estimated that targeting high-risk individuals in low- and middle-income countries with evidence-based interventions to prevent cardiovascular disease could avert 17.9 million deaths from cardiovascular disease over 10 years (Lim et al. 2007).

An essential part of a risk-benefit model approach to individualizing healthcare is having access to evidence-based information that allows clinicians to estimate absolute risk. A key part of this process is the development of valid prognostic models (variously referred to by other names including prognostic index, prediction rule, risk, or clinical prediction model and predictive model). These models can be turned into scoring systems or 'risk calculator' tools that are increasingly web-based and allow clinicians to estimate a patient's absolute risk of a particular disease or outcome based on a number of predictive clinical features and test results.

Methodological standards in prognosis research studies have lagged behind those of other areas such as RCTs and to this end the PROGnosis RESearch Strategy (PROGRESS) partnership has proposed a framework for

research themes in prognosis research as well as guidance in terms of methodology. There are three main phases essential to development of prognostic models which include: development of the model, external validation of the model (ideally in a different setting), and assessment of the impact of using the model on clinical practice. Unfortunately, it is very common for only the first step to be carried out. The importance of external validation in other settings is emphasized by a study examining the performance of four commonly used cardiovascular risk prediction models in Malaysia where the performance of one of the models was found to be poor (Selvarajah et al. 2014).

The final step of the process considers the values of individual patients as they consider potential benefits and harms. This is an essential step given existing evidence for substantial variation in patient preferences for particular treatment approaches and the varying importance placed on particular benefits and harms. For example, in a systematic review of patient preferences regarding antithrombotic therapy, included studies found substantial variation in how study participants considered the trade-off between benefits such as stroke prevention and the harm of bleeding. There was also variability in the decisions patients made when provided with evidence about the benefits and harms of treatment options (MacLean et al. 2012).

Applying evidence to individual settings

The approach used to optimize the care of individual patients by considering benefit to harm trade-off can readily be adapted to how public health evidence can be applied to different settings. Benefits and harms of implementing a public health intervention could be considered at the country level where, in addition to benefits and harms related to the outcomes of the specific intervention, the impact on resource use also needs consideration given the constraint of fixed resources and opportunity cost. Assessing the likelihood of effect modification of the effectiveness of an intervention is a critical consideration, particularly in low- and middle-income countries. There is potential for variation in benefits and harms due to population characteristics such as socio-economic status and nutrition as well as potential variation in effectiveness due to staff training and availability and differences in infrastructure. In terms of baseline risk, rather than considering the baseline risk of an individual patient, the prevalence of disease in a particular country would provide a firm basis for determining the absolute benefit for the country in terms of health improvement that would be gained from the intervention. In the presence of a constant relative measure of effectiveness, a country with a high burden of disease will stand to receive a greater absolute benefit from an intervention targeting this disease. Finally, different values in a particular country or setting could impact on the importance placed on the potential benefits and harms of a given public health intervention, thereby influencing the benefit to harm trade-off.

Many of these ideas are covered in the GRADE EtD framework discussed in the section ‘GRADE and public health guidelines’ in this chapter. For example, the priority given to a public health problem would be closely tied to the burden of disease. Considering the impact of the proposal on equity and considering the feasibility of the public health intervention would be considering the potential for variability in effectiveness or effect modification of the intervention. And considering the acceptability of the public health approach would involve considering the values of the population including how these impact on the benefit to harm trade-off.

Evaluation of the health system

Clinical epidemiology methods can contribute to improved health system design. As well as optimization of resource allocation, they are also useful in health system evaluation. Analogous to methods in evaluating effectiveness of treatments for individuals, these same methods can be extended to assess system wide interventions. These methods are also frequently used in the assessment of many public health interventions that due to their complexity and breadth of focus can be difficult to evaluate using the traditional randomized controlled trial approach.

Cluster RCTs

There are situations where a health system may wish to evaluate or randomize interventions on a work unit basis (by ‘clusters’ such as clinic, area health service or school), but also measure outcomes based on individual participant/patient outcome. These cluster RCTs may be the study design of choice because (1) the intervention applies to the entire cluster (e.g. whole-school health promotion programmes), or (2) the intervention is likely to ‘overflow’ or ‘contaminate’ non-intervention units (e.g. an educational intervention for handwashing may affect

this behaviour in both the targeted health workers, as well as their colleagues who were not allocated to the intervention). An important consideration for cluster RCTs is that individual participant outcomes are likely to be correlated within each cluster. For example, patients attending the same clinic may have more similar health outcomes than patients across different clinics. This needs to be taken into account for in the planning and analysis of the study, and there are established statistical methods for doing so (Hemming et al. 2017). Protopopoff et al. is an example where the cluster randomized controlled trial design was used to evaluate insecticidal nets and indoor residual spraying, randomized to Tanzanian villages as clusters, with childhood malaria as the outcome measure (Protopopoff et al. 2018).

Stepped wedge design

One variation to the cluster randomized trial is the stepped wedge cluster randomized trial design. This is a useful design when a health system has already decided to roll out an intervention programme, but for pragmatic reasons may only be able to roll out at one site/cluster at a time. This presents an opportunity to collect evaluation data. The stepped wedge design involves collecting outcome data at regular intervals (steps) for all clusters, starting to do so before any of the clusters has started on the intervention programme. Then at each subsequent 'step', a cluster is randomly assigned to commence on the intervention programme, thus adding a cluster at each step until all clusters are in the intervention programme. Analysis for stepped wedge cluster randomized trials is complex but good introductions are available (Hemming et al. 2015). Betran et al. is an example where the stepped wedge design was used to evaluate whether provision of medical supply kits improved quality of antenatal care, across 10 antenatal care clinics in Mozambique (Betran et al. 2018).

Population screening

Population screening is defined as offering a test to all individuals in a target group in the population (e.g. a mammogram every two years, for all women aged 50–74 years) to identify people with pre-symptomatic, early disease, and offer early intervention. Recommendations for screening should be based on RCT evidence demonstrating that offering such early intervention improves outcomes (Barratt et al. 1999). The RCT design avoids two common biases that can occur in observational studies of screening: lead-time bias and length-time bias. Lead-time bias refers to the apparent longer survival of patients identified through screening due to their identification at an earlier point in their disease with a resulting longer period of follow-up. Length-time bias refers to the bias that arises due to the greater propensity of screening to identify slowly progressive disease, which will be associated with better survival. In addition to potential benefits in terms of prolonged life and improved quality of life, screening can also be associated with harms including complications from investigations, adverse effects of treatments, and psychological harms. Another harm of screening, both at the individual and population level, is the potential for overdiagnosis. This refers to the identification of disease that is inconsequential and would not have resulted in any harm if left untreated. Randomized controlled trial data may also help indicate the possibility of overdiagnosis in the screening setting (Moynihan et al. 2012).

Challenges in low- and middle-income country settings

In this final section, some key challenges faced in terms of using the methodological approaches highlighted in this chapter are discussed along with current strategies being used to alleviate them.

Access to evidence

An important barrier to the efficient location of research evidence in low- and middle-income settings is poor access to such evidence. This is the result of a combination of infrastructure issues around the availability of reliable internet and electricity supply in addition to reduced availability of evidence sources that are subscription based. Hinari is a programme established by the WHO that provides low- and middle-income countries (LMICs) with access to a large number of journals, ebooks, and information resources including sources of synthesized evidence such as the Cochrane Library. Hinari facilitates either full or partial access to over 250 publishers to more than 5,000 different institutions in LMICs. There are also a number of freely available sites that provide synthesized evidence for policy makers such as SUPPORT summaries that provide policy makers with structured summaries of systematic reviews of ways to organize, finance, and govern the delivery of healthcare interventions

in LMIC settings and PDQ-Evidence that provides systematic reviews, syntheses of reviews, evidence-based policy briefs, primary studies, and structured summaries.

Research capacity building

Training in clinical epidemiology provides clinicians and clinical researchers with the skills they need to effectively use and produce high-quality evidence to guide clinical decision-making. As illustrated in this chapter, training in clinical epidemiological skills also provides public health practitioners with important skills and approaches that can be used in evidence-based public health decision-making. Improving research capacity within a country or region not only increases the volume and quality of research produced but can also facilitate the setting of research priorities that are driven by local rather than global priorities.

One important strategy to strengthen capacity in clinical epidemiology and research training in LMICs has been the establishment of the International Clinical Epidemiology Network (INCLEN) in 1982. This network has established Clinical Epidemiology Units (CEUs) and Clinical Epidemiology Research Training Centres (CERTCs) in 89 academic institutions across 39 countries. The aim of the network has been to develop and upskill patient-orientated researchers and clinician investigators, enabling them to undertake high-quality clinical research. INCLEN scholars in CEUs received training in the fields of Clinical Epidemiology, Biostatistics, Health Social Sciences, and Health Economics. Initial training was conducted at established clinical epidemiology training centres such as McMaster University in Canada, the University of Pennsylvania, and University of North Carolina in the United States, and the University of Newcastle in Australia. However, many of the originally established CEUs have now gone on to become training centres themselves. These units and centres have played a vital role in strengthening clinical research capacity, the creation of a network of researchers, and increased research output including the development of CPGs.

The Cochrane and Campbell Collaborations also have a major focus on improving capacity in LMICs. Both have established a large network of skilled researchers and centres for producing high-quality systematic reviews that can be used to support and mentor researchers in LMICs. The Cochrane effective practice and organization of care (EPOC), equity, and public health groups are all quite involved in capacity building of LMIC researchers. For example, the Norwegian satellite of EPOC provides support to Cochrane authors in LMICs in the form of training stipends for review authors and the provision of support and instruction from the editorial team throughout the systematic review process. This satellite also helps build the capacity of policy makers to appropriately interpret research evidence for decision-making

Indirectness of evidence and transferability of guidelines

At the present time, the majority of research evidence continues to be produced in high income settings and not all of the results will be readily applicable to LMICs. As the guideline development process is both time consuming and resource intensive, a more efficient approach would be to adapt existing guidelines to different settings. The GRADE EtD framework provides a good basis for understanding what additional pieces of evidence would be required to ensure an evidence-based adaptation process. This would include burden of disease studies that are essential for priority setting, cost-effectiveness data to quantify the impact on health resources, pilot study data examining issues of feasibility and an understanding of values in terms of the importance of particular benefits and harms.

Owalabi et al. (2016) provides a useful example of how cardiovascular disease recommendations can be adapted for LMICs by presenting both an implementation cycle and translatable scale that is a useful approach to ranking recommendations according to how readily they may be implemented according to considerations of cost and personnel.

Conclusion

Based on methods developed within public health, clinical epidemiology has extended these methods into new areas aimed at improving both the use of evidence in clinical decision-making and the production of high-quality evidence. Despite the clinical focus of clinical epidemiology, many of the methods are of value to public health decision-making and are being increasingly adopted by public health professionals.

REFERENCES

AGREE Enterprise (2018). *The Appraisal of Guidelines for Research and Evaluation Instrument Evaluates the Process of Practice Guideline Development and the Quality of Reporting*. [Online]. Available at: https://www.agreetrust.org/

Alper, B.S. and Haynes, R.B. (2016). EBHC pyramid 5.0 for accessing preappraised evidence and guidance. *Evidence-based Medicine*, 21, 123–5.

Andrews, J.C., et al. (2013). Grade guidelines: 15. Going from evidence to recommendation—determinants of a recommendation's direction and strength. *Journal of Clinical Epidemiology*, 66, 726–35.

Barratt, A., et al. (1999). Users' guides to the medical literature: XVII. How to use guidelines and recommendations about screening. Evidence-Based Medicine Working Group. *The Journal of the American Medical Association*, 281, 2029–34.

Betran, A.P., et al. (2018). Provision of medical supply kits to improve quality of antenatal care in Mozambique: a stepped-wedge cluster randomised trial. *Lancet Glob Health*, 6, e57–65.

Brownlee, S., et al. (2017). Evidence for overuse of medical services around the world. *Lancet*, 390, 156–68.

Brownson, R.C., Fielding, J.E. and Maylahn, C.A. (2009). Evidence-based public health: a fundamental concept for public health practice. *Annual Review of Public Health*, 30, 175–201.

Burford, B.J., et al. (2012). Assessing evidence in public health: the added value of GRADE. *Journal of Public Health (Oxford)*, 34, 631–5.

European Medicines Agency. (2018). *Medicines—download*. [Online] Available at: https://www.ema.europa.eu/en/medicines/download-medicine-data

Glasziou, P., et al. (2017). Evidence for underuse of effective medical services around the world. *The Lancet*, 390, 169–77.

Glasziou, P.P. and Irwig, L.M. (1995). An evidence based approach to individualising treatment. *British Medical Journal*, 311, 1356–9.

Grol, R. and Wensing, M. (2004). What drives change? Barriers to and incentives for achieving evidence-based practice. *Medical Journal of Australia*, 180, S57–60.

Guyatt, G.H., et al. (2011a). GRADE guidelines–6. Rating the quality of evidence—imprecision. *Journal of Clinical Epidemiology*, 64, 1283–93.

Guyatt, G.H., et al. (2011b). GRADE guidelines: 8. Rating the quality of evidence—indirectness. *Journal of Clinical Epidemiology*, 64, 1303–10.

Guyatt, G.H., et al. (2011c). GRADE guidelines: 7. Rating the quality of evidence—inconsistency. *Journal of Clinical Epidemiology*, 64, 1294–302.

Guyatt, G.H., et al. (2011d). GRADE guidelines: 9. Rating up the quality of evidence. *Journal of Clinical Epidemiology*, 64, 1311–16.

Guyatt, G.H., et al. (2008). GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. (Grading of Recommendations Assessment, Development and Evaluation). *British Medical Journal*, 336, 924.

Hemming, K., Eldridge, S., Forbes, G., Weijer, C., and Taljaard, M. (2017). How to design efficient cluster randomised trials. *British Medical Journal*, 358, j3064.

Hemming, K., Haines, T.P., Chilton, P.J., Girling, A.J., and Lilford, R.J. (2015). The stepped wedge cluster randomised trial: rationale, design, analysis, and reporting. *British Medical Journal*, 350, h391.

Hopewell, S., Clarke, M., Stewart, L., and Tierney, J. (2007). Time to publication for results of clinical trials. *Cochrane Database of Systematic Reviews*, (2), MR000011.

Kwag, K.H., Gonzalez-Lorenzo, M., Banzi, R., Bonovas, S., and Moja, L. (2016). Providing doctors with high-quality information: an updated evaluation of web-based point-of-care information summaries. *Journal of Medical Internet Research*, 18, e15.

Lee, P.N., Forey, B.A., and Coombs, K.J. (2012). Systematic review with meta-analysis of the epidemiological evidence in the 1900s relating smoking to lung cancer. *BMC Cancer*, 12, 385.

Lim, S.S., et al. (2007). Prevention of cardiovascular disease in high-risk individuals in low-income and middle-income countries: health effects and costs. *Lancet*, 370, 2054–62.

Lumley, T. (2002). Network meta-analysis for indirect treatment comparisons. *Statistics in Medicine*, 21, 2313–24.

Maclean, S., et al. (2012). Patient values and preferences in decision making for antithrombotic therapy: a systematic review: *Antithrombotic Therapy and Prevention of Thrombosis*, 9th ed: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines. *Chest*, 141, e1S–23S.

McDonagh, M.S., Peterson, K., Balshem, H., and Helfand, M. (2013). US Food and Drug Administration documents can provide unpublished evidence relevant to systematic reviews. *Journal of Clinical Epidemiology*, 66, 1071–81.

Merlin, T., Lehman, S., Hiller, J.E., and Ryan, P. (2013). The 'linked evidence approach' to assess medical tests: a critical analysis. *International Journal of Technology Assessment in Health Care*, 29, 343–50.

Moberg, J., et al. (2018). The GRADE Evidence to Decision (EtD) framework for health system and public health decisions. *Health Research Policy and Systems*, 16, 45.

Moynihan, R., Doust, J., and Henry, D. (2012). Preventing overdiagnosis: how to stop harming the healthy. *British Medical Journal*, 344, e3502.

National Health and Medical Research Council (NHMRC) (n.d.). *Clinical Practice Guidelines Portal*. [Online] Available at: https://www.clinicalguidelines.gov.au/portal

National Institute for Health and Care Excellence (NICE) (2012). *Methods for the Development of NICE Public Health Guidance* (3rd ed.). London: NICE.

National Institute for Health and Care Excellence (NICE) (n.d.). *NICE Guidance Database*. [Online] London, UK. Available at: https://www.nice.org.uk/

Owolabi, M., Miranda, J.J., Yaria, J., and Ovbiagele, B. (2016). Controlling cardiovascular diseases in low and middle income countries by placing proof in pragmatism. *BMJ Glob Health*, 1, e000105.

Protopopoff, N., et al. (2018). Effectiveness of a long-lasting piperonyl butoxide-treated insecticidal net and indoor residual spray interventions, separately and together, against malaria transmitted by pyrethroid-resistant mosquitoes: a cluster, randomised controlled, two-by-two factorial design trial. *Lancet*, 391, 1577–88.

Rehfuess, E.A. and Akl, E.A. (2013). Current experience with applying the GRADE approach to public health interventions: an empirical study. *BMC Public Health*, 13, 9.

Sackett, D.L. (2002). Clinical epidemiology: what, who, and whither. *Journal of Clinical Epidemiology*, 55, 1161–6.

Pharmaceutical Benefits Scheme (2018). *Public Summary Documents*. [Online] Available at: http://www.pbs.gov.au/info/industry/listing/elements/pbac-meetings/psd

Schunemann, H.J., et al. (2007). WHO Rapid Advice Guidelines for pharmacological management of sporadic human infection with avian influenza A (H5N1) virus. *The Lancet Infectious Diseases*, 7, 21–31.

Selvarajah, S., et al. (2014). Comparison of the Framingham Risk Score, SCORE and WHO/ISH cardiovascular risk prediction models in an Asian population. *International Journal of Cardiology*, 176, 211–18.

Thornton, J., et al. (2013). Introducing GRADE across the NICE clinical guideline program. *Journal of Clinical Epidemiology*, 66, 124–31.

Woolf, S.H. (2008). The meaning of translational research and why it matters. *The Journal of the American Medical Association*, 299, 211–13.

World Health Organization (WHO) (2007). *Prevention of Cardiovascular Disease*. Geneva: WHO.

World Health Organization (WHO) (2014). *WHO Handbook for Guideline Development*. (2nd ed.). Geneva: WHO.

5.11

Validity and bias in epidemiological research

Sander Greenland and Tyler J. VanderWeele

Introduction to validity and bias in epidemiological research

Some of the major concepts of validity and bias in epidemiological research are outlined in this chapter. The contents are organized in four main sections: validity in statistical interpretation, validity in prediction problems, validity in causal inference, and special validity problems in case–control and retrospective cohort studies. Familiarity with the basics of epidemiological study design and a number of terms of epidemiological theory, among them risk, competing risks, average risk, population at risk, and rate, is assumed. A number of textbooks provide more background and depth than can be given here; see, for example, Kelsey et al. (1996), Koepsell and Weiss (2003), Checkoway et al. (2004), and Rothman et al. (2008).

Despite similarities, there is considerable diversity and conflict among the classification schemes and terminologies employed in various textbooks. This diversity reflects that there is no unique way of classifying validity conditions, biases, and errors. It follows that the classification schemes employed here and elsewhere should not be regarded as anything more than convenient frameworks for organizing discussions of validity and bias in epidemiological inference. Many types of bias can be qualitatively illustrated with causal diagrams, which reveal the relationships among confounding and selection bias; for example, see Greenland et al. (1999), Cole and Hermán (2002), Hernán et al. (2004), Glymour and Greenland (2008), and Pearl (2009).

Several important study designs, including randomized trials, prevalence (cross-sectional) studies, and ecological studies, are not discussed in this chapter. Such studies require consideration of the validity conditions

mentioned earlier and also require special considerations of their own. Further details of these and other designs can be found in the general textbooks cited in the preceding paragraphs. For discussions of the problems of ecological studies, see Greenland (2001, 2002a, 2004a) and Morgenstern (2008). Meta-analytic methods are discussed by Eddy et al. (1992), Greenland and O'Rourke (2008), and Borenstein et al. (2009). A number of central problems of epidemiological inference are also not covered, including choice of effect measures, problems of induction, and causal modelling. For critical discussions of effect measures by the present authors, see Greenland et al. (1986, 1991, 1999), Greenland (1987, 1999, 2002b), Greenland and Robins (1988), Rothman et al. (2008, chapter 4), and VanderWeele (2012). Greenland (1998a, 1998b) and Rothman et al. (2008, chapter 2) discuss problems of inductive and probabilistic inference, Greenland and Brumback (2002) review causal modelling in epidemiology, and Rothman et al. (2008, chapter 2) discuss broader issues in causal inference and philosophy.

Among the deeper problems not discussed here are the failure of conventional statistical methods to account for non-random sources of uncertainty, and the tendency of people (including scientists) to make overconfident and biased inferences in the face of uncertainty (Gilovich et al. 2002). Analytical approaches to these problems are discussed by Eddy et al. (1992), Greenland (2005), Greenland and Lash (2008), and Lash et al. (2009, 2014).

Inference and validity

Epidemiological inference is the process of drawing inferences from epidemiological data, such as prediction of disease patterns or identification of causes of diseases or epidemics. These inferences must often be made without the benefits of direct experimental evidence or established theory about disease aetiology. Consider the problem of predicting the risk and incubation (induction) time for acquired immunodeficiency syndrome (AIDS) among persons infected with type 1 human immunodeficiency virus (HIV-1). Unlike an experiment, in which the exposure is administered by the investigator, the date of HIV-1 infection cannot be accurately estimated in most cases; furthermore, the mechanism by which 'silent' HIV-1

infection progresses to AIDS is not known with certainty. Nevertheless, some prediction must be made from the available data if one is to prepare effectively for future healthcare needs.

As another example, consider the problem of estimating how much excess risk of coronary heart disease (if any) is produced by coffee drinking. Unlike an experimental exposure, coffee drinking is self-selected; it appears that persons who drink coffee are more likely to smoke than non-drinkers and probably tend to differ in many other behaviours as well (Greenland 1993). As a result, even if coffee use is harmless, we should not expect to observe the same pattern of heart disease in users and non-users. Thus, small effects of coffee drinking are very difficult to disentangle from the effects of other behaviours. Nevertheless, because of the high prevalence of coffee use and the high incidence of heart disease, any effect of coffee on heart disease risk may be of considerable public health importance.

In both these examples, and in general, inferences will depend on evaluating the validity of the available studies; that is, the degree to which the studies meet basic logical criteria for absence of bias. When (as is usually the case) aspects of validity cannot be guaranteed by successful design strategies, such as blinded random assignment (which ensures that baseline non-comparability is random), we must evaluate the magnitude of bias that would arise from plausible violations of validity conditions.

Because biases due to misinterpretation of statistics are often neglected in textbooks yet pervade the health-sciences literature, the next subsection will describe common misinterpretations to be alert for when examining reports or analysing data. The remaining subsections will outline and illustrate major concepts of validity and bias in epidemiological research as applied in three settings: prediction from one population to another, causal inference from cohort studies, and causal inference from case–control and retrospective cohort studies. Parallel aspects of each application will be emphasized. In particular, each problem requires the consideration of comparison bias (better known as confounding), follow-up bias (i.e. loss-to-follow-up or censoring bias), bias due to mismeasurement, bias due to erroneous statistical models, and bias due to erroneous statistical interpretation. (The term ‘bias’ is, here, used in the informal epidemiological sense, and corresponds to the formal statistical concept of inconsistency.) Case–control studies require the additional consideration of

case-selection bias and control-selection bias, and are often subject to additional sources of measurement error beyond those occurring in prospective cohort studies. Similar problems arise in retrospective cohort studies.

Misinterpretations of statistics

Severe consequences can arise from failure to properly interpret statistical outputs, and failure to recognize when the output is the result of a statistical method breaking down. Unfortunately, few textbooks discuss these problems in detail, even though they can result in inferential errors as severe as those from the more familiar biases discussed in the following subsections.

Two problems very common in the health-sciences literature are briefly described here. Both problems as well as many others can be avoided with certain Bayesian statistical approaches. Detailed arguments for Bayesian approaches can be found in Berger and Berry (1988), Greenland (1998b, 2006), and Thomas et al. (2007). For an introduction to Bayesian methods that can be carried out with ordinary software, see Greenland (2007) and Sullivan and Greenland (2013). The remainder of this subsection concerns only conventional 'frequentist' methods of significance tests, confidence intervals (CIs), *p*-values, and model fitting.

Misinterpretations of statistical significance

Many authors have noted that significance testing is very difficult to interpret correctly, so difficult that otherwise expert epidemiologists and even statisticians routinely make dramatic mistakes (Altman et al. 2000; Greenland 2004b, 2011, 2012a; Greenland and Poole 2011). Perhaps the largest source of misreporting and misjudgement in the scientific literature, however, is the simple confusion of presence or absence of an association with presence or absence of statistical significance (Altman and Bland 1995; Rothman et al. 2008, chapter 10; Greenland 2011). A common consequence of this confusion is the mistaken reporting of non-significant results as 'confirming' the null, when in reality the results are completely ambiguous. For example, a study that results in 95% confidence limits for a

relative risk (RR) of 0.80 and 2.0 is non-significant. Even if perfectly valid, however, it would not support the null hypothesis any more than it would support the hypothesis that the true RR is 1.5 (a 50% increase in risk). Misinterpreting non-significant results as null is easily avoided by paying attention to the confidence limits (Rothman 1978; Altman et al. 2000; Poole, 2001; Rothman et al. 2008, chapter 10). The limits of 0.80 and 2.0 show that the results are not 'significantly' different at the 0.05 level from any RR between 0.80 and 2.0. Unfortunately, mistaken claims that non-significant results support the null hypothesis of no association still pervade the field, especially in important policy and legal applications of statistics to epidemiology (Greenland 2004b, 2011, 2012a; Greenland and Poole 2011).

A converse misinterpretation is to say that two groups showed different results because they had different statistical significance. As an example, one article abstract (Park et al. 2007) reported that an inverse association with colorectal cancer was 'seen for total vitamin D intake in men (RR = 0.72, 95% CI: 0.51, 1.00; *p* for trend = 0.03) but not in women' (the estimate compares the highest quintile of vitamin D intake to the lowest). Examination of table 3 of the paper reveals, however, that the corresponding RR for women was 0.89 with 95% confidence limits of 0.63 and 1.27, which means that the association for women was in fact inverse, just as in men (0.89 = 11% lower risk in the highest quintile than in the lowest). Although it was not significantly different at the 0.05 level from 1 (the null), it was also not significantly different from the RR of 0.72 seen for men; nor was it significantly different from any RR from 0.63 to 1.27. One way to summarize these results would have been to state that an inverse association was observed in both men and women, but that it was weaker and non-significant for women. An even better interpretation is that the study had too few subjects to detect whether men and women actually differ with respect to the relation of vitamin D to colorectal cancer. Such honest reports of study weaknesses are not common, however. This issue is especially problematic in the interpretation of subgroup analyses in randomized trials (Rothwell 2005; Wang et al. 2007).

An opposite misinterpretation of the one just illustrated is to claim that two results are not significantly different because their confidence intervals overlap. It is true that they will not be significantly different if the interval from either group contains the point estimate from the other group (as in the preceding example, in which the men's point estimate of 0.72 falls within

the women's limits of 0.63 and 1.27), and that the two groups will be significantly different if their confidence intervals do not overlap. Nonetheless, if the intervals overlap but neither interval contains both point estimates, one cannot determine significance without conducting a direct test of the difference in results.

Misinterpretations such as just described arise in part from the conventional use of a cut-off value such as 0.05 to declare presence or absence of significance, and the corresponding use of 95% to determine the width of a confidence interval (which is just the collection of all values that are not significant at the 0.05 level). In reality, there is a continuum of degrees of significance, as measured by the *p*-value ('significance level'), and any cut-off value is both arbitrary and distortive (Rothman et al. 2008, chapter 10). Thus, some authors have encouraged a movement away from classifying associations 'significant' or 'non-significant', towards presenting instead the actual *p*-value for the tested hypothesis and for other values of interest (see Poole 1987a, 1987b; Goodman 1993; Greenland and Poole 2013a, 2013b).

Even if one presents *p*-values, however, the complexities and subtleties of determining and interpreting significance become even worse when multiple tests are involved (the 'multiple comparisons' problem). Here, adjustments for the multiplicity of tests can worsen certain problems even as they address others, for example, by worsening power even as it protects the validity (alpha level) of the tests. They may also shift the null hypothesis being tested to one that, if understood, might be seen as irrelevant to the scientific question at hand. For a discussion of these issues and further references, see Rothman et al. (2008, chapter 10) and Greenland and Poole (2011). Many of these problems can be avoided by shifting to a multilevel (hierarchical) approach for the analysis (Greenland 2000a, 2000b).

Sparse-data bias

Most epidemiological studies collect data on so many variables that any thorough stratification (tabulation) of the data on those variables will result in 'sparse data', in which many if not most cell counts may be very small. Classical epidemiological methods such as standardization break down in the face of this sparsity. Although modelling was developed to address

problems of sparse data, it is not a sure fix. Conventional methods for fitting risk and rate models such as maximum likelihood (ML), weighted least squares, and generalized estimating equations are ‘large-sample’ (asymptotic) techniques, meaning that they have sample-size requirements for valid performance. In the context of say logistic regression (the most common method in epidemiology), this means that among other things there are sufficient numbers of cases and non-cases observed at all levels of categorical variables in the model.

A serious problem is that when there are too few cases or non-cases in a given category a coefficient estimate based on that category may ‘blow up’, taking on a value that is much too large in absolute magnitude. This problem can arise in both unconditional and conditional logistic regression (Greenland et al. 2000). In a similar fashion, certain estimates may become inflated when there are too many variables in the model relative to the number of cases and non-cases. This hazard is especially severe when the data are divided into subgroups for separate analysis, for example, separating men and women instead of modelling effects with a variable for gender, or when ‘interaction’ (product) terms are used in the model. In almost all reports, the inflated estimates if ‘statistically significant’ are misinterpreted as representing strong effects rather than being recognized as statistical artefacts.

To spot these problems, investigators should check tables of the variables in the model against the outcome, and beware of estimates obtained when the observed numbers in each category fall below 4 or 5, or when the ratios of the number of cases and non-cases to the number of variables in the model fall below 4 or 5. Various methods for shrinkage estimation (including random-coefficient, hierarchical, multilevel, ridge, penalized, and Bayesian regression) can be used as an alternative to collapsing categories or deleting variables from the model to avoid sparse-data bias (Greenland 2000b, 2007; Sullivan and Greenland 2013).

Validity and bias in prediction problems

Turning now to more traditional topics, the following prediction problem will be used to illustrate several basic concepts of validity. A health clinic for homosexual men is about to begin enrolling HIV-l-negative men in an

unrestricted programme that will involve retesting each participant for HIV-1 antibodies at 6-month intervals. We can expect that, in the course of the programme, many participants will seroconvert to positive HIV-1 status. Such participants will invariably ask difficult questions, such as: What are my chances of developing AIDS over the next 5 years? How many years do I have before I develop AIDS? In attempting to answer these questions, it will be convenient to refer to such participants (i.e. those who seroconvert) as the target cohort. Even though membership of this cohort is not determined in advance, it will be the target of our predictions. It will also be convenient to refer to the time from HIV-1 infection until the onset of clinical AIDS as the AIDS incubation time. We could provide reasonable answers to a participant's questions if we could accurately predict AIDS incubation times, although we would also have to estimate the time elapsed between infection and the first positive test.

There might be someone who responds to the questions posed here with the following anecdote: 'I've known several men just like the ones in this cohort, and they all developed AIDS within 5 years after a positive HIV-1 test.' No trained scientist would conclude from this anecdote that all or most of the target cohort will develop AIDS within 5 years of seroconversion. Of course, one reason is that the men in the anecdote cannot be 'just like' men in our cohort in every respect: they may have been older or younger when they were infected; they may have experienced a greater degree of stress following their infection; they may have been heavier smokers, drinkers, or drug users; and so on. In other words, we know that the anecdotal men and their post-infection life events could not have been exactly the same as the men in our target cohort with respect to all factors that affect AIDS incubation time, including measured, unmeasured, and unknown factors. Furthermore, it may be that some or all of the men referred to in the anecdote had been infected long before they were first tested, so that (unlike men in our target cohort) the time from their first positive test to AIDS onset was much shorter than the time from seroconversion to AIDS onset.

Any reasonable predictions must be based on observing the distribution of AIDS incubation times in another cohort. Suppose that we obtain data from a study of homosexual men who underwent regular HIV-1 testing, and we then assemble from these data a study cohort of men who were observed to seroconvert. Suppose also that most of these men were followed for at

least 5 years after seroconversion. We cannot expect any member of this study to be ‘just like’ any member of our target cohort in every respect. Nevertheless, if we could identify no differences between the two cohorts with respect to factors that affect incubation time, we might argue that the study cohort could serve as a point of reference for predicting incubation times in the target cohort. Thus, we shall henceforth refer to the study cohort as our reference cohort.

Note that our reference and target cohorts may have originated from different populations; for example, the clinic generating the target cohort could be in New York, but the study that generated the reference cohort may have been in San Francisco. Of course, for both the target and reference cohorts, the actual times of HIV-1 infection will have to be imputed, based on the dates of the last negative and the first positive tests.

Suppose that our statistical analysis of data from the reference cohort produces estimates of 0.05, 0.25, and 0.45 for the average risk of contracting AIDS within 2, 5, and 8 years of HIV-1 infection. What conditions would be sufficient to guarantee the validity of these figures as estimates or predictions of the proportion of the target cohort that would develop AIDS within 2, 5, and 8 years of infection? If by ‘valid’ we mean that any discrepancy between our predictions and the true target proportions is unbiased or purely random (unpredictable in principle), the following conditions would be sufficient:

Comparison validity—the distribution of incubation times in the target cohort will be approximately the same as the distribution in the reference cohort.
Follow-up validity—within the reference cohort, the risk of censoring (i.e. follow-up ended by an event other than AIDS) is not associated with risk of AIDS.
Measurement validity—all measurements of variables used in the analysis closely approximate the true values of the variables. In particular, each imputed time of HIV-1 infection closely approximates the true infection time, and each reported time of AIDS onset closely approximates a clinical event defined as AIDS onset.
Specification validity—the distribution of incubation times in the reference cohort can be closely approximated by the statistical model used to compute the estimates. For example, if one employs a lognormal distribution to model the distribution of incubation times in the reference cohort, this model should be approximately correct.

The first condition concerns the external validity of making predictions about the target cohort based on the reference cohort. The remaining conditions concern the internal validity of the predictions as estimates of average risk in the reference cohort. The following subsections will explore the meaning of these conditions in prediction problems.

Comparison validity

Comparison validity is probably the easiest condition to describe, although it is difficult to evaluate. Intuitively, it simply means that the distribution of incubation times in the target cohort could be almost perfectly predicted from the distribution of incubation times in the reference cohort, if the incubation times were observed without error and there was no loss to follow-up. Other ways of stating this condition are that the two cohorts are comparable or exchangeable with respect to incubation times, or that the AIDS experience of the target cohort can be predicted from the experience of the reference cohort.

Confounding

If the two cohorts are not comparable, some or all of our risk estimates for the target cohort based on the reference cohort will be biased as a result. This bias is sometimes called *confounding*. There has been much research on methods for identifying and adjusting for such bias; see the textbooks cited earlier.

To evaluate comparison validity, we must investigate whether the two cohorts differ on any factors that influence incubation time. If so, we cannot reasonably expect the incubation time distributions of the two cohorts to be comparable. A factor responsible for some or all of the confounding in an estimate is called a confounder or confounding variable, the estimate is said to be confounded by the factor, and the factor is said to confound the estimate.

To illustrate these concepts, suppose that men infected at younger ages tend to have longer incubation times and that the members of the reference cohort are on average younger than members of the target cohort. If there were no other differences to counterbalance this age difference, we should then expect that members of the reference cohort will on average have longer incubation times than members of the target cohort. Consequently, unadjusted predictions of risk for the target cohort derived from the reference cohort would be biased (confounded) by age in a downward direction. In other words, age would be a confounder for estimating risk in the target cohort, and confounding by age would result in underestimation of the proportion of men in the target cohort who will develop AIDS within 5 years.

Now suppose that we can compute the age at infection of men in the reference cohort, and that within 1-year strata of age, for instance, the target and reference cohorts had virtually identical distributions of incubation times. The age-specific estimates of risk derived from the reference cohort would then be free of age confounding and so could be used as unconfounded estimates of age-specific risk for men in the target cohort. Also, if we wished to construct unconfounded estimates of average risk in the entire target cohort, we could do so via the technique of age standardization.

To illustrate, let Px denote our estimate of the average risk of AIDS within 5 years of infection among members of the reference cohort who become infected at age x. Let Wx denote the proportion of men in the target cohort who are infected at age x. Then the estimated average risk of AIDS

within 5 years of infection, standardized to the target cohort's age distribution, is simply the average of the age-specific reference estimates *Px* weighted by the age distribution (at infection) of the target cohort; algebraically, this average is the sum of the products *WxPx* over all ages and is denoted by *ΣxWxPx*. Considered as an estimate of the overall proportion of the target cohort that will contract AIDS within 5 years of HIV-1 infection, the standardized proportion *ΣxWxPx* will be free of age confounding.

The preceding illustration brings forth an important and often overlooked point: when one employs standardization to adjust for potential biases, the choice of standard distribution should never be considered arbitrary. In fact, the standard distribution should always be taken from the target cohort or the population about which inferences will be made. If inferences are to be made about several different groups, it may be necessary to compute several different standardized estimates.

Methods for removing bias in estimates by taking account of variables responsible for some or all of the bias are known as adjustment or covariate control methods. Standardization is perhaps the oldest and simplest example of such a method; methods based on multivariate models are more complex. See also Cole and Stuart (2010) for an alternative approach to such adjustment using weighting.

Unmeasured confounders

If all confounders were measured accurately, comparison validity could be achieved simply by adjusting for these confounders (although various technical problems might arise when attempting to do so). Nevertheless, in any non-randomized study, we would ordinarily be able to think of a number of possible confounders that had not been measured or had been measured only in a very poor fashion. In such cases, it may still be possible to predict the direction of uncontrolled confounding by examining the manner in which persons were selected into the target and reference cohorts from the population at large. If the cohorts are derived from populations with different distributions of predictors of the outcome, or the predictors themselves are associated with admission differentially across the cohorts, these predictors will become confounders in the analysis.

To illustrate this approach, suppose that HIV-1 infection via an intravenous route (e.g. through needle sharing) leads to shorter incubation times than HIV-1 infection through sexual activity. Suppose also that the reference cohort had excluded all or most intravenous drug users, whereas the target cohort was non-selective in this regard. Then incubation times in the target cohort will on average be shorter than times in the reference cohort owing to the presence of intravenously infected persons in the target cohort. Thus we should expect the results from the reference cohort to underestimate average risks of AIDS onset in the target cohort.

Random sampling and confounding

Suppose, for the moment, that our reference cohort had been formed by taking a random sample of the target cohort. Can predictions about the target made from such a random sample still be confounded? With the definition of confounding given here, the answer is yes. To see this, note for example that by chance alone men in our sample reference cohort could be younger on average than the total target; this age difference would in turn downwardly bias the unadjusted risk predictions if men had longer incubation times at younger ages.

Nevertheless, random sampling can help to ensure that the distribution of the reference cohort is not too far from the distribution of the target cohort. In essence, the probability of severe confounding can be made as small as necessary by increasing the sample size. Furthermore, if random sampling is used, any confounding left after adjustment will be accounted for by the standard errors of the estimates, provided that the correct statistical model is used to compute the estimates and standard errors. We shall examine the latter condition under the subsection on specification validity.

Follow-up validity

In any cohort study covering an extended period of risk, subjects will be followed for different lengths of time. Some subjects will be lost to follow-up before the study ends. Others will be removed from the study by an event that precludes AIDS onset, which in this setting is death before AIDS onset from fatal accidents, fatal myocardial infarctions, and so on. Because subjects come under study at different times, those who are not lost to

follow-up or who die before developing AIDS will still have had different lengths of follow-up when the study ends; traditionally, a subject still under follow-up at the end of study is said to have been ‘withdrawn from study’ at this time.

Suppose that we wish to estimate the average risk of AIDS onset within 5 years of infection. The data from a member of the reference cohort who is not observed to develop AIDS but is also not followed for the full 5 years from infection are said to be censored for the outcome of interest (AIDS within 5 years of infection). Consider, for example, a subject killed in a car crash 2 years after infection but before contracting AIDS: the incubation time of this subject was censored at 2 years of follow-up.

Follow-up validity means that over any span of follow-up time, risk of censoring is unassociated with risk of the outcome of interest. In our example, follow-up validity means that over any span of time following infection, risk of censoring (loss, withdrawal, or death before AIDS) is unassociated with risk of AIDS. All common methods for estimating risk from situations in which censoring occurs (e.g. person-years, life table, and Kaplan–Meier methods) are based on the assumption of follow-up validity. Given follow-up validity, we can expect that, at any time *t* after infection, the distribution of incubation times will be the same for subjects lost or withdrawn at *t* and for subjects whose follow-up continues beyond *t*.

Violations of follow-up validity can result in biased estimates of risk; such violations are referred to as follow-up bias or biased censoring. To illustrate, suppose that younger reference subjects tend to have longer incubation times (i.e. lower risks) and are lost to follow-up at a higher rate than older reference subjects. In other words, lower-risk subjects are lost at a higher rate than higher-risk subjects. Then, after enough time, the average risk of AIDS in the observed portion of the reference cohort will tend to be overestimated; that is, higher than the average risk occurring in the full reference cohort (as the latter includes both censored and uncensored subject experience).

The follow-up bias in the last illustration would not affect the age-specific estimates of risk (where age refers to age at infection). Consequently, the age bias in follow-up would not produce bias in age-standardized estimates of risk. More generally, if follow-up bias can be traced to a particular variable that is a predictor of both the outcome of interest and censoring, bias in the estimates can be removed by adjusting for

that variable. Thus, some forms of follow-up bias can be dealt with in the same manner as confounding.

Measurement validity

An estimate from a study can be said to have measurement validity if it suffers from no bias due to errors in measuring the study variables. Unfortunately, there are sources of measurement error in nearly all epidemiological studies, and nearly all sources of measurement error will contribute to bias in estimates. Thus, evaluation of measurement validity primarily focuses on identifying sources of measurement error and attempting to deduce the direction and magnitude of bias produced by these sources.

To aid in the task of identifying sources of measurement error, it can be useful to classify such errors according to their source. Errors from specific sources can then be further classified according to characteristics that are predictive of the direction of the bias they produce. One classification scheme divides errors into three major categories, according to their source:

Procedural error: arising from mistakes or defects in measurement procedures
Proxy-variable error: arising from using a 'proxy' variable as a substitute for an actual variable of interest
Construct error: arising from ambiguities in the definition of the variables.

Regardless of their source, errors can be divided into two basic types, differential and non-differential, according to whether the direction or magnitude of error depends on the true values of the study variables. Two different sources of error may be classified as dependent or independent, according to whether or not the direction or magnitude of the error from one source depends on the direction or magnitude of the error from the other source. Finally, errors in continuous measurements can be factored into systematic and random components. As described in the following subsections, these classifications have important implications for bias.

Procedural error

Procedural error is the most straightforward to imagine. It includes errors in recall when variables are measured through retrospective interview (e.g. mistakes in remembering all medications taken during pregnancy). It also includes coding errors, errors in calibration of instruments, and all other errors in which the target of measurement is well defined and the attempts at measurement are direct but the method of measurement is faulty. In our example, one target of measurement is HIV-1 antibody presence in blood. All available tests for antibody presence are subject to error (false negatives and false positives), and these errors can be considered to be procedural errors of measurement.

Proxy-variable error

Proxy-variable error is distinguished from procedural error in that use of proxies necessitates imputation and hence virtually guarantees that there will be measurement error. In our example, we must impute the time of HIV-1 infection. For instance, we might take as a proxy the infection time computed as 6 weeks before the midpoint between the last negative test and the first positive test for HIV-1 antibodies. Even if our HIV-1 tests are perfect, this measurement incorporates error if (as is certainly the case) time of infection does not always occur 6 weeks before the midpoint between the last negative and first positive tests.

Construct error

Construct error is often overlooked, although it may be a major source of error. Consider our example in which the ultimate target of measurement is the time between HIV-1 infection and onset of AIDS. Before attempting to measure this time span, we must unambiguously define the events that mark the beginning and end of the span. While it may be reasonable to think of HIV-1 infection as a point event, the same cannot be said of AIDS onset. Symptoms and signs may gradually accumulate, and then it is only by convention that some point in time is declared the start of the disease. If this convention cannot be translated into reasonably precise clinical criteria for diagnosing the onset of AIDS, the construct of incubation time (the time span between infection and AIDS onset) will not be well defined let alone accurately measurable. In such situations, it may be left to various clinicians

to improvise answers to the question of time of AIDS onset, and this will introduce another source of extraneous variation into the final 'measurement' of incubation time.

Differential and non-differential error

Errors in measuring a variable are said to be differential when the direction or magnitude of the errors tend to vary across the true values of other variables. Suppose, for example, that recall of drug use during pregnancy is enhanced among mothers of children with birth defects. Then, a retrospective interview about drug use during pregnancy will yield results with differential error, as false-negative error will occur more frequently among mothers whose children have no birth defects.

Another type of differential error occurs in the measurement of continuous variables when the distribution of errors varies with the true value of the variable. Suppose, for example, that women more accurately recall the date of a recent cervical smear (Papanicolaou) test than the date of a more distant test. Then, a retrospective interview to determine length of time since a woman's last cervical smear test would tend to suffer from larger errors when measuring longer times.

Errors in measuring a variable are said to be non-differential with respect to another variable if the magnitudes of errors do not tend to vary with the true values of the other variable. Measurements are usually assumed to be non-differential if neither the subject nor the person taking the measurement knows the values of other variables. For example, if drug use during pregnancy is measured by examining pre-partum prescription records for the mother, it would ordinarily be assumed that the error will be non-differential with respect to birth defects discovered postnatally. Nevertheless, such 'blind' assessments will not guarantee non-differential error if the measurement scale is not as fine as the scale of the original variable (Flegal et al. 1991; Wacholder et al. 1991) or if there is a third uncontrolled variable that affects both the measurement and the other study variables.

Dependent and independent error

Errors in measuring two variables are said to be dependent if the direction or magnitude of the errors made in measuring one of the variables is associated with the direction or magnitude of the errors made in measuring the other variable, conditional on the true variables. If there is no such association of errors, the errors are said to be independent.

In our example, errors in measuring age at HIV-1 infection and AIDS incubation time are dependent. Our measure of incubation time is equal to our measure of age at AIDS onset minus our measure of age at infection; hence, overestimation of age at infection will contribute to underestimation of incubation time, and underestimation of age at infection will contribute to overestimation of incubation time. In contrast, in the same example, it is plausible that the errors in measuring age at infection and age at onset are independent.

Systematic and random components of error

For well-defined measurement procedures on continuous variables, measurement errors can be subdivided into systematic and random components. The systematic component (sometimes called the bias of the measurement) measures the degree to which the procedure tends to underestimate or overestimate the true value on repeated application. The random component is the residual error left after subtracting the systematic component from the total error.

To illustrate, suppose that in our study HIV-1 infection time was unrelated to time of antibody testing and that the average time of HIV-1 seroconversion was 8 weeks after infection. Then, even if one used a perfect HIV-1 test, a procedure that estimated infection time as 6 weeks before the midpoint between the last negative and first positive test would on average yield an estimated infection time that was 2 weeks later than the true time. Thus, the systematic component of the error of this procedure would be +2 weeks. Because AIDS incubation time is AIDS onset time minus HIV-1 infection time, use of this procedure would add –2 weeks (i.e. a 2-week underestimation) to the systematic component of error in estimating incubation time.

Each of the components of an error, systematic and random, may be differential (i.e. may vary with other variable values) or non-differential, and may or may not be independent of the error components in other

variables. We shall not explore the consequences of the numerous possibilities. However, one important (but semantically confusing) fact is that, for certain quantities, independent and non-differential systematic components of error will not harm measurement validity in that they will produce no bias in estimation (Poole 1985).

To illustrate, suppose that in our example we wish to estimate the degree to which AIDS incubation time depends on age at HIV-1 infection. Suppose also that the systematic components of the measurements of incubation time and age of infection are –2 weeks and +2 weeks (as mentioned), and do not vary with true incubation time or age at measurement (i.e. the systematic components are non-differential). Then, the systematic components, being equal, will cancel out when we compute differences in incubation time and differences in age at infection. Because only these differences are used to estimate the association, the observed dependence of incubation time on age at infection will not be affected by the systematic components of error (although it may be biased by the random components of error).

Specification validity

All statistical techniques, including so-called ‘distribution-free’ or ‘non-parametric’ methods as well as basic contingency table methods, are derived by assuming the validity of a sampling model or error distribution. A common example is the binomial model, which is discussed in all the textbooks cited in the introduction. For parametric methods, the sampling model is a mathematical formula that expresses the probability of observing the various possible data patterns as a function of certain unknown constants (parameters). Although the parameters of this model may be unknown, the mathematical form of this model incorporates only known or purely random aspects of the data-generation process; unknown systematic aspects of this process (such as most follow-up and selection biases) will not be accounted for by the model.

All parametric statistical techniques also assume a structural model, which is a mathematical formula that expresses the parameters of the sampling model as a function of study variables. A common example is the logistic model (Kelsey et al. 1996; Hosmer and Lemeshow 2000; Checkoway et al. 2004; Jewell 2004; Greenland 2008a, 2008b). The structural model is most often incorporated into the sampling model, and

the combination is referred to as the statistical model. An estimate can be said to have specification validity if it is derived using a statistical model that is correct or nearly so.

If either the sampling model or the structural model used for analysis is incorrect, the resulting estimates may be biased. Such bias is sometimes called specification bias, and the use of an incorrect model is known as model misspecification or specification error. Even when misspecification does not lead to bias, it can lead to invalidity of statistical tests and confidence intervals.

The true structural relation among the study variables is almost never known in studies of human disease. Furthermore, in the absence of random sampling and randomization, the true sampling process (i.e. the exact process leading people to enter and stay in the study groups) will also be unknown. It follows that we should ordinarily expect some degree of specification error in an epidemiological analysis. Minimizing such error largely consists of contrasting the statistical model against the data and against any available information about the processes that generated the data, such as prior information on demographic patterns of incidence.

Many statistical techniques in epidemiology are based on assuming some type of logistic model. Examples include all the popular adjusted odds ratios, such as the Woolf, ML, and Mantel–Haenszel estimates, as well as tests for odds ratio heterogeneity. Classical ‘indirect’ adjustment of rates and other comparisons of standardized morbidity ratios depend on similar multiplicative models for their validity (Breslow and Day 1987; VanderWeele and Vansteelandt 2010).

The degree of bias in traditional epidemiological analysis methods when the model assumptions fail has not been extensively studied. A few traditional methods, such as directly standardized comparisons and the Mantel–Haenszel test, remain valid under a wide variety of structural models. In addition, risk regression has been extended to situations involving more general models than assumed in classical theory (Breslow and Day 1987; Hastie and Tibshirani 1990). Leamer (1978) and White (1993) give more details on the effects of specification error in multiple regression problems, and Maldonado and Greenland (1994) and Greenland and Maldonado (1994) examine the implications of specification error in epidemiology.

Summary of prediction example

The example in this subsection provides an illustration of the most common threats to the validity of predictions. The unadjusted estimates of AIDS risk may be confounded if the target and reference cohorts differ in composition, and may also be biased by losses to follow-up or use of an incorrect statistical model. Finally, our predictions are likely to be compromised by errors in measurements. These sources of error should be borne in mind in any attempt to predict AIDS incidence.

Validity and bias in causal inference

All the bias problems in prediction arise in studies of causation; confounding, follow-up bias, measurement errors, and specification errors must be considered. In fact, as we shall see, problems of causal inference can be viewed as a special type of prediction problem, namely prediction of what would happen (or what would have happened) to a population if certain characteristics of the population were (or had been) altered (Greenland 2012b).

To illustrate the validity issues in causal inference, we shall consider the hypothesis that coffee drinking causes acute myocardial infarction. This hypothesis can be operationally interpreted in a number of ways, such as:

There are people for whom the consumption of coffee results in their experiencing a myocardial infarction sooner than they might have, had they avoided coffee.

Although this hypothesis is appealingly precise, it offers little practical guidance to an epidemiological researcher. The problem lies in our inability to recognize an individual whose myocardial infarction was caused by coffee drinking. It is quite possible that myocardial infarctions precipitated by coffee use are clinically and pathologically indistinguishable from myocardial infarctions due to other causes. If so, the prospect of testing this hypothesis with purely physiological evidence is not good.

We could overcome this impasse by examining a related epidemiological hypothesis; that is, a hypothesis that refers to the distribution of disease in populations. One of many such hypotheses is:

Among five-cup-a-day coffee drinkers, cessation of coffee use will lower the frequency of myocardial infarction.

This form not only involves a population (five-cup-a-day coffee drinkers) but also asserts that a mass action (coffee cessation) will reduce the frequency of the study disease. Thus, the form of the hypothesis immediately suggests a strong test of the hypothesis: conduct a randomized intervention trial to examine the impact of coffee cessation on myocardial infarction frequency. This solution has some profound practical limitations, not least of which would be persuading anyone to give up or take up coffee drinking to test a speculative hypothesis.

Having ruled out intervention, we might consider an observational cohort study. In this case, our epidemiological hypothesis must refer to natural conditions, rather than intervention. One such hypothesis is:

Among five-cup-a-day coffee drinkers, coffee use has elevated the frequency of myocardial infarction.

There have been many conflicting cohort and case–control studies of coffee and myocardial infarction. The present discussion will be confined to the issues arising in the analysis of a single study. For a review of issues arising in the analysis of multiple studies (meta-analysis) see Greenland (1994) and Greenland and O'Rourke (2008).

Consider a cohort study of coffee and first myocardial infarction. At baseline, a cohort of people with no history of myocardial infarction is assembled and classified into subcohorts according to coffee use (e.g.

never-drinkers, ex-drinkers, occasional drinkers, one-cup-a-day drinkers, two-cup-a-day drinkers, etc.). Other variables are measured as well: age, gender, smoking habits, blood pressure, and serum cholesterol. Suppose that at the end of 10 years of monitoring this cohort for myocardial infarction events, we compare the five-cup-a-day and never-drinker subcohorts, and obtain an unadjusted estimate of 1.22 for the ratio of the person-time incidence rates of first myocardial infarction among five-cup-a-day drinkers and never-drinkers (with 95% CIs of 1.00 and 1.49). In other words, it appears that the rate of first myocardial infarction among five-cup-a-day drinkers was 1.22 times higher than the rate among never-drinkers. (Hereafter, myocardial infarction means first myocardial infarction, risk means average risk, and rate means person-time incidence rate.)

The estimated rate ratio of 1.22 may not seem large. Nevertheless, if it accurately reflects the impact of coffee use on the five-cup-a-day subcohort, this estimate implies that persons drinking five cups a day at baseline suffered a 22% increase in their myocardial infarction rate as a result of their coffee use. Given the high frequency of both coffee use and myocardial infarction in many populations, this could represent a substantial health impact. Therefore, we would want to perform a careful evaluation of possible bias in the estimate.

As in the AIDS example, we can proceed by examining a series of conditions sufficient for validity of the estimate as a measure of the effect of coffee:

Comparison validity (no confounding)—if the members of the five-cup-a-day subcohort had instead never drunk coffee, their distribution of myocardial infarction events over time would have been approximately the same as the distribution among the never-drinkers.
Follow-up validity—within each subcohort, the risk of censoring (i.e. follow-up ended by an event other than myocardial infarction) is not associated with the risk of myocardial infarction.
Measurement validity—all measurements of variables used in the analysis closely approximate the true values of the variables.
Specification validity—the distribution of myocardial infarction events over time in the subcohorts can be closely approximated by the statistical model on which the estimates are based.

These four conditions are sometimes called *internal validity conditions* because they pertain only to estimating effects within the study cohort rather than to generalizing results to other cohorts. They are sufficient but not necessary for validity, in that certain violations of the conditions will not produce bias in the effect estimate (although most violations will produce some bias).

The meaning of these conditions for an observational cohort study of a causal hypothesis is explored in the following subsections. An important phenomenon known as *effect modification*, which is relevant to both internal validity and generalizability, will also be discussed following comparison validity.

Comparison validity

In our example, comparison validity means that the distribution of myocardial infarctions among never-drinkers accurately predicts what would have happened in the coffee drinking groups had the members of these groups never drunk coffee. Another way of stating condition C is that the five-cup-a-day and never-drinker subcohorts would be comparable or exchangeable with respect to myocardial infarction times if no one had ever drunk coffee, or that there is no confounding (Greenland et al. 1991; Rothman et al. 2008, chapter 4).

Despite its simplicity, comparison validity depends on the hypothesis of interest in a very precise way. In particular, the research hypothesis (hypothesis 3 in the preceding subsection) is a statement about the impact of coffee among five-cup-a-day drinkers. Thus, this subcohort is the target cohort, and never-drinkers serve as the reference cohort for making predictions about this target.

To illustrate further the correspondence between comparison validity and the hypothesis at issue, suppose for the moment that our research hypothesis was:

> 4. Among never-drinkers, five-cup-a-day coffee use would elevate the frequency of myocardial infarction.

In examining this hypothesis, the never-drinkers would be the target cohort and the coffee drinkers would be the reference cohort. Thus, the comparison validity condition would have to be replaced by a condition such as

(C') If the never-drinkers had drunk five cups of coffee per day, their distribution of myocardial infarctions would have been approximately the same as the distribution among five-cup-a-day drinkers.

Other ways of stating condition C' are that the five-cup-a-day and never-drinker subcohorts would be comparable or exchangeable with respect to time to myocardial infarction if everyone had been five-cup-a-day drinkers, and that the myocardial infarction experience of five-cup-a-day drinkers accurately predicts what would have happened to the never-drinkers if the latter had drunk five cups a day.

Confounding

Failure to meet condition C results in a biased estimate of the effect of five-cup-a-day coffee drinking on five-cup-a-day drinkers, a condition sometimes referred to as confounding of the estimate. Similarly, failure to meet condition C' results in a biased estimate of the effect that five-cup-a-day drinking would have had on never-drinkers.

To evaluate potential confounding, we must check whether the subcohorts differed at baseline on any factors that influence time to myocardial infarction. If so, we could not reasonably expect the myocardial infarction distributions of the subcohorts to be comparable, even if the subcohorts had the same level of coffee use. In other words, we could not expect condition C (or C') to hold, and so we should expect our estimates to suffer from confounding. This is so, regardless of whether adjustment appears to change the association of coffee use and myocardial infarction (Greenland et al. 1999).

Some studies have found a positive association between cigarette smoking (an established risk factor for myocardial infarction) and coffee use (Greenland 1993). It also seems a priori sensible that a person habituated to a stimulant such as nicotine would be attracted to coffee use as well. Thus, we should expect to see a higher prevalence of smoking among coffee users in our study.

Suppose then that, in our cohort, smoking is more prevalent among five-cup-a-day subjects than never-drinkers. This elevated smoking prevalence should have led to elevated myocardial infarction rates among five-cup-a-day drinkers, even if coffee had no effect. More generally, we should expect the myocardial infarction rate among never-drinkers to underestimate the

myocardial infarction rate that five-cup-a-day drinkers would have had if they had never drunk coffee. The result would be an inflated estimate of the impact of coffee on the myocardial infarction rate of five-cup-a-day drinkers. Similarly, we should expect the myocardial infarction rate among five-cup-a-day drinkers to overestimate the myocardial infarction rate that never-drinkers would have had if they had drunk five cups a day.

Adjustment for measured confounders

As in the prediction problem, we can stratify the data on potential confounders with the objective of creating strata within which confounding is minimal or absent. We can also employ standardization to remove confounding from estimates of overall effect. Again, some care in the selection of the standard is required.

To illustrate, let R_{xz} denote the estimated rate of myocardial infarction among cohort members who drank x cups of coffee per day and smoked z cigarettes per day at baseline, with R_{0z} denoting the estimated rate among never-drinkers. Let W_{xz} denote the proportion of person-time among x-cup-per-day drinkers that was contributed by z-cigarette-per-day smokers. Finally, let R_{xc} be the unadjusted rate observed among cohort members who drank x cups per day at baseline, with R_{0c} denoting the estimated unadjusted rate among never-drinkers.

Suppose for the moment that any change in coffee-use patterns would have negligible impact on the person-time distribution of smoking in the cohort. The predicted (i.e. expected) rate among five-cup-a-day drinkers had they never drunk coffee, adjusted for confounding by smoking, is the average of the smoking-specific estimates from the never-drinker (reference) subcohort weighted by the smoking distribution of the five-cup-per-day (target) cohort. Algebraically, this average is the following sum (over z):

This sum is commonly termed the rate in the never-drinkers standardized to the distribution of smoking among five-cup-a-day drinkers. Such

terminology obscures the fact that the sum is a prediction about the five-cup-a-day drinkers, not the never-drinkers.

Given the last computation, a smoking-standardized estimate of the increase in myocardial infarction rate produced by coffee drinking among five-cup-per-day drinkers is the rate ratio standardized to the five-cup-per-day smoking distribution:

$$\sum_z W_{5z} R_{5z} / \sum_z W_{5z} R_{0z},$$

This formula reveals a property common to a simple standardized rate ratio: the same weights W_{xz} must be used in the numerator and denominator sums. Some insight into this formula can be obtained by noting that the unadjusted rate R_{5c}, among the five-cup-a-day drinkers, is equal to:

$$\sum_z W_{5z} R_{5z},$$

so that the standardized rate ratio can be rewritten as:

$$R_c \sum_z W_{5z} R_{0z} / \sum_z W_{5z} R_{0z},$$

This version shows that the ratio is a classical observed unadjusted rate over the rate expected without exposure, or standardized morbidity ratio (SMR). Another standardized rate ratio is:

$$\sum_z W_{0z} R_{5z} / \sum_z W_{0z} R_{0z},$$

This differs from the previous standardized ratio in that the weights are taken from the never-drinkers (W_{0z}) instead of five-cup-a-day drinkers (W_{5z}). Insight into this formula can be obtained by noting that the numerator sum is simply a prediction (expectation) of what would have happened to the never-drinkers if they had been five-cup-a-day drinkers, and the denominator sum is equal to the unadjusted rate R_{0c}, among never-drinkers. Thus, the last standardized ratio is a smoking-standardized estimate of the

increase in the myocardial infarction rate that five-cup-a-day drinking would have produced among the never-drinkers.

Standardization is appealingly simple in both justification and computation. Unfortunately, if the number of cases occurring within the confounder categories tends to be small (under five or so), the technique will be subject to various technical problems including possible bias. These problems can be avoided by broadening confounder categories or by not adjusting for some of the measured confounders. Unfortunately, both these strategies are likely to result in incomplete control of confounding. To avoid having to adopt these strategies, many researchers attempt to control confounding by using a multivariate model. This remedy has problems of its own, some of which we shall address in the subsection on specification validity.

Another problem is that standardized procedures (as well as typical modelling procedures) take no account of potential exposure effects on the adjustment variables or their distribution. Thus, in the preceding example, to justify use of the fixed weights W_{xz} we had to invoke the dubious assumption that changes in coffee use would only negligibly affect the smoking distribution. We shall briefly discuss this issue in the subsection on intermediate variables.

Unmeasured confounders

Among the possible confounders not measured in our hypothetical study are diet and exercise. Suppose that ‘health conscious’ subjects who exercise regularly and eat low-fat diets also avoid coffee. The result will be a concentration of these lower-risk subjects among coffee non-users and a consequent overestimation of coffee’s effect on risk.

Confounding by unmeasured confounders can sometimes be minimized by controlling variables along the pathways of their effect. For example, if exercise and a low-fat diet lowered myocardial infarction risk only by lowering serum cholesterol and blood pressure, control of serum cholesterol and blood pressure would remove confounding by exercise and dietary fat. Unfortunately, such control may also generate bias if the controlled variables are intermediates between our study variable and our outcome variable.

If external information is available to indicate the relationship in our study between an unmeasured confounder and the study variables, we can attempt to use an indirect method to adjust for the confounder; even if external information is unreliable or unavailable, we can examine the sensitivity of our results to unmeasured confounders (Greenland and Lash 2008; VanderWeele and Arah 2011). Sometimes we can also predict the direction of the bias resulting from confounding (VanderWeele 2008; VanderWeele et al. 2008). Finally, we can evaluate the impact that uncertainty about unmeasured confounders should have on our final inferences about the effect (Greenland and Lash 2008; Lash et al. 2009).

Randomization and confounding

Suppose, for the moment, that the level of coffee use in our cohort had been assigned by randomization and that the participants diligently consumed only their assigned amount of coffee. Could our estimates of coffee effects from such a randomized trial still be confounded? By our earlier definition of confounding, the answer is yes. To see this, note for example that by chance alone the five-cup-a-day drinkers could be older on average than the never-drinkers; this difference would in turn result in an upward bias in the unadjusted estimate of the effect of five cups a day, because age is an important risk factor for myocardial infarction.

Nevertheless, randomization can help to ensure that the distributions of confounders in the different exposure groups are not too far apart. In essence, the probability of severe confounding can be made as small as necessary by increasing the size of the randomized groups. Furthermore, if randomization is used and subjects comply with their assigned treatments, any confounding left after adjustment will be accounted for by the standard errors of the estimates, provided that the correct statistical model is used to compute the effect estimates and their standard errors (Robins 1988; Greenland 1990).

Intermediate variables

In effect estimation, we must take care to distinguish intermediate variables from confounding variables. Intermediate variables represent steps in the causal pathway from the study exposure to the outcome event. The

distinction is essential, for control of intermediate variables can increase the bias of estimates.

To illustrate, suppose that coffee use affects serum cholesterol levels (as suggested by the results of Curb et al. 1986). Then, given that serum cholesterol affects myocardial infarction risk, serum cholesterol is an intermediate variable for the study of the effects of coffee on this risk. Now suppose that we stratify our cohort data on serum cholesterol levels. Some coffee drinkers will be in elevated cholesterol categories because of coffee use and so will have an elevated risk of myocardial infarction, yet these subjects will be compared with never-drinkers in the same stratum who are also at elevated risk due to their elevated cholesterol. Therefore, the effect of coffee on the risk of myocardial infarction via the cholesterol pathway will not be apparent within the cholesterol strata, and so cholesterol adjustment will contribute to underestimation of the effect of coffee on risk of myocardial infarction. Analogously, if coffee affected the risk of myocardial infarction by elevating blood pressure, blood pressure adjustment will contribute to underestimation of the effect of coffee. Such underestimation can be termed overadjustment bias.

Intermediate variables may also be confounders and thus present the investigator with a severe dilemma. Consider that most of the variation in serum cholesterol levels is not due to coffee use and that much (perhaps most) of the association between coffee use and cholesterol is not due to the effects of coffee, but rather to factors associated with both coffee and cholesterol (such as exercise and dietary fat). This means that serum cholesterol may also be viewed as a confounder for the coffee–myocardial infarction study and that estimates unadjusted for serum cholesterol will be biased unless they are also adjusted for the factors contributing to the coffee–cholesterol association.

Suppose that a variable is both an intermediate and a confounder. It will usually be impossible to determine how much of the change in the effect estimate produced by adjusting for the variable is due to introduction of overadjustment bias and how much is due to removal of confounding. Nevertheless, a qualitative assessment may be possible in some situations. For example, if we know that the effects of coffee on serum cholesterol are weak and that most of the association between coffee and serum cholesterol is due to confounding of this association by uncontrolled factors (such as exercise and diet), we can conclude that the cholesterol-adjusted estimate is

probably the less biased of the two. Alternatively, if we have accurately measured all the factors that confound the coffee–cholesterol association, we can control these factors instead of cholesterol to obtain an estimate free of both overadjustment bias and confounding by cholesterol. Finally, if we have multiple measurements of coffee use and cholesterol over time, techniques are available that adjust for the confounding effects of cholesterol but do not introduce overadjustment (Hernán et al. 2000; Robins et al. 2000).

Direct and indirect effects

Often, one may wish to estimate how much of the effect under study is indirect relative to an intermediate variable (in the sense of being transmitted through the intermediate), or how much of the effect is direct relative to the intermediate (i.e. not mediated by the intermediate). For example, we might wish to estimate how much of the effect of coffee on myocardial infarction risk is due to its effect on serum cholesterol, or how much is due to the effects of coffee on cardiovascular variables other than cholesterol.

One common approach to this problem is to adjust the coffee–myocardial infarction association for serum cholesterol level via ordinary stratification or regression methods and then use the resulting estimate as the estimate of the direct effect of coffee. This procedure is potentially biased as it may introduce new confounding by determinants of serum cholesterol, even if these determinants did not confound the total (unadjusted) association (Robins and Greenland 1992, 1994; Cole and Hermán 2002; Glymour and Greenland 2008). Nonetheless, given sufficient data, it is possible to obtain separate estimates for direct and indirect effects using special stratification or modelling techniques (Robins and Greenland 1994; Kaufman et al. 2005; Vansteelandt 2012; Valeri and VanderWeele 2013), but again these techniques make much stronger assumptions than those for total effects. These assumptions can be assessed to a certain extent using sensitivity analysis (VanderWeele 2010).

Effect modification (heterogeneity of effect)

Estimation of effects usually requires consideration of effect-measure modification, which is also known as effect modification, effect variation, or heterogeneity of effect. As an example, suppose that drinking five cups of coffee a day elevated the myocardial infarction rate of men in our cohort by a factor of 1.40 (i.e. a 40% increase), but elevated the myocardial infarction rate of women by a factor of only 1.10 (i.e. a 10% increase). This situation would be termed modification (or variation or heterogeneity) of the rate ratio by gender, and gender would be called a modifier of the coffee–myocardial infarction rate ratio.

As another example, suppose that drinking five cups of coffee a day elevated the myocardial infarction rate in men in our cohort by a factor of 400 cases per 100,000 person-years but elevated the rate in women by a factor of only 40 cases per 100,000 person-years. This situation would be termed modification of the rate difference by gender, and gender would be called a modifier of the coffee–myocardial infarction rate difference.

As a final example, suppose that drinking five cups of coffee per day elevated the myocardial infarction rate in our cohort by a factor of 1.22 in both men and women. This situation would be termed homogeneity of the rate ratios across gender.

Effect modification and homogeneity are not absolute properties of an effect, but instead are properties of the way that the effect is measured. For example, suppose that drinking five cups of coffee per day elevated the myocardial infarction rate in men from 1,000 cases per 100,000 person-years to 1,220 cases per 100,000 person-years, but elevated the rate in women from 400 cases per 100,000 person-years to 488 cases per 100,000 person-years. Then the gender-specific rate ratios would both be 1.22, homogeneous across gender. In contrast, the gender-specific rate differences would be 220 cases per 100,000 person-years for males and 88 cases per 100,000 person-years for females, and so are heterogeneous or ‘modified’ by gender. Examples such as this one show that one should not equate effect modification with biological concepts of interaction such as synergy or antagonism (Greenland and O’Rourke 2008; VanderWeele 2009).

Effect modification can be analysed by stratifying the data on the potential effect modifier under study, estimating the effect within each stratum, and comparing the estimates across strata. There are several potential problems with this approach. The number of subjects in each stratum may be too small to produce stable estimates of stratum-specific

effects, particularly after adjustment for confounder effects. Estimates may fluctuate wildly from stratum to stratum owing to random error. A related problem is that statistical tests for heterogeneity in stratified data have extremely low power in many situations, and therefore are likely to miss much if not most of the heterogeneity when used with conventional significance levels (such as 0.05). Finally, the amount of bias from confounding, loss to follow-up, measurement error, and other sources may vary from stratum to stratum, in which case the observed pattern of modification will be biased.

Effect modification and generalizability

Suppose that we succeed in obtaining approximately unbiased (internally valid) estimates from our study. We can then confront the issues of generalizability (external validity) of our results. For example, we can ask whether they accurately reflect the effect of coffee on myocardial infarction rates in a new target cohort. We can view such a question as a prediction problem in which the objective is to predict the strength of the effects of coffee in the new target cohort. From this perspective, generalizability of an effect estimate involves just one additional validity issue, namely confounding of the predicted effect by effect modifiers.

Suppose that the rate increase (in cases per 100,000 person-years) produced by coffee use is 400 for males and 40 for females among five-cup-a-day drinkers in both our study cohort and the new target. If our study cohort is 70% male, while the new target is only 30% male, the average increase among five-cup-a-day drinkers in our study cohort would be $0.7 \times 400 + 0.3 \times 40 = 292$, whereas the average increase in the new target would be only $0.3 \times 400 + 0.7 \times 40 = 148$. Thus, any valid estimate of the average increase in our study cohort will tend to overestimate greatly the average increase in the new target. In other words, modification of the effect of coffee by gender confounds the prediction of its effect in the new target. This bias can be avoided by making only gender-specific predictions of effect or by standardizing the study results to the gender distribution of the new target population.

Follow-up validity

In our example, follow-up validity means that follow-up is valid within every subcohort being compared. In other words, over any span of time during follow-up, risk of myocardial infarction within a subcohort is unassociated with censoring risk in the subcohort. Given follow-up validity, we can expect that, at any follow-up time *t*, the myocardial infarction rates in a subcohort will be the same for subjects lost or withdrawn at *t* and subjects whose follow-up continues beyond *t*.

In fact, we should expect follow-up to be biased by cigarette smoking: smoking is associated with mortality from myocardial infarction and from many other causes; the association of smoking with socioeconomic status might also produce an association between smoking and loss to follow-up. The result would be elevated censoring among high-risk (smoking) subjects. As a consequence, unadjusted estimates of the risks of myocardial infarction will underestimate those risks in the complete subcohorts (as the latter includes both censored and uncensored subject experience). If the degree of underestimation varies across subcohorts, bias in the relative-risk estimates will result.

In fact, the degree of underestimation should vary in this example because of the variation in the prevalence of smoking across subcohorts. Nevertheless, this variation is not necessary for smoking-related censoring to produce biased estimates of absolute effect. For example, if smoking-related censoring produced a uniform 15% underestimation of the myocardial infarction rate in each subcohort, all rate differences would also be underestimated by 15%.

Analogous to control of confounding, any bias produced by the association of smoking with myocardial infarction and censoring can be removed by smoking adjustment. As before, if adjustment is by standardization, the standard distribution should be chosen from the target subcohort.

Because the same correction methods can sometimes be applied, some authors classify follow-up bias as a form of confounding. Nevertheless, the two phenomena are reversed with respect to the causal ordering of the third variable responsible for the bias: confounding arises from an association of the study exposure (coffee use) with other exposures (such as smoking) that affect outcome risk; in contrast, follow-up bias arises from an association between the risk of the study outcome (myocardial infarction) and risks of other end points (such as other-cause mortality or loss to follow-up) that are

affected by exposure, and thus is classified as a form of selection bias by many authors (Kelsey et al. 1996; Hernán et al. 2004). In this regard, note that certain forms of follow-up bias cannot be removed by adjustment, and thus do not resemble confounding (Hernán et al. 2004; Glymour and Greenland 2008). These problems are discussed in the statistics literature under the topic of dependent competing risks (Slud and Byar 1988) and resemble the selection-bias problems of case–control studies, which are discussed later.

Measurement validity

Unlike gender, the continuous variables of coffee use, cigarette use, blood pressure, cholesterol, and age are time-dependent covariates. With the exception of age (whose value at any time can be computed from birth date), this fact adds considerable complexity to measuring these variables and estimating their effects.

Consider that we cannot reasonably expect a single baseline measurement, no matter how accurate, to summarize adequately a subject's entire history of coffee drinking, smoking, blood pressure, or cholesterol. Even if the effect of a subject's history could be largely captured by using a single summary number (e.g. total number of cigarettes smoked), the baseline measurement may well be a poor proxy for this ideal and unknown summary. For these reasons, we should expect proxy-variable errors to be very large in our example.

Proxy-variable error in the study variables

The degree of proxy-variable error in measuring the study variables depends on the exact definitions of the variables that we wish to study. In turn, this definition should reflect the hypothesized effect that we wish to study. To illustrate, consider the following acute-effect hypothesis:

> Drinking a cup of coffee produces an immediate rise in short-term myocardial infarction risk. In other words, coffee consumption is an acute risk factor.

This hypothesis does not exclude the possibility that coffee use also elevates long-term risk of myocardial infarction, perhaps through some other mechanism; it simply does not address the issue of chronic effects.

One way to examine the hypothesis would be to compare the myocardial infarction rates among person-days in which one, two, three, or more cups were drunk with the rate among person-days in which no coffee was drunk (adjusting for confounding and follow-up bias). If we had only baseline data, baseline daily consumption would have to serve as the proxy for consumption on every day of follow-up. This would probably be a poor proxy for daily consumption at later follow-up times where more outcome events occur. A 'standard' analysis, which only examines the association of baseline coffee use with myocardial infarction rates, is equivalent to an analysis that uses baseline consumption as a proxy for consumption on all later days. Thus, estimates from a standard analysis would suffer large bias if considered as estimates of the acute effects of coffee.

The proxy-variable error in this example could easily be differential with respect to the outcome: person-days accumulate more rapidly in early follow-up, where the error from using baseline consumption as the proxy is relatively low; in contrast, myocardial infarction events accumulate more rapidly in later follow-up, where the error is probably higher. This difference in accumulation illustrates an important general point: errors in variables can be differential, even if the variables are measured before the outcome event. Such phenomena occur when errors are associated with risk factors for the outcome; in our example, the error is associated with follow-up time and hence age. In turn, such associations are likely to occur when measurements are based on proxy variables.

Suppose now that we examine the following chronic-effect hypothesis:

> Each cup of coffee drunk eventually results in a long-term elevation of myocardial infarction risk.

This hypothesis was suggested by reports that coffee drinking produces a rise in serum lipid levels (Curb et al. 1986); it does not address the issue of acute effects. One way to examine the hypothesis would be to compare the myocardial infarction rates among person-months with different cumulative doses of coffee (perhaps using a lag period in calculating dose; e.g. one might ignore the most recent month of consumption). If we had only baseline data, however, baseline daily consumption would have to be used to construct a proxy for cumulative consumption at every month of follow-up. This construction could be done in several different ways. For example, we could estimate the subjects' cumulative doses up to a particular date by

multiplying their baseline daily consumption by the number of days that they had lived between age 18 and the date in question. This estimate assumes that coffee drinking began at age 18 and the baseline daily consumption is the average daily consumption since that age. We should expect considerable error in such a crude measure of cumulative consumption.

The degree of bias in estimating chronic effects could be quite different from the degree of bias in estimating acute effects. Furthermore, as discussed in the following, the errors in each proxy will make it virtually impossible to discriminate between acute and chronic effects.

Misclassification and bias towards the null

Although measurement errors can bias effect estimates, in some cases one may be able to predict the direction of the bias. Measurement of a binary (dichotomous) variable is called better than random if, regardless of the true value, the probability that the measurement yields the true value is higher than the probability that it does not. In other words, the measurement is better than random if it is more likely to be correct than incorrect, no matter what the true value is. Given two binary variables, better-than-random measurements with independent non-differential errors cannot inflate or reverse the association observed between the variables. In other words, any bias produced by independent non-differential error in better-than-random measurements can only be towards the null value of the association (which is one for a relative-risk measure) and not beyond.

If either variable has more than two levels, then (contrary to assertions in most pre-1990 literature) the preceding conditions are not sufficient to guarantee that the resulting bias will only be towards the null and not beyond (Dosemeci et al. 1990). Despite this insufficiency, knowing that errors are independent and non-differential can increase the plausibility that any resulting bias is towards the null, although it should not increase the plausibility that the observed association is in the correct direction (Weinberg et al. 1994). For further discussions of measurement error and bias towards the null, see Flegal et al. (1991), Wacholder et al. (1991), Weinberg et al. (1994), Rothman et al. (2008, chapter 9), VanderWeele and Hernán (2012), and Gustafson and Greenland (2017).

There is one important situation in which the assumption of independent non-differential measurement error and hence bias towards the null have particularly high plausibility: in a double-blind clinical trial with a dichotomous treatment and outcome, successful blinding of treatment status during outcome evaluation should lead to independence and non-differentiality of treatment and outcome measurement errors. Successful blinding thus helps to ensure (although it does not guarantee) that any bias produced by measurement error contributes to underestimation of treatment differences (conservative bias).

Measurement error and confounding

If a variable is measured with error, estimates adjusted for the variable as measured will still be somewhat confounded by the variable. This residual confounding arises because measurement error prevents construction of strata that are internally homogeneous with respect to the true confounding variable (Greenland 1980).

To illustrate, consider baseline daily cigarette consumption. This variable can be considered a proxy for consumption on each day of follow-up or can be used to construct an estimate of cumulative consumption (analogous to the cumulative coffee variable discussed in the preceding). Suppose that we stratify the data on a cumulative smoking index constructed from the baseline smoking measurement. Within any stratum of the index, there would remain a broad range of cumulative cigarette consumption. For example, two subjects who were age 40 and smoked one pack a day at baseline would receive the same value for the smoking index and so end up in the same stratum. However, if one of them stopped smoking immediately after baseline, while the other continued to smoke a pack a day, after 10 years of follow-up the former subject would have ten fewer pack-years of cigarette consumption than the continuing smoker.

Suppose now that cumulative cigarette consumption is positively associated with cumulative coffee consumption. Then, even within strata of the smoking index, we should expect subjects with high coffee consumption to exhibit elevated myocardial infarction rates simply by virtue of having higher levels of cigarette consumption. As a consequence, the estimate of the effect of coffee adjusted for the smoking index would still be confounded by cumulative cigarette consumption.

In some cases, a study variable may appear to have an effect (or no effect) only because of poor measurement of an apparently unimportant confounder. This can occur, for example, when an important confounding variable is measured with a large amount of non-differential error. Such an error would ordinarily reduce the apparent association of the variable with the exposure, and would also make the variable appear to be a weak risk factor, perhaps weaker than the study exposure. This in turn would make the variable appear to be only weakly confounding, in that adjustment for the variable as measured would produce little change in the result. However, this appearance would be deceptive because adjustment for the variable as measured would eliminate little of the actual confounding by the variable.

As an example, suppose that coronary proneness of personality was measured only by the baseline yes or no question: Do you consider yourself a hard-driving person? Such a crude measure of the original construct would be unlikely to show more than a weak association with either coffee use or myocardial infarction, and adjusting for it would produce little change in our estimate of the effect of coffee. Suppose, however, that coronary-prone personalities have an elevated preference for coffee. Such a phenomenon would lead to a concentration of coronary-prone persons (and hence, a spuriously elevated myocardial infarction rate) among coffee drinkers, even after stratification on the response to the question.

One would ordinarily expect adjustment for a non-differentially misclassified confounder to produce an estimate lying somewhere between the unadjusted estimate and the estimate adjusted for the true values of the confounder (Greenland 1980). Unfortunately, if the true confounder has more than two levels, it is possible for adjustment by the misclassified confounder to be more biased than the unadjusted estimate (Brenner 1993); this can even happen with confounders with two levels if the confounder affects the exposed and unexposed subjects in opposite directions (Ogburn and VanderWeele 2012). It is also possible for adjustment by factors that affect misclassification to worsen bias (Greenland and Robins 1985). If however the effect of the confounder is in the same direction regardless of exposure, and its misclassification is independent and non-differential with respect to the exposure and outcome, adjustment for the misclassified confounder will on average produce an estimate between the unadjusted

estimate and the estimate adjusted for the true confounder values (Ogburn and VanderWeele 2012, 2013).

Measurement error and separation of effects

Because of their impact on the effectiveness of adjustment procedures, measurement errors can severely reduce our ability to separate different effects of the study variable. Suppose in our example that we wished to estimate the relative strength of acute and chronic effects of coffee. To do so, we must take account of the fact that acute and chronic effects will be confounded. When examining acute effects, person-days with high coffee consumption will occur most frequently among persons with high cumulative coffee consumption. As a consequence, if cumulative coffee consumption is a risk factor, it will be a confounder for estimating the acute effects of coffee consumption. By similar arguments, if coffee consumption has acute effects, these will confound estimates of the chronic effects of cumulative consumption.

Unfortunately, both cumulative and daily consumption are measured with considerable error. As a result, any effect observed for one may be wholly or partially due to the other, even if the other has little or no apparent effect.

Repeated measures

One costly but effective method for reducing the degree of proxy-variable error in measuring time-dependent variables is to take repeated (serial) measurements over the follow-up period and ask subjects to report their pre-baseline history of such variables at the baseline interview. In our example, subjects could be asked about their age at first use and level of consumption at different ages for coffee and cigarettes; they could then be recontacted every year or two to assess their current consumption. Of course, not all subjects may be willing to cooperate with such active follow-up, but the penalties of some extra loss may be far outweighed by the benefit of improved measurement accuracy.

Errors in assessing incidence

An important form of measurement error in assessing incidence is misdiagnosis of the outcome event. In the AIDS example, a false-positive diagnosis of AIDS would result in underestimation of incubation time, whereas a false-negative diagnosis would result in overestimation. In the present example, false-positive errors would result in overestimation of myocardial infarction rates, whereas false-negative errors would result in underestimation. These errors will be of particular concern when the study depends on existing surveillance systems or records for detection of outcome events.

There are special cases in which the errors will induce little or no bias in estimates, provided the errors have little effect on the person-time observed. If the only form of misdiagnosis is false-negative error, the proportion of outcome events missed in this fashion is the same across cohorts, and if there is no follow-up bias, then the relative-risk estimates will not be distorted by the underdiagnosis. Suppose in our example that all recorded myocardial infarction events are true myocardial infarctions, but that in each subcohort 10% of myocardial infarctions are missed. The myocardial infarction rates in each subcohort will then be underestimated by 10%; nevertheless, if we consider any two of these rates, say R_0 and R_5, the observed rate ratio will be:

$$\frac{0.9R_5}{0.9R_0} = \frac{R_5}{R_0}$$

which is undistorted by the underdiagnosis of myocardial infarction. Nonetheless, if coffee primarily induced ‘silent’ myocardial infarctions and these were the most frequently undiagnosed events, the effect of coffee would be underestimated.

Analogously, if the only form of misdiagnosis is false-positive error, the rate of false positives is the same across cohorts, and if there is no follow-up bias, then rate differences will not be distorted by the overdiagnosis. Suppose that the rate of false positives in our example is R_f in all subcohorts; then, if we consider any two true rates, say R_0 and R_5, the observed rate difference will be:

$$(R_c + R_f) - (R_u + R_f) = R_c + R_u$$

which is undistorted by the overdiagnosis of myocardial infarction. However, if there is non-differential underdiagnosis of myocardial infarction, as is probably the case in our example, the rate difference will be underestimated.

Specification validity

As noted earlier, the use of a statistical method based on an incorrect model (specification error) can lead to bias in estimates and improper performance of statistical tests and interval estimates. All statistical techniques, including non-parametric methods, must assume some sort of model for the process generating the data; however, in the absence of randomization or random sampling, it will rarely be possible to identify a 'correct' sampling model. In addition, structural assumptions are rarely (if ever) exactly satisfied. Thus, some specification error should be expected. As before, minimization of specification error must rely on checking the model against the data and against background information about the processes generating the data.

Recall that the unadjusted rate ratio estimates for five-cup-a-day versus never-drinkers is 1.22 in the present example, with 95% confidence limits of 1.00 and 1.49, and a *p*-value of 0.05. Suppose that these figures were obtained by conventional person-time methods (Rothman et al. 2008, chapter 14). These methods are based on a binomial sampling model for the number of cases who drank five cups a day at baseline, given the combined total number of cases among five-cup-a-day and never-drinkers. In our example, the validity of this model depends on the assumption that the myocardial infarction rate remains constant within subcohorts over the follow-up period. It follows that the model (and hence the statistics given earlier) cannot be valid in our example; the subcohort members grow older over the follow-up period, and hence the myocardial infarction rates must increase with follow-up time.

The invalidity just noted can be rectified by stratifying either on follow-up time or the variable responsible for the change in rates over follow-up time (here, age). The stratification need only be fine enough to ensure that the myocardial infarction rate change within strata is negligible over follow-up. As noted earlier, however, we must also adjust for smoking and

perhaps other factors responsible for confounding or follow-up bias. If we stratify finely enough to remove all the bias from these sources, the resulting estimates would be undefined or so unstable that they would tell us nothing about the association of coffee and myocardial infarction.

The standard solution to such problems is to compute adjusted estimates using regression models. These are structural models representing a set of assumptions (usually rather strong ones) about the joint effects of the study variables. Such models allow estimates and tests to be extracted from what would otherwise be hopelessly sparse data, at a cost of a greater risk of bias arising from violations of the assumptions underlying the models (Robins and Greenland 1986). For further details of cohort modelling, see Breslow and Day (1987), Hosmer and Lemeshow (2000), Kelsey et al. (1996), Checkoway et al. (2004), Jewell (2004), and Greenland (2008a, 2008b).

Summary of cohort example

The example used in this subsection provides an illustration of the most common threats to the validity of effect estimates from cohort studies. The unadjusted estimates of the effect of coffee on myocardial infarction will be confounded by many variables (such as smoking), and there will be follow-up bias. As a result, the number of variables that must be controlled is too large to allow adequate control using only stratification. The true functional dependence of myocardial infarction rates on coffee and the confounder is unknown, so that estimates based on multivariate models are likely to be biased. Even if this bias is unimportant, our estimates will remain confounded because of our inability to measure the key confounders accurately. Finally, our inability to summarize coffee consumption accurately would further bias our estimates, making it impossible to separate acute and chronic effects of coffee use reliably.

Given that there are several sources of bias of unknown magnitude and different directions, it would appear that no conclusions about the effect of coffee could be drawn from a study such as the one described in the preceding, other than that coffee does not appear to have a large effect. This type of result—inconclusive, other than to rule out very large effects—is common in thorough epidemiological analyses of observational data. In particular, inconclusive results are common when the data being analysed

were collected for purposes other than to address the hypothesis at issue, for such data often lack accurate measurements of key variables.

Special issues in case–control studies

The practical difficulties of cohort studies have led to extensive development of case–control study designs. The distinguishing feature of such designs is that sampling is intentionally based on the outcome of individuals. In a population-based or population-initiated case–control study, one first identifies a population at risk of the outcome of interest, which is to be studied over a specified period of time or risk period. As in a cohort study, one attempts to ascertain outcome events in the population at risk. Nevertheless, unlike a cohort study, one selects persons experiencing the outcome event (cases) and a 'control' sample of the entire population at risk for ascertainment of exposure and covariate status.

In a case-initiated case–control study, one starts by identifying a source of study cases (e.g. a hospital emergency room is a source of myocardial infarction cases). One then attempts to identify a population at risk such that the source of cases provides a random or complete sample of all cases occurring in this population. Study cases recruited from the source occur over a risk period; controls are selected in order to ascertain the distribution of exposure in the population at risk over that period. Case–control studies may also begin with an existing series of controls (Greenland, 1985). Regardless of how a case–control study is initiated, evaluation of validity must ultimately refer to a population at risk that represents the target of inference for the study.

Relative-risk estimation in case–control studies

The control sample may or may not be selected in a manner that excludes cases. If persons who become cases over the risk period are ineligible for inclusion in the control group (as in traditional case–control designs), a 'rare disease' assumption may be needed to estimate RRs from the case–control data. In contrast, if persons who become cases over the risk period are also eligible for inclusion in the control group (as in newer case–control designs), the rare disease assumption can be discarded. These points are discussed in the textbooks cited at the beginning of this chapter.

The basics of case–control estimation will be illustrated using the following example. We wish to study the effect of coffee drinking on rates of first myocardial infarction and we have selected a population for study (e.g. all residents aged 40–64 in a particular town) over a 1-year risk period. At any point during the risk period, the population at risk comprises persons in this selected population who have not yet had a myocardial infarction.

Suppose that the average number of never-drinkers in the population at risk was 20,000 over the risk period, the average number of five-cup-a-day drinkers was 10,000, there were 120 first myocardial infarctions among never-drinkers, and there were 90 first myocardial infarctions among five-cup-a-day drinkers. Then, if one observed the entire population without error, the estimated rates among never-drinkers and five-cup-a-day drinkers would be:

$$\frac{120}{20{,}000\text{Person-years}} = \frac{90}{10{,}000\text{Person-years}}.$$

Thus, if we observed the entire population, the estimated rate ratio would be:

$$\frac{90/10{,}000\text{Person-years}}{20{,}000\text{Person-years}} = \frac{90/120}{10{,}000/20{,}000\text{Person-years}} = 1.50.$$

This estimate depends on only two figures: the relative prevalence of five-cup-a-day versus never-drinkers among cases (90/120), and the same relative prevalence in the person-years at risk (10,000/20,000). These two relative prevalences are often called the case exposure odds and the population exposure odds.

The first relative prevalence (numerator) could be estimated by interviewing an unbiased sample of all the new myocardial infarction cases that occur over the risk period, and the second relative prevalence (denominator) could be estimated by interviewing an unbiased sample of the population at risk over the risk period. The ratio of relative prevalences from the case– and control–sample interviews would then be an unbiased estimate of the population rate ratio of 1.50. This estimate is called the sample odds ratio.

Three points about the preceding argument should be carefully noted. First, no rare disease assumption was made. Second, the control sample of the population at risk was accumulated over the entire risk period (rather than at the end of the risk period); such sampling is called density sampling (Rothman et al. 2008, chapter 8) or risk-set sampling (Breslow and Day 1987). Third, because of the density sampling, someone may be selected for the control sample, and yet have a myocardial infarction later in the risk period and become part of the case sample as well. Methods for carrying out density sampling can be found in the textbooks cited at the beginning of this chapter.

Validity conditions in case–control studies

The primary advantages of case–control studies are their short timeframe and the large reduction in the number of subjects needed to achieve the same statistical power as a cohort study. The primary disadvantage is that more conditions must be met to ensure their validity (in addition to the four listed in the cohort study example), and hence, there are more opportunities for bias to arise.

Suppose that our case–control study data yield an unadjusted rate ratio estimate (odds ratio) of 1.50, with 95% confidence limits of 1.00 and 2.25. The following series of conditions would be sufficient for the validity of this figure as an estimate of the effect of drinking five cups of coffee a day (versus none) on the myocardial infarction rate:

Comparison validity—if five-cup-a-day drinkers in the population at risk had instead drunk no coffee, their distribution of myocardial infarction events over time would have been approximately the same as the distribution among never-drinkers.
Follow-up validity—within each subpopulation defined by coffee use, censoring risk (i.e. population membership ended by an event other than myocardial infarction, such as emigration or death from another cause) is not associated with myocardial infarction risk.
Selection validity—this has two components:

1. *Case-selection validity*—if one studies only a subset of the myocardial infarction cases occurring in the population over the risk period (e.g. because of failure to detect all cases), this subset provides unbiased estimates of the prevalence of different levels of coffee use among all cases occurring in the population over the risk period.
2. *Control-selection validity*—the control sample provides unbiased estimates of the prevalences of different levels of coffee use in the population at risk over the risk period.

Specification validity—the distribution of myocardial infarction events over time in the subpopulations can be closely approximated by the statistical model on which the estimates are based.
Measurement validity—all measurements of variables used in the analysis closely approximate the true values of the variables.

Issues of comparison validity, follow-up validity, specification validity, effect modification, and generalizability in case–control studies parallel those in follow-up studies, and so will not be discussed here. Case–control studies are vulnerable to certain problems of measurement error that are less severe or do not exist in prospective cohort studies. We shall discuss these problems first, and then examine selection validity and modelling. Finally, we shall briefly discuss analogous issues in retrospective cohort studies.

Retrospective ascertainment

A special class of measurement errors arises from the retrospective ascertainment of time-dependent variables; that is, attempting to measure past values of the variables. Retrospective ascertainment must be based on

individual memories, existing records of past values, or some combination of the two. Therefore, such ascertainment usually suffers from faulty recall, missing or mistaken records, or lack of direct measurements in existing records.

Retrospective ascertainment may be an important component of a cohort study. For example, the cohort study of coffee and myocardial infarction discussed earlier could have been improved by asking subjects about their coffee use and smoking prior to the start of follow-up. This information would allow one to construct better cumulative indices than could be constructed from baseline consumption alone, although the resulting indices would still incorporate error due to faulty recall.

Unless records of past measurements are available for all subjects, measurements on cases and controls must be made after the time period under study as subjects are not selected for study until after that period. Thus, unlike cohort studies, most case–control studies of time-dependent variables depend on retrospective ascertainment. Considering our example, there may be much more error in determining daily coffee consumption ten years before interview than one month before interview; one might then expect case–control studies to be more accurate for studying acute effects than for studying chronic effects. Nonetheless, if acute and chronic effects are heavily confounded, the elevated inaccuracies of long-term recall will make it impossible to disentangle short-term from long-term effects. As illustrated earlier, this confounding can arise in a cohort study. Nevertheless, in a cohort study, such confounding can be minimized by taking repeated measurements. In contrast, such confounding would be unavoidable in a case–control study based on recall, even if detailed longitudinal histories were requested from the subjects.

The preceding observations should be tempered by noting that some case–control studies have access to exposure measurements of the same quality as found in cohort studies and that the exposure measurements in some cohort studies may be no better than those used in some case–control studies. For example, a cohort study in which measurements are derived by abstracting routine medical records would suffer from no less measurement error than a case–control study in which measurements are derived by abstracting the same records.

Outcome-affected measurements

One common potential problem in case–control studies is outcome-affected recall, often termed *recall bias*. This term refers to the differential measurement error that originates when the outcome event affects recall of past events. Examples arise in case– control studies of birth defects, for instance. If the trauma of having an affected child either enhances recall of prenatal exposures among case mothers or increases the frequency of false-positive reports among them, estimates of RR will be upwardly biased by effects of the outcome on case recall. This bias may be counterbalanced by other biases, such as recall bias among controls, making the final direction and magnitude of bias due to faulty measurements hard to predict (Drews and Greenland 1990).

One method commonly proposed for preventing bias due to outcome-affected recall is to restrict controls to a group believed to have recall similar to the cases. Unfortunately, one usually cannot tell to what degree this restricted selection corrects the bias from outcome-affected recall. Even more unfortunately, one usually cannot tell if the selection bias produced by such restriction is worse than the recall bias one is attempting to correct (Swan et al. 1992; Drews et al. 1993).

A problem similar to outcome-affected recall can occur when the outcome event affects a psychological or physiological measurement. This is of particular concern in case–control studies of nutrient levels and chronic disease. For example, if colon cancer leads to a drop in serum retinol levels, the RR for the effect of serum retinol will be underestimated if serum retinol is measured after the cancer develops. Errors of this type can be viewed as proxy-variable errors in which the post-outcome value is a poor proxy for the pre-outcome value of interest.

Selection validity

Selection validity is straightforward to understand but can be extraordinarily difficult to verify. A violation of the selection validity conditions is known as selection bias. Many case–control designs and field methods are devoted to avoiding such bias (Kelsey et al. 1996; Koepsell and Weiss 2003; Rothman et al. 2008, chapter 8).

In some instances, it may be possible to identify a factor or factors that affect the chance of selection into the study. If in such instances we have accurate measurements of one of these factors, we can stratify on (or

otherwise adjust for) the factor and thereby remove the selection bias due to the factor. Because of this possibility, some authors classify selection bias as a form of confounding. Nevertheless, there are some forms of selection bias that cannot be removed by adjustment. These points will be illustrated in the following subsections.

Case-selection validity

Case-selection bias can be minimized (although may still be large due to non-participation) if one can identify every case that occurs in the population at risk over the risk period. This requires a surveillance system for the outcome of interest, such as a population-based disease registry. In our coffee–myocardial infarction example, we would probably have to construct a myocardial infarction surveillance system from existing resources, such as emergency room admission records, ambulance service records, and paramedic records.

Even if all cases of interest can be identified, selection bias may arise from failure to obtain information on all the cases. In our example, many cases would be dead before interview was possible. For such cases, there are only two alternatives: attempt to obtain information from some other source, such as next of kin or co-workers, or exclude such cases from the study. The first alternative increases measurement error in the study. The second alternative will introduce bias if coffee affects risk of fatal and non-fatal myocardial infarction differently, or if coffee affects risk of myocardial infarction survivorship. To illustrate, suppose that coffee drinking reduced one's chance of reaching the hospital alive when a myocardial infarction occurred. Then, the prevalence of coffee use among myocardial infarction survivors would under-represent the prevalence among all myocardial infarction cases. Underestimation of the rate ratio would result if fatal myocardial infarction cases were excluded from the study.

It might seem possible to remove the case-selection bias in this example by redefining the study outcome as non-fatal myocardial infarction. This does not remove the bias, however; it only leads to its reclassification as a bias due to differential censoring (here classified as a form of follow-up bias). In a study of non-fatal myocardial infarction, fatal myocardial infarction is a censoring event associated with risk of non-fatal myocardial infarction; if fatal myocardial infarction is also associated with coffee use,

the result will be underestimation of the rate ratio for non-fatal myocardial infarction. More generally, it is usually not possible to remove bias by placing restrictions on admissible outcomes.

Unfortunately, exclusion is the only alternative for cases that refuse to participate or cannot be located. In our example, if such cases tend to be heavier coffee users than others, underestimation of the rate ratio would result. However, suppose that, within levels of cigarette use, such cases were no different from other cases with respect to coffee use. Then, adjustment for smoking would remove the selection bias induced by refusals and failures to locate cases. (Of course, such adjustment would require accurate smoking measurement, which is a problem in itself.)

Bias that arises from failure to detect certain cases is sometimes called detection bias. If our surveillance system used only hospital admissions, many out-of-hospital myocardial infarction deaths would be excluded, and a detection bias of the sort described here could result.

Control-selection validity

Control-selection bias can be minimized (although may still be large due to non-participation) if one can potentially identify every member of the population at risk at every time during the risk period. In such a situation, one could select controls with one of many available probability sampling techniques, using the entire population at risk as the sampling frame. Unfortunately, such situations are exceptional.

Many studies attempt to approximate the ideal sampling situation through use of existing population lists. An example is control selection by random digit dialling; here, the list (of residential telephone numbers) is not used directly but nevertheless serves as a partial enumeration of the population at risk. This list excludes people without telephone numbers. In our example, if people without telephones drink less coffee than people with telephones, a control group selected by random digit dialling would over-represent coffee use in the population at risk. The result would be underestimation of the rate ratio.

One could redefine the population at risk in the previous example so that the telephone-related selection bias did not exist by restricting the study to persons with telephones. This would require excluding persons without telephones from the case series. The resulting relative-risk estimate would

suffer no selection bias. The only important penalty from this restriction is that the resulting estimate might apply only to the population of persons with telephones, which is a problem of generalizability rather than a problem of selection validity. In a similar fashion, it is often possible to prevent confounding or selection bias by placing restrictions on the population at risk (and hence, the control group). In such instances, however, one must take care to apply the same restrictions to the case series and avoid using restrictions based on events that occur after exposure (Poole 1999; Rothman et al. 2008, chapter 8).

Even if all members of the population at risk can be identified, selection bias may arise from failure to obtain information on all people selected as controls. The implications are the reverse of those for case-selection bias. In our example, if controls who refuse to participate or cannot be located tend to be heavier coffee users than other controls, overestimation of the rate ratio would result. This should be contrasted with the underestimation that results from the same tendency among cases.

More generally, one might expect an association of selection probabilities with the study variable to be in the same direction for both cases and controls. If so, the resulting case- and control-selection biases would be in opposite directions and so, to some extent, they would cancel one another out, although not completely. To illustrate, suppose that among cases the proportions who refuse to participate are 0.05 for five-cup-a-day drinkers and 0.02 for never-drinkers, and among controls the analogous proportions are 0.20 and 0.10. These refusals will result in the odds of five-cup-a-day versus never-drinkers among cases being underestimated by a factor of 0.95/0.98 = 0.97; this in turn results in a 3% underestimation of the rate ratio. Among controls, the odds will be underestimated by a factor of 0.80/0.90 = 0.89; this results in a 1/0.89 = 1.12, or a 12% overestimation of the rate ratio. The net selection bias in the rate ratio estimate will then be 0.97/0.89 = 1.09, or 9% overestimation.

For further discussions of control-selection validity, see the textbooks cited in the beginning of this chapter, and also Savitz and Pearce (1988), Schlesselman (1992), Swan et al. (1992), and Wacholder et al. (1992).

Matching

In cohort studies, matching refers to selection of exposure subcohorts in a manner that forces the matched factors to have similar distributions across the subcohorts. If the matched factors are accurately measured and the proportion lost to follow-up does not depend on the matched factors, cohort matching can prevent confounding by the matched factors, although there are statistical reasons to control the matched factors in the analysis (Weinberg 1985).

In case–control studies, however, matching refers to selection of subjects in a manner that forces the distribution of certain factors to be similar in cases and controls. Because the population at risk is not changed by case–control matching, such matching does not by itself prevent confounding by the matched factors. In fact, it is now widely recognized that case–control matching is a form of selection bias (Glymour and Greenland 2008; Rothman et al. 2008, chapter 11). This bias can be removed by adjusting for the matching factor; to the extent the factor has been closely matched and accurately measured, this adjustment also controls for confounding by the factor.

As an example, suppose that our population at risk is half male, that the men tend to drink less coffee than the women, and that about 75% of our cases are men. Unbiased control selection should yield about 50% men in the control group. However, if we matched controls to cases on gender, about 75% of our controls would be men. Because men drink less coffee than women and men would be over-represented in the matched control group, the matched control group would under-represent coffee use in the population at risk. Note, however, that matching does not affect the gender-specific prevalence of coffee use among controls, and so the gender-specific and gender-adjusted estimates would be unaffected by matching. In other words, the selection bias produced by matching could be removed by adjustment for the matching factor.

The conclusion to be drawn is that matching can necessitate control of the matching factors. Thus, in order to avoid increasing the number of factors requiring control unnecessarily, one should limit matching to factors for which control would probably be necessary anyway. In particular, matching is usually best limited to known strong confounders, such as age and gender in the preceding example (Rothman et al. 2008, chapter 11).

More generally, the primary theoretical value of matching is that it can sometimes reduce the variance of adjusted estimators. However, there are

circumstances in which matching can facilitate control selection and so is justified on practical grounds. For example, neighbourhood controls may be far easier to obtain than unmatched general population controls. In addition, although neighbourhood matching would necessitate use of a matched analysis method, the neighbourhood-matched results would incorporate some control of confounding by factors associated with the neighbourhood (such as socioeconomic status and air pollution).

Special control groups

It is not unusual for investigators to select a special control group that is clearly not representative of the population at risk if they can argue that: (1) the group will adequately reflect the distribution of the study factor in the population at risk; or (2) that the selection bias in the control group is of the same magnitude of (and so will cancel with) the selection bias in the case group. The first rationale is common in case–control studies of mortality in which persons dying of other selected causes of death are used as controls; in such studies, selection validity can be assured only if the causes of death of controls are unrelated to the study factor. The second rationale is common in studies using hospital cases and controls; in particular, selection validity can be assured in such studies if the control conditions are unrelated to the study factor, and the study disease and the control conditions have proportional exposure-specific rates of hospital admission (Schlesselman 1992).

Selection into a special control group usually requires membership in a small and highly select subset of the population at risk. Thus, use of a special control group requires careful scrutiny for mechanisms by which the study factor may influence entry into the subset. See Schlesselman (1992) and Kelsey et al. (1996) for discussions of practical issues in evaluating special control groups, and Rothman et al. (Rothman et al. 2008, chapter 8) for validity principles in mortality case–control studies (so-called proportionate mortality studies).

Case–control modelling

The most popular model for case–control analysis is the logistic model. Details of logistic modelling for case–control analysis are covered in many

textbooks, including Breslow and Day (1980), Schlesselman (1992), Kelsey et al. (1996), Hosmer and Lemeshow (2000), and Jewell (2004), and Greenland (2008b).

One important aspect of case–control modelling is that matched factors require special treatment. For example, suppose that matching is done on age in 5-year categories and age is associated with the study exposure. To control for the selection bias produced by matching, one must either employ conditional logistic regression with age as a stratifying factor, or else enter indicator variables for each age-matching category into an ordinary logistic regression (the latter strategy has the drawback of requiring about ten or more subjects per age stratum to produce valid estimates). Simply entering age into the model as a continuous variable may not adequately control for the matching-induced bias.

Summary of case–control example

The example in this subsection provides an illustration of the most common threats to validity in case–control studies (beyond those already discussed for cohort studies). After adjustments for possible confounding and follow-up bias (along the lines described for the cohort study), there may still be irremediable selection bias, especially if we use only select case groups (e.g. myocardial infarction survivors) or control groups (e.g. hospital controls). In addition, retrospective ascertainment will lead to greater measurement error than prospective ascertainment, and some of this additional error may be differential.

Given the even greater number of potential biases of unknown magnitude and different directions, it would appear that (as in the cohort example) no conclusions about the effect of coffee could be drawn from a study such as the one described, other than that coffee does not have a large effect. Again, this is a common result in thorough epidemiological analyses of observational data.

Special issues in retrospective cohort studies

Two major types of cohort studies can be distinguished depending on whether members of the study cohort are identified before or after the

follow-up period under study. Studies in which all members are identified before their follow-up period are called *concurrent* or *prospective* cohort studies, and studies in which all members are identified after their follow-up period are called *historical* or *retrospective* cohort studies.

Similar to case–control studies, retrospective cohort studies often require special consideration of retrospective ascertainment and selection validity. In particular, retrospective cohort studies that obtain exposure or covariate histories from post-event reconstructions are vulnerable to bias from outcome-affected measurements. Suppose, for example, that a study of cancer incidence at an industrial facility had to rely on company personnel to determine the location and nature of various exposures in the plant during the relevant exposure periods. If these personnel were aware of the locations at which cases worked (as when a publicized ‘cluster’ of cases has occurred), biased exposure assessment could result. Such problems can also occur in a prospective cohort study if exposure or covariate histories are based on post-event reconstructions.

Retrospective cohort studies can also suffer from selection biases analogous to those found in case–control studies. Suppose, for example, that a retrospective cohort study relied on company records to identify members of the cohort of plant employees. If retention of an employee’s records (and hence identification of the employee as a cohort member) were associated with both the exposure and outcome status of the employee, the exposure–outcome association observed in the incomplete study cohort could poorly represent the exposure–outcome association in the complete cohort of plant employees.

Conclusion

Uncertainty about validity conditions is responsible for most of the inconclusiveness inherent in epidemiological studies. This inconclusiveness can be partially overcome when multiple complementary studies are conducted; that is, when new studies are conducted under conditions that effectively limit bias from one or more of the sources present in earlier studies. Ideally, after enough complementary studies have been conducted, each known or suspected source of bias might have been rendered unimportant in at least one study. If at this point all the study results appear

consistent with one another (which is not the case for coffee and myocardial infarction, although the studies of smoking and lung cancer provide a good example), the epidemiological community may reach some consensus about the existence and strength of an effect.

Even in such ideal situations, however, one should bear in mind that consistency is not validity. For example, there may be some unsuspected source of bias present in all the studies, so that they are all consistently biased in the same direction. Alternatively, all the known sources of bias may be in the same direction, so that all the studies remain biased in the same direction if no one study eliminates all known sources of bias. Such problems can be addressed tentatively via sensitivity analysis or the more elaborate methods of risk analysis and bias analysis (Eddy et al. 1992; Greenland and Lash 2008; Lash et al. 2009; VanderWeele and Arah 2011; Greenland 2014). For these and other reasons, many authors have warned that all causal inferences should be considered tentative, at least if drawn from observational epidemiological data alone, although such tentativeness does not in any way rule out interventions when it is clear that the probable cost of failing to act far outweighs the probable cost of intervention (Rothman 1988; Rothman et al. 2008, chapter 2).

ACKNOWLEDGEMENTS

The authors wish to thank I. Hertz-Picciotto, K. Hoggatt, P. Kass, J. Kelsey, G. Maldonado, S. Norrell, J. Schlesselman, and A. Walker for their helpful comments on earlier versions of this chapter.

REFERENCES

Altman, D.G. and Bland, J.M. (1995). Absence of evidence is not evidence of absence. *British Medical Journal*, 311, 485.

Altman, D.G., et al. (eds.) (2000). *Statistics with Confidence*. London: BMJ Books.

Berger, J.O. and Berry, D.A. (1988). Statistical analysis and the illusion of objectivity. *American Scientist*, 76, 159–65.

Borenstein, M., Hedges, L.V., Higgins, J.P.T., and Rothstein, H.R. (2009). *Introduction to Meta-analysis*. New York: Wiley.

Brenner, H. (1993). Bias due to nondifferential misclassification of a polytomous confounder. *Journal of Clinical Epidemiology*, 46, 57–63.

Breslow, N.E. and Day, N.E. (1980). *Statistical Methods in Cancer Research. I: The Analysis of Case Control Studies*. Lyon: IARC.

Breslow, N.E. and Day, N.E. (1987). *Statistical Methods in Cancer Research. II: The Analysis of Cohort Data*. Lyon: IARC.

Checkoway, H., Pearce, N., and Kreibel, D. (2004). *Research Methods in Occupational Epidemiology* (2nd ed.). New York: Oxford University Press.

Cole, S.R. and Hermán, M.A. (2002). Fallibility in estimating direct effects. *International Journal of Epidemiology*, 31, 163–5.

Cole, S.R. and Stuart, E.A. (2010). Generalizing evidence from randomized clinical trials to target populations: the ACTG-320 trial. *American Journal of Epidemiology*, 172, 107–15.

Curb, J.D., et al. (1986). Coffee, caffeine, and serum cholesterol in Japanese men in Hawaii. *American Journal of Epidemiology*, 123, 648–55.

Dosemeci, M., Wacholder, S., and Lubin, J.H. (1990). Does nondifferential misclassification of exposure always bias a true effect towards the null value? *American Journal of Epidemiology*, 132, 746–48.

Drews, C.D. and Greenland, S. (1990). The impact of differential recall on the results of case-control studies. *International Journal of Epidemiology*, 19, 1107–12.

Drews, C.D., Greenland, S., and Flanders, W.D. (1993). The use of restricted controls to prevent recall bias in case-control studies of reproductive outcomes. *Annals of Epidemiology*, 3, 86–92.

Eddy, D.M., Hasselblad, V., and Schacher, R. (1992). *Meta-analysis by the Confidence Profile Method*. New York: Academic Press.

Flegal, K.M., Keyl, P.M., and Nieto, E.J. (1991). Differential misclassification arising from nondifferential errors in exposure measurement. *American Journal of Epidemiology*, 134, 1233–44.

Gilovich, T., Griffin, D., and Kahneman, D. (2002). *Heuristics and Biases: The Psychology of Intuitive Judgment*. New York: Cambridge University Press.

Glymour, M.M. and Greenland, S. (2008). Causal diagrams. In: Rothman, K.J., Greenland, S., and Lash, T.L. (eds.) *Modern Epidemiology* (3rd ed.). Philadelphia, PA: Lippincott Williams & Wilkins, pp. 183–211.

Goodman, S.N. (1993). P-values, hypothesis tests, and likelihood: implications for epidemiology of a neglected historical debate. *American Journal of Epidemiology*, 137, 485–96.

Greenland, S. (1980). The effect of misclassification in the presence of covariates. *American Journal of Epidemiology*, 112, 564–9.

Greenland, S. and Robins, J.M. (1985). Confounding and misclassification. *American Journal of Epidemiology*, 122, 495–506.

Greenland, S. (1985). Control initiated case-control studies. *International Journal of Epidemiology*, 14, 130–4.

Greenland, S., and Robins, J.M. (1988). Conceptual problems in the definition and interpretation of attributable fractions. *American Journal of Epidemiology*, 128, 1185–97.

Greenland, S., Schlesselman, J.J., and Criqui, M.H. (1986). The fallacy of employing standardized regression coefficients and correlations as measures of effect. *American Journal of Epidemiology*, 123, 203–8.

Greenland, S. (1987). Interpretation and choice of effect measures in epidemiologic analyses. *American of Journal of Epidemiology*, 125, 761–8.

Greenland, S. (1990). Randomization, statistics, and causal inference. *Epidemiology*, 1, 421–9.

Greenland, S., et al. (1991). Standardized coefficients: a further critique and a review of alternatives. *Epidemiology*, 2, 387–92.

Greenland, S. (1993). A meta-analysis of coffee, myocardial infarction, and coronary death. *Epidemiology*, 4, 366–74.

Greenland, S. (1994). A critical look at some popular meta-analytic methods. *American Journal of Epidemiology*, 140, 290–6.

Greenland, S. and Maldonado, G. (1994). The interpretation of multiplicative model parameters as standardized parameters. *Statistics in Medicine*, 13, 989–99.
Greenland, S. (1998a). Induction versus popper: substance versus semantics. *International Journal of Epidemiology*, 27, 543–8.
Greenland, S. (1998b). Probability logic and probabilistic induction. *Epidemiology*, 9, 322–32.
Greenland, S. (1999). The relation of the probability of causation to the relative risk and the doubling dose: a methodologic error that has become a social problem. *American Journal of Public Health*, 89, 1166–9.
Greenland, S., Pearl, J., and Robins, J.M. (1999a). Causal diagrams for epidemiologic research. *Epidemiology*, 10, 37–48.
Greenland, S., Robins, J.M., and Pearl, J. (1999b). Confounding and collapsibility in causal inference. *Statistical Science*, 14, 29–46.
Greenland, S., Schwartzbaum, J.A., and Finkle, W.D. (2000). Problems from small samples and sparse data in conditional logistic regression analysis. *American Journal of Epidemiology*, 151, 531–9.
Greenland, S. (2000a). Principles of multilevel modelling. *International Journal of Epidemiology*, 29, 158–67.
Greenland, S. (2000b). When should epidemiologic regressions use random coefficients? *Biometrics*, 56, 915–21.
Greenland, S. (2001). Ecologic versus individual-level sources of confounding in ecologic estimates of contextual health effects. *International Journal of Epidemiology*, 30, 1343–50.
Greenland, S. and Brumback, B.A. (2002). An overview of relations among causal modelling methods. *International Journal of Epidemiology*, 31, 1030–7.
Greenland, S. (2002a). A review of multilevel theory for ecologic analyses. *Statistics in Medicine*, 21, 389–95.
Greenland, S. (2002b). Causality theory for policy uses of epidemiologic measures. In: Murray, C.J.L., et al. (eds.) *Summary Measures of Population Health*. Cambridge, MA: Harvard University Press/WHO, pp. 291–302. (Reprinted as Greenland, S. (2005). Epidemiologic measures and policy formulation: lessons from potential outcomes (with discussion). *Emerging Themes in Epidemiology*, 2, 1–4.)
Greenland, S. (2004a). Ecologic inference problems in studies based on surveillance data. In: Stroup, D.F. and Brookmeyer, R. (eds.) *Monitoring the Health of Populations: Statistical Principles and Methods for Public Health Surveillance*. New York: Oxford University Press, pp. 315–40.
Greenland, S. (2004b). The need for critical appraisal of expert witnesses in epidemiology and statistics. *Wake Forest Law Review*, 39, 291–310.
Greenland, S. (2005). Multiple-bias modeling for analysis of observational data (with discussion). *Journal of the Royal Statistical Society, Series A*, 168, 267–308.
Greenland, S. (2006). Bayesian perspectives for epidemiologic research. I: foundations and basic methods. *International Journal of Epidemiology*, 35, 765–8.
Greenland, S. (2007). Bayesian methods for epidemiologic research. II: regression analysis. *International Journal of Epidemiology*, 36, 195–202.
Greenland, S. (2008a). Introduction to regression models. In: Rothman, K.J., Greenland, S., and Lash, T.L. (eds.) *Modern Epidemiology* (3rd ed.). Philadelphia, PA: Lippincott Williams & Wilkins, pp. 381–417.
Greenland, S. (2008b). Introduction to regression modeling. In: Rothman, K.J., Greenland, S., and Lash, T.L. (eds.) *Modern Epidemiology* (3rd ed.). Philadelphia, PA: Lippincott Williams & Wilkins, pp. 418–57.
Greenland, S. (2011). Null misinterpretation in statistical testing and its impact on health risk assessment. *Preventive Medicine*, 53, 225–8.

Greenland, S. (2012a). Nonsignificance plus high power does not imply support for the null over the alternative. *Annals of Epidemiology*, 22, 364–8.
Greenland, S. (2012b). Causal inference as a prediction problem: assumptions, identification, and evidence synthesis. In: Berzuini, C., Dawid, A.P., and Bernardinelli, L. (eds.) *Causal Inference: Statistical Perspectives and Applications*. John Wiley and Sons, Chichester, UK, pp. 43–58.
Greenland, S. (2014). Sensitivity analysis and bias analysis. In: Ahrens, W. and Pigeot, I. (eds.) *Handbook of Epidemiology* (2nd ed.). New York: Springer, pp. 1087–159.
Greenland, S. and Lash, T.L. (2008). Bias analysis. In: Rothman, K.J., Greenland, S., and Lash, T.L. (eds.) *Modern Epidemiology* (3rd ed.). Philadelphia, PA: Lippincott Williams & Wilkins, pp. 345–80.
Greenland, S. and O'Rourke, K. (2008). Meta-analysis. In: Rothman, K.J., Greenland, S., and Lash, T.L. (eds.) *Modern Epidemiology* (3rd ed.). Philadelphia, PA: Lippincott Williams & Wilkins, pp. 652–82.
Greenland, S. and Poole, C. (2011). Problems in common interpretations of statistics in scientific articles, expert reports, and testimony. *Jurimetrics*, 51, 113–29.
Greenland, S. and Poole, C. (2013a). Living with P-values: resurrecting a Bayesian perspective. *Epidemiology*, 24, 62–8.
Greenland, S. and Poole, C. (2013b). Living with statistics in observational research. *Epidemiology*, 24, 73–8.
Gustafson, P. and Greenland, S. (2017). Misclassification. In: Ahrens, W. and Pigeot, I. (eds.) *Handbook of Epidemiology* (2nd ed.). New York: Springer.
Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*. New York: Chapman and Hall.
Hernán, M.A., Brumback, B., and Robins, J.M. (2000). Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology*, 11, 561–70.
Hernán, M.A., Hernandez-Diaz, S., and Robins, J.M. (2004). A structural approach to selection bias. *Epidemiology*, 15, 615–25.
Hosmer, D.W. and Lemeshow, S. (2000). *Applied Logistic Regression* (2nd ed.). New York: Wiley.
Jewell, N. (2004). *Statistics for Epidemiology*. Boca, FL: Chapman and Hall/CRC.
Kaufman, S., et al. (2005). Improved estimation of controlled direct effects in the presence of unmeasured confounding by intermediate variables. *Statistics in Medicine*, 24, 1683–702.
Kelsey, J.L., et al. (1996). *Methods in Observational Epidemiology* (2nd ed.). New York: Oxford University Press.
Koepsell, T.D. and Weiss, N.S. (2003). *Epidemiologic Methods*. New York: Oxford University Press.
Lash, T.L., Fox, M.P., and Fink, A.K. (2009). *Applying Quantitative Bias Analysis to Epidemiologic Data*. Boston, MA: Springer Publishing Company.
Lash, T.L., et al. (2014). Good practices for quantitative bias analysis. *International Journal of Epidemiology*, 43, 1–17.
Leamer, E.E. (1978). *Specification Searches*. New York: Wiley.
Maldonado, G. and Greenland, S. (1994). A comparison of the performance of model-based confidence intervals when the correct model form is unknown. *Epidemiology*, 5, 171–82.
Morgenstern, H. (2008). Ecologic studies. In: Rothman, K.J., Greenland, S., and Lash, T.L. (eds.) *Modern Epidemiology* (3rd ed.). Philadelphia, PA: Lippincott Williams & Wilkins, pp. 511–31.
Ogburn, E.L. and VanderWeele, T.J. (2012). On the nondifferential misclassification of a binary confounder. *Epidemiology*, 23, 433–9.
Ogburn, E.L. and VanderWeele, T.J. (2013). Bias attenuation results for nondifferentially mismeasured ordinal and coarsened confounders. *Biometrika*, 100(1), 241–8.
Park, S.Y., et al. (2007). Calcium and vitamin D intake and risk of colorectal cancer: the Multiethnic Cohort Study. *American Journal of Epidemiology*, 165, 784–93.

Pearl, J. (2009). *Causality: Models, Reasoning and Inference* (2nd ed.). Cambridge: Cambridge University Press.
Poole, C. (1985). Exceptions to the rule about nondifferential misclassification (abstract). *American Journal of Epidemiology*, 122, 508.
Poole, C. (1987a). Beyond the confidence interval. *American Journal of Public Health*, 77, 197–9.
Poole, C. (1987b). Confidence intervals exclude nothing. *American Journal of Public Health*, 77, 492–3.
Poole, C. (1999). Controls who experienced hypothetical causal intermediates should not be excluded from case-control studies. *American Journal of Epidemiology*, 150, 547–51.
Poole, C. (2001). Low P-values or narrow confidence intervals: which are more durable? *Epidemiology*, 12, 291.
Robins, J.M. (1988). Confidence intervals for causal parameters. *Statistics in Medicine*, 7, 773–85.
Robins, J.M. and Greenland, S. (1986). The role of model selection in causal inference from nonexperimental data. *American Journal of Epidemiology*, 123, 392–402.
Robins, J.M. and Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3, 143–5.
Robins, J.M. and Greenland, S. (1994). Adjusting for differential rates of prophylaxis therapy for PCP in high- versus low-dose AZT treatment arms in an AIDS randomized trial. *Journal of the American Statistical Association*, 90, 737–49.
Robins, J.M., Hernán, M.A., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5), 550–60.
Rothman, K.J. (1978). A show of confidence. *The New England Journal of Medicine*, 299, 1362–63.
Rothman, K.J. (ed.) (1988). *Causal Inference*. Chestnut Hill, MA: Epidemiology Resources.
Rothman, K.J., Greenland, S., and Lash, T.L. (eds.) (2008). *Modern Epidemiology* (3rd ed.). Philadelphia, PA: Lippincott Williams & Wilkins.
Rothwell, P.M. (2005). Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. *The Lancet*, 365, 176–86.
Savitz, D.A. and Pearce, N. (1988). Control selection with incomplete case ascertainment. *American Journal of Epidemiology*, 127, 1109–17.
Schlesselman, J.J. (1992). *Case-Control Studies: Design, Conduct, Analysis*. New York: Oxford University Press.
Slud, E. and Byar, D. (1988). How dependent causes of death can make risk factors appear protective. *Biometrics*, 44, 265–70.
Sullivan, S. and Greenland, S. (2013). Bayesian regression in SAS software. *International Journal of Epidemiology*, 42, 8–17.
Swan, S.H., Shaw, G.R., and Schulman, J. (1992). Reporting and selection bias in case-control studies of congenital malformations. *Epidemiology*, 3, 356–63.
Thomas, D.C., Witte, J.S., and Greenland, S. (2007). Dissecting complex mixtures: who's afraid of informative priors? *Epidemiology*, 18, 186–90.
Valeri, L. and VanderWeele, T.J. (2013). Mediation analysis allowing for exposure-mediator interactions and causal interpretation: theoretical assumptions and implementation with SAS and SPSS macros. *Psychological Methods*, 18(2), 137–50.
VanderWeele, T.J. (2008). The sign of the bias of unmeasured confounding. *Biometrics*, 64, 702–6.
VanderWeele, T.J. (2009). Sufficient cause interactions and statistical interactions. *Epidemiology*, 20,–6–13.

VanderWeele, T.J. (2010). Bias formulas for sensitivity analysis for direct and indirect effects. *Epidemiology*, 21, 540–51.
VanderWeele, T.J. (2012). Confounding and effect modification: distribution and measure. *Epidemiologic Methods*, 1, 55–82.
VanderWeele, T.J. and Arah, O.A. (2011). Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments and confounders. *Epidemiology*, 22, 42–52.
VanderWeele, T.J. and Hernán, M.A. (2012). Results on differential and dependent measurement error of the exposure and the outcome using signed DAGs. *American Journal of Epidemiology*, 175, 1303–10.
VanderWeele, T.J., Hernán, M.A., and Robins, J.M. (2008). Causal directed acyclic graphs and the direction of unmeasured confounding bias. *Epidemiology*, 19, 720–8.
VanderWeele, T.J. and Vansteelandt, S. (2010). Odds ratios for mediation analysis with a dichotomous outcome. *American Journal of Epidemiology*, 172, 1339–48.
Vansteelandt, S. (2012). Estimation of direct and indirect effects. In: Berzuini, C., Dawid, A.P., and Bernardinelli, L. (eds.) *Causal Inference: Statistical Perspectives and Applications*. Chichester: John Wiley and Sons, pp. 126–50.
Wacholder, S., Dosemeci, M., and Lubin, J.H. (1991). Blind assignment of exposure does not always prevent differential misclassification. *American Journal of Epidemiology*, 134, 433–7.
Wacholder, S., et al. (1992). Selection of controls in case-control studies. *American Journal of Epidemiology*, 135, 1019–50.
Wang, R., Lagakos, S.W., Ware, J.H., Hunter, D.J., and Drazen, J.M. (2007). Statistics in medicine—reporting of subgroup analyses in clinical trials. *The New England Journal of Medicine*, 357, 2189–94.
Weinberg, C.R. (1985). On pooling across strata when frequency matching has been followed in a cohort study. *Biometrics*, 41, 103–16.
Weinberg, C.R., Umbach, D., and Greenland, S. (1994). When will non-differential misclassification preserve the direction of the trend? *American Journal of Epidemiology*, 140, 565–71.
White, H. (1993). *Estimation, Inference, and Specification Analysis*. New York: Cambridge University Press.

5.12

Causation and causal inference

Katherine J. Hoggatt, Tyler J. VanderWeele, and Sander Greenland

Introduction to causation and causal inference

This chapter offers an introduction to causal inference theory for public health research. Causal inference can be viewed as a prediction problem, addressing the question of what the likely outcome will be under one action versus an alternative action. Although asking these types of questions is very natural, useful answers require a high degree of precision in both the statement of the causal questions and the description of analysis techniques and limitations. This chapter focuses on formulation of causal hypotheses and considerations in answering them. More thorough coverage of these issues can be found in chapters 2, 4, 9, and 19 of Rothman et al. (2008) and in the book by Hernán and Robins (2020).

We first review considerations for evaluating evidence of causality, then describe the potential-outcome (counterfactual) framework for cause and effect, showing how measures of effect and association are distinguished in that framework. The potential-outcome framework illustrates problems inherent in attempts to quantify the changes in health expected under different actions or interventions. We conclude with a discussion of how research findings may be translated into policy.

The study of cause and effect

The emergence of modern science has led to increasingly sophisticated methods to distinguish causal effect from non-causal association, aided by formal theories of causation. A formal theory of causation facilitates application of deductive reasoning to the process of causal inference. Of

great importance, it allows us to define precisely the difference between causation and association and gives guidance on how to pose causal questions and test causal hypotheses. It does not, however, provide a basis for 'proving' causality, nor can it model the entire process of causal inference (Greenland 2017a).

That process begins in early childhood, in which we acquire a notion of cause and effect that is based on the observation of one event always or often following another. We observe that flipping a switch consistently results in a light going on and off, and readily assign a cause and effect interpretation to these relations. Nonetheless, such naïve equation of 'variation in tandem' with causation will too often conflate causation with mere association. The key observation cementing our inference that a given switch controls the light is that we see how flipping the switch is followed immediately by the light change—in other words, we repeatedly witness that an *intervention*, or an event we believe shares no important causes with the light, is followed by the light.

A key feature of modern theories of causation is that they formalize these early beginnings by providing comparative predictions about the consequences of alternative interventions (e.g. dietary supplementation versus none) rather than singular predictions of events assuming no intervention (e.g. forecasts of flu epidemics) (Greenland 2012). Nonetheless, as Hume (1739) realized and discussed at length, there is no deductive way to prove causality. Thus, causal inference can only be a speculative, inductive process whose output is a theory of cause and effect. Indeed, the history of science shows that all theories, including causal ones, are necessarily provisional and that future evidence may lead to rejecting previously 'established' relations (Popper 1968; Kuhn 1970).

Those working in public health or policy must often make decisions on how to proceed in the face of equivocal evidence. The question then arises of how to weigh the evidence in favour of cause and effect to determine, say, whether smoking should be banned in bars to prevent lung cancer in bar workers, whether pregnant women should be told to take folate to prevent neural tube defects in their offspring, or whether postmenopausal women should be prescribed hormone replacement therapy (HRT) to prevent heart disease. In each of these cases, the final decision by health or policy professionals was made in the absence of definitive proof of

causality, relying instead on the imperfect evidence gleaned from available studies.

Because of its central importance, causal inference is often confused with the more complex process of rational decision-making. While causal inference is a crucial input to decision-making, decision-making also depends on the costs and benefits of the outcomes. For example, policy makers believed that a ban on smoking in bars and pubs was likely to have health benefits that outweigh any negative economic impacts. This perceived imbalance favouring benefits over costs would be a key basis for a proposed ban. In contrast, recommendations that doctors prescribe HRT for the prevention of heart disease must acknowledge that HRT could have serious side effects on women's health. The recommendations are easily challenged if the costs are shifted according to individual preferences and values.

Such examples show that multiple effects should be considered when policymakers evaluate the evidence of cause and effect based on epidemiological studies. Nonetheless, most attempts to formalize the process of causal inference focus on single endpoints. This single-endpoint view must be understood thoroughly before considering multiple endpoints.

Hill's considerations

Many authors have attempted to list considerations that should be weighed when advancing causal arguments. Best known are those by Sir Austin Bradford Hill (Hill 1965), which are closely related to the considerations found in the first US Surgeon General's report on smoking and health (United States Department of Health, Education and Welfare 1964). Hill's article has been widely misinterpreted as providing necessary conditions ('criteria') for causation to be inferred; see Phillips and Goodman (2004) for a discussion of this and other misinterpretations. In Hill's own words, the considerations were *not* meant to be a checklist or set of criteria, but, rather, simply 'points to consider', namely: (1) strength, (2) consistency, (3) specificity, (4) temporality, (5) biological gradient, (6) plausibility, (7) coherence, (8) experimental evidence, and (9) analogy; we shall discuss each of these in turn. Of these considerations or 'viewpoints' (as Hill called them), the only one necessary for inferring causation is temporality (time

order). The other considerations have been debated at length, with various calls for amendment and caution, e.g. see Weiss (1981, 2002), Maclure (1985); Weed (1986); Susser (1991), Kaufman and Poole (2000), Phillips and Goodman (2004, 2006), Höfler (2005, 2006), and Rothman et al. (2008, ch. 2).

Strength

The strength of an association may be considered important because it is harder to explain away strong effects as complete artefacts of biases such as confounding. There is no general rule for how large an association needs to be to provide compelling evidence for an effect, although a risk or rate ratio of 2 is sometimes taken as a boundary below which bias as a complete explanation of an association starts to become plausible. Nonetheless, there are many smaller associations that are generally agreed to reflect causal effects, in part because they have been replicated in a variety of populations using different designs and in part because of considerations other than strength, such as dose–response (biological gradient). As an example, the rate ratio associating environmental tobacco smoke with lung cancer is below 2, but its causal nature is supported by interpolation between the low rates seen in unexposed non-smokers and the much higher rates seen among direct smokers.

On the other hand, relatively strong non-causal associations have been observed. One example is the association between birth order and Down syndrome: maternal ageing, which is strongly associated with infant birth order, has since been accepted as the actual cause of Down's syndrome and the source of the association. In the absence of information on actual causes, sensitivity and bias analysis can be used along with external information to estimate or quantitatively speculate on the extent to which possible bias sources have affected the strength of an observed association (Greenland and Lash 2008; Lash et al. 2009; VanderWeele and Arah 2011; Greenland 2014; Hernán and Robins 2020).

Consistency

A consistent finding is an association reported across populations, time, and study designs. Although the presence of a consistent result is often taken as

a compelling argument for causality, it can be misleading; for example, an inverse association of beta-carotene with cancer was seen across epidemiological studies but not in subsequent randomized trials. Conversely, the absence of consistency does not imply the absence of a causal effect. For example, acute effects may be submerged in studies that examine only chronic exposure, chronic effects may be submerged in studies that involve only short follow-up or one-time exposure measurement, and some agents may operate only in highly-susceptible subpopulations. Nonetheless, consistency across studies helps rule out chance as an explanation for an observed association. Consistency across studies of different designs can provide important evidence against specific biases that have been controlled by design; for example randomization allows for statistical control of unmeasured confounders, and complete follow-up prevents important bias from non-random losses.

Specificity

In Hill's formulation, specificity can refer either to a cause having a single effect or an effect having a single cause. The limitations of this consideration are apparent even in the case of smoking and lung cancer. Although the smoking and lung cancer association is stronger (on the ratio scale) than the association between smoking and heart disease, there is general agreement today that smoking can increase the risk of many diseases, including heart disease and other cancers.

Specificity can however strengthen a causal inference if a competing non-causal hypothesis would predict a non-specific association (Weiss 2002). For example, some have argued that screening for certain cancers is ineffective in and of itself for reducing cancer mortality, and that the apparent protective effect after adjustment for lead-time reflects the fact that those people who choose to undergo regular screenings are generally more health conscious and therefore at a lower risk for cancer due to their 'healthy lifestyle'. If a specific screening instrument, say a mammogram, were shown to be associated with lower mortality from breast cancer but *not* lower mortality from cancers in other sites, it could strengthen the inference that mammography is indeed useful in preventing deaths from breast cancer; this approach is sometimes described as using 'negative controls' (Lipsitch et al. 2010).

Temporality

Temporality means that a cause must necessarily precede its effect. This positive assertion is inarguable, but it is also possible that in some instances a putative cause (X) may be observed to follow the hypothesized effect (Y), whereas in other instances the putative cause (X) precedes and does in fact cause the effect (Y). Thus an observation that Y preceded X in a specific case (reverse order) is not an argument against causality in cases where Y follows X.

Reverse time order is thus a definitive argument against causality only if we know that X *cannot* come before Y. Establishing the correct temporal order between X and Y would, however, be a necessary condition for inferring causality in a given case. If X and Y vary over time, it is also possible that there is feedback between them such that X and Y both affect each other. If data are available on multiple measurements of these variables, longitudinal causal modelling can help evaluate such feedback (Hernán and Robins 2020).

Biological gradient (dose–response)

A biological gradient is a dose–response relation between an exposure and an outcome. Though most epidemiologists think of a ‘linear’ or monotonic dose–response association as strengthening a causal inference, this need not be the case. There may be a sharp increase in risk for specific outcomes at low to moderate doses of a given exposure that tapers off at higher doses, or there may be no adverse effect until saturation of detoxification mechanisms is reached, resulting in increased risk only at higher doses.

Many substances show non-monotonic trends with outcomes in a manner consistent with their hypothesized biological effects. For example, both very low and very high doses of certain vitamins may increase the risk of death, a fact consistent with what we know about the biological properties of these vitamins. In addition, an apparent monotonic dose–response relation (risk always increasing with greater exposure or risk always decreasing with greater exposure) does not prove the observed associations are genuine rather than artefacts of confounding. If an uncontrolled confounder is monotonically associated with the exposure and monotonically associated with the outcome, an apparent dose–response

relation between exposure and outcome could simply reflect the dose–response relation between the confounder and outcome (Weiss 1981). Sensitivity analysis can clarify the extent to which an observed dose–response relation contributes to evidence of causality (Rosenbaum 2003).

Plausibility

A plausible association is one that conforms to the current scientific understanding of the relation between a putative cause and its effect. This understanding can be informed by other epidemiological studies, animal studies, biology, toxicology, etc. As Hill noted, a limitation of plausibility in assessing causality is that current scientific understanding can be incomplete or wrong. For example, the higher malaria rates seen in lowland and swampy areas made plausible the idea that the disease was caused by 'bad air' (Medieval Italian *mala aria*), but was instead simply an artefact of those areas supporting higher densities of *Anopheles* mosquitos. Conversely, it was once deemed implausible or even impossible that genes could 'jump' across genome locations, yet this phenomenon is now well documented.

Coherence

Hill used the term 'coherence' to mean that the hypothesized causal relation of exposure to outcome is not in conflict with current scientific understanding of the disease process. Hill emphasized that an absence of coherence was not a strong argument against causality. If an observed association apparently conflicts with current scientific understanding of a disease process, it may be that the understanding is mistaken rather than that the association is non-causal. For example, at the time it was made, the observation that shallow inhalers of cigarette smoke had increased rates of lung cancer relative to deep inhalers seemed to contradict the understanding of lung cancer aetiology. Subsequent research on lung cancer, however, found that the cells affected by cigarette smoke tended to be in the upper lung and respiratory tract, which was in fact consistent with the observation of an elevated association for shallow inhalers (Wald 1985).

Conversely, coherence alone does not supply strong support of causality. The theory that the inverse association of beta-carotene with cancers

observed in epidemiological studies represented a preventive effect was quite coherent with contemporary theories of antioxidants and cancer, but the theory that this association reflected a causal effect of beta-carotene was nonetheless refuted by subsequent randomized trials.

Experimental evidence

In modern discussions, ‘experimental evidence’ often refers to either human or animal experiments, which are themselves very different categories of evidence. In practice, experimental evidence for a specific causal relation may be limited or absent: for many exposures, human experimentation is impractical or unethical, and, for many disease processes, no suitable animal model exists.

Hill seemed to have a notion of ‘experiment’ that was broader than randomized trials, however, and more akin to the notions of classical writers such as Mill (1862) in which an experiment involves an intervention that largely determines who is exposed. Hill discussed the evidence obtained by examining the change in disease frequency when the potentially harmful exposure has been reduced or removed, as in certain natural experiments. These types of before-and-after comparisons may provide compelling evidence, although they can mislead if there are uncontrolled time-varying confounders that drive the observed associations. For example, the steep rise in lung-cancer rates seen in Western countries in the first half of the 20th century coincided with many changes including the spread of electrical service, indoor plumbing, and telephones, as well as the increased use of cigarettes.

Randomized experiments are sometimes claimed as necessary or even sufficient evidence for or against an effect. Such extreme views are simply wrong (Hill 1965; Pearl 2009; Deaton and Cartwright 2018). Failure to detect an effect in a randomized experiment may reflect only lack of sensitivity or power, and effects generally agreed to have been ‘detected’ may still be controversial. The history of screening mammography (Biller-Adorno and Jüni 2014) shows that even randomized trial evidence need not be definitive or settle issues. Among the many reasons is that trial populations can be quite different from those exposed in everyday life (e.g. trial population are often specially selected for being thought likely to respond positively to treatment, without severe side effects). Then too, trials

can have their own defects and limitations such as noncompliance and drop-out that diminish the advantages conferred by initial randomization (although these problems can be partially adjusted for using special techniques; Greenland 2002, 2005; Hernán and Robins 2020). In addition, because of their much higher expense per participant, randomized trials tend to have smaller sample sizes than other studies, thus limiting their precision and power for detecting effects. Finally, for studies of medical procedures such as mammography the treatment itself may change considerably over time with improved equipment and training, or may be used more poorly in practice than in trials with strictly enforced protocols; in either case, results from even the best trials may be misleading for current practice.

Analogy

Analogy refers to drawing inferences about the association between a given exposure and disease based on what is known about other exposure–disease relations. From the known effects of cigarette smoking on lung cancer, we might expect that inhalation of other combustibles (e.g. cannabis) would have similar effects, even in the absence of studies on the subject. Analogy can be a useful scientific tool for generating new hypotheses about disease processes; however, its utility for assessing causality is limited by the understanding, imagination, and knowledge of the scientist.

Statistical significance

Notably absent from Hill's list is 'statistical significance' or the like. Hill (1965) prefaced his considerations with the supposition that '[o]ur observations reveal an association between two variables, perfectly clear-cut and beyond what we would care to attribute to the play of chance'. This initial statement led some to claim that Hill considered 'statistical significance' a necessary condition. However, such claims ignore Hill's failure to mention any criterion for a 'statistical significance' condition and his devoting a whole section to expressing scepticism about significance tests. Notably, Hill concluded that 'too often I suspect we waste a deal of time, we grasp the shadow and lose the substance, we weaken our capacity to interpret data and to take reasonable decisions whatever the value of P.

And far too often we deduce ‘no difference’ from ‘no significant difference’.’ It appears, then, that Hill was wary of making too much of the absence as well as the presence of ‘statistical significance’ when forming inferences.

Regardless of Hill’s precise views, today it is widely recognized that ‘statistical significance’ is a controversial concept whose presence or absence should never be confused with presence or absence of an effect: if a statistical hypothesis test is desired one should present the actual resulting p-value for consideration, not whether the p-value crossed a threshold (Greenland et al. 2016; Wasserstein and Lazar 2016; Amrhein et al. 2017, 2019; Greenland 2017b; Wassertein et al. 2019; McShane et al. 2019; Rafi and Greenland 2020). Furthermore, in presenting such p-values for associations or effects, is essential that they be accompanied by a p-value for an alternative hypothesis or an interval estimate, with thorough recognition of the many assumptions on which all statistics depend (Greenland 2017b; Rafi and Greenland 2020).

Summary

For reasons such as those just discussed, Hill and other critical thinkers have discouraged taking his ‘viewpoints’ as a checklist for causality. Despite the ongoing popularity of Hill’s considerations—sometimes in modified form—there is no general agreement on which of them should be used or how they should be used in making causal inferences (Weed and Gorelic 1996; Kaufman and Poole 2000; Holman et al. 2001; Höfler 2005, 2006; Rothman et al. 2008). Some have argued that such aids to inference amount to refined common sense, which if taken as rigid criteria may harm the inferential process (Lanes and Poole 1984; Phillips and Goodman 2004, 2006; Rothman et al. 2008, ch. 2). Others have used the considerations to develop general deductive formats for evaluating causal hypotheses (Maclure 1985; Weed 1986; Susser 1991). Finally, researchers may use different considerations or deploy them differently based on the specifics of the inferences being made, suggesting that, in practice, most of the considerations reflect values rather than objective inferential tools (Poole 2001).

Formal approaches to causal inference

Implicit in the discussion of causal considerations is the idea that there is a well-defined notion of what constitutes cause and effect. Perhaps the oldest non-circular set of definitions arises from the *potential-outcome* or *counterfactual* framework. The potential-outcome or counterfactual model (Rosenbaum 2002; Morgan and Winship 2007; Rothman et al. 2008 ch. 4; Pearl 2009; VanderWeele 2015; Hernán and Robins 2020) specifies what happens to individuals or populations under alternative actions or interventions and also makes explicit the problems in operationally defining causes (Greenland 2002, 2005; Hernán 2005). This model of causation is also useful in defining methodological concepts such as confounding (as discussed in Rothman et al. 2008, ch. 4). Causal diagrams are a useful and closely related tool that enables quick visualization of the assumptions used in causal models; for basic introductions see Glymour and Greenland (2008), Pearl (2009), Elwert (2013), and Pearl and Mackenzie (2018).

Potential outcomes and counterfactual causality

To see how causal inference can be viewed as a prediction problem, consider that we may wish to predict how the rates of lung cancer among bar workers would change if an indoor smoking ban was enacted. The potential-outcome model allows us to formalize questions such as 'Will a smoking ban decrease the rates of lung cancer in 10 years beyond what can be expected in the absence of a ban?'

In discussions of potential outcomes, three components must be defined:

A target of interest that will be the subject of the study of causation. This target may be an individual, a group, or any other unit of study.
A list of possible interventions or actions x_0, x_1,… that could have been applied to the target over some time span of interest.
An outcome measure (Y) taken after implementation of the intervention or action.

If x_a is the action taken from among possible actions x_0 and x_1, a potential-outcome model posits that there is a well-defined outcome, $Y(x_a)$, that would have followed from that specific action x_a. This outcome is often termed a 'potential outcome'.

As an example, suppose the treatment of interest is sending to members of an insurance plan, on their fiftieth birthday, a mail-in home sampling kit to screen for stool blood, a possible indicator of colon cancer. Suppose $X = 1$ if the kit is supplied to a person, and $X = 0$ if not. Ordinary analyses might examine one outcome variable, such as $Y = 1$ if the person dies of colon cancer by age 60 and $Y = 0$ if the person does not die by this age. In contrast, a potential-outcome analysis imagines a separate outcome variable for every treatment being compared: for $X = 1$, we imagine the outcome $Y(1) = 1$ if the person dies of colon cancer by age 60 *when sent the kit* and $Y(1) = 0$ if the person does not die *when sent the kit*. Similarly, for $X = 0$ we consider the outcome $Y(0) = 1$ if the person dies of colon cancer by age 60 *when not sent the kit* and $Y(0) = 0$ if the person does not die when not sent the kit. Offering the kit is then beneficial for the person with respect to these outcomes if $Y(1) = 0$ but $Y(0) = 1$, which is to say if the person would not die of colon cancer by age 60 if sent the kit but *would so die* if the kit *was not sent*. Presumably, such a benefit would arise because the kit, if sent, would be received, used, and would lead to detection of and intervention for an otherwise fatal tumour.

The potential-outcome model is widely used in statistical discussions of causality and has become increasingly popular in epidemiology and other fields. Yet it is not without controversy, largely because *at most* one potential-outcome variable can be observed for each person. That observable potential outcome is the one corresponding to the treatment actually received; all the other potential outcomes are unobserved and hence are ‘missing data’. Understandably, the incompleteness of the data under the model is upsetting to practitioners, and the inevitable missingness requires reconsideration of many conventional statistical methods. Such reconsiderations have, however, led to innovative methods for previously intractable problems, such as the estimation of effects of time-varying treatment regimens (Hernán and Robins 2020).

There are also a number of aspects of the model that are often misrepresented as limitations but are not. For example, the model does *not* require that outcomes be well defined for any possible intervention; it simply requires that the outcomes be well defined for those interventions of interest. If, in a study of motorcycle crash fatalities, we wish to estimate the effect of a helmet law that fines riders who do not use helmets, versus no law, we do not need to specify the outcomes any other actions (such as a

law mandating jail time for non-use). Also, contrary to some misunderstandings, the potential-outcome model is not inherently deterministic: although the model is quite often introduced and taught using deterministic language, a potential outcome could instead be the parameter of a probability distribution, for example the probability of disease (Rothman et al. 2008, ch. 4).

Causes and effects

The term ‘effect’ is sometimes used to specify the outcome of a causal process. A researcher may describe lung cancer or heart disease as two possible effects of cigarette smoking. A quantitative definition of an effect is a numerical contrast in outcome measures corresponding to different actions or interventions, for example, $Y(x_1)$ and $Y(x_0)$; the numerical contrast is then called a *measure of effect*. In the earlier example, if x_1 is the presence of a helmet law and x_0 is the absence of a helmet law, then $Y(x_1)$—$Y(x_0)$ could be the fatality-rate difference for motorcycle crashes and $Y(x_1)/Y(x_0)$ could be the fatality-rate ratio.

Thus, under a potential-outcome model, an effect is a contrast in outcomes corresponding to *two* different actions in *one* single target population (where the population could comprise a single individual). Effects must be defined with respect to a clearly specified alternative or reference condition. To say that smoking one pack of cigarettes a day increases the risk for lung cancer and heart disease would be incorrect without specifying the alternative: smoking one pack a day may be harmful relative to not smoking, the implicit reference condition, but may be protective relative to smoking two packs a day. Statements about the effects of causes may be ambiguous or meaningless unless both the index and reference actions are stated explicitly.

It follows that an action cannot be considered inherently causal—it can only be deemed causal when the outcome associated with it is compared to the outcome associated with an alternative action. Consider a person with depression who is prescribed an antidepressant medication. The weeks following initiation of treatment are thought to be periods of increased risk for suicide. If this person commits suicide in the period following treatment initiation, some would say the antidepressant caused the suicide. Given that

the person was depressed to begin with, it is possible that they would have committed suicide even if they had not been prescribed the medication, in which case the antidepressant did not cause the suicide in reference to the alternative of no prescription. If instead of being prescribed an antidepressant, however, the person had been hospitalized, they may not have committed suicide, in which case the antidepressant treatment may be considered a cause of their suicide, but only in reference to the alternative action of hospitalization.

A second important point is that an effect is a contrast in outcomes for one target population, but, for that single population, we can observe *at most* one of the actions and its corresponding outcome. The other actions, which are not observed, are *counterfactual* (contrary to fact), and as a consequence their corresponding potential outcomes are not observed. For example, $Y(x0)$ would be unobserved (missing data) for all persons who in fact received treatment $x1$. Thus, potential outcomes for counterfactual actions can only be estimated, not observed.

If the index and reference actions lead to identical outcomes, then the index action is said to have no effect on the target population. When an action has an effect, describing it as causal or preventive depends on which outcome we are referring to. If smoking a pack a day for 20 years (versus none) causes an individual to die at age 70, then this smoking also prevents survival past age 70. In parallel, if not smoking prevents death at age 70 relative to smoking a pack a day for 20 years, it causes that individual to survive past age 70.

Confounding and measures of association

Our inability to observe the potential outcome of more than one action fundamentally limits our ability to make causal inferences. In any attempt to estimate a causal effect, at least one potential outcome, corresponding to a counterfactual action, must be estimated or predicted rather than measured. In practice, this is accomplished by measuring outcomes in other individuals who in fact experience the reference action. In experiments and in cohort studies, these individuals are often called the control subjects.

As an example, on 1 July 2000, Florida weakened its existing motorcycle helmet use law to exempt individuals over the age of 21 who had at least US$10,000 in insurance in case of an accident. To estimate the effect of

weakening the helmet law in Florida, we could compare (contrast) the rates of motorcycle crash fatalities before and after the change. Specifically, we use the fatality rate in 1999 as a substitute for the unobserved (counterfactual) fatality rate in 2001 had the law not been changed. The resulting contrast is no longer a measure of effect, because it is not a contrast of the outcomes of two different actions in a single target population. Instead, it is a *measure of association*, which is a contrast of outcomes across groups exposed to different actions. We use this measure of association as a substitute for or estimate of the corresponding measure of effect. The amount of *confounding* can be quantified as the difference or ratio between the measure of association and the measure of effect that the association is intended to estimate.

Although confounding is a serious threat to our ability to make causal inferences, we can often determine factors that contribute to the disparity between the measure of association and the measure of effect. In our example, these factors, commonly referred to as *confounders*, would be the reasons why the motorcycle fatality rates in 1999 would differ from those in 2001 had the helmet law not changed. When confounders are measured, analytic methods can be used to minimize the amount of confounding due to those factors and thus strengthen our causal inferences. The most basic methods for confounding control are based on stratifying the observations on values of the confounders. Summary comparisons can be obtained, e.g. via *standardization*, which averages outcomes in the index and reference groups using a shared weighting scheme or standard (Rothman et al. 2008, ch. 4). Other methods for obtaining summary comparisons make special assumptions to combine measures of association across confounder levels (e.g. logistic regression on exposure and confounders assumes the exposure-outcome odds ratios are constant across levels of the confounders). In all these approaches, the resulting summary comparison across confounder levels is said to be *adjusted* for the confounders.

An adjusted measure of association will be a good estimate for the desired measure of effect only insofar as we have controlled for (e.g. stratified on or modelled) the important confounders. If there are other factors that contribute to important differences between the counterfactual (alternative) outcome and the observed outcome used as its substitute, then the measure of association will remain a biased estimate of the measure of effect. In the case of the helmet law and the motorcycle crash fatalities, one

would have to ask if all the important factors contributing to changes in the rate of crash fatalities *other than the change to the helmet law* had been measured and properly controlled. To the extent we are uncertain or doubtful about our measurement and control of important confounders, we would have to doubt the validity of the adjusted association as an estimate of the measure of effect.

Practical uses of causal inference

Academic discussions of cause and effect often focus on the 'effects' of changes that have no realistic interpretation as actions (e.g. elimination of lung cancer) or actions that are not realistic options (e.g. achieving a 'smoke-free society'). Even when discussions focus on the impact of feasible interventions, they are often based on unrealistic assumptions about their consequences, for example, assuming 100% compliance with a given treatment or behaviour-modification programme.

Discussions of the expected health impact if a given disease or outcome could be removed are irrelevant and often misleading for policy formulation. For example, consider the 'burden of disease due to lung cancer'. A rough ranking of actions and questions from irrelevant and unrealistic to most relevant would be:

Directly eliminating lung cancer (e.g. 'How many deaths could be prevented if we could eliminate lung cancer?').
Removing risk factors for lung cancer (e.g. 'How many deaths could be prevented if we could eliminate cigarette smoking?').
Applying an intervention to change risk behaviours for an entire population (e.g. 'If we could implement a given smoking-cessation programme to the entire smoking population of a nation, how many deaths could be prevented?').
Applying an intervention that will change in risk behaviours in a specially targeted segment of a population (e.g. 'If we could implement a given smoking-cessation programme in a targeted group of motivated individuals, how many deaths could be prevented?').

We will discuss each of these actions in turn.

Removing outcomes

When the outcomes are not directly manipulable, any prediction based on the hypothesized removal of the outcomes needs to consider not just the fact of the outcome removal but also the method used to remove it (Greenland 2002, 2005; Kalbfleisch and Prentice 2002; Hernán 2005; Hernán and VanderWeele 2011). Even if we can define a realistic intervention for removal of the outcome, we cannot assume that removal of the outcome will leave our population with a risk profile similar to that for another population that did not experience the intervention or outcome. Unfortunately, standard statistical procedures for analysing these types of data make the assumption that the risk profiles across these populations (intervention and no outcome vs. no intervention and no outcome) are similar (an assumption of 'independent competing risks'). Thus, it is reasonable to expect that even if we can conceive of an intervention that will prevent a specific outcome, estimates of effect in light of this outcome removal will likely be biased if based on the experience of those who do not experience the intervention or the outcome.

To estimate the effects of outcome removal, one must compare successfully treated individuals. To illustrate, suppose Y represents cause of death, with $Y = y_1$ for death from myocardial infarction (MI) and $Y = y_0$ for death from other causes ('competing risks'). An individual will eventually experience one but not both of these outcomes, so at least one must be counterfactual. Now suppose T represents years of survival past age 50. The value T will take depends not only on the value of Y but also on how that value (y_1 or y_0) is brought about. Suppose a man with a sedentary lifestyle and poor diet died from an MI at age 54, so that we observe $T(y_1) = 4$ (the number of years lived past the age of 50). The value of $T(y_0)$ would then be the number of years lived past age 50 *if the MI death had been prevented*. This value would depend critically on how the MI death was prevented. If the man had taken up a regimen of physical activity and a healthy diet in his twenties, $T(y_0)$ could be considerably larger than $T(y_1)$ because the regimen would lower his risk of many causes of death besides MI. If instead the MI death was prevented by medical care at the time the MI occurred, $T(y_0)$ and $T(y_1)$ may not differ by much because the man would remain at an increased risk for stroke, cancer, and other potentially fatal events.

Removing risk factors

Consider the debate on obesity and mortality. Numerous studies have reported increases in mortality associated with having a body mass index (BMI) over 30 (the cutpoint for obesity) relative to a 'normal' BMI of 20–24.9. Some have used these findings to assert that obesity causes early mortality and that getting people to lose weight will save lives. But suppose not all individuals who go from a BMI of over 30 to a BMI of (say) 24 lose weight in the same way. Some might start exercising and eating more healthfully, in which case lower mortality may in fact result. Others might undergo gastric bypass surgery or become chronic users of amphetamines, each of which are known to confer some risk for early mortality.

Unfortunately, studies that examine only the association of BMI with mortality tell us nothing about the impact of different interventions to change BMI. To make a relevant causal inference, one must specify an intervention aimed at reducing BMI and then compare the mortality outcomes under that intervention to outcomes under a reference intervention (such as no intervention). The resulting measure of association, as an estimate of a measure of effect, would *not* be a contrast in outcomes of persons at different levels of BMI; it would be a contrast in outcomes under *specific interventions* among persons who started with the same BMI.

The very idea of the 'health effects of BMI' suffers from confusion of the effects of interventions (e.g. a diet and exercise regimen) with the downstream results of intermediate outcomes (e.g. changes in BMI). At best, mortality rates across BMI groups are averages taken across a range of *unknown* interventions that brought about the observed BMI levels. As a consequence, comparisons of these rates correspond to comparing average outcomes under differing distributions of unknown interventions (VanderWeele and Hernán 2013), and even this limited interpretation holds only under very strong assumptions. A change in BMI is not an intervention, thus one cannot unambiguously define the causal effect of a change in BMI. A similar problem arises if one considers the effect on lung cancer incidence if more people quit smoking. As with BMI, a researcher would need to specify an intervention (e.g. nicotine patches, hypnotherapy, medication, etc.) before discussing the effect of getting smokers to quit.

When conceptualizing exposures in observational studies as interventions, as opposed to protocols in clinical trials, there are limitations

to how detailed the description can be, making vagueness about the hypothetical intervention unavoidable (Robins and Greenland 2000). Nevertheless, moving away from thinking of outcomes such as BMI as potential causes—as opposed to intermediates—is necessary for making policy-relevant causal inferences from observational data. A related problem arises in discussing the causal effects of personal characteristics. For example, considerable debate has surrounded the question of whether it is meaningful to speak about factors such as race or sex as being possible causes of effects (Holland 1986; Greenland 2005; Kaufman 2008; VanderWeele and Hernán 2012).

Although it might seem natural to ask whether the risk of a disease like hypertension is affected by a person's race, studies that contrast rates of hypertension in different racial groups do not adequately address this question. To ameliorate the racial disparities in hypertension, we would first need to identify feasible interventions. For this goal, discussing race as a cause is of no help, because changing the race of individuals is not a clearly defined action, let alone a feasible or desirable intervention. Observed associations between race and hypertension indicate the disparities that exist, but these associations generally do not pinpoint underlying causes and are thus insufficient to inform efforts to mitigate disparities (Kaufman 2008).

Finally, when considering effects of baseline characteristics it is important to remember that later events may be affected by those characteristics and thus adjustment for those events will bias the resulting effect estimate. This is important to bear in mind as modern estimates may involve adjustment for many variables, some of which should not be included based on causal analysis (Rothman et al. 2008 chapter 9; Glymour and Greenland 2008; Westreich and Greenland 2013).

Identifying and evaluating feasible interventions

Even when an exposure is an intervention, the discussion of its effects can be complicated if there is treatment non-compliance, the intervention is not randomized, or if the researcher wishes to generalize effect estimates to different target populations.

In practice, apart from occasional mandates, treatments are only prescribed or recommended, therefore prescription or recommendation is all

that can be unambiguously studied and implemented. What a subject actually chooses to do in response to the assigned treatment is a function not only of the assignment but also of other factors, most of which are either infeasible to modify (such as the subject's age) or for which modification introduces further treatment (e.g. side effects). If the researcher wishes to understand the effects of received treatment rather than treatment assignment, it is important to realize that received treatment is an *outcome*, not an intervention, and is thus subject to the problems described earlier for intermediate outcomes such as BMI. Special techniques are available for studying the effects of received treatment (e.g. instrumental variables, g-estimation; see Hernán and Robins (2020) for introductions to these topics). Nonetheless, effects estimated by these techniques are those expected under ideal conditions and will likely overestimate the effects that will be seen in field implementation.

The most relevant causal hypotheses for policymakers and health practitioners concern the changes in outcome to be expected under the implementation of feasible interventions, subject to all their compliance problems. Estimating field effects of feasible interventions is difficult, but doing so at least avoids the unrealism and biases that arise from pretending that subject characteristics or behaviour are under direct or perfect control of health practitioners or society. In studying these types of interventions, the effect measure of interest will be a contrast in outcomes corresponding to different levels of the intervention (e.g. whether or not to fund a given programme), rather than the 'received treatment' discussed earlier. Thus the question is not what would happen if people were to comply with the intervention they were assigned, but rather what would result if people were offered a given treatment at a given cost.

Studies that address causal hypotheses about feasible interventions face profound methodological difficulties. If the intervention is not assigned at random (e.g. if individuals who participate in a given intervention are volunteers, and their outcomes are compared to those who chose not to participate), the resulting measure of association is likely to be confounded and will therefore be a poor estimate of the desired measure of effect. The association can also be distorted due to refusals, loss to follow-up, and measurement error, among other things. These threats to validity are discussed in epidemiological methods texts such as Rothman et al. (2008).

Problems in generalizing or transporting the results from a particular study to a target population of interest are often not given detailed attention, yet they are of great concern for policymaking. The target populations for policy are always different from study populations, for at the very least the target population is defined in terms of future experience, whereas study results are always based on the past. Even if there are no time trends, however, factors such as the compliance rate and response to the intervention depend on the characteristics of the individuals involved in a given study. The results from a study may therefore not be applicable to other populations, each of which may be composed of a different mix of individuals.

This problem of population heterogeneity of effect is sometimes framed as an issue of 'effect measure modification', meaning that the measure of the intervention effect depends on characteristics of the individual or the population (see Rothman et al. 2008, ch. 4; Hernán and Robins 2020). In theory, the solution is to identify factors responsible for the heterogeneity, stratify on these factors, estimate effect measures within the stratified groups, and predict the likely impact of an intervention in a target population by standardizing the stratified estimates based on the distribution of the factors in the target population. Extensions based on sample reweighting are also available (Cole and Stuart 2010; Westreich et al. 2017). Nonetheless, it can be difficult to identify all the important sources of heterogeneity, and there are always severe data limitations on the number of factors that can be studied in this fashion.

From research to policy and beyond

This chapter has reviewed some of the considerations necessary in making causal inferences relevant to public health. If the explicit goal of the study is to report associations and thus provide data for subsequent evaluations, rather than to estimate causal effects, these considerations can be deferred to research-synthesis reports (Greenland et al. 2004). In fact, causal inferences from single epidemiological studies are rarely justified (with exceptions typically characterized by enormous risk or rate ratios), and epidemiology is almost never the sole basis for policy decisions, although epidemiological results are important inputs for those decisions.

To name just a few examples, studies on the link between smoking and lung cancer have led to labelling of cigarette packs and bans on indoor smoking; results linking air pollutants to disease and mortality have informed emissions standards; and findings on the connection between vitamin deficiencies and adverse health outcomes have led to fortification of milk, cereals, and flour. In all these examples, decisions were based not only on multiple epidemiological studies but also on evidence from laboratory and clinical studies and considerations of mechanistic theories.

It is thus important for health researchers as well as policymakers to understand the fundamentals of causal inference, especially the often neglected issues of defining meaningful and feasible interventions. It is equally important for researchers to understand that research is done to inform policymaking, not to set policy, and that causal inference represents a synthesis of multiple evidence streams. Apart from perhaps outbreak investigations or large, well-executed randomized trials, no single study is capable of establishing a causal relation. Furthermore, randomized trial populations must be reasonably representative of the intended target populations in order to fully inform clinical or policy decisions, yet trial and target populations are usually quite different. Thus, decisions should be based on a careful consideration of the entire relevant scientific and policy literature and must address cost–benefit considerations as well as purely scientific issues.

REFERENCES

Amrhein, V., Korner-Nievergelt, F., and Roth, T. (2017). The earth is flat (p> 0.05): significance thresholds and the crisis of unreplicable research. *Peer J*, 5, e3544.

Biller-Adorno, N. and Jüni, P. (2014). Abolishing mammography screening programs? A view from the Swiss Medical Board. *New England Journal of Medicine*, 370, 1965–7.

Cole, S.R. and Stuart, E.A. (2010). Generalizing evidence from randomized clinical trials to target populations: the ACTG 320 trial. *American Journal of Epidemiology*, *172*(1), 107–15.

Deaton, A. and Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials (with discussion). *Social Science & Medicine*, 210, 2–90.

Elwert, F. (2013). Graphical causal models. In: Morgan, S.L. (ed.) *Handbook of Causal Analysis for Social Research*. Dordrecht: Springer, pp. 245–72.

Glymour, M.M. and Greenland, S. (2008). Causal diagrams. In: Rothman, K.J., Greenland, S., and Lash, T.L. (eds.). *Modern Epidemiology* (3rd ed.). Philadelphia, PA: Lippincott Williams & Wilkins, pp. 183–209.

Greenland, S. (2002). Causality theory for policy uses of epidemiologic measures. In: Murray, C.J., et al. (eds.) *Summary Measures of Population Health*. Cambridge, MA: Harvard University Press/World Health Organization, pp. 291–301.

Greenland, S. (2005). Epidemiologic measures and policy formulation: lessons from potential outcomes (with discussion). *Emerging Themes in Epidemiology*, 2, 1–4.
Greenland, S. (2012). Causal inference as a prediction problem: assumptions, identification, and evidence synthesis. In: Berzuini, C., Dawid, A.P., and Bernardinelli, L. (eds.) *Causal Inference: Statistical Perspectives and Applications*. Chichester: John Wiley and Sons, pp. 43–58.
Greenland S. (2014). Sensitivity analysis and bias analysis. In: Ahrens, W. and Pigeot, I. (eds.) *Handbook of Epidemiology* (2nd ed.). New York: Springer, pp. 685–706.
Greenland, S. (2017a). For and against methodology: some perspectives on recent causal and statistical inference debates. *European Journal of Epidemiology*, 32, 3–20,
Greenland, S. (2017b), The need for cognitive science in methodology. *American Journal of Epidemiology*, 186, 639–45.
Greenland, S., Gago-Domiguez, M., and Castellao, J.E. (2004). The value of risk-factor ('black-box') epidemiology (with discussion). *Epidemiology*, 15, 519–35.
Greenland, S. and Lash, T.L. (2008). Bias analysis. In: Rothman, K.J., Greenland, S., and Lash, T.L. (eds.) *Modern Epidemiology* (3rd ed.). Philadelphia, PA: Lippincott, pp. 345–80.
Greenland, S., et al. (2016), Statistical tests, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology*, 31, 337–50.
Hernán, M.A. (2005). Hypothetical interventions to define causal effects—afterthought or prerequisite? *American Journal of Epidemiology*, 162, 618–20.
Hernán, M.A., and Robins, J.M. (2020). *Causal Inference: What If?* New York: Chapman & Hall.
Hernán, M.A. and VanderWeele, T.J. (2011). Compound treatments and transportability of causal inference. *Epidemiology*, 22, 368–77.
Hill, A.B. (1965). The environment and disease: association or causation? *Proceedings of the Royal Society of Medicine*, 58, 295–300.
Höfler, M. (2005). The Bradford Hill considerations on causality: a counterfactual perspective. *Emerging Themes in Epidemiology*, 2(1), 1–9.
Höfler, M. (2006). Getting causal considerations back on the right track. *Emerging Themes in Epidemiology*, 3(1), 1–3.
Holland, P.W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81, 945–60.
Holman, C.D.J., Arnold-Reed, D.E., de Klerk, N., McComb, C., and English, D.R. (2001). A psychometric experiment in causal inference to estimate evidential weights used by epidemiologists. *Epidemiology*, 12, 246–50.
Hume, D. (1739). *A Treatise of Human Nature*. Selby-Bigge L.A. (ed.) (1888). Oxford: Oxford University Press.
Kalbfleisch, J.D. and Prentice, R.L. (2002). *The Statistical Analysis of Failure-Time Data* (2nd ed.). New York: Wiley.
Kaufman, J.S. (2008). Epidemiologic analysis of racial/ethnic disparities: some fundamental issues and a cautionary example (with discussion). *Social Science & Medicine*, 66, 1659–80.
Kaufman, J.S. and Poole, C. (2000). Looking back on causal thinking in the health sciences. *Annual Review of Public Health*, 21, 101–19.
Kuhn, T.S. (1970). *The Structure of Scientific Revolutions* (2nd ed.). Chicago, IL: University of Chicago Press.
Lanes, S.F. and Poole, C. (1984). 'Truth in packaging?' The unwrapping of epidemiologic research. *Journal of Occupational Medicine*, 26, 571–4.
Lash, T.L., Fox, M.P., and Fink, A.K. (2009). *Applying Quantitative Bias Analysis to Epidemiologic Data*. Boston, MA: Springer Publishing Company.
Lipsitch, M., Tchetgen, E.J., and Cohen, T. (2010). Negative controls: a tool for detecting confounding and bias in observational studies. *Epidemiology*, 21, 383–8.

Maclure, M. (1985). Popperian refutation in epidemiology. *American Journal of Epidemiology*, 121, 343–50.
McShane, B.B., Gal, D., Gelman, A., Robert, C., and Tackett, J.L. (2019). Abandon statistical significance. *The American Statistician*, 73(suppl 1), 235–45.
Mill, J.S.A. (1862). *System of Logic, Ratiocinative and Inductive* (5th ed.). London: Parker, Son and Bowin. (Cited in: Clark, D.W., and MacMahon, B. (eds.) (1981). *Preventive and Community Medicine* (2nd ed.). Boston, MA: Little, Brown.
Morgan, S.L. and Winship, C. (2007). *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. New York: Cambridge University Press.
Pearl, J. (2009). *Causality: Models, Reasoning and Inference* (2nd ed.). Cambridge: Cambridge University Press.
Pearl, J. and Mackenzie, D. (2018). *The Book of Why: The New Science of Cause and Effect*. New York: Basic Books.
Phillips, C.V. and Goodman, K.J. (2004). The missed lessons of Sir Austin Bradford Hill. *Epidemiologic Perspectives and Innovation*, 1, 3.
Phillips, C.V. and Goodman, K.J. (2006). Causal criteria and counterfactuals; nothing more (or less) than scientific common sense. *Emerging Themes in Epidemiology*, 3, 5.
Poole, C. (2001). Causal values. *Epidemiology*, 12, 139–41.
Popper, K.R. (1968). *The Logic of Scientific Discovery*. London/New York: Hutchinson.
Rafi, Z. and Greenland, S. (2020). Semantic and cognitive tools to aid statistical science: replace confidence and significance by compatibility and surprise. *BMC Research Methodology*. [Online]. Available at: https://arxiv.org/pdf/1909.08579.pdf
Robins, J.M. and Greenland, S. (2000). Comment on 'Causal inference without counterfactuals' by A.P. Dawid. *Journal of the American Statistical Association*, 95, 477–82.
Rosenbaum, P.R. (2002) *Observational Studies* (2nd ed.). New York: Springer.
Rosenbaum, P.R. (2003). Does a dose–response relationship reduce sensitivity to hidden bias? *Biostatistics*, 4, 1–10.
Rothman, K.J., Greenland, S., and Lash, T.L. (2008). *Modern Epidemiology* (3rd ed.). Philadelphia, PA: Lippincott Williams & Wilkins.
Susser, M. (1991). What is a cause and how do we know one? A grammar for pragmatic epidemiology. *American Journal of Epidemiology*, 133, 635–48.
United States Department of Health, Education and Welfare (1964). *Smoking and Health: Report of the Advisory Committee to the Surgeon General of the Public Health Service.* PHS Publ No. 1103. Washington, DC: Government Printing Office.
VanderWeele, T.J. (2015). *Explanation in Causal Inference: Methods for Mediation and Interaction*. Oxford: Oxford University Press.
VanderWeele, T.J. and Arah, O.A. (2011). Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments and confounders. *Epidemiology*, 22, 42–52.
VanderWeele, T.J. and Hernán, M.A. (2012). Causal effects and natural laws: towards a conceptualization of causal counterfactuals for non-manipulable exposures with application to the effects of race and sex. In: Berzuini, C., Dawid, A.P., and Bernardinelli, L. (eds.) *Causal Inference: Statistical Perspectives and Applications*. Chichester: John Wiley and Sons, pp. 101–13.
VanderWeele, T.J. and Hernán, M.A. (2013). Causal inference under multiple versions of treatment. *Journal of Causal Inference*, 1(1), 1–20.
Wald, N.A. (1985). Smoking. In: Vessey, M.P. and Gray, M. (eds.). *Cancer Risks and Prevention*. New York: Oxford University Press.
Wasserstein, R.L. and Lazar, N.A. (2016). The ASA statement on p-values: context, process, and purpose. *The American Statistician*, 70(2), 129–33.

Wasserstein, R.L., Schirm, A.L., and Lazar, N.A. (2019). Moving to a world beyond ‘p< 0.05’. *The American Statistician*, 73(Suppl 1), 1–19.
Weed, D.L. (1986). On the logic of causal inference. *American Journal of Epidemiology*, 123, 965–79.
Weed, D.L. and Gorelic, L.S. (1996). The practice of causal inference in cancer epidemiology. *Cancer Epidemiology, Biomarkers & Prevention*, 5, 303–11.
Weiss, N. (1981). Inferring causal relationships: elaboration of the criterion of ‘dose–response’. *American Journal of Epidemiology*, 113, 487–90.
Weiss, N.S. (2002). Can the specificity of an association be rehabilitated as a basis for supporting a causal hypothesis? *Epidemiology*, 13, 6–8.
Westreich, D. and Greenland, S. (2013). The table 2 fallacy: presenting and interpreting confounder and modifier coefficients. *American Journal of Epidemiology*, 177(4), 292–8.
Westreich, D., Edwards, J.K., Lesko, C.R., Stuart, E., and Cole, S.R. (2017). Transportability of trial results using inverse odds of sampling weights. *American Journal of Epidemiology*, 186(8), 1010–14.

5.13

Systematic reviews and meta-analysis

Nandi Siegfried and Lawrence Mbuagbaw

Introduction

Systematic reviews use systematic, explicit, and reproducible methods to identify relevant studies, assess the risk of bias in these studies, extract information, and synthesize findings. Systematic reviews often employ statistical methods (known as meta-analysis) to combine data from similar studies in order to obtain more precise estimates than those derived from the individual studies included in the review.

Systematic reviews play a critical role in guiding practitioners, policymakers, and consumers to make well-informed decisions and choices about healthcare. The massive expansion in the volume of biomedical literature, beginning in the middle of the twentieth century, has made it increasingly difficult for healthcare decision-makers to keep up with the findings of individual studies. Reviews of the literature are thus important for taking stock of existing knowledge and making informed choices about healthcare and resource allocation (Chalmers et al. 2014). Poorly conducted reviews drawing incorrect conclusions from the literature have the potential to cause serious harm to patients and populations, and waste resources (Chalmers and Glasziou 2009). This is especially important in situations where health problems are severe and the scarcity of resources makes it vital that they are not wasted (Chinnock et al. 2005).

In this chapter we focus our discussions on the methodology of systematic reviews to evaluate the effectiveness of healthcare interventions. Reviews of effectiveness preferentially include randomized controlled trials (RCTs) and as such, most of our descriptions and examples relate to systematic reviews of RCTs. The utility of the systematic review method is increasingly applied to synthesize results from observational studies and more recently,

qualitative data. The specificities related to the methods required to conduct such reviews are not covered in any detail in this chapter, and readers interested in the broader applicability of reviews are advised to expand their reading beyond what is included here. In the following pages, we outline the key steps involved in planning and executing a typical systematic review.

What is a systematic review?

An up-to-date synthesis of the results from relevant primary studies should provide healthcare decision-makers with a strong foundation on which to base their policy and practice decisions. If due care is not paid to the methods used during the collection, identification, and analysis of results from individual studies, the final synthesis may be at risk of bias. Bias includes any distortions that occur during the investigation process and may lead to overestimating or underestimating the strength of the observed effect. The methods of a systematic review aim to reduce the risk of biases that may be present in a traditional literature review. The main biases that can affect a review are outlined next and the differences between systematic and traditional literature reviews are presented in Table 5.13.1.

Bias that may affect reviews

Selection bias

Selection bias may occur in a review when the collection of primary studies systematically favours the identification of some studies over others. The main types of selection bias encountered in reviews are publication and language bias.

There is robust research evidence to indicate that studies with positive results are more likely to be published than those with neutral or negative results (Hopewell et al. 2009). If unpublished, negative studies are not identified and included in a review, the synthesized results will be at risk of publication bias and may yield a false positive result (Hopewell et al. 2007). A well-conducted systematic review will attempt to seek out unpublished studies, within resource and time constraints, to reduce potential publication bias.

English is the dominant language of publication in the biomedical literature and many authors seek to publish their work first in English-language journals. There is evidence that trials published in English are more likely to have statistically significant, positive results than studies published in German-language journals (Egger et al. 1997). Failing to include studies which are written in languages other than English, be they published or not, may lead to language bias which may affect the final synthesized results of the review, although this has not been shown consistently (Morrison et al. 2012). A well-conducted systematic review will not limit the search to a single language and review authors should recognize the need to budget for translation of articles where necessary.

Indexing bias

Access to large volumes of literature via electronic databases has streamlined the conduct of reviews and greatly reduced the time required to conduct a systematic review of high quality. In order to accurately identify a primary study of a specific study design, the terms of the search strategy must match that of the database record. For example, for a search strategy to find a randomized controlled trial in a database, the original record must have been clearly identified or indexed as such when the record was created. An initiative to increase the precision of searches for RCTs in MEDLINE was initiated by the Cochrane Collaboration in the 1990s leading to greater accuracy when searching for RCTs in MEDLINE and Embase (Glanville et al. 2006). Systematic review team authors should note that indexing bias is likely to remain a concern when searching across other databases and when searching for observational studies and other non-randomized designs.

Information bias

A review is dependent on the quality of the reporting of the included studies. Where reporting of a study is poor or data is lacking, a review may be at risk of information bias. During the conduct of a systematic review, the author team is encouraged to actively seek out the missing data by contacting the authors of the primary study (Young and Hopewell 2011). Where this is not possible, the systematic review team will be required to transparently report any assumptions made about the missing data and should indicate to the reader the possible effects of the missing data on the synthesized results.

Table 5.13.1 A comparison of the typical characteristics of systematic versus more traditional literature reviews

Systematic review	Literature review
States the methods of the review in a protocol	Often does not state the methods of the review
Answers a focused question or tests a hypothesis; usually narrow in scope	May have a clear question but it often involves general discussion with no stated hypothesis; often very broad in scope
Attempts to locate all relevant studies to minimize publication and other biases	Usually does not attempt to locate all relevant studies
Uses explicit eligibility criteria to determine which studies will be included in the review to prevent selection bias on behalf of the reviewer	Usually does not specify why some studies are included and others excluded
Systematically appraises the methods of the included studies to determine possibility of bias in these studies and sources of heterogeneity between the study results	Often does not assess the risk of bias in the studies or acknowledge differences in their methods
Bases results on studies that are methodologically sound	Often does not distinguish between methodologically sound and unsound studies
May combine study results statistically in a meta-analysis	Often does not pool results in a meta-analysis
Results illustrate level of heterogeneity of data, allowing quantification of uncertainty	Results often presented in certain terms without indicating uncertainty or variability

Developing a protocol

A systematic review is classified as secondary research, but it is nevertheless an original study which collects, appraises, and combines results from primary studies. For this reason, a research protocol is developed prior to the conduct of the review to ensure that the study objectives and methods are stated explicitly and to provide a template for the team to follow.

The protocol should contain the following sections:

Background to the clinical problem and intervention under study
Formulation of the review question (using the PICO structure, see next)
Inclusion and exclusion criteria
Search strategy
Methods of study selection, data extraction, and quality appraisal
Methods of analysis
Methods of overall quality assessment (e.g. Grading of Recommendations Assessment, Development, and Evaluation, GRADE)

Ideally the protocol should undergo peer review to ensure that it is rigorous and meets acceptable standards.

Registration of a review on a publicly accessible platform is encouraged and is mandatory for some funders. This improves transparency and allows detection of any changes to the protocol that may have occurred during the process of the review. It may also reduce the likelihood of unnecessary duplication of effort (Straus and Moher 2010). Review protocols can be registered publicly on the freely available platform, PROSPERO (https://www.crd.york.ac.uk/prospero/). PROSPERO includes protocol details for systematic reviews relevant to health and social care, welfare, public health, education, crime, justice, and international development, where there is a health-related outcome. Systematic review protocols on PROSPERO can include any type of any study design. Reviews of reviews and reviews of methodological issues that contain at least one outcome of direct patient or clinical relevance are also accepted.

Key point: The methods of a good systematic review should be described in a protocol which is registered publicly in advance of the conduct of the review.

Formulating the review question

The first step in conducting a systematic review is to clearly formulate the research question. A focused review question guides the selection of terms to include in the search strategy and narrows the scope of the inclusion criteria. The PICO-ST structure provides a useful framework for developing the review question:

P Populations
I Intervention
C Comparison
O Outcome
S Study design
T Timeframe

The minimum structure of a PICO question is as follows:

> [**Intervention**] compared with [**Comparison**] for preventing/treating [**Outcome**] in [**Populations**]

The study design (S) and timeframe (T) are optional additional components to the PICO. These allow the reader to immediately discern the design(s) of the studies included in the review (e.g. if the included studies are limited to RCTs). The timeframe component refers either to the time-point at which the primary outcome is measured, e.g. death at 1-year follow-up, or the time span of the studies included in the review, e.g. RCTs published between 1990 and 2010.

The O in the PICO refers to the outcome designated as the primary outcome(s) in the review question. The review team should carefully consider which outcome is the most meaningful to measure for answering the clinical question and denote this as the primary outcome. Note that this may differ from the outcome(s) designated as primary outcomes across the included studies.

Key point: The systematic review question is best formulated using the PICO structure.

Searching for studies

The search strategy forms the foundation of a systematic review and is best prepared by a trained librarian or information specialist. A comprehensive search should be highly sensitive, aiming to identify and locate as many of the eligible studies as is feasible within the timeframe and available resources. A search should also be precise in order to reduce the need to screen vast quantities of irrelevant material. Precision is a measure of the number of records that require screening to identify a single eligible study.

In practice the balance between sensitivity and precision is weighted towards high sensitivity to reduce selection bias (including language and publication bias), with low precision. Systematic review teams should generally expect to screen large numbers of records even if they will only a few studies.

Data sources to search

Published reports

An experienced information specialist will develop a search strategy using database-specific syntax including the keywords related to the review PICO. Examples of databases are listed next. At a minimum a well-conducted review of the effects of healthcare interventions should include the first three databases.

General electronic journal databases of published reports:

- European and North American focus:
  - MEDLINE (http://www.ncbi.nlm.nih.gov/pubmed)
  - EMBASE (http://www.elsevier.com/online-tools/embase)
  - CENTRAL (Cochrane Central Register of Controlled Trials) (http://www.cochranelibrary.com)
- South American focus
  - LILACS (Latin American Caribbean Health Sciences Literature) (http://lilacs.bvsalud.org/en/)
- African focus
  - African Journals Online (https://www.ajol.info/)

Subject-specific electronic journal databases:

- Databases will be dependent on the review topic, e.g. a review of behavioural interventions may include PsycINFO (https://www.apa.org/pubs/databases/psycinfo/index.aspx)

Unpublished reports

The so-called grey literature, such as conference abstracts, research reports, policy documents, dissertations, book chapters, and personal

correspondence, should be sought by contacting authors of identified papers, experts in the field, and relevant organizations, including pharmaceutical and medical device companies, and by searching pertinent websites. These sources should be identified at the protocol stage.

Ongoing studies

There are several prospective trials registries available which publish trial protocols and status updates of trial progress. These provide a useful source to identify ongoing studies and studies which are recently completed but not yet published. The choice of registry will be dependent on the topic.

Global focus (supported by World Health Organization)

- http://apps.who.int/trialsearch/AdvSearch.aspx

American and global focus

- https://www.clinicaltrials.gov

African focus

- http://www.pactr.org

Australasian focus

- http://www.anzctr.org.au/

Within the field of observational studies, there is increasing recognition that prospective registries are important, and several small registries do exist. Review teams are encouraged to seek these out in their chosen field should they wish to include observational studies.

Methods to screen for studies

At least two members of the review team should independently screen the titles and abstracts of the records yielded by the search and select all potentially eligible studies. The full-text article or report should then be obtained for each agreed potentially eligible study. Where disagreement or uncertainty occurs, it is advisable to adopt a conservative approach and obtain the full-text report.

Two investigators then apply an eligibility rule based on the PICO to each full-text manuscript and determine if the inclusion criteria are met. A third member of the team can act as arbiter where there is disagreement or uncertainty. The number of records at each stage of the screening, eligibility assessment, and inclusion process must be recorded in order to populate the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flow diagram in the final report (see section 'Reporting a systematic review', later in this chapter). Reasons for excluding studies after reading the full text should also be recorded and provided in the final report.

It is worth noting that a single study may be reported in more than one article or conference proceeding. For example, a trial protocol and a paper outlining the preliminary results of a trial may be published or presented at a conference prior to the publication of a final report of the trial results. This is viewed as legitimate publication practice. However, sometimes results are published in different articles with no cross-referencing to other relevant reports and can lead to double counting of results in meta-analyses (Abraham 2000). This phenomenon has been labelled 'salami publication' and is to be discouraged. The final number of included studies should therefore refer to the number of discrete studies and not the number of articles or reports identified.

Key point: The search strategy of a systematic review requires the expertise of an information specialist and should be sufficiently comprehensive to ensure that as many eligible studies as possible are included in the review.

Data extraction

What data to collect

The specific items of data to be collected from included studies will be influenced by the objectives of the systematic review. In general the following items should be collected:

Eligibility of the study for inclusion or reason for exclusion.
Study methods, such as study design and duration, as well as items that will be used to assess the risk of bias, such as sequence generation, allocation concealment, and blinding.
Data regarding participants (e.g. sex, age, ethnicity, and education level) and setting (e.g. institution and geographic location) which may influence the size of the effect estimate or assist in assessing the applicability of results.
The number of intervention groups and the details of the specific interventions involved. Any information relating to the fidelity with which the intervention was administered (i.e. the extent to which it was implemented as planned) should be obtained. This is particularly important in preventive or complex interventions, which are usually implemented under ‘real-world’ conditions.
Information regarding the outcomes such as their definitions, time points at which they were measured, and units of measurement should be collected. Usually, this is done only for those outcomes prespecified in the review protocol. However, listing all other outcomes reported in the included study may enable analysis of risk of bias due to selective outcome reporting.
Results should be collected only for those outcomes prespecified in the protocol. For a meta-analysis it is also necessary to collect the numerical data for analysis.
Other information to be collected could include funding source, details of ethics approval, name or contact details of authors, and citations.

How to collect the data

Once the data to be collected have been determined, it is necessary to design appropriate data collection forms. These may be electronic or paper-based, depending on the preference of reviewers.

Data extraction refers to the process of recording all the relevant data from the included studies onto data collection forms. Piloting data collection forms using a representative sample of studies is recommended. To reduce the risk of errors during data extraction, it is advisable that at least two reviewers, working independently, are involved in extracting the data. Any disagreements between the reviewers should be resolved through consensus or arbitration by a third person.

Key point: Selection of what data to extract is guided by the objectives of the review question. Two independent investigators should extract data in duplicate to avoid errors.

Evaluation of the risk of bias

Interpretation of the results of any type of study is dependent on how well the study was conducted. For example, a randomized controlled trial may yield a significant beneficial result, but if a large proportion of the trial participants failed to complete the trial and outcomes were not available for those lost to follow-up, we may be less likely to believe the result. Similarly, if results from several poorly conducted studies are synthesized in a systematic review, we should be concerned about the effect of the potential risk of bias on the pooled results. An evaluation of the risk of bias present in individual included studies is therefore a key part of the systematic review method.

Risk of bias is assessed based on the information in the reports of studies. Sometimes, the reported information may be insufficient to make critical judgements about the methods of a study. In these cases, it is important to contact the study authors to clarify any issues regarding study design or missing data that are not described clearly in the report.

There are several methods for assessing the risk of bias related to study design as outlined next.

Randomized controlled trials

For RCTs, the Risk of Bias domain-based evaluation tool developed by the Cochrane Collaboration is in widespread use (Higgins et al. 2011). The tool requires a judgement regarding the level of risk of bias (low, high, or unclear) for each of the following six domains:

Sequence generation
Allocation concealment
Blinding or masking of participants, providers, and assessors
Size of, and differential loss, to follow-up
Selective outcome reporting
Other biases such as publication bias

In the past, quality scales which provided an overall score or quality rating were in widespread use. The limitations of such scales include unreliable assessments of validity and difficulty in assigning weights to items in the scale. We mention the use of scales only to provide a complete picture for readers, but would discourage their use.

Non-randomized and observational studies

For observational studies of interventions, the Cochrane ROBINS-I tool (Risk Of Bias In Non-randomized Studies—of Interventions) is available for the appraisal of risk of bias in estimates of the effectiveness or safety of an intervention from studies that did not use randomization to allocate interventions (Sterne et al. 2016). Studies may include observational studies such as cohort studies and case-control studies in which intervention groups are allocated during the course of usual treatment decisions, and so-called quasi-randomized studies where allocation is not fully randomized. It addresses bias due to confounding, selection of participants, misclassification of the intervention, deviations from the intervention, missing data, measurement of the outcomes, and selection of the reported result. Assessment of bias can be rated as low, moderate, serious, or critical.

The Newcastle-Ottawa Scale (NOS) is another popular tool which can be used for appraising longitudinal and cross-sectional studies (Wells et al. 2018). Risk of bias is assessed in three domains: selection of participants, comparability of the cohorts, and outcomes. The scale is based on a star system and a total of nine stars can be allocated to a study with no concerns.

Key point: The risk of bias is essential to evaluate in the studies included in a systematic review to indicate the robustness of the results. Tools should employ a domain approach and scores from scales are to be avoided.

Data synthesis

A good systematic review includes a detailed synthesis of the results from included studies. Synthesis may be narrative, meta-analytical, or a combination of both. The protocol of the review must identify a priori the threshold(s) for determining if a meta-analysis should be undertaken. In addition, the team should indicate in the protocol if there are important

subgroups expected between studies and within the data which warrant separate analyses.

Meta-analysis

Meta-analysis is the statistical pooling of estimates of effect from individual studies to create an overall estimate of effect from the combined studies.

There are several benefits of doing a meta-analysis if it is appropriate so to do. It provides a more precise estimate of the treatment effect, helping one to be more confident about the size of any effect. There is more statistical power to detect small effects, which may be clinically significant; and it helps in the evaluation of the applicability of the results to other settings.

Meta-analysis should only be conducted when the included studies are considered sufficiently similar (or homogenous) to combine together. This is a matter of judgement and systematic review teams familiar with the data should ask the following:

Are the interventions, populations, and outcomes similar across studies or do they differ markedly? (Clinical heterogeneity)
Are the methods of the studies similar or different across studies? (Methodological heterogeneity)
Is there a small or large degree of variation between the results of the individual studies? (Statistical heterogeneity)

If the studies are indeed considered sufficiently similar both methodologically and clinically, then further testing for the presence of statistical heterogeneity can be conducted during the meta-analytical procedure to assess the degree of statistical heterogeneity present. Statistical heterogeneity is quantified with the I^2 statistic, which ranges from 0 to 100% (Higgins et al. 2003). The I^2 statistic represents the proportion of variability between the studies that cannot be explained by chance. Higher values of I^2 indicate more statistical heterogeneity.

Meta-analyses would require the use of either the fixed effects or the random effects model, each with its own assumptions and approaches to weighting studies and generating confidence intervals.

Measures of effect

The decision on which effect measure to use should, ideally, be taken when the systematic review is being planned. It should be based on the mathematical appropriateness of the effect measure for the type of data to be analysed. Review teams should select effect measures that are meaningful to clinicians and patients and are commonly used. For example, in a meta-analysis of studies of people living with HIV, the dichotomous outcome measure of undetectable viral load is a preferred measure to change in viral load, which requires log transformation and may not translate into a clinical benefit.

The team should avoid conducting multiple exploratory analyses using different effect measures and then focusing on the analysis that generates their preferred result (Bender et al. 2008). However, in rare instances changing the effect measure can reduce some heterogeneity.

For effectiveness data, combined estimates of dichotomous data are generally expressed as relative risks or odds ratios with associated 95% confidence intervals. The risk difference is an alternative which might be more appealing to a lay audience.

Combined estimates of continuous data are generally expressed as mean differences and associated confidence intervals. Where different scales are used to measure outcomes across studies, the means and standard deviations from the different scales can be combined into a standardized mean difference (SMD) and associated confidence interval. However, the latter can be difficult to interpret clinically and review teams should endeavour to provide the reader with some indication of the clinical significance of the SMD.

Time-to-event data (hazard ratios) can also be pooled using the generic inverse variance (GIV) method which accommodates other forms of non-standard data. The GIV uses the log transformed versions of the data to place them on a similar and symmetric scale.

At times, data may not be usable in the reported format and it may be necessary to contact the authors of a study to obtain disaggregated data, or data may require transformation (for example, using logs, estimating means from medians and confidence intervals from p-values) to facilitate meta-analysis. In such cases, transparent reporting of changes to the planned analysis is required in the final report.

Subgroup analysis

When statistical heterogeneity is present, this should be interrogated further according to the possible subgroups identified in the protocol. Conducting a subgroup analysis may reveal true differences in the data and provide an explanation for the observed heterogeneity. If no explanation is forthcoming, caution is required in interpreting the pooled results and a meta-analysis may not be appropriate.

Even if there is an a priori reason to expect a subgroup analysis to show something different to the overall result, this is no guarantee that a statistically significant difference is reliable clinically. This is because the more analyses are done, the more likely it is that statistically significant results will be found, even when there is truly no difference between the subgroups. Subgroup analyses in a systematic review should be regarded as a way of showing that the direction of effect is the same across different types of patient or as a generator of a hypothesis for testing in future research. The results of subgroup analyses should be treated with caution because the effect of chance will be larger for the subgroup than for the overall result.

It is recommended that the sources of clinical heterogeneity be determined a priori, based on clinical expertise and scientific rationale, and only focusing on variables related to the research question (participants, interventions, comparisons, outcomes). There is extensive guidance on how to conduct and interpret subgroup analyses (Higgins et al. 2002; Gagnier et al. 2013; Sun et al. 2014). Most statistical packages incorporate a formal test for subgroup effects or test of interaction. This test tells you how likely it is that the difference in effect estimates from the subgroups is due to chance.

Sensitivity analysis

Like all statistical analyses, meta-analysis is dependent on underlying assumptions. Once the decision to conduct a meta-analysis has been made, investigators select a statistical model to underpin the analysis. Both the random and fixed effects models are widely used, with the use of a Bayesian approach also recommended for addressing heterogeneity and missing data (Higgins et al. 2009; Moreno et al. 2018). At the protocol stage, reviewers should identify the preferred model and their reasons for its selection.

At analysis stage, a sensitivity analysis can be conducted comparing the results obtained from different models. Consistent results between models provide greater confidence in the overall pooled result; where results differ

between models, investigators need to consider the underlying reasons for this and exhibit caution in interpreting the final result.

Sensitivity analyses can also be conducted on other parameters, such as study quality. A meta-analysis of all eligible studies can be compared with a meta-analysis limited to those studies which are rated at low risk of bias. Again, investigators should think carefully about the interpretation of inconsistent results and how the inclusion of studies at high risk of bias affects the results.

Unit of analysis errors

Trials can employ randomization at an individual, cluster, or crossover level. Combining studies with different levels of randomization in a meta-analysis can lead to unit of analysis errors and yield potentially spurious results. Analytical approaches for such data must account for the correlation between participants in the same cluster and for participants who contribute control and intervention data on crossover designs (Donner et al. 2001; Elbourne et al. 2002).

Meta-regression

Meta-regression is another way of exploring or explaining heterogeneity in meta-analyses (Thompson and Higgins 2002). It allows us to investigate the effects of more than one categorical or continuous variable on the outcome. It is an extension of simple regression in which the outcome variable is the measure of effect and the explanatory variables are the covariates that might influence the measure of effect. Meta-regression is not recommended when there are fewer than ten studies (Thompson and Higgins 2002). The findings from meta-regression are observational and should be treated as such.

Publication bias (Funnel plots)

Despite reviewers' efforts to conduct exhaustive and comprehensive searches of published and unpublished data, some relevant data sources may be missed (Sedgwick 2015). The presence of publication bias plays an important role in how we appraise the completeness of a body of evidence and its quality. The presence of publication bias can be investigated using a graphical test called a funnel plot, or be statistically tested using Egger's and Begg's significance tests (Thornton and Lee 2000).

The funnel plot is a plot of sample size against the effect size of the included studies. If publication bias is absent, the points on the plot would look like a funnel, with the smaller studies spread widely at the base of the funnel and the larger studies spread narrowly closer to the top. Asymmetry in the funnel plot is indicative of publication bias (Song et al. 2013) but observed asymmetry may be present for other reasons. Egger's and Beggs's significance tests investigate whether the study estimate is related to its size and are regression- and correlation-based, respectively (Thornton and Lee 2000).

In Figure 5.13.1 we illustrate a forest plot of invented data. This meta-analysis includes ten studies conducted in low-income and high-income settings comparing text messaging (SMS) and cash transfers to a text messaging alone on the outcome of treatment adherence.

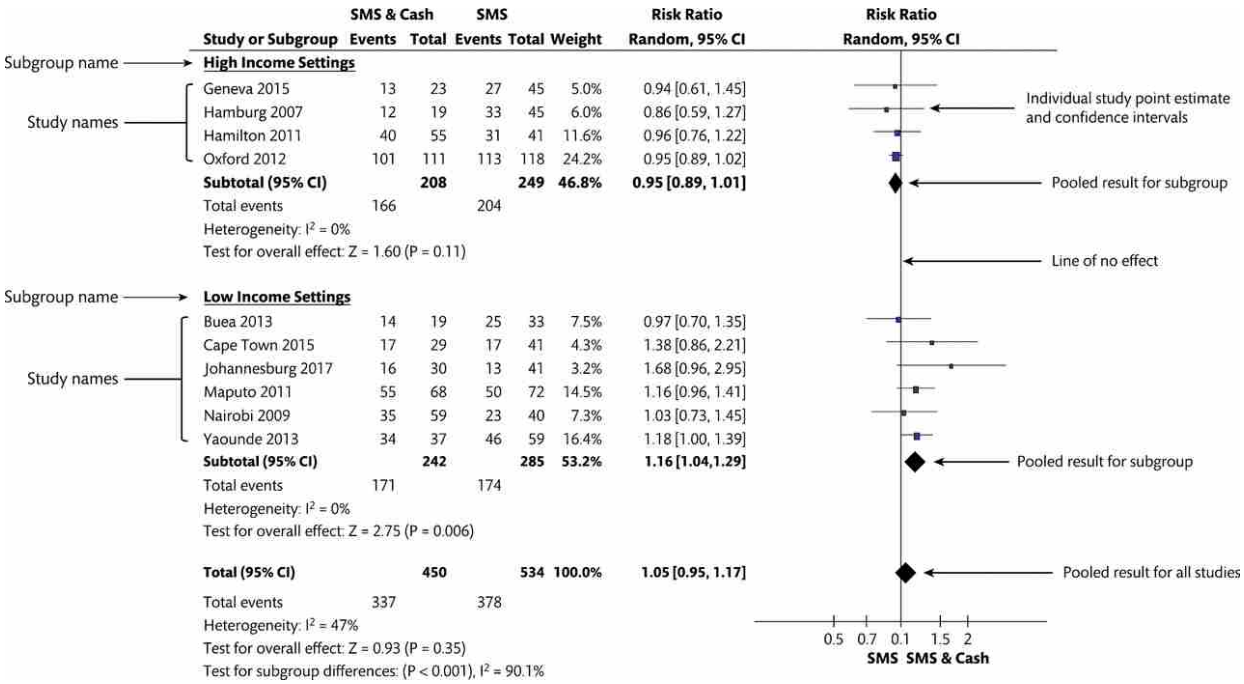


Fig. 5.13.1 Forest plot of invented data, from a meta-analysis of ten studies in high- and low-income settings, comparing text messaging and cash transfers to text messaging alone on the outcome of virologic success.

The outcome is dichotomous and measured as a relative risk. The data are pooled using the random effects approach and 95% confidence intervals are reported.

The vertical line is the line of no difference or a relative risk = 1.

A subgroup analyses is conducted to compare the relative merits of SMS & cash to the SMS alone, separately in the high-income and low-income settings.

The squares represent the point estimates and the horizontal line going across them represents the breadth of the confidence interval for each estimate.

The size of each square is proportional to the weight allocated to the study in the meta-analysis. The weight is derived from the inverse of the variance. In other words, the less precise, smaller studies with wider confidence intervals would have smaller squares.

The diamond represents the pooled estimate. For a positive outcome, like adherence, a diamond to the right of the line of no effect would indicate that the outcome is more likely to occur in the participants who receive SMS and Cash.

The amount of heterogeneity between all ten studies is reported as 47%; between the subgroups as 90.1%; in the high-income subgroup 0% and in the low-income subgroup 0%. The test for subgroup differences is statistically significant ($p<0.001$).

Narrative syntheses

When a meta-analysis is judged to be inappropriate due to significant heterogeneity, authors are advised to present a comprehensive narrative synthesis. Presentation of the study characteristics and the data should be given the same consideration as if a numerical synthesis were conducted. Note that a narrative synthesis is not a synonym for a ‘traditional literature review’. A narrative synthesis refers to the reporting of results in a systematic review where a meta-analysis is not appropriate or possible, and employs the same rigorous systematic review methods as a review which includes a meta-analysis (Popay et al. 2006).

Methodological and clinical heterogeneity should be reported with risk of bias and consistency in direction of estimates described. It is important to recognize that formulation of a single overall statement of the direction of benefit (or harm, or no effect) may not be possible, and caution is advised against counting studies with significant findings versus those without. Narrative syntheses contribute to the delineation of the current evidence base

and can inform and influence primary studies to provide more standardized reporting of outcomes to allow for future meta-analysis.

Key point: Meta-analysis should only be conducted when included studies are judged to be sufficiently similar, both methodologically and clinically. During meta-analysis, tests for statistical heterogeneity indicate whether the pooled results are statistically homogenous. If not, consider whether subgroup differences identified a priori explain the observed heterogeneity. Where meta-analysis is not appropriate, a detailed narrative synthesis should be reported.

Summarizing the overall quality of the evidence

Once the results of a systematic review are available, decision-makers need to know how much confidence to place in the evidence arising from the review. This need to integrate the review results with a quality measure led to the development of several methods to present review evidence graded by a measure of strength. Current hierarchical evidence rating systems include the A to D approach used by the United States Preventive Services Task Force (U.S. Preventive Services Task Force), the Oxford Centre for Evidence-based Medicine Levels 1 to 5 (Centre for Evidence-based Medicine 2016), and the GRADE approach (Guyatt et al. 2011a). We report here in detail on the GRADE method for interventions.

GRADE (Grading of Recommendations Assessment, Development, and Evaluation) provides a method to classify the overall quality of evidence arising from a systematic review as *high*, *moderate*, *low*, or *very low* (Guyatt et al. 2011a). Reviews which include several RCTs are initially graded as *high* quality given that the randomized design provides the highest level of evidence to support effectiveness. This level of quality may then be downgraded dependent on limitations in the included studies which include the presence of risk of bias, inconsistency, indirectness, imprecision, and/or publication bias. Down-grading by a single level takes place when the limitations are judged as ‘serious’, and by two levels when judged as ‘very serious’. On the other hand, a review of observational studies is initially graded as providing *low* quality evidence and can in turn be downgraded for the same reasons as RCTs, or upgraded if circumstances permit (Guyatt et al. 2011e).

Results for a maximum of seven critical outcomes are presented in a GRADE Summary of Findings table. Critical outcomes refer to those outcomes which consumers of the review—those who will be directly impacted by the results—consider most important to their health. For example, trial investigators may measure a biomarker to evaluate efficacy of a medication, but patients may be more interested in the impact of potential adverse effects of the medication on their quality of life. The systematic review team may wish to consult with consumer associations to best reflect their needs in the selection of critical outcomes.

For each of the seven critical outcomes, the systematic review team will assess whether there is an overall risk of bias across the included studies. If a risk of bias is judged to be present, the overall quality will be downgraded (Guyatt et al. 2011f).

When grading indirectness, the team must judge whether or not the included studies match the PICO of the systematic review, either directly or indirectly. For example, evidence may need to be extrapolated from adult studies to a PICO focused on children, where no trials of children exist. In such a case, the results would be downgraded for indirectness (Guyatt et al. 2011c).

Results are downgraded for inconsistency where heterogeneity is present and it cannot be explained e.g. by subgroup differences. The presence of heterogeneity alone is insufficient to down-grade the results; it is the presence of unexplained heterogeneity that drives a decision to down-grade (Guyatt et al. 2011d).

When considering imprecision, authors need to consider the width of the confidence interval around the effect estimate and determine whether this is precise or not by considering the sample size and event rate (Guyatt et al. 2011b).

Note that judgement is required for all of the aforementioned domains, and the review team must record the rationale for each judgement in a footnote in the GRADE Summary of Findings table. This ensures transparency and allows the reader to understand the reasons for each decision.

Once all the domains are graded, an overall quality of evidence per outcome is generated. The levels of quality are interpreted as follows:

High: We are very confident that the true effect lies close to that of the estimate of the effect.
Moderate: We are moderately confident in the estimate of effect: the true effect is likely to be close to the estimate of effect, but possibly to be substantially different.
Low: Our confidence in the effect is limited: the true effect may be substantially different from the estimate of the effect.
Very low: We have very little confidence in the effect estimate: any estimate of effect is very uncertain.

It is important to recognize that the results for different outcomes may arise from pooling the results from different studies and as a result, the quality of evidence may differ across outcomes.

Key point: GRADE is a system of quality assessment which incorporates the risk of bias, indirectness, imprecision, and inconsistency to determine the overall quality of evidence arising in a systematic review. Quality is graded as high, moderate, low, or very low.

Reporting a systematic review

The PRISMA for reporting systematic reviews is the equivalent of CONSORT for reporting randomized trials (Moher et al. 2010) and was first published in 2009 (Liberati et al. 2009) building on the earlier QUOROM statement (Moher et al. 1999). PRISMA is an evidence-based minimum set of items for reporting in systematic reviews and meta-analyses. Although PRISMA focuses on the reporting of reviews evaluating randomized trials, it can be extended to reporting systematic reviews of other types of research, particularly evaluations of interventions.

PRISMA provides a checklist for authors to ensure that the following items are comprehensively and transparently reported:

Title
Abstract
Introduction
Methods
Results
Discussion
Funding

The PRISMA checklist is explained in a detailed Explanation and Elaboration document which has been published in several journals (Liberati et al. 2009). A mandatory component of reporting in a systematic review is the PRISMA flow diagram which documents the stages and resultant yield of records of the search strategy.

Several extensions to the PRISMA Statement have been developed and include PRISMA for Network Meta-analysis and PRISMA for Diagnostic Test Accuracy (see http://prisma-statement.org). Authors of systematic reviews are encouraged to report their methods and results according to the relevant PRISMA Statement.

Key point: Systematic reviews should be reported using the structure recommended in the PRISMA Statement.

Appraising the quality of a systematic review

A systematic review which conforms to the methods outlined in this chapter would be considered well-conducted. However, similar to the quality assessment of included studies, the quality of a systematic review can also be categorized according to whether or not a risk of bias is present. It is particularly helpful for healthcare decision-makers to know how much confidence to place in the results of a systematic review.

Several tools exist to assess the quality of the conduct of a systematic review. The AMSTAR-2 critical appraisal tool provides a 16-item checklist incorporating the key methodological components of a review (Shea et al. 2017). The AMSTAR-2 tool does not provide a score but a rating of confidence in the results of the review based on the presence of weaknesses in critical items. It can be applied to reviews of randomized and non-randomized studies of healthcare interventions.

The ROBIS tool was developed in 2016 to evaluate the risk of bias in reviews of most types of research questions including diagnosis, prognosis, and treatment (Whiting et al. 2016). It provides a similar judgement-based domain approach to the Cochrane Risk of Bias tool for trials. Four domains include multiple questions related to the methods and the appropriateness of the study eligibility criteria, identification, and selection of studies, data collection and study appraisal, and synthesis and findings. An overall risk of bias (high, low, or unclear) is then derived from an assessment of whether the concerns raised in the four domains are addressed in the interpretation of findings, combined with an evaluation of the relevance of the included studies to the review question, and the degree to which the reviewers avoid emphasizing statistical significance.

AMSTAR-2 and ROBIS are designed to provide a structured approach to evaluating risk of bias in systematic reviews which requires judgement. Many may find this difficult to do, and it is recommended that evaluation is completed independently by at least two investigators who have methodological and content expertise.

Key point: The quality of a systematic review can be evaluated using a tool such as AMSTAR-2 or ROBIS, to indicate the robustness of the synthesized results.

Related approaches to evidence synthesis

Scoping reviews

Scoping reviews are variably defined in the literature. In 2005, Arksey and O'Malley described four aims of scoping reviews viz. (1) to examine the extent, range and nature of research activity; (2) to determine the value of undertaking a systematic review; (3) to summarize and disseminate research findings; and (4) to identify research gaps in the literature (Arksey and OMalley 2005). A scoping review will generally have a broad focus compared with a systematic review which adopts a narrow focus defined by a PICO question. Given that a scoping review aims to 'map' the current research base on a topic, the type of included study will not be limited to RCTs but may include all forms of studies to the topic. Clearly, use of robust systematic methods would enhance the quality of a scoping review, but a scoping review may follow a more traditional review approach and

categorize the results by topic area rather than present a quantitative summary or an assessment of the study quality. A 2014 scoping review of 344 scoping reviews described variability in the methodological process, terminology, and reporting of scoping reviews. The authors identified a need for methodological standardization to increase the use and the relevance of findings derived from scoping reviews (Pham et al. 2014).

Overviews of reviews

An overview of reviews aims to identify and collate the results from relevant reviews on a specific topic (Thomson et al. 2010). An overview, also known as an umbrella review, may be conducted for several reasons. Firstly, when several systematic reviews on distinct interventions exist for the same condition, decision-makers may wish to summarize the evidence base on all relevant treatment modalities and present this in a single report. Secondly, when the number and focus of reviews on a relevant topic is not known, an overview can aid in identifying the scope of the current evidence which has undergone review.

In an overview, the review is the type of included study design. Note that an overview may include only systematic reviews, or may include both systematic and literature reviews. Regardless of the type of included review, the methods of an overview mirror those of a systematic review with the exception that the risk of bias assessment uses the ROBIS (or similar) tool, instead of the ROB tool. There is currently no consensus regarding the application of GRADE or other ranking systems to produce an overall quality assessment in an overview.

Individual patient or participant data analysis

This is a form of meta-analysis in which trial investigators collaborate to share individual level data from their trials, instead of aggregated data, in order to ensure that the pooled results mimic a single large trial. This allows for standardization of analyses across studies independent of how the aggregated analysis was conducted (although clustering of each individual trial must be retained rather than treated as coming from one large trial) (Riley et al. 2010). Individual participant data analysis can be very resource-intensive and is dependent on close collaboration of trial investigators, but

arguably produces a more reliable estimate than a meta-analysis of aggregated data.

Prospective meta-analysis

There is debate in the literature as to whether a single large ‘mega-trial’ conducted in trial sites around the world following an identical protocol provides better evidence than several smaller regional trials combined in a meta-analysis (Shrier et al. 2007). Concerns have been raised about the potential of increasing error when pooling aggregated results from small underpowered trials even if precision is increased. One possible solution is prospective meta-analysis in which trials are identified, evaluated, and determined to be eligible before the results of any included studies are known or published, thereby avoiding some of the potential biases inherent in standard, retrospective meta-analyses (Askie et al. 2011). This methodology provides the same strengths as a single large-scale multicentre randomized study while allowing greater pragmatic flexibility. The combined meta-analytic results using individual patient data from each trial is published when all the trials are complete.

Network meta-analysis

Simple meta-analysis compares one intervention to another. This is a direct pairwise comparison. However, reviewers may be interested in comparing more than two treatment options. Trials may compare two treatments with each other or the treatments may have both been compared to placebo. In this instance a multiple treatment comparison, or a network meta-analysis (NMA), is possible. A NMA has the advantage of being able to provide information on the relative merits of interventions that have never been compared directly. A NMA can also increase the precision of effect estimates by using two sources of data: the direct pairwise comparisons and the indirect comparisons based on the comparisons to placebo (Mills et al. 2011; Mbuagbaw et al. 2017). Three key assumption underlie NMAs. The first is homogeneity, which simply refers to the absence of heterogeneity. The second is transitivity, which examines the validity of making indirect comparisons as all the control groups may not be identical. The third is consistency between the direct and indirect evidence (Mills et al. 2013). The

results may be presented graphically or as ranks that represent the probability of an intervention being first or second and so on. NMA is a fairly novel field and guidance on how to conduct, interpret, appraise the quality of the evidence and report NMAs is becoming more available.

Diagnostic test accuracy reviews

Reviewers may be interested in summarizing the accuracy of a diagnostic test. The searching and screening would be similar to a typical interventional systematic review, but in this case the data extracted would be the sensitivities and specificities of the diagnostic test in question as reported by the included studies (Leeflang 2014; Campbell et al. 2015). Many reviewers use the Quality Assessment of Diagnostic Accuracy Studies (QUADAS) tool to appraise the quality of the included studies. There are two important analytical considerations with diagnostic test accuracy (DTA) reviews. The first is the threshold effect. Tests based on continuous data (e.g. haemoglobin) require a threshold above (or below) which the test is positive. If different studies use different thresholds, this introduces heterogeneity in the review. The second issue is that sensitivity and specificity are correlated and therefore must be modelled together. The hierarchical summary ROC (HSROC) model and the bivariate model random effects model are used to address the first and second issues, respectively. Both models allow for covariate adjustment (Leeflang 2014).

Rapid reviews

No consistent definition for rapid reviews exists in the literature and methodologies differ across published rapid reviews (Tricco et al. 2015). However, the hallmark of a rapid review is acceleration of the systematic review process, usually by simplifying or leaving out methodological stages. This allows evidence to be quickly produced for time-sensitive decision-making. Conduct of a rapid review is dependent on a highly skilled review team and urgent provision of necessary resources. With the availability of electronic database searching, computerized study selection, and shared analytical tools, stages of a review can be sped up significantly. A rapid review need not, and arguably should not, compromise on rigorous methodology, but this remains a risk when urgency prevails. We suggest that

protocol templates for systematic reviews based on categories of interventions be developed in anticipation and utilized when a rapid review is required. Research to delineate the consequences of methodological shortcuts is warranted (Tricco et al. 2015).

Qualitative reviews and CERQual

Systematic review methodology has been successfully applied to identifying and synthesizing qualitative study data. Unlike an interventional review, the aims of a systematic review of qualitative data include providing greater in-depth understanding of an issue, informing theory development, and describing major themes related to a topic. CERQual is an approach to providing a confidence rating in the synthesized qualitative evidence, based on the GRADE approach to judging overall quality (Lewin et al. 2015). Methodological limitations, relevance, coherence, and adequacy of data are considered and an overall summary of the level of confidence regarding the representation of the phenomenon of interest is produced.

Use of systematic reviews in practice guidelines

Guidelines development should be based on evidence derived from systematic reviews. In 2007, the World Health Organization was criticised for failing to base global clinical recommendations on systematic review evidence and for an over-reliance on experts with limited methodological expertise (Oxman et al. 2007). In response, the World Health Organization (WHO) actively developed a rigorous approach to guidelines development based on the GRADE framework and outlined in the WHO Handbook for Guideline Development (WHO 2014). Evidence derived from systematic reviews forms the foundation of the GRADE framework for decision-making, which is further supplemented by values and preferences, and issues of acceptability, feasibility, resource use, and equity. The process is transparent and the quality of evidence underpinning a recommendation is clearly reported as part of the overall recommendation (Andrews et al. 2013).

Subsequent research to evaluate guidelines development within WHO has indicated that recommendations are increasingly based on systematic review evidence (Sinclair et al. 2013) but challenges remain. Active engagement

between guidelines developers and systematic reviewers is to be encouraged and a certain level of tension is anticipated. Critics have identified that at times strong recommendations have been inappropriately based on low quality evidence and advocate for further training in GRADE methods for guideline panellists (Alexander et al. 2014).

What does the future hold for systematic reviews?

Living systematic reviews and artificial intelligence

The concept of a 'living' systematic review grew out of the recognition that the value of a systematic review lies not only in its accuracy, but also in its currency. A living systematic review aims to continually update the evidence as new data becomes available (Elliott et al. 2017). It requires a pre-determined frequency of searching which is predictable, allowing for updating of guidelines in real time. New technologies to speed up the review process are available, for example, through machine-learning of screening for references for eligibility assessment. The accuracy of such screening classifiers can be improved by feedback models from human identification which can further be enhanced by using citizen science via crowd-funding initiatives (Thomas et al. 2017). The potential and limitations of harnessing automation to increase human responsiveness need to be carefully considered and tested to ensure the validity of results from living reviews, but offers great promise.

Mining big data

Big data in the form of biobank development integrated with clinical data (Bouzille et al. 2018) may offer robust longitudinal data which may in time provide a substitute for data derived from RCTs. The utility may be greatest for treatment of rare diseases. However, ethical issues regarding access to the data and consent remain a subject of ongoing debate (Lipworth et al. 2017) and statistical modelling will be required to determine the key criteria for drawing causality inferences from longitudinal data derived from such databanks.

Advancing systematic review methodology

Current efforts are underway to determine the most appropriate ways to conduct systematic reviews. In addition to independently-led methodological reviews of systematic review practices, the Cochrane Methods Groups (https://methods.cochrane.org/) and the Northern Ireland Network for Trials Methodology Research (https://www.qub.ac.uk/sites/TheNorthernIrelandNetworkforTrialsMethodologyResearch/) are actively engaged in this process.

Software and resources for systematic reviews

Several resources exist for the conduct of systematic reviews, some of which are commercial (not reported here) and some are not-for-profit. We present a selection which readers may find useful as starting points.

Cochrane

Cochrane (formerly, the Cochrane Collaboration) is an international voluntary organization which generates and compiles Cochrane systematic reviews, considered the gold standard of reviews. Several helpful resources are available for review authors:

The Cochrane Library

Cochrane reviews are published on an electronic database, *The Cochrane Library.* Many countries provide national subscriptions to medical professionals and the general public. See http://www.cochranelibrary.com/

The Cochrane Handbook for Systematic Reviews of Interventions

This handbook is a step-by-step guide to writing, analysing, and reporting a Cochrane review. See http://handbook-5-1.cochrane.org/

REVMAN

This software is free and provides a text-based review template and built-in meta-analytic software. It also contains links to the Cochrane Handbook. See

http://community.cochrane.org/help/tools-and-software/revman-5

The GRADE Working Group

The GRADE Working Group continues to develop the GRADE approach and software for reviews, health technology assessment and guidelines. See http://www.gradeworkinggroup.org/

GRADEPro

This software is free and produces Summary of Findings tables to present the overall quality of evidence based on meta-analysis. The software interfaces directly with REVMAN. See https://gradepro.org/

Rayyan QCI

Rayyan offers free support for reference management, screening, and storage. See: https://rayyan.qcri.org/welcome

Conclusions

In this chapter, we have provided the reader with a comprehensive approach to systematic review methodology, a guide to understanding and interpreting meta-analysis, and a synopsis of the many applications of systematic reviews and related syntheses. Although the science of systematic reviews is relatively new within epidemiology, the impact of systematic review and meta-analytic methodology has been considerable: evidence from reviews informs the direction of primary research, underpins guidelines, and provides patients, clinicians, and the general public with greater certainty regarding treatment and prevention choices. Yet it is important to remember that systematic reviews, like all studies, are subject to biases. The data generated from a review is secondary to that reported in the included studies and is thus totally dependent on the quality of the individual research studies used in the review. Users of evidence derived from systematic reviews should remember that critical appraisal remains an essential tool for the thoughtful researcher, clinician, and policymaker.

ACKNOWLEDGEMENTS

We are grateful to the original authors of the chapter, Jimmy Volmink and Mike Clarke, for their permission to replicate sections of the original chapter.

REFERENCES

Abraham, P. (2000). Duplicate and salami publications. *Journal of Postgraduate Medicine*, 46, 67–9.

Alexander, P.E., et al. (2014). World Health Organization recommendations are often strong based on low confidence in effect estimates. *Journal of Clinical Epidemiology*, 67, 629–34.

Andrews, J., et al. (2013). GRADE guidelines: 14. Going from evidence to recommendations: the significance and presentation of recommendations. *Journal of Clinical Epidemiology*, 66, 719–25.

Arksey, H. and Omalley, L. (2005). Scoping studies: towards a methodological framework. *International Journal of Social Research Methodology: Theory and Practice*, 8, 19–32.

Askie, L.M., Brocklehurst, P., Darlow, B.A., Finer, N., Schmidt, B., and Tarnow-Mordi, W. (2011). NeOProM: neonatal oxygenation prospective meta-analysis collaboration study protocol. *BMC Pediatrics*, 11, 6.

Bender, R., et al. (2008). Attention should be given to multiplicity issues in systematic reviews. *Journal of Clinical Epidemiology*, 61, 857–65.

Bouzille, G., et al. (2018). Integrating Biobank Data into a Clinical Data Research Network: the IBCB project. *Studies in Health Technology and Informatics*, 247, 16–20.

Campbell, J.M., et al. (2015). Diagnostic test accuracy: methods for systematic review and meta-analysis. *International Journal of Evidence-based Healthcare*, 13, 154–62.

Centre for Evidence-Based Medicine (2016). *OCEBM Levels of Evidence*. Oxford: Oxford Centre for Evidence-based Medicine. [Online] Available at: https://www.cebm.net/2016/05/ocebm-levels-of-evidence/

Chalmers, I., et al. (2014). How to increase value and reduce waste when research priorities are set. *The Lancet*, 383, 156–65.

Chalmers, I. and Glasziou, P. (2009). Avoidable waste in the production and reporting of research evidence. *The Lancet*, 374, 86–9.

Chinnock, P., Siegfried, N., and Clarke, M. (2005). Is evidence-based medicine relevant to the developing world? *PLoS Medicine*, 2, e107.

Donner, A., Piaggio, G., and Villar, J. (2001). Statistical methods for the meta-analysis of cluster randomization trials. *Statistical Methods in Medical Research*, 10, 325–38.

Egger, M., Zellweger-Zahner, T., Schneider, M., Junker, C., Lengeler, C., and Antes, G. (1997). Language bias in randomised controlled trials published in English and German. *Lancet*, 350, 326–9.

Elbourne, D.R., Altman, D.G., Higgins, J.P., Curtin, F., Worthington, H.V., and Vail, A. (2002). Meta-analyses involving cross-over trials: methodological issues. *International Journal of Epidemiology*, 31, 140–9.

Elliott, J.H., et al. (2017). Living systematic review: 1. Introduction-the why, what, when, and how. *Journal of Clinical Epidemiology*, 91, 23–30.

Gagnier, J.J., et al. (2013). Consensus-based recommendations for investigating clinical heterogeneity in systematic reviews. *BMC Medical Research Methodology*, 13, 106.

Glanville, J.M., Lefebvre, C., Miles, J.N., and Camosso-Stefinovic, J. (2006). How to identify randomized controlled trials in MEDLINE: ten years on. *Journal of the Medical Library Association*, 94, 130–6.

Guyatt, G., et al. (2011a). GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. *Journal of Clinical Epidemiology*, 64, 383–94.

Guyatt, G.H., et al. (2011b). GRADE guidelines 6. Rating the quality of evidence--imprecision. *Journal of Clinical Epidemiology*, 64, 1283–93.

Guyatt, G.H., et al. (2011c). GRADE guidelines: 8. Rating the quality of evidence--indirectness. *Journal of Clinical Epidemiology*, 64, 1303–10.

Guyatt, G. H., et al. (2011d). GRADE guidelines: 7. Rating the quality of evidence--inconsistency. *Journal of Clinical Epidemiology*, 64, 1294–302.

Guyatt, G.H., et al. (2011e). GRADE guidelines: 9. Rating up the quality of evidence. *Journal of Clinical Epidemiology*, 64, 1311–16.

Guyatt, G.H., et al. (2011f). GRADE guidelines: 4. Rating the quality of evidence--study limitations (risk of bias). *Journal of Clinical Epidemiology*, 64, 407–15.

Higgins, J., Thompson, S., Deeks, J., and Altman, D. (2002). Statistical heterogeneity in systematic reviews of clinical trials: a critical appraisal of guidelines and practice. *Journal of Health Services Research & Policy*, 7, 51–61.

Higgins, J.P., et al. (2011). The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *British Medical Journal*, 343, d5928.

Higgins, J.P., Thompson, S.G., Deeks, J.J., and Altman, D.G. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal*, 327, 557–60.

Higgins, J.P., Thompson, S.G., and Spiegelhalter, D.J. (2009). A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society Series A (Statistics in Society)*, 172, 137–59.

Hopewell, S., Loudon, K., Clarke, M.J., Oxman, A.D., and Dickersin, K. (2009). Publication bias in clinical trials due to statistical significance or direction of trial results. *Cochrane Database of Systematic Reviews*, (1), MR000006.

Hopewell, S., Mcdonald, S., Clarke, M., and Egger, M. (2007). Grey literature in meta-analyses of randomized trials of health care interventions. *Cochrane Database of Systematic Reviews*, (2), MR000010.

Leeflang, M.M. (2014). Systematic reviews and meta-analyses of diagnostic test accuracy. *Clinical Microbiology and Infection*, 20, 105–13.

Lewin, S., et al. (2015). Using qualitative evidence in decision making for health and social interventions: an approach to assess confidence in findings from qualitative evidence syntheses (GRADE-CERQual). *PLoS Med*, 12, e1001895.

Liberati, A., et al. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. *British Medical Journal*, 339, b2700.

Lipworth, W., Mason, P.H., and Kerridge, I. (2017). Ethics and epistemology of big data. *Journal of Bioethical Inquiry*, 14, 485–8.

Mbuagbaw, L., et al. (2017). Approaches to interpreting and choosing the best treatments in network meta-analyses. *Systematic Reviews*, 6, 79.

Mills, E.J., et al. (2011). Multiple treatment comparison meta-analyses: a step forward into complexity. *Clinical Epidemiology*, 3, 193–202.

Mills, E.J., Thorlund, K., and Ioannidis, J.P. (2013). Demystifying trial networks and network meta-analysis. *British Medical Journal*, 346, f2914.

Moher, D., Cook, D.J., Eastwood, S., Olkin, I., Rennie, D., and Stroup, D.F. (1999). Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement. Quality of Reporting of Meta-analyses. *The Lancet*, 354, 1896–900.

Moher, D., et al. (2010). CONSORT 2010 Explanation and Elaboration: updated guidelines for reporting parallel group randomised trials. *Journal of Clinical Epidemiology*, 63, e1–37.
Moreno, E., Vazquez-Polo, F.J., and Negrin, M.A. (2018). Bayesian meta-analysis: the role of the between-sample heterogeneity. *Statistical Methods in Medical Research*, 27(12), 3643–57.
Morrison, A., et al. (2012). The effect of English-language restriction on systematic review-based meta-analyses: a systematic review of empirical studies. *International Journal of Technology Assessment in Health Care*, 28, 138–44.
Oxman, A.D., Lavis, J.N., and Fretheim, A. (2007). Use of evidence in WHO recommendations. *The Lancet*, 369, 1883–9.
Pham, M.T., Rajic, A., Greig, J.D., Sargeant, J.M., Papadopoulos, A., and McEwen, S.A. (2014). A scoping review of scoping reviews: advancing the approach and enhancing the consistency. *Research Synthesis Methods*, 5, 371–85.
Popay, J., et al. (2006). *Guidance on the Conduct of Narrative Synthesis in Systematic Reviews. A Product from the ESRC Methods Programme.* [Online] Available at: http://www.lancaster.ac.uk/shm/research/nssr/research/dissemination/publications/NS_Synthesis_Guidance_v1.pdf
Riley, R.D., Lambert, P.C., and Abo-Zaid, G. (2010). Meta-analysis of individual participant data: rationale, conduct, and reporting. *British Medical Journal*, 340, c221.
Sedgwick, P. (2015). What is publication bias in a meta-analysis? *British Medical Journal*, 351, h4419.
Shea, B.J., et al. (2017). AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. *British Medical Journal*, 358, j4008.
Shrier, I., Platt, R.W., and Steele, R.J. (2007). Mega-trials vs. meta-analysis: precision vs. heterogeneity? *Contemporary Clinical Trials*, 28, 324–8.
Sinclair, D., Isba, R., Kredo, T., Zani, B., Smith, H., and Garner, P. (2013). World Health Organization guideline development: an evaluation. *PLoS One*, 8, e63715.
Song, F., Hooper, L., and Loke, Y. (2013). Publication bias: what is it? How do we measure it? How do we avoid it?. *Open Access Journal of Clinical Trials*, 5, 71–81.
Sterne, J.A., et al. (2016). ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *British Medical Journal*, 355, i4919.
Straus, S. and Moher, D. (2010). Registering systematic reviews. *Canadian Medical Association Journal*, 182(1), 13–14.
Sun, X., Ioannidis, J.P., Agoritsas, T., Alba, A.C., and Guyatt, G. (2014). How to use a subgroup analysis: users' guide to the medical literature. *The Journal of the American Medical Association*, 311, 405–11.
Thomas, J., et al. (2017). Living systematic reviews: 2. Combining human and machine effort. *Journal of Clinical Epidemiology*, 91, 31–17.
Thompson, S.G. and Higgins, J.P. (2002). How should meta-regression analyses be undertaken and interpreted? *Statistical Medicine*, 21, 1559–73.
Thomson, D., Russell, K., Becker, L., Klassen, T., and Hartling, L. (2010). The evolution of a new publication type: steps and challenges of producing overviews of reviews. *Research Synthesis Methods*, 1, 198–211.
Thornton, A. and Lee, P. (2000). Publication bias in meta-analysis: its causes and consequences. *Journal of Clinical Epidemiology*, 53, 207–16.
Tricco, A.C., et al. (2015). A scoping review of rapid review methods. *BMC Medicine*, 13, 224.
U.S. Preventive Services Task Force. *Grade Definitions*. [Online] Available at: https://www.uspreventiveservicestaskforce.org/Page/Name/grade-definitions#grade-definitions-after-july-2012
Wells, G., et al. (2018). *The Newcastle-Ottawa Scale (NOS) for Assessing the Quality of Nonrandomised Studies in Meta-analyses.* [Online] Available at:

http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp
Whiting, P., et al. (2016). ROBIS: a new tool to assess risk of bias in systematic reviews was developed. *Journal of Clinical Epidemiology*, 69, 225–34.
World Health Organization (WHO) (2014). *WHO Handbook for Guideline Development*. Geneva, Switzerland: World Health Organization.
Young, T. and Hopewell, S. (2011). Methods for obtaining unpublished data. *Cochrane Database of Systematic Reviews*, (11), MR000027.

5.14

Statistical methods

Gail Williams and Robert S. Ware

Introduction to statistical methods

Statistics is the study of mathematically based techniques used to collect, analyse, and interpret quantitative data. In public health, this occurs in a complex of biological, clinical, epidemiological, social, ecological, and administrative systems. Sometimes we want to describe, explore, or summarize data without extending our findings beyond the coverage of the data we have—that is, the sample. More usually, however, we want to extend, or generalize, our findings from the sample to a larger group, sometimes called the 'target' population. In this situation, we need to consider carefully two aspects of our data collection. Firstly, what was the actual source of our data, and how did we select the sample from the population? If we deliberately excluded some people in our sampling, or made it more difficult for some rather than others to be included in our sample, then the sample may not represent the population equitably—we may have a *biased* sample. Secondly, we expect that if we repeat the data collection, even under identical conditions, we will get somewhat different results. This variability, called *sampling variability*, needs to be taken into account in deciding how precisely findings from our sample reflect what is happening in the population.

We now have access to a wider range of statistical techniques than ever before. It is important that analytical methods used are appropriate to the sampling strategy, the type of data collected, and the research questions to be answered. We usually want to generalize our results to a population. The process of using samples to draw conclusions about populations is called *statistical inference*.

In this chapter, we discuss basic concepts underpinning data analysis and statistical inference. The ideas of probability theory are heavily involved in random sampling, which takes place under the assumption that a probability model governs the selection of the sample. Probability models are also involved in helping quantify uncertainty in sample estimates. We will also discuss a range of statistical techniques in common usage, outlining the underlying assumptions, and focusing on interpretation.

Basic concepts

Populations, samples, and sampling

For the purpose of discussing statistical methods, a *population* is the complete set of entities we want to consider in our enquiry. The entities may, for example, be free-living humans, hospital patients, disease vectors, communities, or health services. A *sample* is defined generally as a subset of a population. The sample may not be of interest in itself; its use is to provide information about the population, such as to estimate the incidence of malaria or to compare the risk of lung cancer among smokers and non-smokers. By taking a sample from the population instead of enumerating the entire population, we save resources and time. We may choose to spend some of these savings on achieving greater quality of measurement, or better follow-up of those we select. The price paid, however, is that we cannot make an absolute statement about the population; we have only an estimate —the sample incidence, or the excess of lung cancer in our sample of smokers.

Probability and random sampling

The solution to the concerns of obtaining an unbiased estimate and being able to quantify uncertainty entails the notion of ‘fairness’ in the sample selection process. A ‘fair’ method of sampling does not favour or discriminate against any member of the population. However, if we can quantify any ‘unfairness’, we can correct for this in the analysis, and so still obtain an unbiased estimate. For example, if males are twice as likely as females to be selected from a population that contains equal numbers of each sex, we can compile the data so that females have twice the weight. Thus,

what actually matters is that we know how likely it is that each individual in a population is selected for the sample—i.e. the *probability of selection* of each person.

We measure probability with a single number, which has particular properties, some of which are as follows:

Probabilities lie between 0 and 1 (or 0 and 100%).
If the probability of the event is 0, there is no chance of the event occurring.
If the probability of the event is 1, the event is certain to occur.
The combined probability of a complete collection of events is 1.

A process whose outcome is unpredictable and governed by rules of probability is termed *random*. If we are drawing random lots to select one person from 100 people, we can say (before sampling) that the probability of inclusion of an individual is 1 in 100, but we cannot predict exactly which individuals will be included. Other classic random processes are outcomes of coin tosses or rolls of a die. We know that in the long run, the chance or probability of a head is 0.5 or 50% on any one toss of an unbiased coin, but the outcome of any particular toss is unpredictable.

The word ‘random’ as used in statistics has a specialized meaning. It is used to describe a process whose outcome (in this case, getting selected for the sample or not) is governed by the play of chance. Random does not mean ‘haphazard’ or ‘arbitrary’.

A sample in which each person in the population has the same chance of selection is called a *simple random sample*. This can be achieved by listing all the members of the population (this list is called a *sampling frame*), applying an unbiased random process of selection, such as drawing lots or using computer-generated random numbers, and obtaining a sample of the required size. In practice, compiling a sampling frame can be difficult. More generally, a sample for which we know the probability of selection (whether the same or not) for each person in the population is called a *probability* or *random sample*.

Recommended sampling techniques usually require identifying a population as the source of our sample. Thus, to obtain a sample of 7-year-old children to estimate the prevalence of childhood asthma, we might begin by obtaining school class lists. It is important to be clear about the population for which we want to draw conclusions. Is it the population of 7-year-old school children or is it all 7-year-olds? It may only be feasible to

sample formally a restricted population (school children), but we really want to make conclusions about a wider population (all school-aged children). This wider population is sometimes referred to as the *target population*. Whether conclusions can be extrapolated to a wider population than actually sampled is then a matter of judgement.

If someone has no chance of being selected in the sample, and they are different with respect to the characteristics we are measuring, or if individuals vary in both their chances of selection and the characteristics of interest, we expect to have a biased sample. We often have no way of predicting how individuals vary in the characteristics being measured, but if we can control, and therefore know, the probability of selection, we can control the potential for bias. Sampling which does not follow these rules is termed *non-probability sampling*. Here we do not know the probability of inclusion of population units. Sometimes we cannot define the population sampled. These methods include convenience samples (people waiting in outpatient departments) or haphazard samples (‘people in the street’).

Critically, a sample is termed a random sample because of its method of selection. We cannot determine whether a sample is random by examining it, we must know the process by which it was obtained. Examining a sample for observations which are deviant in some respect or comparing its characteristics with those of a target population cannot ascertain randomness.

Thus, ‘random’ does not equal ‘representative’ (although using a random method of selection, in some sense, decreases our chances of non-representativeness). In summary, the reasons for using random sampling are to:

Enable quantification of the sampling variation using probability theory.
Ensure that a statistically unbiased estimator of the quantity of interest is obtained.

Complex sampling

Complex sampling is the term used to describe probability sampling which is not simple. A variety of methods, or their combination, may be used. *Stratified sampling* is used when the population consists of a number of subpopulations or strata which vary in the quantity of interest. We sample

randomly within each stratum separately, in order to control the stratum-specific numbers, or to take into account large variability across strata. For example, typically (but not universally) urban areas have larger populations than rural areas. For sampling such a population we might first wish to stratify by rural/urban status, to achieve predetermined numbers or sampling fractions in each region. If sampling fractions differ for each stratum, the overall sample will not be a simple random sample, because the probability of selection will vary across strata; it will, however, be a probability or random sample because we can, knowing the sampling fractions within strata, calculate the probability of selection of any unit. If we control the stratum sizes to exactly reflect those in the population we can achieve greater precision than a method which ignores strata. In designing a stratified sampling scheme, we need to be clear about our priorities—do we want to disproportionately increase the numbers in a particular stratum, or do we want to control optimally for interstrata variability?

If we select a simple random sample from a large city, individuals selected would be widely scattered and costly to reach if we have to conduct face-to-face interviews, for example. *Cluster sampling* involves selecting groups of individuals, rather than individuals themselves, to increase logistical efficiency. A random sample of people in a city could be achieved by first selecting a simple random sample of census collection districts, and including all people living within the selected districts. The disadvantage is that the sample may be artificially ‘homogeneous’; the selected people may be more alike than if we had used a simple random sample. This statistical inefficiency is measured by what is known as the *design effect*, which must be estimated in the analysis phase.

Two-stage sampling is similar to cluster sampling except that a sample of the cluster is taken instead of the whole cluster. This helps overcome the problem of ‘homogeneity’ to some extent, but creates a more complex process, as there are two sampling stages.

In some cases, the sampling fractions for selecting clusters and selecting from them are carefully manipulated so that the overall probability of each population unit being selected is the same. One such method is called *sampling proportional to size*—clusters are selected with a probability proportional to their size (larger clusters have larger chance of selection) and then a fixed number of individuals is selected randomly from each selected cluster. Another method is *proportional sampling* in which clusters are

selected with equal probability, and a fixed sampling fraction is used to select individuals from each selected cluster. Cluster sampling and two-stage sampling do not require an explicit sampling frame for the whole population. In two-stage sampling we need a list of first-stage units (e.g. census districts) and then a list within only each first-stage unit actually selected. This is an advantage, particularly if sampling poorly documented populations. Two-stage sampling can be extended to three or more stages, for example, districts, towns, households, and individuals.

Stratified, cluster, or multistage sampling strategies provide opportunities to measure entities at all levels. Thus, factors may be relevant at the village level (presence of a health centre, degree of remoteness), the household level (family income, access to piped water), or the individual level (age, sex).

Complex sampling allows individuals to have different probabilities of inclusion in a sample. It is imperative that we can calculate each person's probability of inclusion, otherwise we cannot make the necessary analytical adjustments. All information relevant to probability of inclusion must be documented in the course of the sample selection.

Complex sampling schemes require expertise and care to design and analyse. It is usually necessary to seek advice from a professional statistician. Most statistical software has appropriate analytical procedures, and these may have particular requirements in the construction of the data set.

We next consider the issue of characterizing information collected on our sample. This involves categorizations of variable types, crucial in deciding how to present or analyse data.

Describing variables

A primary distinction is made between *quantitative* and *categorical* variables, because these are analysed using different techniques. The distinction is based on the kinds of *values* a variable may take. For the purpose of statistical analysis, what matters is the type of measurement of the variables, specifically whether they are measured on a numerical scale (such as age) or in categories (male or female). Sometimes a numerical variable (e.g. blood pressure in mmHg), is subsequently categorized (hypertensive or not hypertensive), but it is important to realize that it is the final form (numerical or categorical) of the variable that matters for analysis.

Categorical variables

A categorical variable is one whose values are expressed in mutually exclusive categories (e.g. survival status, social class, sex, religion, extent of disability). These are further subdivided into *nominal* and *ordinal*. A nominal variable has values that are categories, with no implicit ordering of the categories (e.g. sex, religion). A variable with two levels only (e.g. sex) is referred to as a *binary*, or *dichotomous* variable. Categories of an *ordinal variable* have an underlying order (e.g. social class, extent of disability). Categorical variables are usually described by *frequencies* and *percentages* in each value category, and presented in tabular or graphical form, such as bar charts or pie charts. For example, a study of child mortality examined the types of birth attendant at delivery of children in the study. This can be presented in a table (Table 5.14.1), or graphically in various ways (Figure 5.14.1).

Table 5.14.1 Community-based sample of children under 5 years of age in Bohol Province, the Philippines, 1984–1986, by birth attendant at delivery

Birth attendant	Number of deliveries (*n* = 511)	Percentage of deliveries (%)
Untrained TBA	8	1.6
Trained TBA	224	43.8
Midwife	186	36.4
Nurse	3	0.6
Physician	34	6.7
Unknown	56	11.0
Total	511	100.0

TBA, traditional birth attendant.

Source: data from *Acute Respiratory Infection Research Project*, funded by the Australian International Development Assistance Bureau, Bohol, Philippines, Copyright © 1986–1992.

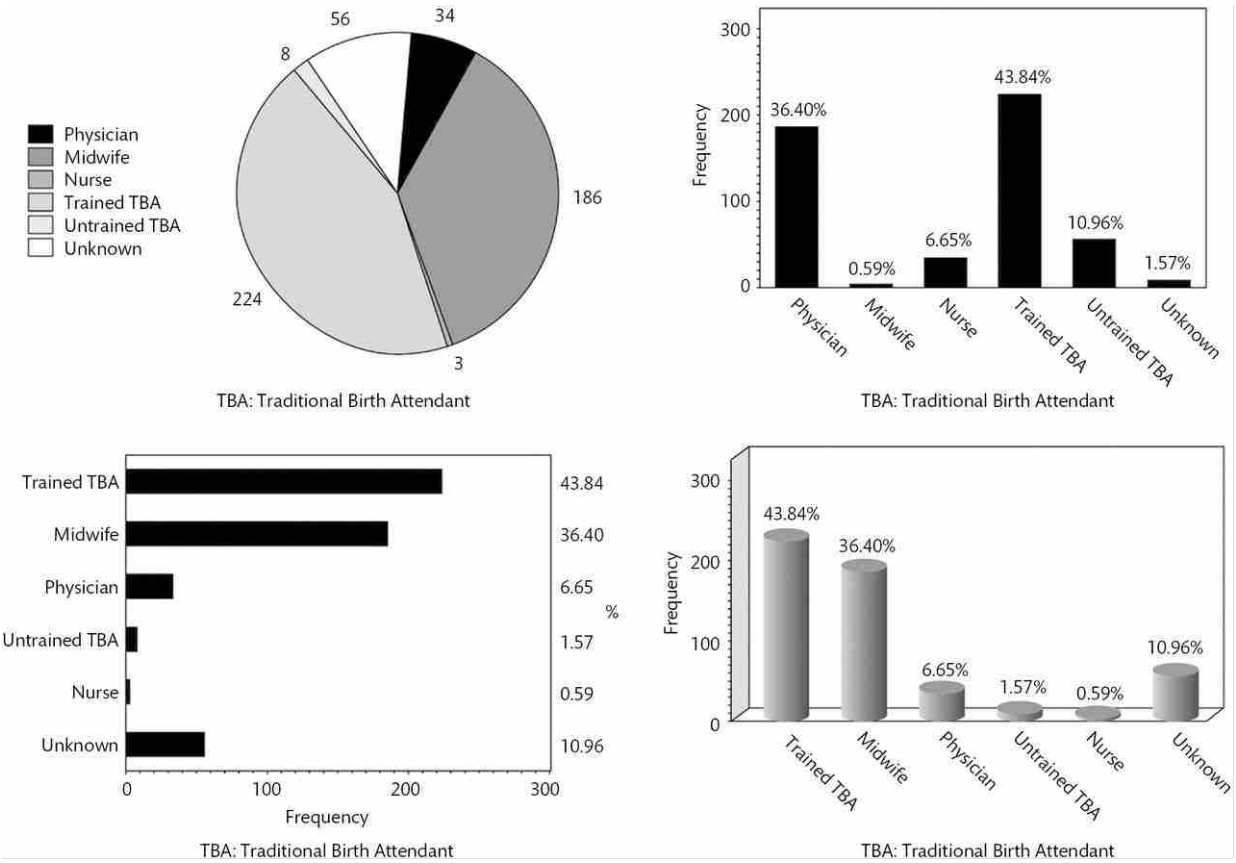


Fig. 5.14.1 Community-based sample of children under 5 years of age in Bohol Province, the Philippines, in 1984–86, by birth attendant at delivery. Alternative graphical methods of presentation. Although the choice of graphical presentation is partly a matter of taste, bar charts tend to be used for ordinal variables (as the bars can be ordered appropriately) and pie charts used for nominal variables (to de-emphasize order).

Quantitative variables

Quantitative variables express values in numerical measurement (e.g. height, age, number of births), and are further subdivided into *discrete* and *continuous* variables. A discrete variable takes only certain distinct values, usually whole numbers (e.g. number of births). A continuous variable takes all values within the observable range (e.g. age, birthweight, serum cholesterol). It is essential to specify units of measurement.

Frequency distributions

Quantitative variables are usually described, at least initially, in a frequency distribution, which tabulates, in mutually exclusive and contiguous groups, the number or percentage in each group. For example, a study documents birthweights (kg) of children born at a major obstetric hospital in Brisbane, Australia, using a sample of 500 full-term babies (gestation of at least 39 weeks). A frequency distribution portrays this (Table 5.14.2), after cut-off points for groups according to birth weight are chosen.

Choice of cut-off points for groups is to some extent arbitrary. We usually want at least eight groups, and prefer them to be evenly spaced, although this is not strictly necessary. Some recommend using $\sqrt{N}$ (square root of N) groups, where N is the sample size. We may use cut-offs defined for other purposes (e.g. using 2.5 kg as a cut-off identifies 'low' birthweight babies, as commonly defined). If we want to compare with a frequency distribution from another study, we use the same cut-off points. A frequency distribution is often graphed in a histogram (Fig. 5.14.2a). This usually requires a large number of values to provide a picture of the shape of the distribution, which may be important when we are choosing a method of analysis. Not all distributions are symmetric. A variable may have a negatively (Fig. 5.14.2b) or positively (Fig. 5.14.2c) skewed distribution, when the peak is pushed to the right or left, respectively.

Table 5.14.2 Sample of births in Brisbane, Australia, 1981, by birthweight

Birthweight (kg)	Number of babies (n = 500)	Percentage of babies (%)
2.3–2.49	6	1.2
2.5–2.69	10	2.0
2.7–2.89	19	3.8
2.9–3.09	63	12.6
3.1–3.29	71	14.2
3.3–3.49	94	18.8
3.5–3.69	78	15.6
3.7–3.89	65	13.0
3.9–4.09	41	8.2
4.1–4.29	26	5.2
4.3–4.49	20	4.0
4.5–4.69	6	1.2
4.7–4.89	1	0.2
Total	500	100.0

Source: data from Najman, J.M. et al., Cohort profile: the Mater University of Queensland Study of Pregnancy (MUSP), *International Journal of Epidemiology*, Volume 34, Issue 5, pp. 992–7, Copyright © 2005.

Measures of location

The measure most commonly used to summarize the magnitude of a quantitative variable is the *mean* (the usual arithmetic mean, or average). When some values need to be given more or less weight, a *weighted* mean may be used. When distributions are positively skewed, *geometric means* are typically used. The *median* of a variable is the middle value of its distribution; that is, half the distribution is below, and half is above, the median. The mean for the distribution of 500 birthweights given in Figure

5.14.2a is 3.49 kg, and the median is 3.47 kg; these are very similar because of the symmetry of the distribution.

When data are 'skewed', as shown in Figure 5.14.2b, c, the median is usually considered preferable to the mean (i.e. a better representative of the data), because the 'extreme' values may unduly influence the mean, whereas the median is little affected. In positively skewed data, the mean is greater than the median; in negatively skewed data, the reverse is true. The *mode* is the most frequently occurring value; a variable will have two local modes if the distribution has two peaks (this can be checked using histogram plots as in Fig. 5.14.2). The mode has few useful applications in health statistics, as well as being difficult to handle mathematically. A *quantile* is a value cutting off a specified proportion of values of a random variable. Common particular cases are tertiles (two cut-off points, dividing the distribution into three equal areas of 33.3% each), quartiles (25%), quintiles (20%), deciles (10%), and percentiles (1%). Note that the median is also the second quartile, the fifth decile, and the fiftieth percentile. An extremely useful method of graphical display of a continuous variable is the *box plot*. These compactly incorporate information about location, variability, and shape of a distribution, as well as about outliers.

Measures of variability

From a statistical viewpoint, describing the variability of a quantity is just as important as describing its average value. The most useful measures of variability depend on the difference between individual values and the mean value (deviations). If all values equal the mean (no variability at all), all deviations are zero; large deviations indicate greater variability.

One way of combining deviations in a single measure is to first square the deviations and then average the squares. Squaring is done because we are equally interested in negative deviations and positive deviations; if we averaged without squaring, negative and positive deviations would 'cancel out'. This measure is called the *variance* of the set of observations. It is 'the average squared deviation from the mean'. Because the variance is in 'square' units, a second measure is derived by taking the square root of the variance. This is the *standard deviation* (SD), and is the most commonly used measure of variability in practice. The SD for the distribution of 500 birthweights given in Figure 5.14.2a is 0.45 kg. The *range* of a variable is

defined as the difference between the largest and the smallest values. For the birthweight data in Figure 5.14.2a, the range is (from the raw data) 4.71—2.34 = 2.37 kg. The range is susceptible to extreme values and depends on the number of observations; in general, the more observations, the greater the range. In practice, reporting the smallest to the largest values (2.34 to 4.71 kg) is more useful than the range itself.

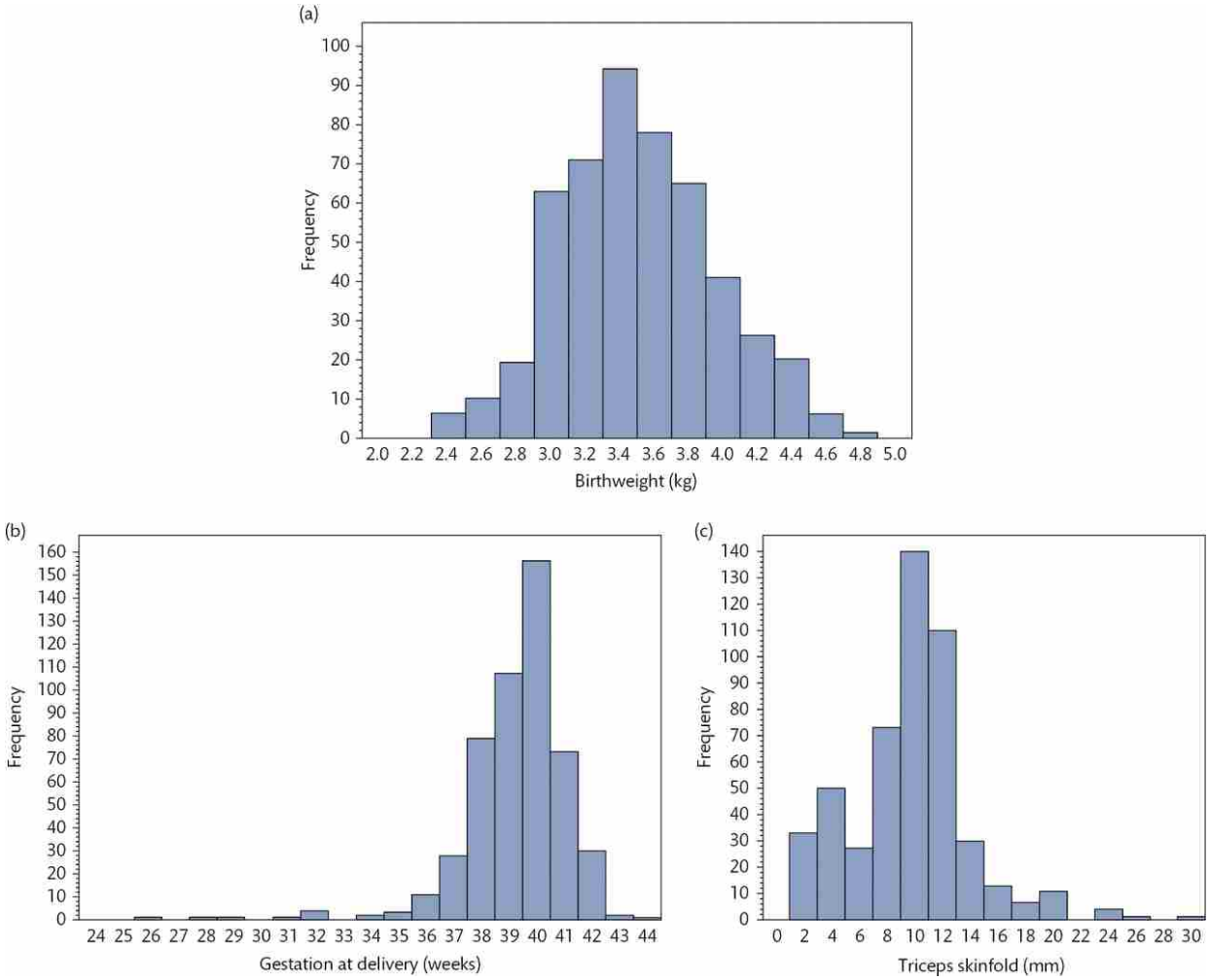


Fig. 5.14.2 Frequency distributions of characteristics of 500 babies born in Brisbane, Australia, 1981. (a) birthweight (kg), (b) gestation at delivery (weeks) for 500 babies, and (c) triceps skinfold thickness for 500 5-year-olds.

The *coefficient of variation* is the ratio of the SD to the mean, usually written as a percentage. It thus expresses variation in relation to location or

size. For the birthweight data, the coefficient of variation is 0.45 kg/3.49 kg = 0.129 or 12.9%: the SD is 12.9% of the mean birthweight.

Probability distributions

Just as a sample mean or SD (if the sample is randomly drawn) reflect the population mean or SD, so the observed distribution in the sample reflects the distribution of values in the population. In practice, it is found that certain distribution shapes re-occur. It turns out to be extremely useful to give a particular mathematical form to these characteristic shapes, which can then serve as 'models' to enable further conclusions from samples.

The normal distribution

The distribution of birthweight for a sample of 500 full-term babies born in Brisbane, Australia, followed a roughly symmetric and bell-shaped curve. The empirical observation that distributions of continuous variables often follow this shape prompted mathematicians to work on finding a mathematical or idealized representation to describe this.

The smooth curve so discovered has the mathematical formula:

$$f(X) = \frac{1}{\sqrt{2\pi}(SD)} \exp\left(-\frac{1}{2 \times (SD)^2} (X - Mean)^2 \right),$$

where 'exp' is the exponential (or antilog) function, and π ('pi') = 3.14159. *X* is the symbol given to the value of the variable in question (e.g. birthweight). Note the dependence on the mean of the population and its SD. This formula is known as the normal distribution and is a specific instance of a probability distribution; a mathematical description of the behaviour of a random variable (a variable whose values cannot be predicted with certainty). It can be thought of as an 'idealized' description of data, and can serve as a model upon which we base an analytical approach. The purpose of having an idealized model may be to 'smooth' the data or acquire some insight into the processes that underlie the generation of observations, by allowing us to test data against a given model. Once we have sample

estimates for the mean and SD, we can fit this distribution to the sample. The mean controls the location of the central peak and the SD controls the extent of variability.

Many statistical procedures depend upon assuming an underlying probability model. Checking actual observations against these predictions can provide us with insight: if the probability model 'fits', then we can use features of the model to make further inferences or predictions about the population. If the model does not 'fit', then we need to find a more appropriate one. We would not regard the normal distribution a 'good' model for all distributions, such as those shown in Figure 5.14.2b, c. Other probability distributions may be used in such instances; they may be more complex, because they incorporate additional information about the extent and type of skewness.

The usefulness of the normal model lies in the fact that its properties have been well studied. Specifically, for example, we know that approximately two-thirds of the distribution is enclosed by 1 SD either side of the mean, and approximately 95% of the distribution is enclosed by 2 SDs either side of the mean. Figure 5.14.3 shows the normal model fitted to the mean and SD of the sample of birthweight.

Within this normal model, 2 SDs either side of the mean takes us from 3.49—(2 × 0.45) kg = 2.59 kg to 3.49 + (2 × 0.45) kg = 4.39 kg. The area (or the area under the curve) between 2.59 and 4.39 kg in Figure 5.14.3 (the outer set of vertical lines) is approximately 95% of the total area.

In summary, the descriptive properties of the normal distribution are as follows:

It describes a continuous variable across the entire numerical scale, although in practice values from an assumed normal distribution will be confined to a particular range.
It is bell-shaped, and symmetrical about its mean value; that is, half of the distribution lies on either side of the mean.
It is determined completely by its mean and SD.
Limits sets by the number of SDs either side of the mean predict the proportion of the distribution enclosed by these limits—for 1 SD, the area is 68% of the distribution; for 1.96 SDs (≈2 SDs), it is 95%; and for 2.56 SDs, it is 99%. The number of SDs corresponding to different proportions is calculated from the standard normal distribution, which has zero mean and a SD of 1.

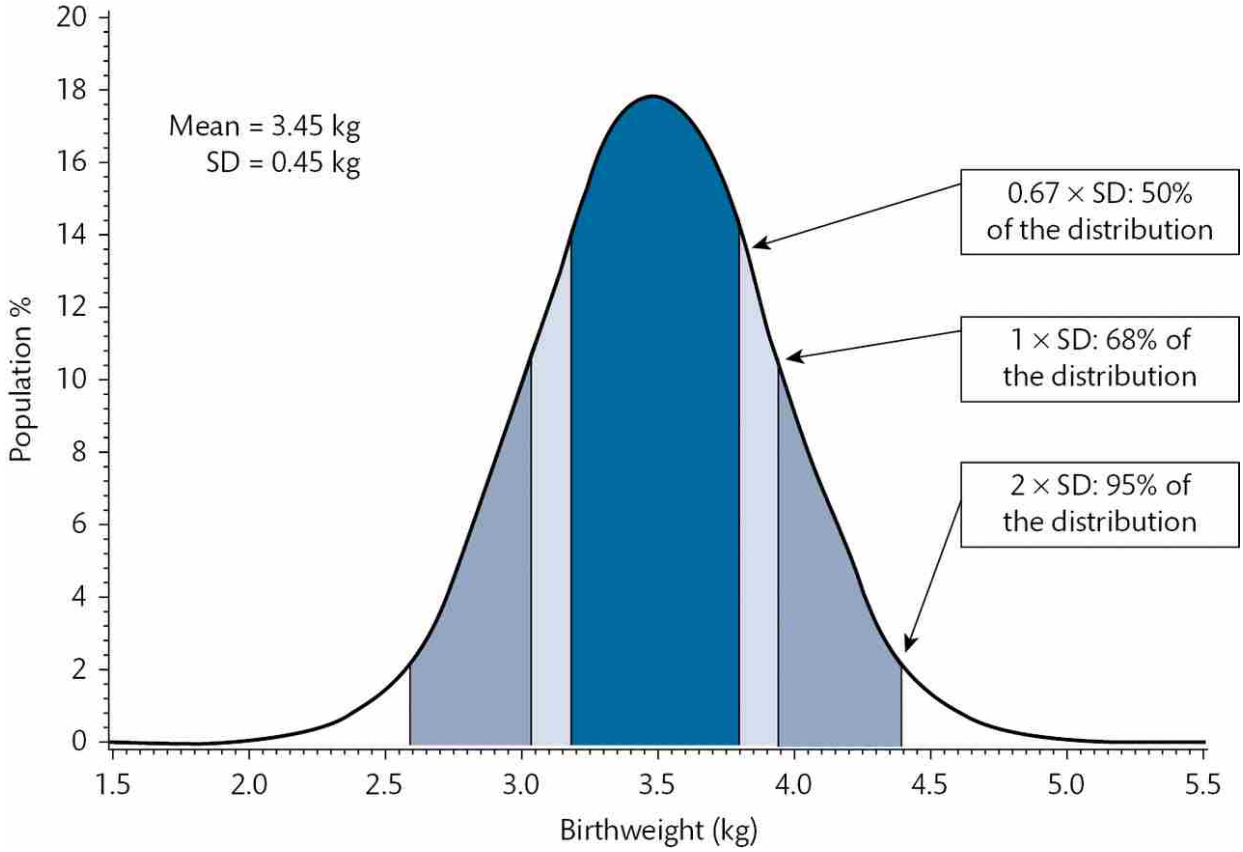


Fig. 5.14.3 A normal distribution fitted using the sample of 500 birthweights (mean = 3.49 kg, SD = 0.448 kg) showing areas enclosed by 0.67, 1, and 2 standard deviations (SDs) either side of the mean. The purpose of fitting a distribution in this way may simply be to obtain a 'smoothed' picture of a distribution. But we can use this 'smoothed' representation as a model for further calculations, such as estimating: (1) the proportion of babies whose birth weights are within a particular range or (2) the value of birth weight that cuts off a certain proportion of birth weights.

Sampling variability

The concept of sampling variability lies at the heart of what is known as *statistical inference*. Using a sample to estimate a mean or a proportion relating to a population, as opposed to examining the entire population, results in a considerable reduction in cost. This, however, results in some uncertainty about our estimate. Mathematical rules have been established which enable us to predict, within limits of probability, the sampling variability inherent in an estimation process based on random sampling.

These rules contribute to the well-known practices of estimating confidence intervals (CIs), testing hypotheses, and fitting statistical models.

Sampling distributions

The *sampling distribution* of an estimate describes the chance variability we would expect to see in the distribution of sample estimates if we were to repeat our sampling process a large number of times. In practice we only have one sample and one estimate (of the mean, for example) and so must rely on mathematical theory to tell us the properties of the sampling distribution of the mean and therefore its precision.

Consider the example of taking a random sample of births, in order to estimate mean birthweight. As random sampling favours neither high nor low birthweights (i.e. it is unbiased), we expect that any particular sample mean will be as likely to be above the mean as below it. So, the first property governing the behaviour of random sampling is:

Irrespective of sample size, the sampling distribution of a mean has the same mean as the population mean.

Next, consider the size of the sample. Given an estimate is unbiased, how close is it likely to be to the true population value (i.e. the value we would obtain if we measured everyone in the population)? Intuitively, we would expect a random sample of 1,000 members of a population to provide a 'better' or more precise estimate than a sample of 100. This results in the second property governing the behaviour of random sampling: governing the behaviour of random sampling:

The sampling distribution of a mean has variability (i.e. standard deviation) which decreases with the size of the sample.

The practical implication of (1) and (2) is that as the sample size N becomes larger, our sample estimate is more likely to be closer to the population mean.

Both properties of the sampling distribution of the mean have been mathematically demonstrated and further quantified. An important mathematical discovery related the expected SD of a sample mean to the

sample size and the SD of the variable within the population. The SD of a mean is more usually referred to as the *standard error of the mean* (SE).

The variability in the sample mean is defined by

$$\text{SD of the means Standard error of the mean} = \frac{\text{SD of the population}}{\sqrt{N}},$$

where N is the size of the sample.

This shows that the larger the sample size, the less variable are the means, and that the greater the variability in the original measurements in the population, the more variable are the means.

In a remarkable mathematical result (the Central Limit Theorem), Pierre-Simon Laplace showed, in 1810, that the sampling distribution of the mean follows a normal distribution, once N becomes large enough, even when the original distribution itself does not follow a normal distribution! This puts us in a powerful position to predict the behaviour of sampling variability.

Confidence intervals

Confidence interval for a mean

The sample mean provides us with an estimate of the unknown population mean. Using the previous theory, we now know what to expect of its sampling variability.

Using our example of birthweights, and $N = 64$, we have an expected SE of 0.07 kg. Using the assumption of a normal distribution for sample means (the Central Limit Theorem), there is a 95% chance that the error of our estimate is less than 0.14 kg (1.96 × 0.07). Another way of putting this is to say that there is a 95% chance that:

Sample mean—population mean <0.14 kg, if the sample mean is greater than the population mean; and
Population mean—sample mean <0.14 kg, if the population mean is greater than the sample mean.

Reorganizing these, with 95% probability, equivalent statements are:

Sample mean—0.14 kg < population mean, if the sample mean is greater than population mean; and
Sample mean + 0.14 kg > population mean, if the population mean is greater than the sample mean.

Combining (1) and (2), we can say, with 95% probability, that the limits (sample mean—0.14 kg, sample mean + 0.14 kg) enclose the true (unknown) population mean. This is known as the *95% CI for the mean*. We use other values (from the standard normal distribution) to obtain the CIs corresponding to different percentages. Generally:

$$\text{CI for a mean} = \text{sample mean} - k \times \text{SE},\ \text{sample mean} + k \times \text{SE}$$
$$\text{where SE} = \text{standard error of mean} = \frac{SD}{\sqrt{N}},$$

where k = 1.28, 1.645, 1.96, and 2.56 for a 80%, 90%, 95%, and 99% CI, respectively.

In expressing these results we say ‘there is a 95% probability that the interval (3,284 g, 3,516 g) contains the true mean birthweight’, *not* that ‘there is a 95% probability that the population mean lies within (3,284 g, 3,516 g)’.

In practice, we do not know the SD of the population and usually replace it with the sample SD. This introduces very little error if the sample size is 50 or more. If the sample is small, we use a different multiplier for the standard error (i.e. a different k); this multiplier is obtained from the *t*-distribution (i.e. not the standard normal distribution), available in statistical software. We need to specify another parameter called *degrees of freedom* (df), which depends on the sample size: for sample size N, df = N—1. Therefore, for N = 30, df = 29; for a 95% CI, we use k = 2.045, instead of k = 1.96.

Note that a CI combines information regarding the size (location) of the parameter, the variability of values within a population, and the sample size used to estimate the parameter. It is thus a very useful quantity to summarize sample results.

Confidence interval for a proportion

We can also obtain a CI for proportions. The concept of a CI for a mean extends to a proportion, or indeed other epidemiological measures. Let us return to the birth attendant data from the Philippines (Table 5.14.1). We estimate that 224 of the 511 babies (43.8%) were delivered by trained traditional birth attendants (TBAs). The principles underlying consideration of sampling errors for proportions are the same as for the mean. The Central Limit Theorem again assures us that, provided samples are large enough, sample proportions will have, approximately, a normal distribution.

The only different feature of the process of obtaining a CI for a proportion is the formula for the standard error of the estimate. The standard error of a proportion is

$$\text{Standard error of a proportion (SEP)} = \sqrt{\frac{p \times (1-p)}{N}},$$

where p is the sample proportion, and N is the sample size. In our example, we have $N = 511$, $p = 0.438$:

$$\text{Standard error of a proportion (SEP)} = \sqrt{\frac{0.438 \times (1-0.438)}{511}} = 0.022.$$

The formula for a CI for a proportion is then completely analogous to that for a mean. The 95% CI for the proportion of babies delivered by TBAs is then 0.438—(1.96 × 0.022), 0.438 + (1.96 × 0.022), or (0.395, 0.481). Expressing the result in percentages, an estimated 43.8% of babies are delivered by TBAs, with a 95% CI of (39.5%, 48.1%).

Basic principles and methods of comparative analysis

This section deals with some of the basic methods of analysis. Firstly, we consider situations in which we want to compare a set of mutually exclusive groups. This comparison may be on a number of different variable types, such as categorical or continuous health outcome, or the prevalence or mean of a risk factor. Secondly, we introduce models, aimed at examining associations of variables, or patterns of risk factors.

Group comparisons

Many research questions in public health involve comparing groups. This may be in the context of an experimental or intervention study, a designed analytical study such as a cohort, case– control, or cross-sectional studies. Such analyses proceed in two phases: first we describe the pattern of the data *within* each group, using methods such as those described earlier, and then we evaluate the comparisons *between* groups.

The comparison between groups may initially be descriptive, but then we usually want an evaluation of the 'significance' of the differences we see; that is, are the differences compatible with sampling variability alone, or are there 'real' differences between the groups, which might be due to biological or other effects? If the differences are not attributed to chance, are they important from a clinical or public health perspective? We also often want to determine whether differences are due to 'causal' factors, and whether these causal factors are modifiable and so provide a basis for prevention or management. The judgement of causality involves additional evidence other than statistical analysis, such as biological plausibility, magnitudes of effects, and considerations of confounding. Although statistical analysis, particularly multivariable analysis (see later), can help provide some information, the final judgement of causality is a complex one.

Some examples of research questions, how they might be framed, and the nature of the variables involved are shown in Table 5.14.3.

If a relationship is being examined, it is usually important to consider whether variables have a role (or potential role) as 'outcome', 'explanatory', or 'predictor' variables. In Table 5.14.3, birthweight, length of labour, and stage of disease are outcome variables; mother's height, number of previous births, and distance from the health centre are potentially explanatory variables ('potentially' because if it turns out there is no relationship, then one variable can hardly 'explain' the other). A variable may take on different roles in different questions; for example, number of years of schooling is an outcome variable in question 6 and an explanatory variable in question 7. So, for the purpose of classifying variables as outcome or explanatory, what matters is the orientation of the question (which event comes first?).

Analyses of relationships depend on the types of variables involved, and there may be several valid methods from which to choose.

Comparing proportions or percentages

The first step in examining the relationship between two categorical variables is usually to cross-classify them; that is, break down one variable by the values of another. This is usually displayed in a contingency table, which shows all combinations of values. Consider an example (Table 5.14.4) taken from a study of skin cancer in Queensland, Australia (Green et al. 1999).

Table 5.14.3 Some research questions that involve relationships between two variables

Question	Variable combination
1. Does **height** influence the **birthweight** of her child?	Continuous–Continuous
2. Is there an association between **length of labour** and the **number of previous births** a woman has had?	Continuous–Discrete
3. Do people present at a later **stage of disease** if they live a greater **distance from the health centre**?	Ordinal–Continuous
4. Do different **ethnic groups** have different **calcium intake**?	Nominal–Continuous
5. Does the **number of sunburns** a person has experienced predict the **number of skin cancers** they have?	Discrete–Discrete
6. Does the **number of years of schooling** a child acquires depend on his or her **birth order** in the family?	Discrete–Ordinal
7. Does the **number of years of schooling** a girl acquires influence her **paid employment status** in later life?	Discrete–Nominal
8. Do different **socioeconomic groups** present with different degrees of **disease severity**?	Ordinal–Ordinal
9. Does **diagnosis at presentation** depend on **socioeconomic group**?	Nominal–Ordinal
10. Does susceptibility to **skin cancer** depend on **skin colour**?	Nominal–Nominal

Table 5.14.4 Numbers of people with a non-melanoma skin cancer (NMSC) diagnosed on a sun-exposed site in the period 1993–2006, by skin type, as reported in 1992

	No NMSC	One NMSC lesion only	Multiple NMSC lesions	Total
Always burn	236 (69.4)	50 (14.7)	54 (15.9)	340
Burn, then tan	866 (78.8)	112 (10.2)	121 (11.1)	1,099
Tan only	148 (82.7)	16 (8.9)	15 (8.4)	179
Total	1,250 (77.3)	178 (11.0)	190 (11.7)	1,618

Figures in parentheses are percentages within each skin type.

Source: data from *The Lancet*, Volume 354, Issue 9180, Green, A. et al., Daily sunscreen application and betacarotene supplementation in prevention of basal-cell and squamous-cell carcinoma of the skin; a randomised controlled trial, pp. 72–5, Copyright © 1999 Elsevier Ltd. All rights reserved.

A contingency table should include *all* combinations of *all* categories of each variable. The overall incidence of skin cancer in the follow-up period is 22.7% (combining single and multiple lesions), but incidence is higher in those reporting a skin type that sunburns, with or without tanning (21.2% and 30.6%, respectively), than in those whose skin tans without burning (17.3%).

If we repeat the skin cancer study with another sample of 1,600 or so participants, we will obtain slightly different results, even if we follow exactly the same procedures. This is due to sampling variability: within our study population, individual people with the same characteristics will have different skin cancer outcomes. How do we take this into account in our interpretation of findings? To simplify the example, consider the comparison of the three skin types, combining single and multiple lesions. The first step is to calculate what we would expect to see if people with different skin types produced on average, the *same distribution of responses*. In other words, what would we expect to see if sampling variability was the *only* factor operating to produce variability in responses? The answer is: the same distribution of skin cancer within each skin type. The results of calculating these ‘expected’ results are shown in Table 5.14.5.

The figures in parentheses in the table are called *expected frequencies*; that is, the numbers expected if skin type is unrelated to skin cancer. Note that the marginal totals of observed and expected frequencies are the same. We have observed more skin cancer in the ‘Always burn’ group (about 27 more cases) than expected on the basis of no real difference between skin types and less in the ‘Tan only’ group (10 fewer cases). These differences relate either to:

Sampling variability alone (i.e. no additional risk of skin cancer associated with skin type).
Sampling variability *plus* an additional risk of skin cancer associated with skin type.

The statement that, in the study population, skin type is unrelated to skin cancer is called the *null hypothesis*, and is the first step in any formal evaluation of the role of chance.

The Pearson chi-squared test

The statistic called the *chi-squared test statistic*, developed by Karl Pearson, enables us to summarize the differences between observed and expected frequencies, so we can make a decision as to the significance of these. Calculations are shown in Table 5.14.6.

Table 5.14.5 Numbers of people with a non-melanoma skin cancer (NMSC) diagnosed on a sun-exposed site in the period 1993–2004, by skin type, as reported in 1992, and numbers expected (in parentheses) if the distributions of skin types are the same in each group

	No NMSC lesions	One or more NMSC lesions	Total
Always burn	236 (262.7)	104 (77.3)	340
Burn, then tan	866 (849.0)	233 (250.0)	1,099
Tan only	148 (138.3)	31 (40.7)	179
Total	1,250	368	1,618

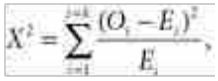
Source: data from *The Lancet*, Volume 354, Issue 9180, Green, A. et al., Daily sunscreen application and betacarotene supplementation in prevention of basal-cell and squamous-cell carcinoma of the skin; a randomised controlled trial, pp. 72–5, Copyright © 1999 Elsevier Ltd. All rights reserved.

Table 5.14.6 Calculation of the chi-square statistic for data in Table 5.14.5

	No lesions	One or more lesions
Always burn	236 262 7. ^{2} (−) = 271. 262. 7	104 77 3. ^{2} (−) = 922. 77. 3
Burn, then tan	866 849 0. ^{2} (−) = 034. 849. 0	233 250 0. ^{2} (−) = 116. 250. 0
Tan only	148 138 3. ^{2} (−) = 068. 138. 3	31 40 7. ^{2} (−) = 231. 40 7.
Sum of all (Observed Frequency − Expected Frequency)^2 Expected Frequency = 271. + 922. + 034. + 116. + 068. + 231. = 16. 42		

Source: data from *The Lancet*, Volume 354, Issue 9180, Green, A. et al., Daily sunscreen application and betacarotene supplementation in prevention of basal-cell and squamous-cell carcinoma of the skin; a randomised controlled trial, pp. 72–5, Copyright © 1999 Elsevier Ltd. All rights reserved.

The formula for the chi-squared statistic is:



where k is the number of cells in the table. The following is observed:

The smallest value of X^2 that can occur is 0, when all individual terms have their smallest value as 0 and corresponding expected and observed frequencies are exactly equal.
The value of X^2 increases as differences between observed and expected frequencies increase.
The larger the sample, the larger the value of X^2, for a given strength of association: if we double the sample size, for example, we double the calculated value of X^2.

So, the measure (X^2) increases as the observed frequencies increasingly depart from the expected frequencies. Thus small values of X^2 are consistent with (1) above, but increases in X^2 suggest that (2) is the underlying explanation.

The calculated value of X^2 is 16.42 (Table 5.14.6). But how large does X^2 need to be before we can decide in favour of (2) over (1)? Is there a cut-off point? The mathematics of this question have been solved: for any given value of X^2 we can find the probability that this value, or one larger than it, will occur if the observed differences are due to chance alone; that is, if (1) is true. The relationship between values of X^2 and the probability of such values occurring is expressed in the *chi-squared distribution*. Tables for the chi-squared distribution are available in standard textbooks, or values can be obtained from statistical software. The probabilities depend on another parameter, the df, which depends on the number of row and column categories in the contingency table. If the table has r rows and c columns, df = (r—1) × (c—1). For Table 5.14.6, there are three rows and two columns: df = (3—1) × (2—1) = 2.

The probability, referred to as the p-value, of getting X^2 = 16.42, or larger, is 0.0003. So if (1)—sampling variability—is the sole explanation for the observed treatment difference, the probability of seeing such a large value of X^2 is 0.0003, or 0.03% or 1 chance in 3,333. Because this is very small, we would feel inclined to reject (1) in favour of (2). In practice, the most common cut-off point to use in order to make this decision is p = 0.05 or 5%.

We then say that there is a *statistically significant difference* between skin types in risk of skin cancer ($p = 0.0003$). From Table 5.14.6, we see that the 'Always burn' skin type contributes by far the most to the value of X^2, so we could conclude that this category is responsible for most of the statistically significant finding.

Limitations of the Pearson chi-squared test

The use of the *Pearson chi-squared test* involves an important underlying assumption—that the observed frequencies are independent. If, for example, instead of skin type we were comparing pre- and post-intervention incidence of skin cancer within the same sample of people, the observations would not be independent, because outcomes for the same person would be expected to be related and the Pearson chi-squared test should not be used.

The chi-squared test also depends on samples being 'large enough' to be valid. For contingency tables, 'large enough' has been found to depend on the size of the expected frequencies. The rule of thumb has emerged that if around 1 in 5 (or 20%), or more, of expected frequencies are less than 5, then the test is not valid. Statistical software provides simulation or permutation tests which give an estimated *p*-value directly, which is then interpreted in the usual way.

We had decided to reject the null hypothesis based on a probability cut-off point in order to distinguish between (1) and (2). This decision might have been incorrect. It is unlikely (about 1 chance in 3,333), not impossible, that a value of $X^2 = 16.42$, or larger, could occur if (1) were true. Rejecting the null hypothesis, concluding there is a significant difference when there is not, is a *type 1 error*. The reverse error, failing to reject the null hypothesis when it is false or concluding there is no difference when there really is, is called a *type 2 error*. This error typically arises when we have a small sample size, and our data are insufficient to detect a difference.

Table 5.14.7 summarizes the four possible outcomes from a statistical test of a hypothesis, according to whether there is a real effect (e.g. a real difference) and whether the test statistic is significant (i.e. the *p*-value is small). A type 1 error is a false-positive finding and a type 2 error is a false-negative finding.

Asserting that a difference, or an association, is ‘significant’ does not mean that the association is ‘strong’ or that it is causal. A statistically significant finding means that the size of the difference or association is larger than would be likely to occur by chance alone (i.e. sampling variability alone) in the absence of any real association. These techniques can be applied to larger contingency tables, although sometimes further examination of the data may be needed to identify the nature of the differences.

Table 5.14.7 Type 1 and type 2 errors

		True effect	
Null hypothesis true	**Null hypothesis not true**		
Statistically significant effect (reject null hypothesis)	No	✓	Type 1 error (false-positive)
	Yes	Type 2 error (false-negative)	✓

The chi-squared test explores all variability in proportions in a contingency table. However, sometimes we want to examine a more specific type of association, and the chi-squared test may not be the best one to use. An example of such a situation occurs when we want to examine trends in proportions.

Suppose we regarded the skin type groups (in order) in Table 5.14.5 as indicative of an increasing inflammatory reaction to sun exposure and wanted to test a hypothesis about increasing risk of skin cancer with increasing inflammatory response (rather than just variability in risk with skin type, whatever the pattern). That is, we want to treat the variable as ordinal rather than nominal—we need a statistical test that specifically looks for a trend pattern. Various options are available, which usually involve assigning an increasing score to the categories of the ordinal variable. Tests for these will be discussed in the subsection for regression modelling of binary variables.

Comparing means

Suppose we want to examine whether a child's birthweight depends upon his or her mother's marital status. We know that the distribution of birthweight, a continuous variable, can be summarized by its mean and standard deviation. So we examine these summary measures within values of the categorical variable (marital status, a nominal variable) to see how the continuous variable varies with the categorical variable. Table 5.14.8 shows the results.

If the frequency distribution of a continuous variable is approximately normal (as it is for birthweight), then it is described fully by its mean and SD, which is why we use these measures in the example. In other cases, such as skewed distributions, we might use the median or the geometric mean as a measure of location and the interquartile range as a measure of variability, and present these instead of, or along with, the mean and standard deviation.

These results show that, compared to 'Married' women, women who have 'Never married' have slightly heavier babies (by 0.07 kg), whereas women 'Living together' and 'Other' have slightly lighter babies (by 0.19 kg and 0.17 kg, respectively). These differences seem comparatively small, but would be regarded as important differences (by comparison with other risk factors that decrease birth weight, e.g. smoking). The mean for married women is very close to the overall mean, because they make up most (363/500 = 72.6%) of the sample. We now need to make an assessment of the role of chance in producing the observed differences. We know that some variability in observed means is due to sampling error, but could sampling error be the complete explanation? The crucial question is whether the means differ more than would be expected by chance alone.

The sampling variability of a mean is assessed by its standard error of the mean (SE) which has earlier been used to derive a CI for the true mean. We can apply these techniques to the data in Table 5.14.8, examining mothers' marital status and birthweight (Table 5.14.9).

Table 5.14.8 Mean, standard deviation (SD), and range of birthweight, by marital status of mother, for a group of Brisbane babies born in 1984–1988

Marital status	*N*	Mean (kg)	SD (kg)	Minimum (kg)	Maximum (kg)
Married	363	3.57	0.42	2.33	4.85
Never married	70	3.64	0.44	2.34	5.06
Living together	53	3.38	0.42	2.27	4.18
Other	14	3.40	0.40	2.86	4.28
Total	500	3.56	0.43	2.27	5.06

Source: data from Najman, J.M. et al., Cohort profile: the Mater University of Queensland Study of Pregnancy (MUSP), *International Journal of Epidemiology*, Volume 34, Issue 5, pp. 992–7, Copyright © 2005.

Table 5.14.9 Mean birthweight (kg), standard errors (SE), and 95% confidence intervals (CI), by marital status of mother, for a sample of Brisbane babies born in 1984–1988

Marital status	*N*	Mean (kg)	SE (kg)	95% CI
Married	363	3.57	0.022	(3.53, 3.62)
Never married	70	3.64	0.052	(3.53, 3.74)
Living together	53	3.38	0.058	(3.27, 3.50)
Other	14	3.40	0.108	(3.16, 3.63)
Total	500	3.56	0.019	(3.52, 3.59)

Source: data from Najman, J.M. et al., Cohort profile: the Mater University of Queensland Study of Pregnancy (MUSP), *International Journal of Epidemiology*, Volume 34, Issue 5, pp. 992–7, Copyright © 2005.

Note the variability in the standard errors due to the variability in the sample size: the largest group (‘Married’ women) has the smallest standard error and the smallest group (‘Other’) has the largest standard error. Correspondingly, the CIs vary greatly in their widths. For ‘Married’ and ‘Never married’ mothers the CIs overlap considerably, but the ‘Married’

group has a very narrow CI, which does not overlap the ‘Living together’ category. We might feel that the difference among ‘Married’ and the ‘Never married’ could be explained by sampling variability, but that the difference between ‘Married’ and ‘Living together’ categories could not. Sometimes overlap or non-overlap of CIs is used in this way, but if we wish to make a formal statement about the difference between means then we must make a formal statistical test.

The two-sample *t*-test

The first step of a formal statistical test is to formulate a null hypothesis. In this case, the null hypothesis is that (true) mean birthweights do not vary among marital status groups. First consider the difference between birthweights of children of ‘Married’ (3.57 kg) and ‘Never married’ (3.64 kg) mothers—a difference of 0.07 kg. The next step involves assessing the role of chance in producing this observed difference if there is no difference in the population means. As before, we need to obtain the probability of getting a difference of 0.07 kg, by chance, if the null hypothesis is true. We evaluate this by first calculating a *t*-statistic. Perhaps not surprisingly, the test statistic depends on:

The size of the difference between the means.
The standard errors of the means.

Therefore:

$$t = \frac{\text{Difference between the means}}{\sqrt{\text{Sum of } (\text{SE})^2}}$$

with degrees of freedom (df) = sum of sample sizes –2.
This is called the *two-sample t-test*. The larger the difference between the two sample means, and the smaller the standard errors (or the larger the samples), the larger the value of *t*. As for the chi-squared test, once we have a calculated value and a value for the degrees of freedom, we obtain a *p*-value using statistical software.

For the comparison of birthweights of children whose mothers are ‘Married’ or ‘Never married’, we have:

$$t = \frac{3.57 - 3.64}{\sqrt{0.002^2 + 0.052^2}} = -1.24,$$

with degrees of freedom (df) = 363 + 70—2 = 431.

From statistical tables, we see that the *p*-value is > 0.10 (using df = infinity (∞)). The probability that such a difference could occur by chance is quite high. We then attribute the observed difference to sampling variability and conclude the difference is not significant.

Analysis of variance

For the example of mothers’ marital status and birthweight, we could carry out more *t*-tests, looking at the difference between other pairs of groups. However, it is considered bad practice to ‘search’ for significance in this way. A more acceptable practice is to carry out a ‘global’ test on the set of four means, to see if there is any significant variability among the means as a set, without specifying any particular pair of means. Pairwise differences may then be examined post hoc, once overall variability has been established. The technique used for this global test on a set of means is called *analysis of variance* (ANOVA). The null hypothesis is that there is no variation in the population means for all groups. Effectively, we assume all mothers, irrespective of marital status, belong to the same population, as far as their babies’ birthweights are concerned. No new conceptual issues arise in this analysis.

Again we work out a test statistic, now called an *F*-statistic. This test statistic depends on:

The variability among the means.
The standard errors of the means.

Now we have 2 df to work out: the so-called ‘numerator’ df is the number of groups minus 1 and the ‘denominator’ df is the total sample size minus the number of groups being compared.

We would usually use statistical software for ANOVA, which will also give us a *p*-value. In this case, the *F*-statistic is 14.01, with df = (3,497), and $p = 0.0029$. This is smaller than our usual cut-off point of 0.05, and we conclude that there are significant differences in birthweight among the marital status groups. These differences obviously arise from the 'Living together' and 'Other' groups.

If we had decided not to reject the null hypothesis, the matter would have ended there. When we do reject the null hypothesis, we might want to examine the pairwise differences using *t*-tests. Although opinions vary, sometimes modified versions of the *t*-test are used to take into account the multiple comparisons that might be involved in such examination. Multiple comparisons increase the overall risk of a type 1 error (i.e. finding at least statistically significant result just by chance).

Assumptions of the two-sample *t*-test and ANOVA

The two-sample *t*-test and ANOVA have important underlying assumptions. These are as follows:

If the sample is small, the mean is based on a variable with a normal distribution; if the variable does not have a normal distribution, the sample is large enough for the Central Limit Theorem to apply to the distribution of the mean.
The SDs within each group are approximately equal.
The samples are independent observations; that is, separate groups of individuals.

We have already established that the distribution of birthweight is quite close to a normal distribution, so the first assumption has been met. What about the second? Inspection of the 'SD' column in Table 5.14.8 confirms this—the SDs vary only slightly.

A statistical test, called Bartlett's test, can be used to determine whether SDs are significantly different. Some statistical software automatically carries out this test when an ANOVA is requested. Alternative tests are available in such a situation.

There are two main options for dealing with non-normal distributions: transformations and non-parametric tests.

Transformations: it is sometimes possible to transform the variable to make a normal distribution, and then carry out the *t*-test or ANOVA on the transformed variable. If a variable has a positively skewed distribution, the log transformation is sometimes used to give a more symmetric distribution, closer to the normal distribution. Many transformations are available, but there is no guarantee that any will work in any particular case. The distribution of the transformed variable needs to be examined to confirm this.

Non-parametric tests: if a transformation cannot be found, it is possible to use a non-parametric test. These do not require the first two assumptions, but still require independence between observations (the third assumption). Instead of a *t*-test, we use the Wilcoxon–Mann–Whitney test to get a *p*-value for comparing two means; instead of ANOVA, we use the Kruskal–Wallis analysis of variance to obtain a *p*-value for comparing two means.

If the assumptions are met and non-parametric tests are used, these tests will have less ability to detect differences between means; that is, they will result in larger *p*-values than the *t*-test or ANOVA and be less likely to reject the null hypothesis. Usually researchers want to make the maximal use of their data, and carry out the most sensitive tests possible, so they prefer to use tests based on the normal distribution if possible.

Repeated measurements of a quantitative variable

The questions in Table 5.14.3 are about relationships between two variables, each of which is measuring a different entity. Sometimes our question relates to whether a single variable changes over time. Suppose we are investigating the effectiveness of an intervention to improve the nutritional status of children. In this case, we might measure the nutritional status of each child within the sample of children before and after the intervention. At first sight, this looks similar to the previous case: there are two variables, nutritional status before intervention and nutritional status after intervention. But this is different because we are really observing one variable at two different times. We expect measurements taken on the same individual at two different times to be related so the measures are not independent. This relationship needs to be taken into account when the effect of a second variable (intervention) is being evaluated.

In a trial to examine the effectiveness of a dietary intervention, 17 participants had vitamin A intake (in µg) assessed at baseline and 6 months later, to see whether the dietary intervention helped increase vitamin A intake over the 6-month period (Table 5.14.10). At first sight, we might see this as a situation for a two-sample *t*-test, comparing mean vitamin A intake at baseline and at 6 months. This is not appropriate because the third assumption for the two-sample *t*-test—independence of the groups being compared—has not been met.

Table 5.14.10 Vitamin A intake levels (micrograms = µg) at baseline and 6 months post-intervention

Person	Baseline	6 months	Change (6 months–baseline)
1	120	124	4
2	122	135	13
3	126	140	14
4	131	130	–1
5	135	169	34
6	157	189	32
7	160	176	16
8	181	196	15
9	195	223	28
10	200	254	54
11	205	267	62
12	211	289	78
13	215	301	86
14	220	227	7
15	246	305	59
16	255	278	23
17	278	299	21
Mean (SD)	185.7 (49.6)	217.8 (65.4)	32.1 (26.3)

Almost all 6-month values are higher than baseline values, although some individual differences are quite small. The appropriate summary statistic to compare paired repeated measurements of a continuous variable is the *mean change* within individuals. For each participant, we subtract baseline vitamin A levels from 6-month values. The mean and SD of these individual changes are then calculated. Hence, for matched data, we finally deal with one variable (i.e. change after intervention) and not two variables. In this

example, the mean change in vitamin A levels was 32.1 µg, with a range of changes from –1 to + 86 µg and a SD of 26.3 µg. Where the distribution of the observed changes is not normal, a more valid summary would be the median.

To get a 95% CI for the mean change, we first calculate the standard error of the mean change in vitamin A intake, which is

$$\mathrm{SE(change)} = \frac{\mathrm{SD(change)}}{\sqrt{N}} = \frac{26.3}{\sqrt{17}} = 6.4\mu g.$$

The 95% CI = (mean change—1.96 × SE(mean change), mean change + 1.96 × SE(mean change)) = (19.5, 44.6).

Notice that the lower limit of this 95% CI is well above zero, so we have considerable confidence that the change is positive, at least.

We can formally test the null hypothesis that the true increase is zero by calculating a *t*-statistic, as follows:

$$t = \frac{\text{mean change}}{\mathrm{SE(change)}}$$

This is referred to as the *paired or one-sample* t-*test*, because it assumes paired measures, and we want to evaluate the significance of the change (or more generally, the difference) within the paired values. The larger the mean change, and the smaller the standard error (or the larger the sample), the larger the value of *t*.

Once *t* and its df are calculated, we obtain a *p*-value from *t*-tables or statistical software. For the paired *t*-statistic on *N* pairs, the df is *N*—1. We calculate $t = 5.02$, with 16 df, for which $p = 0.0001$. It is extremely unlikely that such a difference could occur by chance, or sampling variability, alone, with no effect attributable to the intervention. So we reject the null hypothesis and conclude that the intervention has increased dietary intake by an estimated 32.1 µg per day, with a 95% CI of (19.5 µg, 44.6 µg).

Again, this *t*-test is based on the assumption that the change in vitamin A level is normally distributed. If this assumption cannot be made, a non-parametric counterpart test can be used—the Wilcoxon signed-rank test.

Correlation and regression

We now examine similar questions, when both variables being considered are continuous. This leads to some different measures of data presentation, and different techniques for assessing the significance and size of associations.

Scatter plots

Suppose we are interested in the relationship between birthweight and maternal weight prior to pregnancy, both continuous variables. An efficient method of displaying such a relationship visually is to use a scatter plot in which each point represents a particular mother (weight)–child (birthweight) pair. Figure 5.14.4 shows that heavier mothers tend to have heavier babies. Although such a pictorial representation of the relationship is useful, we usually summarize the relationship more concisely, with a ‘coefficient of association’.

Coefficients of association

Various coefficients of association have been developed for quantitative measures, with some common properties. These include the following:

A value of zero for the coefficient means no association.
A negative value for the coefficient means that one variable increases as the other decreases.
A positive value for the coefficient means that the variables increase or decrease together.
The range of the coefficient is from –1 to +1. Attainment of –1 or +1 is said to be ‘perfect association’.

For the relationship shown in Figure 5.14.4, we would expect a measure of association to be a positive value, somewhat less than 1.

Pearson correlation coefficient

The most commonly used measure of association is the Pearson correlation coefficient (r), used to describe the linear relationship between two continuous variables. If values of both variables increase together, r will be positive. If values of one variable increase as those of the other decrease, r will be negative. In the earlier example, 'Y' is birthweight and 'X' is maternal weight, corresponding to our scatter plot in Figure 5.14.4. The correlation coefficient is estimated as +0.50. Is this a 'high' or 'strong' correlation? As a rough guide, in biostatistics, 0–0.15 is usually regarded as a 'low' correlation, 0.16–0.4 as a 'modest' correlation, 0.41–0.7 a 'moderate' correlation, and above 0.7 a 'high' correlation, although context usually matters.

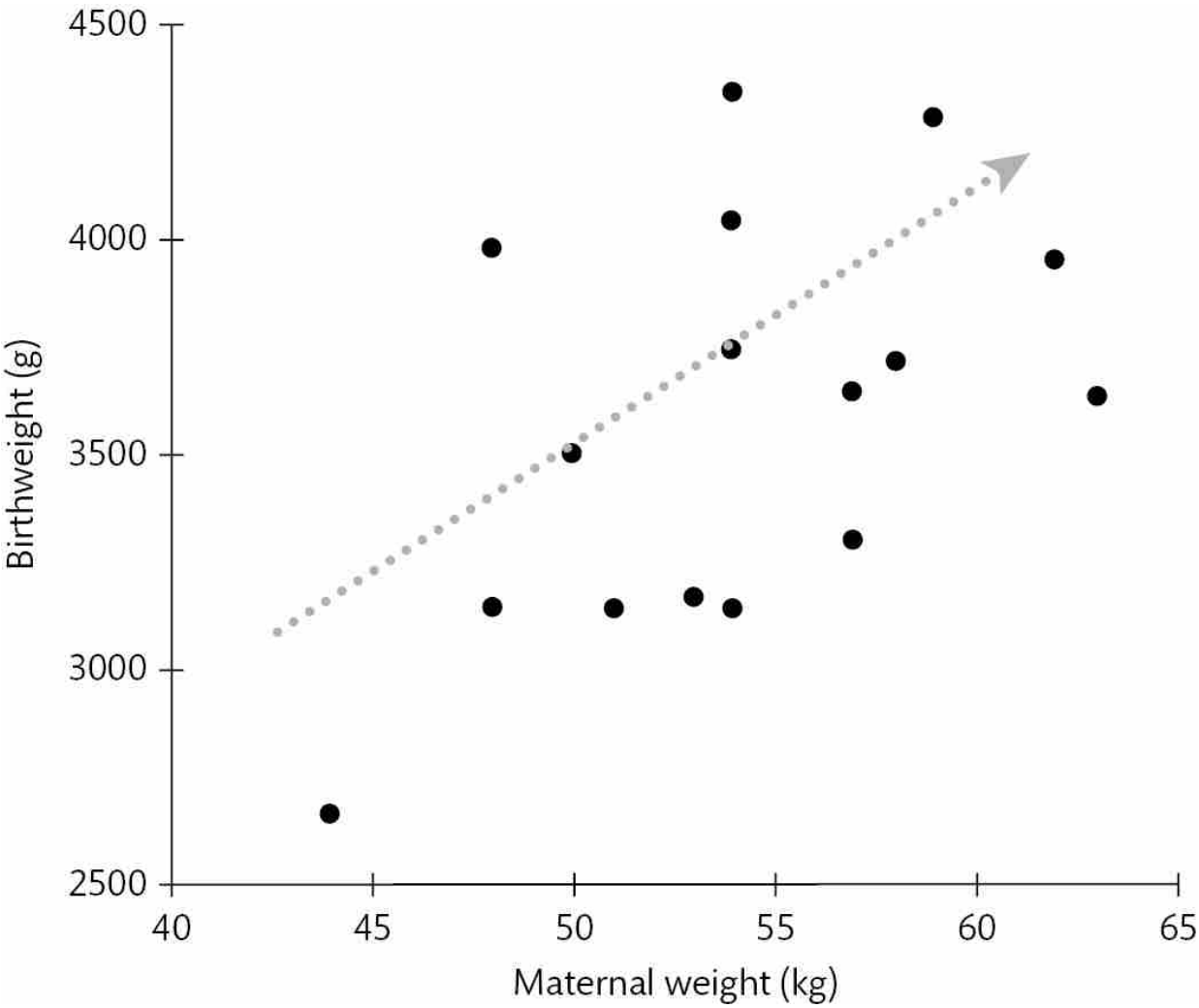


Fig. 5.14.4 Scatter plot and regression line of birthweight (g) and maternal pre-pregnancy weight (kg) fitted by least squares for a group of 16 term Brisbane babies, born in 1984–1988.

Source: data from *The Lancet*, Volume 354, Issue 9180, Green, A. et al., Daily sunscreen application and betacarotene supplementation in prevention of basal-cell and squamous-cell carcinoma of the skin; a randomised controlled trial, pp. 72–5, Copyright © 1999 Elsevier Ltd. All rights reserved.

Significance of the correlation coefficient

The value of 0.5 obtained may be considered moderate in magnitude. A separate issue is that of significance. This examines the role of chance or sampling variability in observing a coefficient of this size. We need to obtain the probability that a coefficient as large as 0.5 would be observed by chance in a sample of this size if maternal weight and birthweight had no true relationship (a true correlation coefficient of 0).

In the case of the Pearson correlation coefficient, the *p*-value is obtained directly from tables or statistical software. As before, we need to calculate the df parameter which depends on the sample size *N*. In this case, df = *N*—2, or 16—2 = 14. This gives $p = 0.048$, just less than the conventional cut-off. So the correlation coefficient could be described as 'significant' or, more completely, as 'significantly different from zero'.

Types of association and correlation

Figure 5.14.5 shows various types of association and how they would be described. It is important to note that the correlation coefficient measures only linear relationships; alternatively, we might say the correlation coefficient only measures that part of the relationship which is linear. A linear relationship is one in which a straight line approximates the relationship.

Note that the lower right graph in Figure 5.14.5 would give a correlation close to zero correlation for the Pearson coefficient. The reason for this is that the data are non-linear, even though there appears to be a relationship between the two variables. Thus, the Pearson coefficient only measures strength of a linear relationship. If we wanted to estimate strength of association for a non-linear relationship, we would have to specify a particular non-linear form.

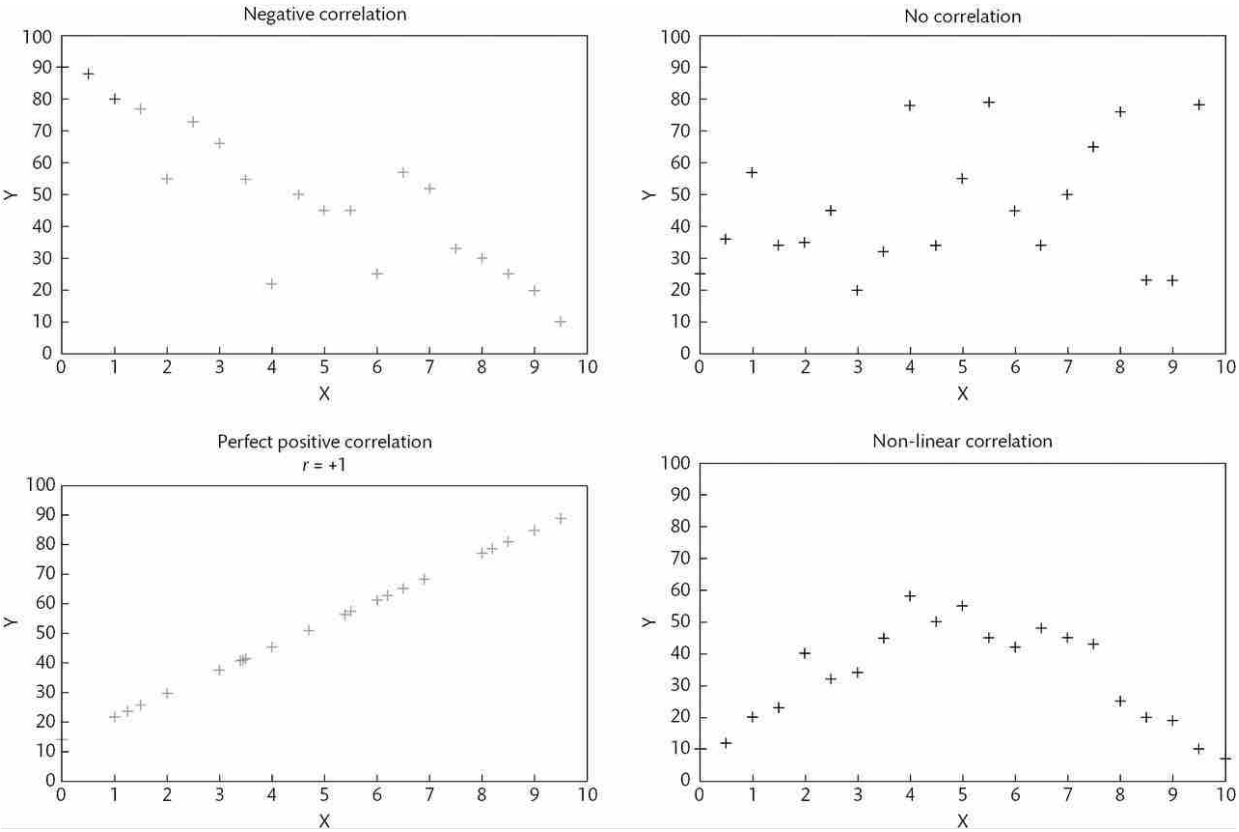


Fig. 5.14.5 Types of association.

Regression

To illustrate the regression approach to analysis, we use the data in Figure 5.14.4. We wish to 'model' the association between the two variables by fitting a straight line to the data, as follows. A linear model involving one dependent and one independent variable is expressed as:

$$Y = \alpha + \beta X + e.$$

In setting up a regression model, one variable, usually denoted Y, called the *dependent variable* (or *outcome*), is the variable to be predicted or modelled. The other variable, usually denoted X, is sometimes called the *independent variable* (or better, the *predictor* or *explanatory variable*) and is the variable used to predict or explain Y. The formula of the model assumes Y has a straight-line relationship with X. The symbols α and β represent the intercept

and the slope, respectively, of the line on a scatter plot of Y versus X. The parameter β is called the *regression coefficient*, and is a measure of how much Y changes for a unit (e.g. 1 kg of maternal weight) change in X; α is usually not of interest. The quantity e measures the 'error'—given that it is very unlikely such a simple model will describe data such as that in Figure 5.14.4, we allow for the 'error' to have a distribution. In this case, for each value of X, the errors are assumed to have a normal distribution about the predicted value $\alpha + \beta X$.

A mathematical algorithm is needed to estimate α and β—the simplest of these is the *method of least squares* which minimizes the squared differences between observed values and those predicted by the model.

For the birthweight data in Figure 5.14.4, we obtain the following model:

$$Y = 1{,}124 + 45.5X.$$

That is, birthweight increases by 45.5 g for every 1 kg of mother's weight. So mothers who differ by, say, 10 kg in their pre-pregnancy weight would expect to have babies, on average, who differ by 455 g in birthweight. Standard errors and CIs can be calculated for a regression coefficient. In this case, the 95% CI for the regression coefficient is (4.5, 86.5). Although this CI is quite wide, it excludes the value 0, which would correspond to no relationship between mother's weight and child's birthweight (a horizontal line on Figure 5.14.4).

We can also formally test for the statistical significance of the regression coefficient, using statistical software. For this example $p = 0.03$, which is small enough to conclude the relationship is significant.

The model can be used for prediction. If a pregnant woman has a pre-pregnancy weight of 50 kg, we can predict that her infant will have a birthweight of $Y = 1{,}124 + (45.5 \times 50) = 3{,}399$ g.

Assumptions of correlation and regression analyses

Correlation and regression coefficients are estimated on an assumption of a linear association. To make inferences (test hypotheses or calculate CIs), however, you need to make additional assumptions. These are as follows:

For correlation analyses, the two variables involved have joint normal distributions.
For regression analyses, the outcome (dependent variable) has a normal distribution.
For correlation and regression analyses, the samples are independent observations; that is, all pairs of observations are independent.

We have already established that the distribution of birthweight is quite close to a normal distribution, so the second assumption needed for regression analysis is met. What about the others? What do we do if the assumptions are not met? Again there are two main options: transformations and non-parametric correlation coefficients.

Transformations: transformations can be used, as described in the subsection on *t*-tests and ANOVA.
Non-parametric correlation coefficients: if a transformation cannot be found, it is possible to use a non-parametric test instead. These do not require either of the two distribution assumptions mentioned here. Instead of a Pearson correlation, we use the Kendall (τ or tau) or Spearman correlation coefficient (r_s).

The successful fitting of a straight line to the data does not establish causality. Although we predict that babies whose mothers' weights differ by 1 kg will have birthweights differing on average by 45.5 g, we cannot predict that increasing a mother's weight by 1 kg will cause an increase of 45.5 g in her child's birthweight. Observing an association between an outcome and a potential risk factor is, however, the first step in examining causality; we then have to consider possible explanations, which include chance, bias, and confounding. The results of a significance test enable us to assess the role of chance, but bias and confounding are crucial to consider. Multivariable statistical modelling is one approach to examining confounding; and is dealt with later in this chapter.

Survival data

In a cohort study, if an event or disease of interest is likely to occur in most participants, then interest centres on the actual time till the event, rather than

the occurrence of the event itself. Examples include time to death following a diagnosis of breast cancer, time to remission of disease following treatment, or time to cessation of breastfeeding, in a cohort of infants. Such data are referred to as *survival data* or *failure time data*. A person's 'survival time' is his or her time from entry in the study until the occurrence of the event. The distribution of survival times is often very skewed and may be bimodal (e.g. if significant postoperative mortality is followed by increased then gradually decreasing prospects of survival).

Survival data may be *censored* if there are participants who do not experience an event because the study ends before this occurs, are 'lost to follow-up' (because of, e.g. migration or death), experience another type of event which precludes the event of interest (e.g. death from another cause), or cease to be eligible for the study (e.g. in a study of under-5s, children are ineligible after their fifth birthday). In such instances, the person's survival time cannot be known precisely, only that it is *at least* a certain amount. *Survival analysis* enables use of all of a person's 'at-risk' time, irrespective of whether an event is experienced. An additional feature is that it allows the incidence rate to vary for the entire period of observation.

The Worcester Heart Attack Study (WHAS) examined factors associated with survival rates following acute myocardial infarction (MI) (Tonne et al. 2005). Data were collected from 1975 to 2001, on 500 MI patients' hospital admissions in the Worcester, Massachusetts Standard Metropolitan Statistical Area. Data analysis focuses on time from admission to death, or time from admission to the last-follow-up, if the patients were still alive. Of the 500 patients, 215 (43%) died within their follow-up period.

The analytical approach is to divide the follow-up time into short periods, called *risk sets*, and count events that occur within each risk set, as a fraction of all people who were in the study for that period. The latter is particularly important: people will drop out of the analysis because they have experienced the event (died) or had finished their follow-up period. These fractions are then manipulated in a process attributed to Kaplan and Meier (1958) to form a curve that shows the estimated probability of survival to that time. Results for fitting this curve to the WHAS data overall and to two subgroups (with and without complications at admission) are shown in Figure 5.14.6, with 95% CIs.

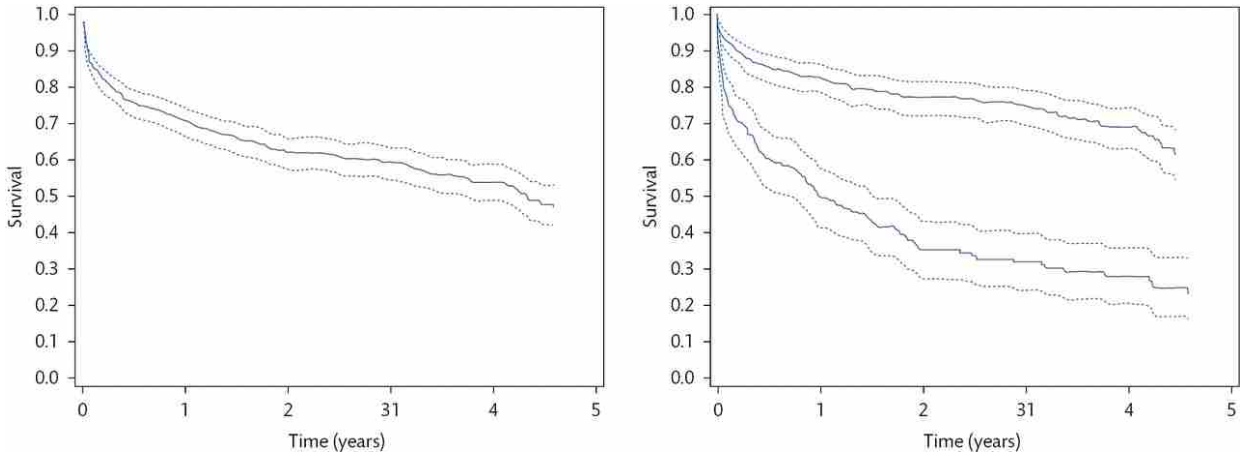


Fig. 5.14.6 Survival curves, with 95% CIs for patients with MI, overall and according to presence or absence of complications at time of admission (Worcester Heart Attack Study). Those without complications shown in the upper graph on the right.

Source: data from Tonne C, et al., Long-term survival after acute myocardial infarction is lower in more deprived neighborhoods, *Circulation*, Volume 111, pp. 3063–70, Copyright © 2005 American Heart Association, Inc. All rights reserved.

Notice that mortality is highest immediately after the initial MI, with only about 70% surviving the first year, but flattens a little at about 2 years. This analysis yields more information than simply examining mortality rates. The survival curve can also be used to estimate the median survival, the length of time corresponding to 50% survival: project from a horizontal line at survival = 0.5 to the horizontal axis to give an estimated 50% of people survive to 5 years or more (i.e. the 5-year survival is 50%). Other quantiles can similarly be estimated. Kaplan–Meier survival curves can be formally compared using the Wilcoxon or logrank tests (Martinez and Naranjo 2010; Doi and Williams 2013).

Additional comments on censoring

Whenever a participant experiences an event which precludes them from experiencing the event of interest, they are censored at that point. For example, if, in the WHAS, the event of interest is death from cardiovascular disease, then a patient who experiences death from cancer is considered censored at the time of death. We also assume that censoring is independent —that is, that people with censored observations have the same future risk as those that are not censored. This is quite a strong assumption. For example,

suppose we are examining survival of patients admitted to an intensive care unit. If follow-up ceases at discharge then, given that those discharged are more likely to have a better prognosis than those not discharged, censored observations would be likely to have a lower risk. If drop-outs differ from non-drop-outs this is also likely to lead to dependent censoring.

A specific example of this occurs in ‘competing risk’ models. In the example given earlier for death from cardiovascular disease, and censoring by death from cancer, we need to consider whether a person who dies from cancer would have the same future risk (had she/he not died from cancer) of death from cardiovascular disease. This may seem plausible. However if, in a series of patients with obesity, the event of interest is death from cardiovascular disease, death from diabetes would be a competing event, since it is likely that risks of these outcomes would be associated. In such instances it is argued that the methods described earlier would result in biased estimates of survival. The appropriate analyses are complex, essentially involving the joint analysis of the cumulative incidence of both the event of interest and any competing event(s)’ (Doi and Williams 2013).

Multivariable analysis

This section deals with the situation where we have more than one variable predicting an outcome variable. The term ‘multivariable’ is used in preference to ‘multivariate’ which is commonly also used to describe this situation. Classical statistical usage for the term ‘multivariate’ is in describing analyses which are required to predict more than one outcome variable simultaneously, such as when one has a set of items purporting to measure a common dimension, for example, depression or anxiety, or when measurement of outcomes is repeated over time. We consider various models which arise, as a consequence of different outcome variables or study designs. As we shall see, these models share a common framework, the generalized linear model, which is a generalization of the simple regression model discussed earlier; models differ only where necessary to incorporate their different assumptions. We begin with models based on the normal distribution.

Normal theory regression

Having established that birthweight and maternal weight are associated, and having estimated the regression coefficient, we might be interested in further exploring the prediction of birth weight. We could perform the same analyses on a range of other variables, such as family income, tobacco, or alcohol consumption. In each case, a regression model provides us with a measure of the size of the effect (the regression coefficient, at a 95% CI) and its statistical significance (*p*-value). In this way, we might identify a collection of predictors or 'risk factors'. It is likely that these predictors will themselves be related: consider the following example, now based on a sample of 500 births.

Regression of birthweight upon maternal pre-pregnancy weight yields:

Birthweight (g) = 2,832 + 11.3 × maternal weight (kg).
Birthweight increases by 11.3 g per kg of maternal weight.

Regression of birthweight upon number of previous pregnancies yields:

Birth weight (g) = 3,453 + 29.5 × number of previous pregnancies.
Birth weight increases by 29.5 g per previous pregnancy.

But if a woman's weight increases with increasing number of pregnancies, then these two findings might be related; for example, does birthweight increase with increasing number of pregnancies because of the increasing weight of the mother, or is the effect separate?

To examine this, we set up a multivariable linear model such as:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + e,$$

where β_1 and β_2 are regression coefficients indicating the dependence of Y upon variables X_1 and X_2, X_1 being maternal weight and X_2 being the number of previous pregnancies. This model is fitted in the same way as before, to obtain:

Birth weight (g) = 2,830 + 11.0 × maternal weight (kg) + 16.9 × number of previous pregnancies.

We see that the maternal weight effect remains virtually the same, but the effect of number of previous pregnancies has about halved, and is no longer significant. To see why this has happened, we use our multivariate model to

assess the effect of a difference in maternal weight of 1 kg on birthweight. Suppose a woman A has a pre-pregnancy weight of 50 kg and one previous pregnancy. Her baby's predicted birthweight (g) is 2,830 + 11.0 × 50 (kg) + 16.9 × 1.

Suppose a second women B has a pre-pregnancy weight of 49 kg and no previous pregnancies. The latter has an effect on birth weight, so B is not comparable to A. The prediction model can make A and B comparable in terms of pregnancy history, by setting B to have one previous pregnancy. The predicted birthweight for B's baby (if she had one previous pregnancy) is 2,830 + 11.0 × 49 (kg) + 16.9 × 1.

Subtracting these two equations to obtain the difference in predicted birthweight for these two women, gives a predicted difference of 11.0 g, the regression coefficient for maternal weight in the multivariable model. Thus, the regression coefficient of 11.0 g represents the effect of an additional 1 kg of maternal weight, after controlling (or adjusting) for number of previous pregnancies.

Reciprocally, the regression coefficient for number of previous pregnancies reflects the association between birthweight and number of previous pregnancies, after adjusting for the effect of maternal weight in the multivariate model. As this effect has been much decreased to the point of non-significance, we would conclude that the significant effect we saw in the first separate analysis was due to confounding by maternal weight.

A *confounding variable* is one that is associated with both the outcome variable and a predictor variable. In summary, a multivariable regression model yields estimates of the separate or 'adjusted' relationships between the outcome variable and the predictor variables.

Essentially, no new principles emerge when this process is extended to more predictors or risk factors. We obtain a series of estimated regression coefficients, and confidence limits for the parameters, and *p*-values, each reflecting the association between the outcome and the predictor, after adjusting for all other predictors included in the fitted model.

If an outcome variable is continuous, but the normal distribution for errors does not apply, the outcome variable can sometimes be transformed, so that the normal assumption is satisfied. Typical transformations include the log or square root transform, which can correct skewness. While these are useful statistical devices, they can sometimes make results more difficult to interpret.

Logistic regression

Logistic regression is used in similar situations to normal theory regression, but where the outcome variable is binary (e.g. dead/alive, having a specific disease or not) and we want to model the risk of this outcome (death, diseased) according to certain predictors or risk factors. The difference is that we use a transformation, called the *logit* transformation, to account for the binary nature of the variable. This is seen on the left hand side of the following formula:

$$\log(p) = \log(p/(1-p)) = \alpha + \beta_1 X_1 + \beta_2 X_2 + e,$$

where log is the logarithm function to base e, p is the predicted proportion of people with the outcome of interest, given particular values of the predictor variables X_1 and X_2. Note that the left hand side is now modelling the log(odds of outcome), where odds of outcome may be estimated by the number with the outcome, divided by the number without the outcome. The error distribution on the right hand side (e) is assumed to follow a binomial distribution, which is a probability distribution describing the number of times an outcome occurs, in a set number of individuals at risk. The regression coefficients β_1 and β_2 now need to be re-interpreted in the light of the transformation. When the model is fitted to data, and the regression coefficients back-transformed by exponentiating (taking anti-logs), we obtain odds ratios for the association between the outcome and the risk factor. The odds ratio is the odds of the outcome (defined earlier) within those who have the risk factor, divided by the odds of the outcome within those who do not have the risk factor. It is a widely used measure of association between a risk factor and an outcome. An odds ratio of one indicates no association between risk factor and outcome; odds ratios of greater than one or less than one indicate an increased or decreased risk of the outcome associated with the risk factor, respectively.

As an example, suppose we are interested in examining the predictors of overweight in adolescent girls, using data on 1,823 14-year-old girls born into a cohort study (Najman et al. 2005). Of these we note that 466, or 26%, are overweight at age 14, and also that being overweight at age 14 is associated with parental weight at the child's birth. A girl was more likely to

be overweight at 14 if her mother was overweight at the start of the pregnancy (51% of daughters of overweight mothers were overweight, compared to 21% of non-overweight mothers). Calculating from first principles, the odds ratio for the association between mothers and daughters being overweight is (0.51/(1 − 0.51))/(0.21/(1 − 0.21)) = 3.9. A similar, but not quite so marked, result was noted for fathers'- weight: 39% of daughters of overweight fathers were overweight, compared to 21% of daughters of non-overweight fathers. These data are displayed in Table 5.14.11.

In addition, girls who were overweight at age 14 had a mean birthweight of 3.4 kg, compared to 3.3 kg for those who were not. We could surmise (and can conform by additional analyses) that the mother's pre-pregnant weight, the father's weight at the time of pregnancy, and the daughter's birthweight are likely to be positively associated, resulting in some confounding of their effects. In order to sort out the separate contributions of these three predictors we use a logistic regression model as follows. We would typically begin with a model which uses one risk factor at a time, then a multivariable model, or series of them which examines the effects of adjusting for other risk factors. As before, effects obtained from multivariable models are adjusted for all other variables (risk factors) in the model. When this is done for this example, we find the various effect estimates and 95% CIs as shown in Table 5.14.12.

Table 5.14.11 Association between maternal pre-pregnant overweight status and daughters' overweight status at age 14

Risk factor		*N*	Daughters overweight	Odds ratio (OR) for daughter overweight and risk factor (1 = reference category)
Mother overweight	No	1,545	324 (21%)	Reference
	Yes	278	142 (51%)	3.9
Father overweight	No	1,381	293 (21%)	Reference
	Yes	442	173 (39%)	2.4

Source: data from Mamun, A.A., Lawlor, D.A., O'Callaghan, M.J. et al. Family and early life factors associated with changes in overweight status between ages 5 and 14 years: findings from the Mater University Study of Pregnancy and its outcomes. *International Journal of Obesity (London)*, 29(5):475–82.

Table 5.14.12 Logistic regression: predicting overweight in adolescent girls from maternal and paternal overweight status prior to pregnancy, and birthweight

Risk factor		Odds ratio (OR) for daughter overweight and risk factor (95% confidence intervals)		
Unadjusted	**Adjusted for other parental overweight**	**Adjusted for other parental weight and birthweight**		
Mother overweight	No	Reference	Reference	Reference
	Yes	3.9 (3.0–5.1)	3.5 (2.7–4.6)	3.4 (2.6–4.5)
Father overweight	No	Reference	Reference	Reference
	Yes	2.4 (1.9–3.0)	2.1 (1.6–2.6)	2.1 (1.6–2.6)
Birth weight	(kg)	1.4 (1.1–1.7)		1.2 (1.0–1.5)

Source: data from Mamun, A.A., Lawlor, D.A., O'Callaghan, M.J. et al. Family and early life factors associated with changes in overweight status between ages 5 and 14 years: findings from the Mater University Study of Pregnancy and its outcomes. *International Journal of Obesity (London)*, 29(5):475–82.

Our multivariable analyses (last two columns) show some reduction in effect estimates after adjustment, indicating that the effects of maternal and paternal overweight are largely separate effects. Adjusting for birthweight makes very little difference to the magnitude of parental effects. Notice that the effect of birthweight, as a continuous variable, is expressed differently to the categorical variables of parental overweight. The odds ratio for birthweight gives the increase in odds of the daughter being overweight at age 14, for a unit (1 kg) increase in birthweight, that is, a child who was 4 kg at birth is predicted to have 40% increase in odds of being overweight at 14, compared to a child who was 3 kg at birth. This same relative increase applies when comparing a 3.5 kg child with a 2.5 kg child. When we include continuous variables in a logistic model we assume that the relative effect, per unit change in the variable, is constant. This is an assumption that should be checked.

Poisson regression

Poisson regression (named after the French mathematician who first described a probability distribution for count data) is used when the outcome variable is a count variable (e.g. number of deaths over a period, or number of incident cases of disease). There two differences, compared to normal theory regression. The first is that we apply a log transformation to our count data, and the second is that we (usually) have to incorporate an 'offset' variable' which measures the period of time at risk, over which the events are accumulated. This model is expressed according to the following formula:

$$\log(\text{number of events}) = \alpha + \beta_1 + X_1 + \beta_2 + X_2 + \log(\text{person} - \text{time at risk}) + e$$

or, equivalently,

$$\log(\text{event rate}) = \alpha + \beta_1 X_1 + \beta_2 X_2 + e.$$

The error distribution on the right hand side (e) is assumed to follow a Poisson distribution, which is a probability distribution describing a 'count' of a number of independent events. As for logistic regression, the regression coefficients β_1 and β_2 need to be re-interpreted in the light of the log transformation. When the model is fitted to data, and the regression coefficients back-transformed by exponentiating (taking anti-logs), we obtain rate ratios for the association between the outcome and the risk factor. The rate ratio in this context is the rate of the outcome (as a person-time rate) within those who have the risk factor, divided by the rate of the outcome within those who do not have the risk factor. Again a rate ratio of one indicates no association between risk factor and outcome; rate ratios of greater than one or less than one indicate an increased or decreased risk of the outcome associated with the risk factor, respectively.

As an example, consider the results of a Poisson regression applied to examining risk factors for the occurrence of a particular form of skin cancer (squamous cell carcinoma (SCC)) in a cohort study. Because participants dropped out of the study at different stages throughout the 12-year period of follow-up, it was necessary to take account of each individual's person-time of follow-up. This leads to the offset variable described earlier. The outcome of this analysis was whether or not a participant had a new SCC lesion

detected in the follow-up period. Note that an individual only appears once in the analysis. Subsequent new SCC lesions did not contribute. This is an important assumption of Poisson regression, that is, that the outcomes are independent of each other—it is likely that SCCs will tend to cluster within individuals, and therefore subsequent events within the same individual will not be independent. An alternative model (e.g. negative-binomial regression) must be used if one wants to consider multiple events within an individual. Variables to be considered in a multivariable analysis for predicting SCC included sex, age, skin colour, history of skin cancer, and skin response to exposure to the sun. The results of this analysis are shown in Table 5.14.13.

The first column of rate ratios indicates the increase in risks associated with each risk factor (and their 95% CIs), taken one at a time. The second column of rate ratios shows the increases in risks when each risk factor is adjusted for the effects of the others, and any associations among risk factors. We see that the general tendency is for rate ratios to reduce, after adjustment for confounding. Most notably, the rate ratios (RR) associated with a previous history of skin cancer drops from 5.08 (3.55–7.20) to 2.50 (1.69–3.71) when adjusted for other risk factors; this is because factors (such as age, skin response to the sun) will be associated both with a past history and future risk of skin cancer. When these factors are taken into account, some of the effect of history of skin cancer, as a single factor is reduced. Note also that the age effect appears modest—RR = 1.06 (1.05–1.08)—but one needs to take into account that this is the increase in risk, per year of age. Over a decade of age this translates to a cumulative rate ratio of 1.84 (1.64–2.23).

Cox proportional hazards regression

The Cox proportional hazards model (Cox 1972) is a widely used method for multivariable analysis of survival data. It can be thought of as a combination of the Kaplan–Meier method which involves dividing the follow-up time into small intervals, and the Poisson regression model approach which formulates a model within each of these time periods.

Consider the WHAS data referred to in Figure 5.14.6. Suppose our follow-up time is divided into intervals determined by the time points t_1, t_2, t_3, etc. We imagine a regression model within each of these intervals (t_1, t_2)

(t_2,t_3) etc., which relates the probability of death within the interval to a risk factor, such as complications of a myocardial infarct. From such a model we can obtain an effect estimate (a relative risk of death in this case, associated with having complications) for that time interval. This relative risk would vary over time. However, we begin by making a simplifying assumption that the risk ratio (usually termed hazard ratio in this context) is constant over time—this is called the proportional hazards assumption.

We now consider a Poisson model', to cover the time period from t to t + 1, as follows (Doi and Williams 2013).

$$\log(\text{event rate at time t}) = \alpha + \beta_1 X_1 + \beta_2 X_2 + e.$$

Conceptually, this model is then fitted for each time period (with the constraint that the estimates β_1 and β_2 are, the same for each period). This occurs by a process called partial likelihood. The resultant parameter estimates the log of the hazard ratio.

Table 5.14.13 Comparison of unadjusted and multivariable Poisson regression for risk factors for SCC: rate ratios

Risk factor		Rate ratios (RR) for incidence of SCC (95% confidence intervals)	
Unadjusted	**Multivariable analysis**		
Sex	Female	Reference	Reference
	Male	1.53 (2.17–5.72)	1.39 (0.98–1.99)
Age	Years	1.08 (1.06–1.10)	1.06 (1.05–1.08)
Skin colour	Olive	Reference	Reference
	Medium	1.33 (0.52–3.40)	1.22 (0.47–3.14)
	Fair	2.11 (0.86–5.21)	1.60 (0.63–4.07)
History of skin cancer	No	Reference	Reference
	Yes	5.08 (3.55–7.20)	2.50 (1.69–3.71)
Skin response to sun	Tans	Reference	Reference
	Burns, then tans	1.13 (0.59–2.20)	1.45 (0.74–2.86)
	Burns, never tans	2.33 (1.17–4.62)	2.12 (1.02–4.39)

Source: data from *The Lancet*, Volume 354, Issue 9180, Green, A. et al., Daily sunscreen application and betacarotene supplementation in prevention of basal-cell and squamous-cell carcinoma of the skin; a randomised controlled trial, pp. 72–5, Copyright © 1999 Elsevier Ltd. All rights reserved.

The usual form of the model expresses the ‘intercept’ as a log of a ‘baseline hazard’, this being the predicted death rate for the reference category of the predictor variable:

$$\log(\text{event rate at time } t) = \log(\text{event rate at time } t \text{ for reference categories for } X_1 \text{ and } X_2) + \beta_1 X_1 + \beta_2 X_2 + e.$$

Thus the hazard ratios β_1 and β_2 give the multipliers of the baseline hazard associated with being in particular risk groups. You can see that the hazard

ratios ‘multiply’, since the aforementioned model, after exponentiating both sides becomes:

$$\text{Event rate at time t} = \text{event rate at time t for reference categories for } X_1 \text{ and } X_2 \times \beta_1 X_1 \times \beta_2 X_2.$$

When we fit a Cox model to the WHAS data, using the predictors of patient’s sex, age (years), and presence of complications, we find the following, as shown in Table 5.14.14.

The ‘unadjusted’ hazard ratios represent separate models, one at a time. The multivariable analysis gives hazard ratios which are mutually adjusted for other variables in the model. Thus we see the sex effect reduces when age and complications are taken into account. The magnitude of the complications effect reduces but remains quite strong.

The proportional hazards assumption and time-dependent variables

As noted earlier, the formulation of the aforementioned model assumes a constant hazard ratio. We can formally test this assumption and alter the model to accommodate some variation in the hazard ratio. We can also include time-dependent risk factors. For example if we want to include alcohol intake as a risk factor, we could include alcohol intake measured at baseline (time constant) or at any point in time, when it might change (time-dependent). While the process for doing this is a little complicated, no new concepts emerge in terms of interpreting results.

Variations and extensions to other probability models

Logistic regression has been extended to the situation where one has a categorical outcome with more than two outcomes. This is known as a multinomial (as opposed to a binomial) model and leads to *multinomial regression*. A standard multinomial model will require more than just a single odds ratio parameter associated with each risk factor level, since now there are more outcome levels to deal with. For example, suppose an

outcome variable has three levels (no disease, moderate-level severity disease, severe disease), and we are evaluating a binary risk factor (history of tobacco use, no history). A multinomial model will involve an odds ratio quantifying the increased risk associated with smoking for moderate-level severity disease, and another quantifying the increased risk for severe disease. If the outcome variable is an ordinal variable (such as in the current example), then the multinomial model can be simplified by assuming that the odds ratios associated with moving from no disease to moderate-level severity disease, and from moderate disease to severe disease are the same. This is referred to as *cumulative logit regression* or *ordinal regression*. The assumption of a common odds ratio is called a proportional odds assumption and should be checked for validity as it is quite a strong assumption. It is sometimes considered that a disadvantage of logistic regression is that it results in odds ratios, which are often not appropriate measures of relative risks, particularly in dealing with outcomes which occur commonly. *Log-binomial regression* has been proposed to deal with this situation, since it assumes a binomial distribution of errors (as does logistic regression) but the use of a log rather than logit distribution results in estimated effects which are interpreted as risk ratios. However, fitting these models can sometimes be technically problematic.

Table 5.14.14 Cox regression for mortality after a myocardial infarct by patient age at MI, sex, and presence/absence of congestive heart failure (CHF) complications

Risk factor		Hazard ratios (RR) for death (95% confidence intervals)	
Unadjusted	**Multivariable analysis**		
Sex	Female	Reference	Reference
	Male	1.46 (1.12–1.92)	0.86 (0.65–1.13)
Age	Years	1.07 (1.06–1.08)	1.06 (1.05–1.08)
Complications	No	Reference	Reference
	Yes	3.31 (2.53–4.34)	2.41 (1.82–3.19)

Source: data from Tonne, C., Schwartz, J., Mittleman, M., Melly, S., Suh, H., and Goldberg, R. (2005). Long-term survival after acute myocardial infarction is lower in more deprived neighborhoods. *Circulation*, 111, 3063–70.

As mentioned earlier, Poisson regression assumes that the outcomes of interest are independent. There are many instances when this is not the case. In the SCC example given earlier, individuals will probably have a propensity to develop SCC which is not fully explained by the risk factors we include in a model, leading to a failure of the assumption of independence. In modelling the occurrence of episodes of an infectious disease, this will also be a problem—while we can 'explain' some of the aggregation of episodes within individuals by our predictor variables (nutritional status, environmental exposure, etc.), we are unlikely to explain this completely, as unmeasured and unknown factors will affect the risk of an episode. One of the most common alternatives to use is *negative-binomial regression* which posits a somewhat more complex mechanism for development of an episode, that is that each individual has a particular propensity for the outcome (e.g. a skin cancer), and that given this propensity, an individual's number of episodes will have a Poisson distribution. Once this more complex process is taken into account, the results of a negative-binomial distribution are very similar to those of a Poisson distribution, that is, one obtains RR for episodes of the outcome of interest. Other alternatives are also used, which include *zero-inflated*

regression models. These are based on models such as the Poisson or negative-binomial but particularly assume that there are more values of zero counts (numbers of episodes) than would be predicted by those models. This would be appropriate in the case of a vector-borne disease, for example, where the population may include a large subpopulation who have no exposure to the relevant vector, and a smaller subpopulation who have varying, but non-zero, exposure.

Study designs which have a hierarchical nature often use *random effect regression* models. These may involve different types of probability models, but have the defining feature that they include factors (covariables) which are considered themselves to have distributions. For example, one may design a study to examine the relationship between air pollution and child health. A number of sites are chosen which are close to air pollution monitoring stations, and a sample of schools is selected within a defined radius of the monitoring sites. Children are chosen from each school and their lung function measured. In this case a regression model with lung function as the outcome would be appropriate, and the primary risk factor would be the air pollution level, with perhaps other covariables (socioeconomic status, age, sex, etc.) also used in a multivariable model. However, variation among schools needs also to be taken into account. Given that we had taken a sample of schools, we would regard this effect as random, in the sense that we want to regard the study schools as representative of all schools. This contrasts with the more usual ‘fixed’ effects, such as for smoking status where we have levels of non-smoking or smoking and there is no sense in which we have a sample of levels of smoking. A consequence of a random effects regression model is that we can allow for effects (such as the mean lung function or the regression coefficient linking lung function and air pollution) to vary by school. Our model than examines the variability of these effects and estimates a mean across schools. Thus one has within-school effects and between-school effects.

The multivariable models so far considered have the common property that they either consider individual records to be independent, or, as in the negative-binomial or the random effects model, incorporate non-independence into the probability model itself. Non-independences occur in many study designs. These include repeated measures designs, hierarchical designs, or spatial designs. Regression models applied to such designs

incorporate a non-independence correlation structure which allows the correlations between measures within the same individuals over time, or within the same subpopulation, or the same spatial unit, to be estimated and taken into account. Depending on the software used, a variety of options is available for this structure, such as a constant correlation, an exponentially decaying correlation over time or space, and others. Complex designs are often fitted using the method of *generalized estimating equations* (GEEs), which is an extremely flexible approach to fitting models, and has an added advantage of coping somewhat better with missing data than classical fitting approaches.

Generalized additive regression models (GAMs) have been used particularly to explore non-linearity in effects. A simple regression model assumes, for example, that the relationship between a continuous outcome variable (birthweight) and a predictor (maternal weight) is linear—the increase in birthweight per unit of maternal weight is constant over the range of maternal weight. This is a strong assumption and will often need to be explored. A GAM is constructed in a very similar fashion to a generalized linear model except that one may specify a predictor as a collection of curvilinear functions fitted over various subranges of the predictor variable. The complexity of the non-linearity may be specified. This is a very sensitive and flexible approach to exploring curvilinearity of effects, or in the case of a confounding variable which has non-linear effects, ensuring minimal residual confounding.

Summary

Modern statistics and software offer a wide array of approaches to statistical analyses. The key to appropriate and efficient analysis lies in clarity around the study design and the questions to be answered, thorough preliminary analyses of the data available and selection of the appropriate techniques to be used, and a critical approach to presentation and interpretation of results. Common errors include applying techniques that are not appropriate to variable types or their distributions, ignoring design features (such as repeated or hierarchical sampling), and misinterpretation of computer output. Unfortunately, the availability of statistical software that can perform sophisticated analyses is likely to make these errors more frequent, rather

than the opposite. There is little substitute for consulting a qualified statistician, particularly with complex study designs.

ACKNOWLEDGEMENTS

Text extracts reproduced from Springer, *Methods of Clinical Epidemiology*, Chapter 10 Modelling Binary Outcomes and Chapter 11 Modelling Time-to-Event Data, pp. 173–7, Gail M. Williams and Robert Ware, Copyright © Springer-Verlag Berlin Heidelberg 2013. With kind permission of Springer Science + Business Media.

REFERENCES

Cox, D.R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society*, 34, 187–220.

Doi, S.A.R. and Williams, G.M. (eds.) (2013). *Methods of Clinical Epidemiology*. Springer Series on Epidemiology and Public Health. Germany: Springer, Verlag.

Green, A., et al. (1999). Daily sunscreen application and betacarotene supplementation in prevention of basal-cell and squamous-cell carcinoma of the skin; a randomised controlled trial. *The Lancet*, 354 (9180), 72–5.

Kaplan, E.L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *American Statistical Association*, 53, 457–81.

Martinez, R.L.M.C. and Naranjo, J.D. (2010). A pretest for choosing between logrank and Wilcoxon tests in the two-sample problem. *METRON—International Journal of Statistics*, LXVIII(2), 111–25.

Najman, J.M., Bor, W., O'Callaghan, M., Williams, G.M., Aird, R., and Shuttlewood, G. (2005). Cohort profile: the Mater University of Queensland Study of Pregnancy (MUSP). *International Journal of Epidemiology*, 34, 992–7.

Tonne, C., Schwartz, J., Mittleman, M., Melly, S., Suh, H., and Goldberg, R. (2005). Long-term survival after acute myocardial infarction is lower in more deprived neighborhoods. *Circulation*, 111, 3063–70.

5.15

Measuring the health of populations

The Global Burden of Disease study methods

Theo Vos, Christopher J.L. Murray and Alan D. Lopez

Introduction to measuring the health of populations

Over the last two decades, the global health landscape has undergone rapid transformation. People around the world are living longer than ever before, and populations are getting older. Many countries have made remarkable progress in preventing child deaths. As a result, disease burden is increasingly defined by disability as opposed to being dominated by premature mortality. The leading causes of death and disability are shifting from communicable diseases in children to non-communicable diseases in adults. These global trends differ across regions. Notably, in sub-Saharan Africa, communicable, maternal, and newborn diseases and nutritional deficiencies continue to dominate. While low- and middle-income countries are tackling this 'unfinished agenda' of largely poverty-related diseases, increasingly they also need to prepare their health services for a growing burden of non-communicable diseases and injuries, which some have called the 'double burden' or 'triple burden' (Bradshaw et al. 2003; Marshall 2004; Boutayeb 2006). In high-income countries, health budgets are steadily increasing relative to gross domestic product due to ageing of the population, an ever-expanding array of medical technologies, and greater demands of consumers for healthcare services (Lorenzoni et al. 2014; Richardson 2014; Keehan et al. 2015). For governments and other healthcare providers to be able to respond to these challenges, high-quality comparable data on the size and trends in mortality and morbidity are essential. The word 'comparable' is of paramount importance. Commercial and professional interests often lead to implausible claims of the number of deaths or cases affected by a

particular disease. The only way to deal with these exaggerated claims is to make a comprehensive assessment of the health of each population by ensuring that all deaths by cause add up to the total number of deaths and a similar rigour is applied to classifying cases of disease. The successive iterations of the Global Burden of Disease study (GBD—a full list of abbreviations is given in Table 5.15.2) (Murray et al. 1996; Mathers et al. 2008; Murray et al. 2012) are the only global effort to estimate health loss across the age spectrum and for all diseases and injuries. After the initial GBD assessment for 1990, updates until 2008 largely continued to use the same methods, with piecemeal improvements. In 2007, the Bill & Melinda Gates Foundation funded the Global Burden of Diseases, Injuries, and Risk Factors Study 2010 (GBD 2010), which endeavoured to rethink methods and assumptions underlying population health measurement while making use of the vastly improved health data and computational resources. GBD 2013, GBD 2015, and GBD 2016 are updates of the GBD 2010 study. Going forward, every year all estimates are updated to reflect the time period from 1990 to the most recent complete calendar year: GBD 2017 is due for release in October of 2018.

About GBD

The GBD approach is a systematic, scientific effort to quantify the comparative magnitude of health loss due to diseases, injuries, and risk factors by age, sex, and geography for specific points in time (Table 5.15.1). Since GBD 2010, every GBD update has been published in a series of papers in *The Lancet*. The information is made available as a global public good that will be useful for informing the design of health systems and the creation of public health policy. GBD 2016 estimated premature death and disability from 333 diseases and injuries, 2,982 sequelae (direct consequences of disease and injury), and 84 risk factors for 23 age groups and both sexes for the years 1990 through to 2016. The study produced many billion estimates of health outcomes.

Table 5.15.1 Number of diseases and injuries, risk factors, sequelae, age groups, sexes, geographical locations, and years of estimation for successive iterations of the Global Burden of Disease study

	1990	2000–04	2010	2013	2015	2016
Diseases and injuries	107	136	291	306	315	333
Risk factors	10	25	67	79	79	84
Sequelae	483	500+	1,160	2,337	2,619	2,982
Age groups	5	8	20	20	20	23
Sexes	2	2	2	2	2	2
Geographical locations	8	211	215	295	590	774
Years of estimation	1	1	21	24	26	27

The Institute for Health Metrics and Evaluation (IHME) at the University of Washington in Seattle has acted as the coordinating centre for the GBD study since 2007. GBD 2010 was a collaborative effort among 488 researchers from 50 countries and 303 institutions, while GBD 2016 involved the work of 2,518 researchers from 133 countries and three territories. The collaboration strengthens both the data-gathering effort and the quantitative analysis by bringing together some of the foremost minds from a wide range of disciplines. The annual updates ensure that the international community can have access to high-quality estimates in the timeliest fashion. Through sound measurement, it provides the foundational evidence that will lead to improved population health. Abbreviations are shown in Table 5.15.2.

Table 5.15.2 Abbreviations

CODEm	Cause of death ensemble model
DALY	Disability-adjusted life year
DFLE	Disability-free life expectancy
DW	Disability weight
GBD	Global Burden of Disease
HALE	Health-adjusted life expectancy or healthy life expectancy
ICD	International Classification of Diseases and Related Health Problems
IHME	Institute for Health Metrics and Evaluation
5q0	Probability of dying before fifth birthday
45q15	Probability of dying between 15th and 60th birthdays
UNAIDS	Joint United Nations Programme on HIV/AIDS
VA	Verbal autopsy
WHO	World Health Organization
YLD	Years lived with disability
YLL	Years of life lost

Value choices and population health metrics

Several metrics that combine mortality and morbidity are in use. They all have in common that health loss or health gain is measured in units of time. There are two classes of population health measures: health expectancies and health gap measures. Health expectancies extend the concept of life expectancy, a composite measure of age-specific mortality rates in a given year in a population, and adjust years in a life table for loss of health from non-fatal conditions. Demographers calculate life expectancy in a life table by applying the currently prevailing mortality rates in a population to a hypothetical birth cohort. In each row of the life table deaths and survivors of the birth cohort are tracked at successive age groups until a top age category (e.g. 80+ or 100+) in which everyone is assumed to die. In each age

group the average number of years lived is calculated and the sum of years lived over all ages divided by the initial size of the assumed birth cohort is the average life expectancy. A common interpretation of life expectancy is ‘the average number of years a child born today can expect to live’. Of course, this interpretation only holds if the currently prevailing mortality rates at each age would continue to apply for the total life span of the birth cohort. In reality, mortality rates are changing considerably over time. As mortality rates are declining at almost all ages in most countries, the true life expectancy in a population is higher than the figures that are frequently used in health statistics. However, it is a popular and frequently used statistic because it is an easy way to summarize in one number a lot of detailed information on mortality rates by age and because users have an intuitive understanding of how to interpret a life expectancy measure (even if that interpretation is not quite right).

Health expectancies are an adaptation of life expectancy taking into account that not all years of remaining life expectancy are lived in good health. The coarsest version of health expectancy is disability-free life expectancy (DFLE), which classifies disability in a dichotomous way: those above the threshold contribute no time to DFLE and those below contribute a full year to the measure (Murray et al. 2002). Health-adjusted life expectancy (HALE), also sometimes referred to as healthy life expectancy, grades the severity of disability across several categories or the full spectrum and adjusts time lived at each age in a life table by the severity-weighted prevalence of disability. The attraction of health expectancies is that they capture health of a population in a single number and that the concept is relatively easy to explain to policymakers. Its main uses are ranking lists by country or subpopulations within a country, and tracking of change over time for the same population. The downside of health expectancies is that it is not easy to attribute a population’s health to underlying diseases, injuries, or risks.

The disability-adjusted life year (DALY) is a health gap measure that was created to allow measurement of a population’s health with the ability to express the contribution of diseases, injuries, and risks as a proportion of the total burden of disease. It is called a health gap measure as it quantifies loss of health from mortality and morbidity in a population against a stated ideal that ‘everyone should live until a ripe old age, free of disease’. A number of social value choices are incorporated into the DALY, all of which were

reconsidered by a panel of ethicists and philosophers for GBD 2010 (Murray et al. 2012).

The first inevitable choice in constructing a health gap measure concerns the choice of the standard for length of life. In earlier versions of GBD, a standard life table was chosen that mirrored life expectancy of Japanese women in the early 1990s. This was Coale and Demeny's West level 26 female model life table with a life expectancy at birth of 82.5 years (Coale and Demeny 1966). A choice was made to set a slightly lower standard for men, postulating a biological difference in life expectancy between men and women. As there was no male model life table in the Coale and Demeny series that came close to the chosen female life table, the West level 25 female model life table with life expectancy at birth of 80 years was chosen for men. Use of a standard life table essentially is a weighting for age at death: infant deaths are considered equivalent to a loss of the number of years of life expectancy at birth; a death at age 20 is equivalent to the loss of the remaining standard life expectancy at that age. It also follows that even deaths at the oldest age groups still account for some years of life lost. The panel of ethicists and philosophers advised that there was no clear rationale for a difference between males and females and that it would be fitting to update the standard to achievable levels of life expectancy in 2010. A choice was made to create a new standard life table consisting of the lowest observed mortality rates in any age group in any population greater than 5 million. The standard life table used in GBD 2016 has a life expectancy at birth of 86.6 years. The full list of life expectancy at each age from the standard life table can be found in the web appendix to one of the GBD 2016 papers published in *The Lancet* (Naghavi et al. 2017).

A second choice is whether to apply the same standard life table to all populations and over time. From a principle of wishing to compare 'like with like' and for the sake of making meaningful comparisons of changes in health status of populations over time, GBD uses a single model life table to measure years of life lost, regardless of whether the death occurred in a population currently experiencing high or low mortality rates.

A third choice that has to be made in a summary measure like the DALY is how to value health loss from non-fatal conditions relative to the loss of health from deaths. The previous two choices determined how counts of death are translated into a number of years of life lost. To translate non-fatal health loss in a population into a comparable time measure of 'lost healthy

years', a severity weighting for each disease or consequence of disease and injury is needed. The original GBD study in the 1990s derived these 'disability weights' (DWs) from a panel of international health experts who valued 22 tracer conditions with person trade-off methods. Subsequently, they were asked to interpolate all remaining health states by indicating a distribution across seven classes of severity after binning the valuations for the 22 tracer conditions (Murray et al. 1996). DWs ranged from a low of 0.005 for mild anaemia and a high of 0.85 for a number of severe congenital anomalies. Since then, national burden of disease studies (Melse et al. 2000; Begg et al. 2008) have continued to use these DWs but supplemented them with a new set of DWs developed for the Dutch burden of disease study (Stouthard et al. 2000), as well as a number of ad hoc derivations of DWs for new conditions. As this had led to a lack of consistency between DWs, and because there had been criticism of depending on the values of a small panel of health experts, for GBD 2010 population surveys were conducted in Peru, Indonesia, Bangladesh, Tanzania, and the United States, as well as an open-access Internet survey. More than 30,000 respondents took part from 167 countries. For GBD 2013, four more surveys were conducted in Italy, Sweden, Hungary, and the Netherlands using the same survey design, which doubled the available data from the previous surveys. To minimize the cognitive burden, a simplified method of pairwise comparisons was adopted. Each time, a respondent was asked to decide, between random pairs of health states, 'who is the healthier?' The health states were described in short lay descriptions capturing the essential domains of loss of function in fewer than 35 words, as pilot testing indicated this was the maximum of information respondents were able to absorb. Valuations were obtained for a parsimonious set of 235 health states covering all sequelae of diseases and injuries. In addition, respondents to the Internet survey were asked to answer a small number of population health equivalence questions to anchor the values from the pairwise comparison on a 0–1 scale, where 0 represents no health loss and 1 represents health loss equivalent to being dead. All DWs were estimated with an uncertainty interval. The least severe DW, of 0.003, was for mild vision loss, while the worst DW, of 0.770, was given to the acute, psychotic phase of schizophrenia. Further methods and details of all lay descriptions and DWs can be found in Salomon et al. (2015).

Two other choices that had been used in previous GBD studies, discounting, and age-weighting, were rejected by the panel of ethicists and

philosophers. Both these choices had been subject to criticism in the academic literature (Anand and Hanson 1997; Tsuchiya 2001; Arnesen and Kapiriri 2004) but also were difficult to explain to policymakers. With discounting a future stream of lost health, greatest value is given to loss of health in the year of interest while years into the future get valued less and less. Previous GBD studies used an instantaneous discount rate of 3%, which meant that a female infant death with a standard life expectancy of 82.5 years would translate into 30.5 years of life lost. Age-weighting was a further complication incorporated into the DALY of previous GBD studies. The age-weighting function intended to give less value to health loss at young and older ages and greater value to health loss in young adults, arguing that the latter have a social duty to look after the young and the old.

A further general choice is whether to generate estimates of non-fatal health loss from incident or prevalent cases. Previous GBD studies used an incidence approach, arguing that it fits better with the measurement of lost years from deaths and that it would suit policymaking about prevention, as the future stream of health lost in an incident case of disease reflects the potential health gain from prevention (Murray et al. 1996). However, it was difficult to explain to policymakers that the burden of disease in a particular year was not reflecting the health loss from current cases of disease. The most important consideration to adopt a prevalence approach to non-fatal health loss measurement in GBD 2010 was that it would have been computationally nigh impossible to correct for comorbidity with an incidence approach, as it would have to deal with a large combination of lost health streams of different durations. A corollary of choosing a prevalence approach is also that by definition you cannot apply discounting, but that was a decision already made.

This was a long introduction about the metrics used in burden of disease and how key assumptions have changed over time. It is important to point out that any change in these value choices means using a different metric that is no longer comparable to another. For that reason, in each cycle of GBD estimates are explicitly re-estimated for the whole time period from 1990 to the most recent year in order to avoid anyone making comparisons with the previous GBD estimates. A list of key terms is found in Box 5.15.1.

Estimation principles

The GBD cause list is a set of mutually exclusive and collectively exhaustive categories. The 333 causes included in GBD 2016 were chosen based on considerations of their relevance to global health policymaking and/or the size of their expected contribution to mortality and/or morbidity burden. At the beginning of each GBD cycle, decisions are made about new additions to cause and risk factor lists by the GBD Scientific Council consisting of prominent members from the collaboration and senior faculty at IHME.

Box 5.15.1 Key terms

Years of life lost (YLLs): years of life lost due to premature mortality.

Years lived with disability (YLDs): years of life lived with any short-term or long-term health loss.

Disability-adjusted life years (DALYs): the sum of YLLs and YLDs. DALYs are also defined as years of healthy life lost.

Healthy life expectancy, or health-adjusted life expectancy (HALE): the number of years that a person at a given age can expect to live in good health if current rates of mortality and disability would remain constant.

Sequelae: consequences of diseases and injuries.

Health states: groupings of sequelae that reflect key differences in symptoms and functioning.

Disability weight: number on a scale from 0 to 1 that represents the severity of health loss associated with a health state.

Uncertainty interval: a range of values that is likely to include the correct estimate of health loss for a given cause.

All quantities of interest are estimated arguing that an uncertain estimate even when data are sparse or not available is preferable to no estimate. No estimate can be interpreted to mean no burden from that condition.

To convey to users the strength of the evidence for each quantity, uncertainty distributions for each quantity are reported in various metrics of uncertainty such as 95% uncertainty intervals.

An important component of the analytical work on GBD is to ensure that the set of results are internally consistent. The sum of cause-specific mortality must equal all-cause mortality: ‘you only die once’. Deaths are assigned following International Classification of Diseases and Related Health Problems (ICD) rules on underlying cause. The sum of cause-specific estimates of impairments such as blindness and anaemia must equal estimates of all-cause blindness and anaemia. For non-fatal disease

outcomes, incidence, prevalence, remission, and duration must be internally consistent.

Burden of disease estimation should be seen as an iterative process. New data and methodological innovation will lead to revision of estimates. Revisions should re-estimate the entire time series so that results are always available over time using consistent data and methods.

Data synthesis

The emphasis in the GBD is on systematically identifying all published and unpublished data sources. Published literature is a good source for some conditions such as epilepsy. For many others, national data systems such as vital registration, hospital discharges, national surveys, and surveillance data are equally important sources. By starting with the universe of all measurements, we can avoid publication or selection bias where different groups pick and choose between data sources.

Making use of all available data requires careful consideration of non-sampling error due to variation in case definition, diagnostic technology, study design, training of data collectors, recall periods (1 month, 12 months, lifetime), etc. For example, review of diabetes prevalence studies found 18 different case definitions in use, and review of myocardial infarction data showed marked differences in rates using heart-related serum creatine kinase fraction or troponin tests. Across studies or measurements, sampling error is a much smaller contributor to observed variance than these sources of non-sampling error. The GBD approach is to identify and correct for bias.

Key areas in cause of death data concern mapping across variants of the ICD, redistribution of intermediate causes (such as sepsis and heart failure) and ill-defined or ‘garbage’ codes (such as unspecified cancer or senility), and validity of verbal autopsy (VA) diagnoses (attribution of cause of death based on interviews with relatives of a deceased). For epidemiological data, statistical methods are used to quantify the relationship between different types of measurement such as variations in case definitions, diagnostic technologies, or recall periods. Setting a ‘desired’ reference measurement definition allows adjustments of all available data (‘crosswalks’) into comparable units, definitions, or categories. Where possible, uncertainty in these mappings is propagated into the uncertainty interval for the measurement. Some measurements may have to be excluded because they

cannot be made comparable to the rest of the measurements or have fundamental problems of validity. For instance, measures of lifetime prevalence were rejected as invalid measures of disease occurrence due to recall bias. A few exceptions to this rule were made for conditions where exclusion of such estimates would leave little other information for making estimates.

After adjusting the data into comparable units, appropriate statistical methods are used that are designed to handle both sampling and non-sampling error. Covariates are added to the models, and we use statistical methods to borrow strength across time or geography to improve predictions where data are sparse. All estimates in GBD are generated with 1,000 draws reflecting uncertainty from sampling error, non-sampling error, methods of correcting data, and where possible, model specification. Ideally, the validity of the data synthesis approach is determined using out-of-sample predictive validity. This has been fully implemented for mortality and causes of death but not for the synthesis of data on disease sequelae. Computational limitations and sparse data have so far limited the extent to which formal out-of-sample predictive validity can be evaluated for disease sequelae.

The analytical steps of GBD

The GBD approach contains 22 distinct components, as outlined in Figure 5.15.1. The components of GBD are interconnected. For example, when new data are incorporated into the age-specific mortality rates analysis (components 2 and 3), other dependent components must also be updated, such as rescaling deaths for each cause (component 9), healthy life expectancy, or HALE (component 17), years of life lost (YLL) (component 18a), and estimation of YLLs attributable to each risk factor (component 22a). A brief description of these key components follows. More detailed descriptions of each component are included in published articles (Gakidou et al. 2017; Hay et al. 2017; Naghavi et al. 2017; Vos et al. 2017; Wang et al. 2017).

Covariates

Where data are sparse, extensive use is made of a database of covariates. The more than 200 covariates cover socioeconomic factors (e.g. gross

domestic product), demography (e.g. population density, total fertility rate), behaviours (e.g. mean consumption of alcohol, tobacco, and a range of foods), and health system characteristics (vaccination coverage or a composite measure of health systems access). Essential is that values are available for all populations and years of analysis. The information largely comes from international databases, national censuses, surveys, and risk estimates generated by GBD (e.g. the prevalence of obesity or stunting). Where information is missing for a time period or particular population, imputation techniques are used to fill in the gaps.

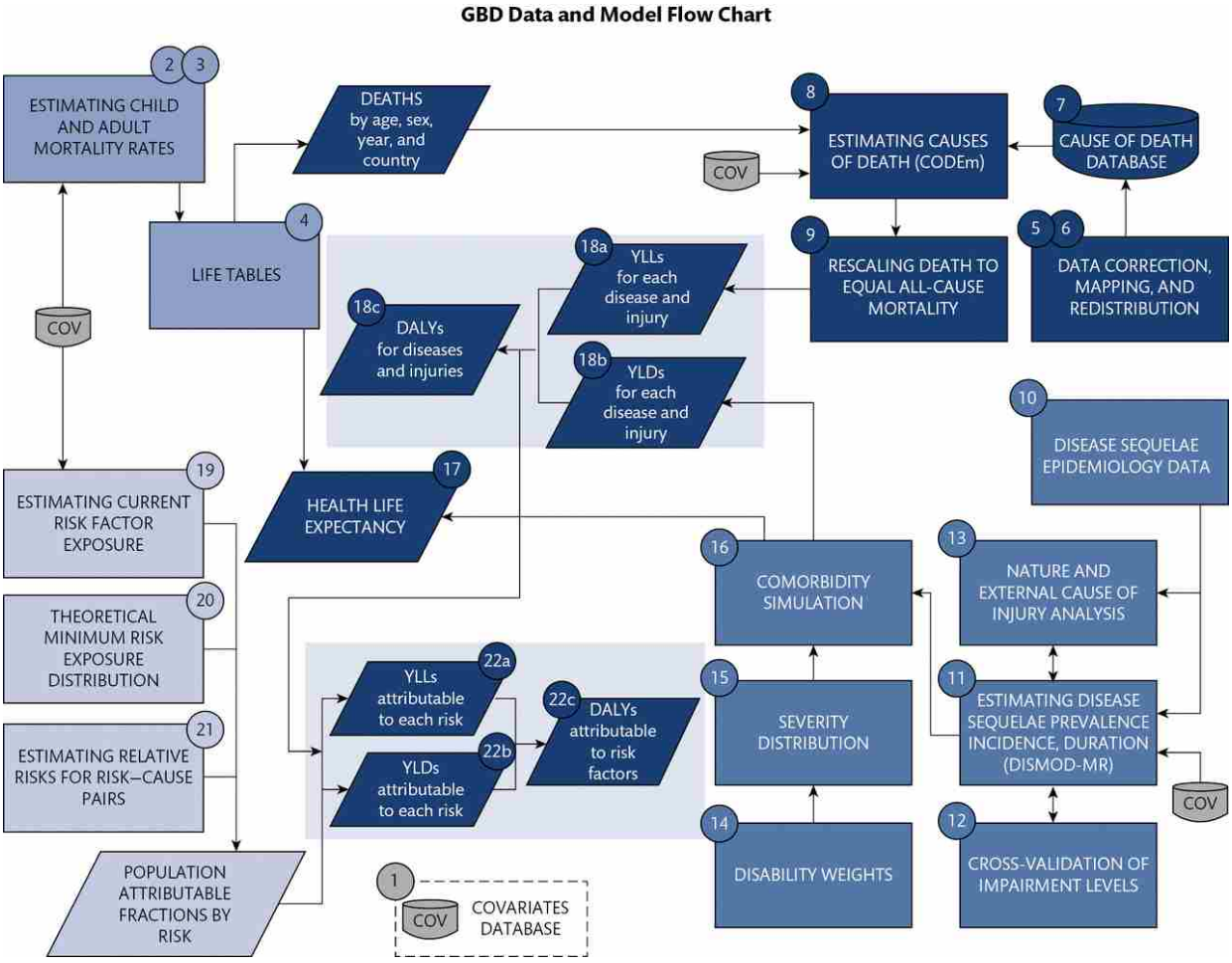


Fig. 5.15.1 The 22 components of the Global Burden of Disease study.

Estimating mortality in children under 5

Sources of under-5 mortality data include vital and sample registration systems as well as surveys and censuses that ask mothers about live births and deaths of their children. All available data are processed to address biases and estimate the probability of death between ages 0 and 5 ($_5q_0$) using Gaussian process regression.

Estimating adult mortality

Sources of adult mortality include vital and sample registration systems as well as surveys and censuses that ask about deaths among siblings. Similar methods to those for childhood deaths are used to estimate the probability of dying between ages 15 and 60 ($_{45}q_{15}$).

Life tables

The estimates of $_5q_0$ and $_{45}q_{15}$ are entered into a model life table system to determine age-specific mortality rates by sex for each country between 1970 and the most recent year of estimation with life expectancy at birth estimates and counts of death by age, sex, country, and year as outputs.

Cause of death data sources

Information on causes of death comes primarily from vital registration systems and VA studies. VA is a means of ascertaining the cause of death in populations with incomplete vital registration systems. With a structured questionnaire, interviewers ask relatives about the signs and symptoms of a recently deceased individual. Validation studies suggest that VA performs reasonably well for many causes of death (Lozano et al. 2011). For selected causes, additional sources such as police records for homicides and road traffic accidents, surveillance systems for maternal mortality, and cancer registries provide further information.

Data correction, mapping, and redistributions

A first task to make cause of death data comparable is the mapping across various revisions and national variants of the ICD. Algorithms have also been developed to deal with the problem of ill-defined coding in vital

registration and VA data where deaths are assigned to unknown causes, intermediate causes, or immediate causes of death rather than the underlying cause of death. For comparisons between countries and over time it is undesirable to have varying numbers of deaths in such 'garbage' code categories. The GBD approach is to reassign such deaths to more specific causes. The allocation of these garbage codes across target causes is done on the basis of published scientific literature, expert judgement, or statistical analysis. In some cases, deaths in garbage codes are reassigned proportionately across all causes. Full details can be found in the web appendix to the GBD paper on causes of death (Naghavi et al. 2017).

Cause of death database

The GBD 2010 cause of death database contains 2,798 site-years of vital registration data from 130 countries and 486 site-years of VA data across 66 countries, of which 10% were nationally representative. For the estimation of mortality from cancer, 2,715 site-years of cancer registry incidence data across 93 countries were included. From registries that also track cancer mortality, mortality-to-incidence ratios were derived. The log of the mortality-to-incidence ratios for each type of cancer was modelled using a statistical model that borrows information over time and between countries with guidance from a covariate on level of development as measured by an indicator combining a country's gross domestic product (GDP), education level, and fertility rate. The ratios were then applied to the incidence figures and the resulting cancer death estimates added to the cause of death database. Multiple demographic and health surveys, other surveys, and censuses provided data on the fraction of deaths in the reproductive age groups that are pregnancy-related. Sibling history data from 1,609 survey years covering 61 countries were identified. Another eight countries had nationally representative maternal mortality surveillance systems.

Estimating causes of death

GBD 2010 used a new statistical approach, named CODEm (Cause of Death Ensemble model), to determine the number of deaths from each cause by age, sex, country, and year. Ensemble modelling was made famous by the recipients of the Netflix Prize, BellKor's Pragmatic Chaos, in 2009, who

engineered the best algorithm to predict how much a person would like a film, taking into account their past movie preferences.

The basis of CODEm is that the analyst takes no firm position on what the best statistical method and choice of predictive covariates are but rather tests a large range of plausible statistical models and determines which single model or combination of models performs best. Four families of statistical models are used: mixed effects linear and spatiotemporal Gaussian process regression models of either the log of the death rate or the logit of the cause fraction. These four families of statistical models are run with multiple combinations of covariates. Based on published studies, plausible relationships between covariates and each cause are identified and tested. All models where the sign on the coefficient for a covariate is in the direction expected based on the literature and the coefficient is statistically significant are retained.

Next, the performance of single models and blends of various models, so-called ensembles, is evaluated using out-of-sample predictive validity tests. This is done by running the models on 70% of the data and comparing predictions for the remaining data withheld from the model-building exercise with the actual observed data. Tests of out-of-sample performance include the root-mean-squared error of the log of the cause-specific death rate, the direction of the trend in the prediction compared to the data, and the validity of the 95% uncertainty interval. The best-performing model or ensemble is selected on the basis of out-of-sample predictive validity.

Typically for each cause of death, several hundred single models are run and ensembles created of five to 20 best-performing individual models, but for some causes almost 100 models contribute to the final ensemble.

CODEm was not used for a number of conditions, including infectious diseases such as HIV/AIDS, measles, whooping cough, congenital syphilis, and malaria in sub-Saharan Africa, which are known to be subject to considerable misclassification bias in cause of death registrations. Instead, natural history models are created based on incidence or prevalence, and case-fatality rates. Estimates of HIV/AIDS deaths are modelled using software programs developed by UNAIDS and adapted for use in GBD. Information on deaths due to conflict, natural disasters, and epidemics from international databases is used to incorporate sudden peaks in mortality from these causes.

Rescaling deaths to equal all-cause mortality

The number of deaths from all causes in a sex, age, country, and year category are scaled to ensure their sum equals the total number of deaths estimated in the GBD demographic analysis of steps 2 to 4.

Systematic review of the epidemiology of disease sequelae

A large effort in GBD is dedicated to the systematic review of published and unpublished data on prevalence, incidence, remission, mortality risk, and duration. In GBD 2016, non-fatal estimates were made for 328 diseases and injuries and their 2,498 sequelae. The systematic reviews are supplemented with analyses of household survey data, antenatal clinic surveillance, reportable disease notifications, disease registries, hospital admissions data, outpatient visit data, medical claims data, population-based cancer registries, active screening data, and other administrative data. As with data on causes of death, a considerable effort goes into documenting and correcting for factors that influence the quality and comparability of all data points, such as different case definitions, diagnostic technologies, and sampling strategies (Vos et al. 2017).

Estimating disease sequelae prevalence by Bayesian meta-regression

For the task of developing estimates for disease prevalence by age, sex, year, and country, a Bayesian meta-regression tool was specifically developed for GBD: DisMod-MR. This tool is designed to process data presented with a wide variety of age intervals; to predict values for countries and regions with sparse data; to correct data from studies with different case definitions, diagnostic technology, or sampling strategies; and to ensure internal consistency between related epidemiological parameters such as incidence, prevalence, excess mortality, and duration. DisMod-MR estimates a log-normal model for all epidemiological data. The model includes the option to select covariates that predict variation in true rates and to correct for study characteristics that are a source of measurement bias. It estimates random effects for seven super-regions, 21 regions, and 195 countries and territories. It uses differential equations governing the relationship between the parameters of incidence, remission, mortality, prevalence, and duration to

enforce consistency. More detail is provided elsewhere on the development and application of DisMod-MR (Vos et al. 2012; Flaxman et al. 2015).

Cross-validation of impairment levels

There are no data sources we can use to set all-cause disability in a population similar to the all-cause mortality rates from demographic data sources that determine the 'envelope' of deaths. However, there are a number of conditions for which there are data collections of the overall prevalence in a population. In GBD we call these impairments. In GBD 2016 this was the case for vision loss, hearing loss, anaemia, infertility, heart failure, intellectual disability, epileptic seizures, pelvic inflammatory disease, and Guillain-Barré syndrome. In these cases, the quality and extent of data on the overall impairment level were considered stronger than the disease-specific estimates of the same impairments.

External cause and nature of injury analysis

Injuries are classified in the cause list according to the external cause such as a road injury, animal bite, or interpersonal violence, whereas the disability from injury is determined by the nature of injury such as brain trauma, hip fracture, or spinal cord lesion. The injury analysis is carried out in six steps. First, household survey, hospital discharge, and emergency department data were analysed using DisMod-MR to generate estimates of incidence for each external cause by age, sex, country, and year. Survey data included recall of injuries that led to hospital admission and other levels of medical attention. The meta-regression includes a covariate for whether an individual was admitted to hospital or not, which is used to generate predictions both for injury warranting hospital admission and injury warranting outpatient care. Second, multinomial regression models were used to analyse hospital data from 28 countries that have dual coding of discharges by external cause and nature of injury after ICD-9 and ICD-10. This produces matrices of external codes and nature of injury categories by age and sex, for high-income and low/middle-income countries and separately for injuries that were treated in emergency departments and those that led to admission. Third, for each nature of injury the prevalence of short-term disability is estimated from incidence and estimates of duration for major causes of injury from a follow-

up study in the Netherlands and supplemented with expert advice from trauma surgeons for the remaining causes. Fourth, the probability of individuals developing long-term disability is estimated in a re-analysis of follow-up data from seven studies in the United States, the Netherlands, and China that documented outcomes of injuries after one year. Fifth, DisMod-MR is used again to estimate the prevalence of individuals in the population who have disability from a previous injury. Prevalence of long-term disability is estimated from incidence assuming zero remission and, for spinal cord lesion, more severe traumatic brain injury, extensive burns, and hip fractures, a relative risk of death compared with the general population-based on follow-up studies of people with such long-term disabilities. For long-term outcomes with a different disability weight for a treated and untreated case, we estimate access to effective treatment on a scale ranging from 20% in the least developed countries to full access in high-income countries. This concerns mainly spinal cord injury and amputations, where access to a functional wheelchair or a prosthesis substantially alters the health loss from those disabling outcomes. In the last step, the years lived with disability (YLD) due to prevalent cases of long-term injury were attributed back to external causes in proportion with the contributions of these causes to every type of injury.

Disability weights

Disability weights used since GBD 2010 have already been discussed in the earlier section on value choices.

Severity distributions

For many diseases, the sequelae of the disease are linked to more than one health state (often classified as mild, moderate, and severe) including stroke, anxiety, major depressive disorder, dementia, and asthma. DWs across these severity levels are measured in population surveys for individuals without comorbidity. Therefore, estimates of the distribution of severity need to be controlled for comorbidity, otherwise the severity distribution would be systematically biased toward more severe symptoms. For example, if individuals with stroke are also likely to have diabetes and kidney disease,

the reported distribution of functional health status would be shifted toward the more severe end.

For some diseases for which data are available for the distribution of severity by age, sex, and region, we pooled proportions in each health state using DisMod-MR or simple meta-analysis methods. Data for severity distributions are often scarcer and of poorer quality than data for prevalence of disorders. Because of the heterogeneity of the available evidence for disease severity, disease-specific reviews had to be supplemented with an analysis of three high-income country data sources which combined a range of diagnostic information and health status data. These sources allowed the assessment of the severity distributions, taking into account comorbidity. With the general lack of comparable data on severity of diseases, it is difficult to account for the differential access over time and by location to treatments that can affect severity. This is clearly an area where new data collection is warranted and in the future may be facilitated by analyses of linked datasets of health encounter data with diagnostic information and health surveys with general health status measures.

Comorbidity simulation

In past GBD studies, YLDs were estimated for each disease and sequela and then simply added, implicitly assuming that disability from different conditions is additive. At the level of an individual this can lead to a person with several severe sequelae having a combined level of DWs exceeding one, that is, being worse than dead. The 2003 Australian Burden of Disease study (Begg et al. 2008) introduced a method to simulate hypothetical populations with a correction for comorbidity using a multiplicative rather than additive function to combine DWs. A similar approach was adopted for GBD 2010 and later iterations.

For each age, sex, country, and year, we use a Monte Carlo simulation of 40,000 individuals to estimate the cooccurrence of sequelae using the prevalence data with uncertainty intervals as probabilities. In simulated individuals with two or more sequelae a combined DW is calculated with a multiplicative function and then parsed back proportionately to each comorbid sequela. The mean reduction across all individuals with a particular sequela is taken as the comorbidity correction.

Tests on real data from the US Medical Expenditure Panel Surveys (Agency for Healthcare Research and Quality 2000) suggest that this multiplicative model is appropriate and accounts for the larger component of comorbidity corrections as long as these are estimated separately for each age and sex category. For several GBD iterations we have had the intention to also take dependent comorbidity (e.g. the common combinations of e.g. diabetes and heart disease or alcohol dependence and anxiety or depression) into account but have been held back by prohibitively long computation times. We are hoping to resolve this by optimizing the simulation code.

Healthy life expectancy

The life tables developed in component 4 for each country and the output of the comorbidity simulations were combined to estimate healthy life expectancy for 195 countries and territories as well as over 500 subnational locations for which GBD 2016 made estimates. Healthy life expectancy is a single summary metric of population health that is not affected by population age structure. It can be thought of as equivalent to the expectation of life at birth in full health given the currently observed age-specific death rates and severity-weighted prevalence of disease and injury sequelae.

Calculation of YLDs, YLLs, and DALYs

YLLs are simply the multiplication of the deaths by cause, age, sex, country, and year with the relevant value from the standard life table for the age at death. YLDs are the product of prevalence of a sequela and the comorbidity-corrected DW for the health state assigned to the sequela. DALYs are the simple addition of YLLs and YLDs. Each quantity in GBD is estimated with uncertainty by generating 1,000 draws. A thousand values of each summary measure are computed by taking the first draw for each quantity, then the second draw and so on until 1,000 estimates are calculated. Thus, the correlation across age and over time for causes of death and prevalence is retained.

Estimating risk factor exposure

In GBD, risk factors are chosen based on their likely importance to disease burden or policy; availability of sufficient data and methods to enable estimation of exposure distributions by country; sufficient evidence for causal effect and to estimate outcome-specific effect sizes; and evidence to support generalizability of effect sizes to populations other than those included in epidemiological studies.

Analogous to the same principles used for work on disease sequelae, a database of published and unpublished sources on the prevalence of exposure was created. For some risks, innovative sources such as satellite imagery for ambient air pollution by particulate matter are used.

Theoretical minimum risk factor level

The tradition in risk factor epidemiology is to assess attributable burden by estimating a counterfactual scenario where exposure to a risk factor in the past is set to a counterfactual distribution. Attributable deaths, YLLs, YLDs, or DALYs are then the difference between observed burden of disease and the counterfactual scenario. Following the framework for comparative risk assessment (Murray and Lopez 1999), counterfactuals are set to the lowest risk level that has been observed or is theoretically possible—the theoretical minimum risk exposure level.

Estimating relative risks for risk factor–disease pairs

For each risk–disease pair, relative risks for specific disease outcomes are taken from published meta-analyses or updated meta-analyses that have been undertaken as part of GBD. Standardized approaches to the attenuation of relative risks with age for cardiovascular disease outcomes are included. Uncertainty in the relative risks for each risk–disease pair by age and sex is propagated into the final estimates. Full details on the risk factor estimation are provided in Gakidou et al. (2017).

Estimating deaths, YLDs, YLLs, and DALYs attributable to risk factors

Risk factor burden is assessed using the distributions of burden for each disease associated with a risk factor. Uncertainty from the prevalence of

exposure, relative risks, and, where appropriate, the theoretical minimum risk exposure level, are all propagated into the final risk factor uncertainty. Final uncertainty reflects both the uncertainty for an age-sex group from the process of disease and injury estimation and the uncertainty in the population attributable fractions. We assume that uncertainty in the population attributable fractions is uncorrelated with uncertainty in disease and injury burden. For the computation of clusters of risk factors (e.g. diet, or environmental and occupational risks), we have assumed that the combined effects of risk factors are multiplicative. Where the impact of one risk is mediated through one or more risks, we derive mediation factors from cohort studies that provide relative risk estimates with and without controlling for these risks that lie in the pathway to a disease outcome. Aggregate estimates of groups of risks take into account the known mediation factors.

Exploring GBD results

The GBD is considered a 'public good'. To guide policymakers, researchers, students, and interested individuals through the large amount of results, IHME has created innovative data visualization tools to explore health trends for different countries and regions. The visualization tools allow people to view GBD estimates through hundreds of different dimensions. To explore the IHME website to use the GBD data visualization tools, visit https://vizhub.healthdata.org.

The *Mortality* and *Causes of Death* visualizations show the results, data inputs, and data transformations of all mortality estimates. *GBD Compare* shows results of deaths, YLLs, YLDs, and DALYs by diseases and risk factors. The main view is a 'tree map', which is basically a square pie chart showing the proportional distribution of the overall burden by underlying causes. Additional views can be selected showing a map of the world, rankings, time plots, or age plots. Each can interactively be explored by country, age, sex, year, cause, or risk factor. Data are available as counts, rates, and proportions of total. Rates can be shown by age for all ages combined and age-standardized to the GBD standard population (Naghavi et al. 2015). The former is most useful to explore the amount of disease burden health systems need to address. The latter helps to understand the underlying epidemiological trends, that is, after taking out population growth and ageing

as drivers of differences in burden estimates. The arrow diagram allows the user to compare ranking lists for countries and world regions between years. Cause patterns is the view of choice to compare country results with any grouping of other countries. It can also be used to explore age patterns. The heatmap is a different method of showing rankings of diseases or risk factors between countries and regions of choice. Top ranking causes are coloured in deep red and successively shaded in lighter hues that change to yellow, green, and blue further down the ranking. This tool quickly shows areas of progress or concern in a particular country in comparison to others. The *Life Expectancy & Probability of Death* tool graphs these two summary measures of mortality and how they have changed over time for each country. Input data sources and detailed results from GBD can be explored and downloaded at the http://ghdx.healthdata.org/gbd-2016/data-input-sources and http://ghdx.healthdata.org/gbd-results-tool sites.

Key GBD results

A comparison of the top 20 causes of DALYs in the world between 1990 and 2016 shows a shift from communicable, maternal, neonatal, and nutritional disorders (dark blue) to non-communicable diseases (blue) and injuries (light blue) (Fig. 5.15.2). Ischaemic heart disease has become the leading cause in 2016, displacing lower respiratory infections and diarrheal diseases which led the rankings in 1990. Among the major infectious diseases, only HIV/AIDS increased in rank and size, but the peak of HIV/AIDS DALYs occurred in 2005, after which the widespread use of antiretrovirals made death and YLL rates drop, though with a consequent rise of YLDs in the survivors. The growing importance of disability contributing to the burden of disease globally is illustrated by the prominent and rising positions of low back pain, migraine, and major depressive disorder, each responsible for large numbers of YLDs but no mortality (Fig. 5.15.2).

The column on the right of Figure 5.15.2 shows the proportion change in DALYs (with a 95% uncertainty interval) between 1990 and 2016. The change is due to a combination of population increase, ageing of the world's populations, and change in the epidemiological trends. Figure 5.15.3 shows the same top-20 ranking of DALYs but based on age-standardized rates. In this figure the column on the right reflects the change in epidemiological rates over time, having removed the demographic drivers of changes in the

numbers of DALYs. Apart from a shift in the ranks of some conditions, the greater difference is in the proportion change column. For instance, the increase in DALYs from ischaemic heart disease of 46.8% was entirely due to demographic changes, while the epidemiological rates actually dropped by 23.4%, as seen in Figure 5.15.3. In contrast, for diabetes, DALY counts, and age-standardized rates increased by 108.3% and 14.9%, respectively. Both measures have their own use. Changes in the volume of disease (as in Fig. 5.16.2) inform policymakers about health service planning. The epidemiological changes tell us something about the impact of prevention and cure on the occurrence of disease.

The shift from communicable diseases largely affecting children to non-communicable disease in adults is further illustrated by the change over time in the ranking order of the main risk factors (Fig. 5.15.4). While child and maternal malnutrition has remained at the top of the ranking list, its size reduced by 55.7%. Dietary risks, high systolic blood pressure, high fasting plasma glucose, and high body mass index increased in rank as well as in size over the 1990 to 2016 estimation period. The small increase in burden from smoking by 5% globally masks an increase in low- and middle-income countries being compensated by a decreasing burden from smoking in high-income countries.

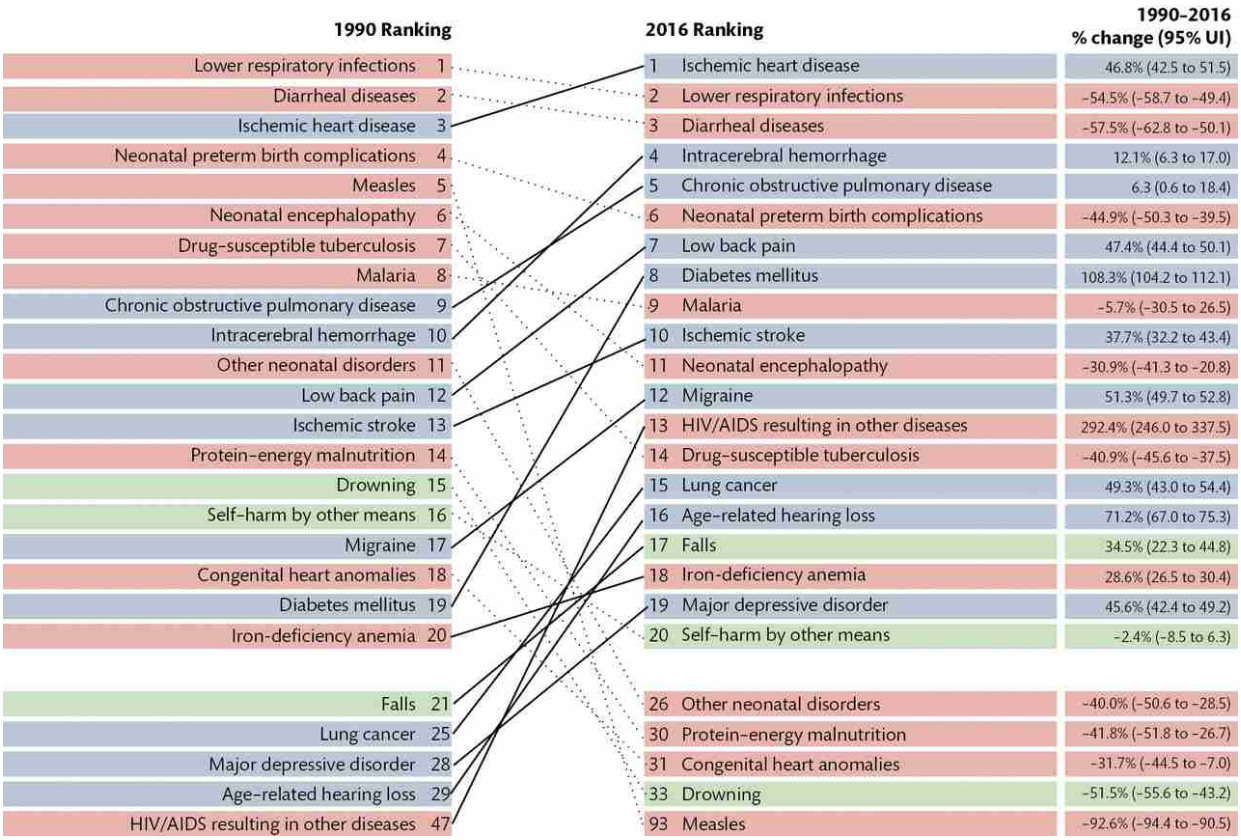


Fig. 5.15.2 Global top-20 ranking of diseases and injuries in DALYs, 1990 and 2016.

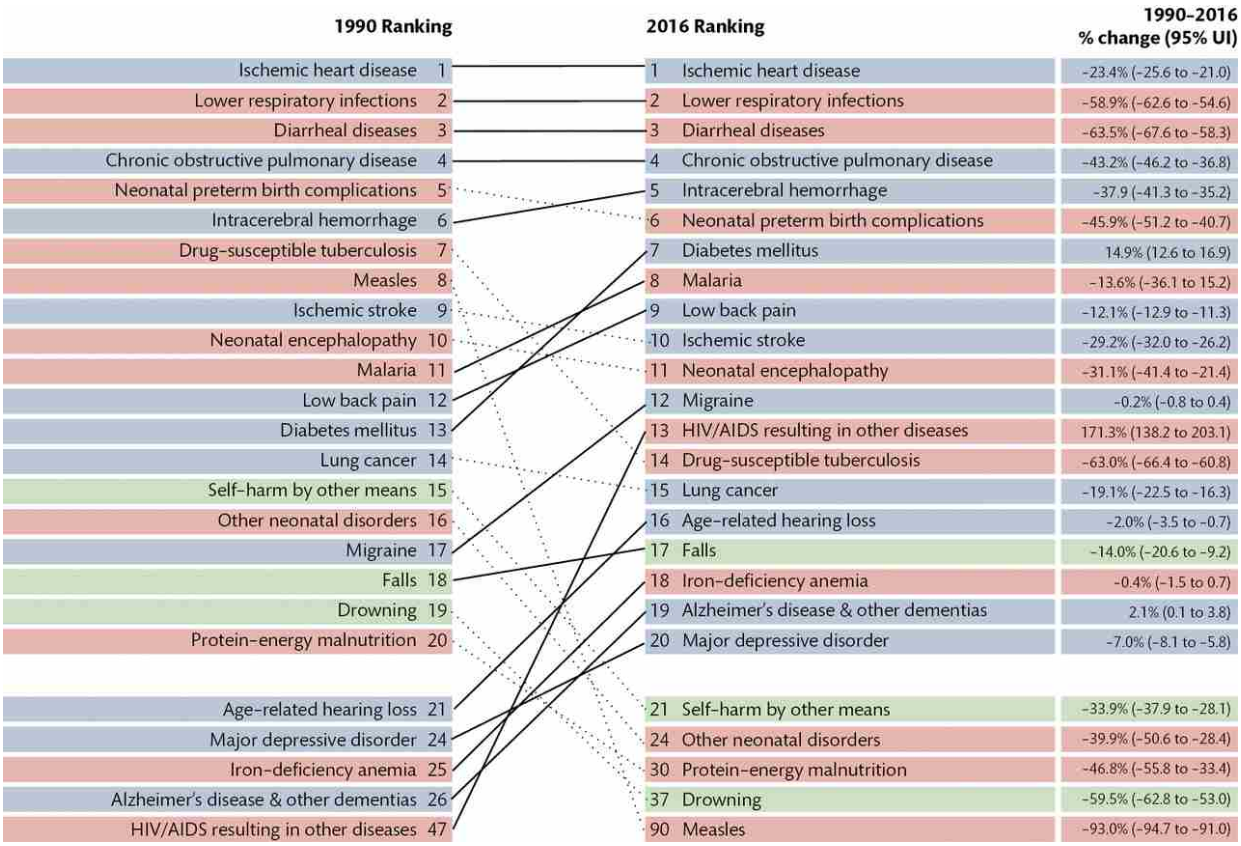


Fig. 5.15.3 Global top-20 ranking of diseases and injuries in age-standardized DALY rates, 1990 and 2016.

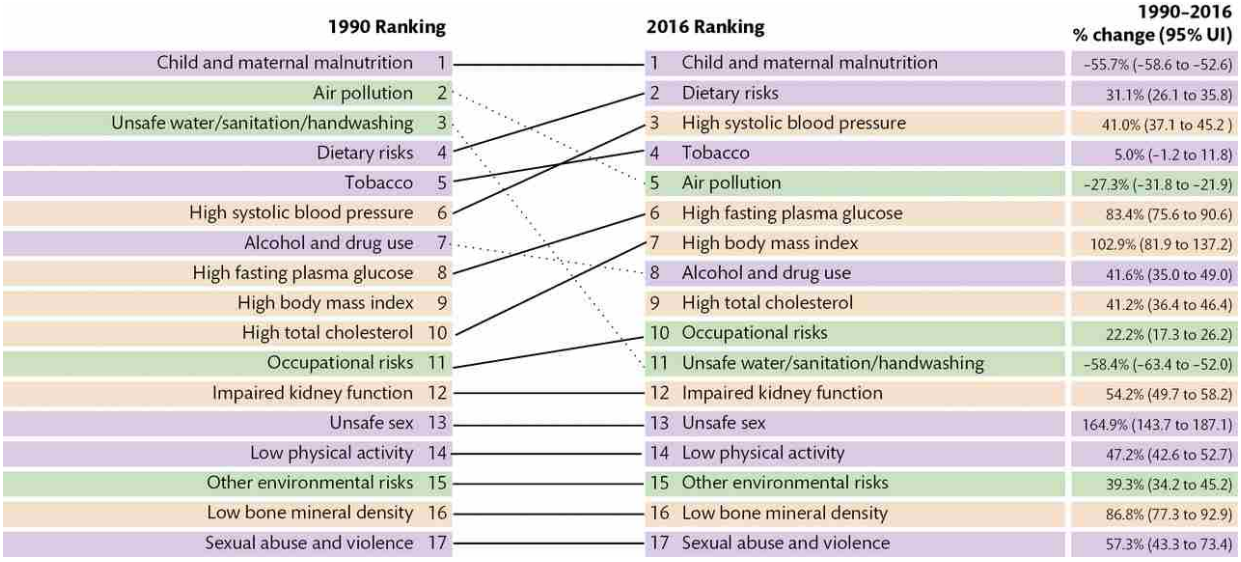


Fig. 5.15.4 Global top-20 ranking of risk factors in DALYs, 1990 and 2016.

The future of GBD

GBD is updated annually. The intention is to continue expanding the network of collaborators and strengthen ties between country researchers, ministries of health, the World Health Organization, and IHME to improve national estimates and to undertake subnational burden of disease studies. Through such partnerships, IHME is helping governments and donors gain insights into localized health trends to inform planning and policymaking. Local epidemiological assessment is crucial for informing local priorities.

The future of GBD is to continue to refine the spatial resolution of estimates, making use of disease mapping techniques developed for malaria ('The Malaria Atlas Project' n.d.). Forecasts of GBD estimates to 2040 will soon become available, first for a baseline scenario, to be followed by analyses of the health impacts of areas of major policy concern, e.g. different levels of global warming, spread of antimicrobial resistance, or the introduction of new vaccines for major infectious diseases.

IHME is also committed to building capacity for GBD analysis in countries around the world, and conducts a variety of training workshops. Information on training can be found at http://www.healthdata.org/gbd/training.

REFERENCES

Agency for Healthcare Research and Quality (2000). Medical Expenditure Panel Survey. [Online] Available at: https://meps.ahrq.gov/mepsweb/

Anand, S. and Hanson, K. (1997). Disability-adjusted life years: a critical review. *Journal of Health Economics*, 16(6), 685–702.

Arnesen, T., and Kapiriri, L. (2004). Can the value choices in DALYs influence global priority-setting? *Health Policy (Amsterdam, Netherlands)*, 70(2), 137–49.

Begg, S.J., Vos, T., Barker, B., Stanley, L., and Lopez, A.D. (2008). Burden of disease and injury in Australia in the new millennium: measuring health loss from diseases, injuries and risk factors. *The Medical Journal of Australia*, 188(1), 36–40.

Boutayeb, A. (2006). The double burden of communicable and non-communicable diseases in developing countries. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 100(3), 191–9.

Bradshaw, D., et al. (2003). Initial burden of disease estimates for South Africa, 2000. *South African Medical Journal = Suid-Afrikaanse Tydskrif Vir Geneeskunde*, 93(9), 682–8.

Coale, A.J., and Demeny, P.-G. (1966). *Regional Model Life Tables and Stable Populations*. Princeton, NJ: Princeton University Press.

Flaxman, A.D., Vos, T., and Murray, C.J.L. (eds.) (2015). *An Integrative Meta-regression Framework for Descriptive Epidemiology*. Washington, DC: University of Washington Press.

Gakidou, E., et al. (2017). Global, regional, and national comparative risk assessment of 84 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990–2016: a systematic analysis for the Global Burden of Disease study 2016. *The Lancet*, 390(10100), 1345–422.

Hay, S.I., et al. (2017). Global, regional, and national disability-adjusted life-years (DALYs) for 333 diseases and injuries and healthy life expectancy (HALE) for 195 countries and territories, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *The Lancet*, 390(10100), 1260–344.

Keehan, S.P., et al. (2015). National health expenditure projections, 2014–24: spending growth faster than recent trends. *Health Affairs*, 34(8), 1407–17.

Lim, S.S., et al. (2012). A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990–2010: a systematic analysis for the Global Burden of Disease study 2010. *The Lancet (London, England)*, 380(9859), 2224–60.

Lorenzoni, L., Belloni, A., and Sassi, F. (2014). Health-care expenditure and health policy in the USA versus other high-spending OECD countries. *The Lancet*, 384(9937), 83–92.

Lozano, R., Lopez, A.D., Atkinson, C., Naghavi, M., Flaxman, AD., and Murray, C.J.L. (2011). Performance of physician-certified verbal autopsies: multisite validation study using clinical diagnostic gold standards. *Population Health Metrics*, 9(32), 32.

Marshall, S.J. (2004). Developing countries face double burden of disease. *Bulletin of the World Health Organization*, 82(7), 556.

Mathers, C., Fat, D.M., and Ties Boerma, J. (2008). *The Global Burden of Disease: 2004 Update*. [Online] Available at: http://www.who.int/healthinfo/global_burden_disease/2004_report_update/en/

Melse, J.M., Essink-Bot, M.L., Kramers, P.G., and Hoeymans, N. (2000). A national burden of disease calculation: Dutch disability-adjusted life-years. Dutch Burden of Disease Group. *American Journal of Public Health*, 90(8), 1241–7.

Murray, C.J.L., Lopez, A.D., World Health Organization, World Bank, and Harvard School of Public Health (1996). *The Global Burden of Disease: A Comprehensive Assessment of Mortality and Disability from Diseases, Injuries, and Risk Factors in 1990 and Projected to 2020*. [Online] Available at: http://apps.who.int/iris/handle/10665/41864

Murray, C.J.L., Salomon, J.A., Mathers, C.D., Lopez, A.D., and World Health Organization (2002). *Summary Measures of Population Health: Concepts, Ethics, Measurement and Applications*. [Online] Available at: http://apps.who.int/iris/handle/10665/42439

Murray, C.J.L., et al. (2012). Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990–2010: a systematic analysis for the Global Burden of Disease study 2010. *The Lancet*, 380(9859), 2197–223.

Murray, C.J.L., et al. (2012). GBD 2010: design, definitions, and metrics. *The Lancet*, 380(9859), 2063–6.

Murray, C.J.L., and Lopez, A.D. (1999). On the comparable quantification of health risks: lessons from the Global Burden of Disease study. *Epidemiology (Cambridge, Mass.)*, 10(5), 594–605.

Naghavi, M., et al. (2017). Global, regional, and national age-sex specific mortality for 264 causes of death, 1980–2016: a systematic analysis for the Global Burden of Disease study 2016. *The Lancet*, 390(10100), 1151–210.

Naghavi, M., et al. (2015). Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990–2013: a systematic analysis for the Global Burden of Disease study 2013. *The Lancet*, 385(9963), 117–71.

Richardson, J.R. (2014). Can we sustain health spending? *The Medical Journal of Australia*, 200(11), 629–31.

Salomon, J.A., et al. (2015). Disability weights for the Global Burden of Disease 2013 study. *The Lancet Global Health*, 3(11), e712–23.

Stouthard, M.E.A., Essink-Bot, M.-L., and Bonsel, G.J. (2000). Disability weights for diseases: a modified protocol and results for a western European region. *European Journal of Public Health*, 10(1), 24–30.

The Malaria Atlas Project. (n.d.). *MAP*. [Online] Available at: https://map.ox.ac.uk/

Tsuchiya, A. (2001). *The Value of Health at Different Ages*. York: Centre for Health Economics, University of York.

Vos, T., et al. (2017). Global, regional, and national incidence, prevalence, and years lived with disability for 328 diseases and injuries for 195 countries, 1990–2016: a systematic analysis for the Global Burden of Disease study 2016. *The Lancet*, 390(10100), 1211–59.

Vos, T., et al. (2012). Years lived with disability (YLDs) for 1160 sequelae of 289 diseases and injuries 1990–2010: a systematic analysis for the Global Burden of Disease study 2010. *The Lancet*, 380(9859), 2163–96.

Wang, H., et al. (2017). Global, regional, and national under-5 mortality, adult mortality, age-specific mortality, and life expectancy, 1970–2016: a systematic analysis for the Global Burden of Disease study 2016. *The Lancet*, 390(10100), 1084–150.

5.16

Mathematical models of transmission and control of infectious agents

Alex Welte and Cari van Schalkwyk

Introduction

Indeed, the ‘heavy lifting’ of healthcare is in the care of patients, the development and distribution of vaccines, drugs and devices, and the conception and implementation of sensible systems and policies. However, in recent decades, spectacular increases in the availability of computational capacity have paved the way for mathematical modelling to play an ever-increasing role in many aspects of public health, by supporting formal analyses at various scales of the processes involved.

The core ideas of mathematical modelling are beautiful, powerful, and simple. However, it can seem like a dark art to outsiders, distracted by the details or daunted by the considerable training and experience that appears to be required for attaining confidence, intuition, and proficiency.

By its nature, mathematics is a broad framework, and the natural maturation of mathematical modelling lies in abstraction and generalization. Under this large and growing umbrella there is room for many mathematical ideas, approaches, and techniques. For biomedical research in general, and public health in particular, most mathematical analysis and modelling takes the form of ‘statistical’ modelling of some kind.

In statistics, the key point is usually to understand how details of data are driven by ‘randomness’, and how, despite the randomness, when there is ‘sufficient’ data, we can learn a great deal about the underlying situations and events from which our data are generated. These ‘statistical’ methods of biomedical and public health research are the underlying mathematics that brings us such things as ‘confidence intervals’ for estimates, ‘p-values’ that express the consistency of a hypothesis with data, ‘correlation coefficients’,

and so on. This health-applied statistical field, usually called biostatistics, is a major industry, and some familiarity with its key concepts is a prerequisite for reading about public health research at any depth beyond the most superficial. While we presume that the reader has some basic statistical knowledge, we will not lean heavily on this assumption.

This chapter explores a particular kind of ‘modelling’—and it is not the common (bio)statistical kind. We focus on what we would call ‘dynamical’ modelling (as opposed to ‘statistical’ modelling, see Table 5.16.1.). This essentially entails the reduction, to mathematics, of key facts and principles inherent in the ‘processes’ or ‘mechanisms’ in an epidemiological situation. We can then manipulate these mathematical constructs, in search of insights that, while ultimately implied in the model construction, are not superficially apparent from our primary data and our intuition.

Table 5.16.1 Basic features and uses of ‘statistical’ and ‘dynamical’ models

Statistical Models	Dynamical Models

• Account for random error and bias in data to find correlations that may imply causality
• Support assessment of relationships
• Assume independence of individuals (sometimes independence within clusters)
• Typical questions:
  ◦ Is HIV status positively associated with the risk of TB infection?
  ◦ Are insecticide-treated bed nets (ITNs) or indoor residual spraying (IRS) more effective for controlling malaria?
	• Systems approach—explicitly model (hypothetically ‘declare’) specific mechanisms, to explore their interactions
	• Encapsulate particular relationships between parts of the system
	• Explicitly focus on interdependence of individuals
	• Typical questions:
	  ◦ Based on increased TB risk due to HIV, how should we expect the TB notification rate to vary with HIV prevalence?
	  ◦ How do we expect the age distribution of malaria incidence to change after implementing ITNs or IRS?

Of course, it’s not just surveys and clinical trial measurements that involve randomness. Many processes that are fundamental to epidemiology are ultimately subject to some randomness. Decapitation in an industrial accident is sure to lead to death, but even being infected with Ebola virus does not have an entirely predictable outcome. There are many levels of

randomness that inevitably creep into any description of a complex situation involving biology and society. Typically, given a fully (as far as possible) specified situation and set of rules, outcomes (like infections) vary (in number), by chance, around some ‘average’, ‘mean’, or ‘expected’ value.

Our focus, however, is on capturing and understanding the average/expected behaviour of an infectious epidemic, given some specific assumptions about the rules that govern its behaviour. This allows us to use methods of quantitative investigation that are not fundamentally dominated or complicated by randomness.

Infections such as HIV, tuberculosis (TB), malaria, measles, dengue, zika, cholera, Ebola, etc. persist as major causes of ill health, psychosocial stress, and premature death—not to mention economic strain and impediments to development more broadly. Among the many details which are specific to each infectious agent and outbreak, an important recurring theme, in the analysis of infection moving through a population, is that there is more going on than just a collection of individuals who happen to share a ‘risk profile’ or ‘environment’. There is a fundamental difference between the risk of heart disease, driven by genetic and lifestyle factors, and the risk of tuberculosis, which depends heavily on how many TB infected people one comes into contact with. This dynamical interaction between people who have acquired, and those who are susceptible to, an infection, is the heart and soul of infectious disease epidemiology. Using mathematics to model this process provides a powerful way of elucidating concepts that can help us understand aspects of disease dynamics that would otherwise be difficult to approach intuitively.

Aside from helping us better understand subtle collective effects, there is another fundamental aspect of the unique power of mathematical modelling: direct observation, and the application of formal experimental techniques, can only take us so far in the context of studying people, illness, care, and social systems. Ideally, scientists prefer to watch a system again and again, varying one ‘thing’ at a time (temperature, voltage, force, infectiousness, etc.) to elucidate the impact various factors have on the behaviour and/or state of a system—whether this be a solid material, a liquid solution hosting chemical reactions, or a gas trapped in an airtight container. When we want to understand a very much more complex system, comprised of many components, such as a society grappling with diseases, it’s almost too obvious to bear spelling out that we cannot just perform arbitrary

experiments. Even if it were ethically/morally defensible, it would simply not be possible to learn enough, sufficiently quickly, and at feasible cost, to provide meaningful actionable insights.

For example, we may wonder:

> *Do we need to vaccinate everyone to prevent measles outbreaks, or is some incomplete, approximately quantifiable, level of coverage sufficient to provide 'herd immunity' by restricting the transmission network so that even those who are not vaccinated will be highly unlikely to encounter infectious individuals, i.e. that a single case will not plausibly lead to a wider outbreak?*

This question cuts to the heart of 'collective' effects in infectious spread, and it clearly cannot be answered by

isolating numerous communities with various well-defined interaction/activity networks
subjecting them to different (precisely measured) levels of vaccine coverage
introducing a case of measles into each isolated community
noting how big the outbreaks tend to be as a function of the vaccination coverage and social network structure

One would have to do it all many times, to explore the many permutations of factors. Even worse, because there is inherent randomness in disease transmission in the real world, one would actually need to probe the *average* behaviours of any particular combinations of factors (as well as other statistical/distributional properties, such as, for instance, the worst 5% of outcomes).

What else can we do, then, to explore such questions, which are forever and fundamentally beyond the scope of controlled experiments? This is where models come into their own. Glossing, for now, over pretty much all the *details*, but not, we hope, all the important *concepts* (which we will also revisit more thoroughly), we can summarize a model-based investigation of the measles question as:

building a *model world* in which we declare specific hypothetical rules about measles transmission—based on biological and health systems insights.
(**Model world** construction is best done by a multidisciplinary team.)
capturing the *states* and *rules* of the model world in precise mathematical statements/equations/computer code
(*i.e. building a* ***mathematical model***
, using particular techniques.)
calibrating/fitting the model to *data* from past measles outbreaks with varying levels of vaccine uptake, including, crucially, no vaccine.
(***Calibration***
is a process that engages heavily with biostatistics.)
running *scenarios* from various *initial conditions* defined by factors like which (or simply *how many*) population members are initially infected/infectious; and, critically, the vaccine coverage.
(*This is sometimes known as '*
driving
' the model.)
Exploring trends in the *realizations* of the simulated epidemics by noting how various indicators of epidemic trajectory/severity/duration depend on the simulated policies (*i.e. conduct an actual* ***investigation.***)

A model-based investigation, then, leverages our current insights, beliefs, data sets, perhaps in conjunction with some interesting 'what if' scenarios, and turns these into their downstream implications (sometimes referred to as 'predictions', causing unnecessary confusion). Some inferences are obvious, like '*Increasing vaccine coverage reduces the size of outbreaks*'. Some may be striking discoveries, like '*essentially complete outbreak suppression can be achieved at less than complete vaccine coverage*'.

The fundamentals of modelling

The recurring central problem facing all attempts to describe the natural world, including the world of human endeavours, is this:

> *How do we describe a particular part of the world, about which we have specific questions, without being drawn into attempting to describe so much of it that we drown in details and complexity?*

This is the recurring challenge, whether we are trying to make sense of basic physical concepts like gravity, or trying to understand why some parents decline to have their children given readily available vaccinations. Mathematical modelling is a powerful tool which can help us to make key aspects of this process explicit and clear, and to be disciplined and consistent in our investigations.

Stepping one step further back: our current goal is to help the reader improve their mental model of *the modelling process itself*, without getting too far into philosophical asides. To this end, we will rely heavily on a modelling framework which, as far as we know, is the brainchild of Tony Starfield. We will call it the '*real-world, model-world, model'* paradigm (Starfield et al. 1990). We would also like to tip our hat to the somewhat older heuristics toolbox contained in the classic *How to Solve It* by George Polya (Polya 1945). Polya's ambition was far broader than our current focus on models of infectious agents, but for anyone intending to wrestle with formal problem construction and solution, it is extremely valuable, and hence paraphrased in Box 5.16.1. Starfield acknowledged Polya's groundwork by choosing the title 'How to Model it' for his own excellent and entertaining guide to, in his words: 'modelling in the computer era' (Starfield et al. 1990).

Box 5.16.1 Polya's 'How to Solve It' heuristic

- **Understand the problem.** Restate the problem in your own words (possibly to someone else); Consider whether you have enough information; Draw a picture.
- **Devise a plan.** Solve a simpler problem (the simplest possible, we recommend); eliminate possibilities; guess (and check); look for patterns; Be creative!
- **Carry out the plan that you have chosen.** At first, persist with your plan some through initial setbacks; If it really does not work, make another plan.
- **Review/extend.** Look back at what you have done—what worked and what didn't; Learn from this to help you solve future problems.

Source: data from Polya, G. (1945) *How To Solve It*. Princeton University Press.

Particularly striking in Starfield's exposition (Box 5.16.2), and the main point which we would like to help make more mainstream, is the habit of taking seriously the middle ground between the external *real world* and *mathematical details*, what he calls the '*model world*'. We have found this

device to be extraordinarily useful, not just pedagogically, but also as a modelling insiders' heuristic that tangibly supports the planning and execution of model-based investigations.

Box 5.16.2 Starfield's three realms

Real world: The world we live in: inordinately complex, only partly knowable, not under our direct control.

Model world: Rules and 'degrees of freedom' are arbitrarily declared and fully known. We also have the power to declare the model world to be in any (internally consistent) particular state, without considering how such a state might have been attained.

Model: Mathematical constructs and computer code that capture aspects of a model world—its states and rules. Models are intended to be manipulated and explored (analysed) to help understand the model world, and by extension, aspects of the real world.

Source: data from Starfield, A. M., Smith, K. A. and Bleloch, A. L. (1990) *How to model it: Problem solving for the computer age*. McGraw-Hill.

The model world

Before any mathematical equations or lines of computer code are written, there is usually some considerable, if not always explicit, mental construction of some simplified analogue of reality—a '*model world*'! In a model world which we create, the possibilities are constrained, the fundamental rules are asserted rather than discovered, and we can press rewind, change the starting point (or even the rules), and let time run again, as often as we want. This is essentially what we always do, at least implicitly, when conceiving of any object, system, or principle.

When a bridge is designed, it does not matter exactly how many turns will be applied to each bolt—what matters is the basic structure and sufficiency of all beams and joins—although the final bridge will manifest countless minute details. All gadgets are manufactured to 'tolerances' which, while acknowledging the impossibility of producing an exact incarnation of a blueprint, provide guidance on the scale of acceptable deviation. When we say the word 'house' we have access to substantial shared meaning with all speakers of English, without having to spell out anything else, because we all carry around with us the *concept*; in other words, a *model*, of a house.

The point of this lengthy conceptual preamble is that the first, **and most important**, step in *practical mathematical modelling* does not require any

substantial detailed knowledge of *the formal subject of mathematical modelling*. It is an *inherently* interdisciplinary process, and one to which professional modellers are but one of several crucial contributing groups. The more this idea becomes accepted, the better will be the models that are implemented, and the better will be the interpretation of modelling investigations into real world action—a process lately popularly known as knowledge translation. What we are calling Starfield's 'three-realms' 'model of modelling', then, emphasizes that before we write down equations, or generate computer code to run simulations, we need to spell out, preferably in plain English (or another natural language) and simple diagrams, which hypothetical world(s) we wish to explore. We are free to create as many new model worlds as we like. One of the most important rules of thumb in model world construction is to start as simply as possible. In other words, start by constructing the simplest model world in which anything resembling your actual question can actually be asked, even/especially if this version of your question is so simple that you can answer it.

For example, if we wish to model the spread of seasonal influenza, we may want to explore a model world characterized by the following characteristics:

People infected with flu virus become infectious immediately, and remain infectious for precisely 7 days.
Any pair of individuals, regardless of infection status, have some given probability per day of coming into contact with each other in a manner that would cause transmission if one of them is infected and the other not.
Once people cease to be infectious after 7 days, they cannot be infected again.
No one dies during the timeframes under consideration (3 months of flu season).

This relatively simple to grasp, if obviously simplistic, set of rules specifies a fully-fledged 'model world': a realm in which we know precisely what the rules are—because we declared them!

The model world is further unlike the real world in that we get to choose, arbitrarily, in what state the population is at some point in time—without regard to questions about how we got there. For example, we can 'start' with one individual who is infected with flu, living among 1,000 who are not, and have never been; and we can do this without having to ask, much less

explain, where this one person came from, or how he/she became infected with flu. Then we can ‘let four months pass’ under the rules specified earlier, and we can investigate how the world unfolds. We can ‘dilute’ the transmission process by declaring that some proportion of the individuals begin the flu season with immunity from a vaccine they took before the outbreak, and we can consider different population sizes and numbers of cases specified in the ‘initial conditions’ from which our ‘model scenarios’ start.

To generate analysable scenarios, we usually need to go beyond conversational thought experiments. Of course, if we make a simple enough model world, we can immediately tell what will happen. Consider that aliens have landed in an isolated town of 100 people, and they exercise the arbitrary power to seize one person every day for their amusement and nourishment. It does not require mathematical skill to conclude that in at most 100 days, the town will be devoid of living people; somewhat earlier if our model world allows for other causes of death, or successful escape—but we have the power to rule these out, if we wish.

Alternatively, perhaps the aliens need a day to recharge their stun gun. They aim at one juicy looking human early in the morning, but, alas for the aliens, the stun gun bullets are flaky and half of them misfire. Then we can no longer be sure that one person dies per day—but every day, there is half a chance of someone dying.

Consider yet another possibility: each alien needs to eat one person each day to survive (the food on board having rotted rapidly after exposure to earth’s bacteria after landing). In fact, if the aliens aren’t doing so well themselves in the face of earthling microbes, then we would need to keep track of the number of aliens who are healthy enough to hunt, if we wish to be able to specify the number of human deaths.

The point of this cartoonish example is to show that full specification of model world rules can take many forms, but, if done carefully, often leads to simple mathematical ideas. In the three examples of aliens hunting for humans, we have human deaths occurring, respectively:

at a steady rate of one per day;
randomly, but, on average, once every other day;
at a rate determined by the current number of healthy aliens.

A key determinant of the satisfactoriness of a story, especially in the realm of science fiction, or fantasy, is no doubt whether the model world created by the author seems internally coherent, and whether the particular storyline (scenario) plays itself out in a manner that is consistent with various explicit and implicit rules we have discerned. Even children want to know whether and how it makes sense that Lord Voldemort was apparently killed by his own death spell, in the final climax of Harry Potter and the Deathly Hallows. Fortunately, the internet provides parents with detailed analyses of the nature of wands, and wand 'allegiance' or 'ownership'. Armed with this knowledge, parents can reassure their children that it's not a case of 'it's complicated', or just some arbitrarily slapped together plot holes. The natural ability of people of all ages to discuss, critique, and savour stories is ample testament to the universal capacity of the human mind to build, upload, analyse and critique model worlds by wielding substantial sophistication, insistence on internal consistency, and intuition.

And just as the naturalistic learning and investigations of children are similar to, but not the same as, formal scientific inquiry, so creating and exploring model worlds in words, pictures, and thought, is intimately related to, if not quite the same as, quantitatively generating 'scenarios' through mathematical devices such as algebra and computer simulation.

Historical overview of disease modelling

Perhaps the first application of dynamical mathematical modelling in public health was carried out in 1760 by Daniel Bernoulli to understand the efficacy of inoculating people with smallpox to prevent their developing the disease later. Bernoulli used what we would now call an 'SR model' (of susceptible and removed elements) to elucidate the variations in the size of susceptible and removed subpopulations over time (Bernoulli 1760; Dietz and Heesterbeek 2002; Bacaër 2010). Bernoulli concluded that, while inoculation carried some risk, it could substantially increase life expectancy, not only by trading the risks and benefits for individuals but (crucially) also by reducing overall transmission in the population.

More than one hundred years after Bernoulli's seminal work, Ronald Ross showed that malaria infects people through the bites of mosquitoes (Ross 1895, 1899), and for this discovery he was awarded the Nobel Prize for Physiology or Medicine in 1902. In fact, Ross may have been a better

mathematician than he was an entomologist. He developed one of the first mathematical models of disease transmission based on differential equations, to support his claim, in the face of considerable scepticism, that malaria could be eradicated by reducing the mosquito population to below a critical level, without the need to eradicate the mosquito population entirely.

The notion that there are critical thresholds on dynamical parameters, below which disease transmission cannot persist, is perhaps the central point of mathematical models of the spread of infectious agents. This is most famously embodied in the concept of R_0, the basic reproduction number, defined next.

Another crucial example of such a 'tipping point', which can only be understood through the collective dynamics of a population, is the attainment of 'herd immunity' at partial coverage of an imperfect vaccine. Even if there are still infected and susceptible individuals in a population, if transmission is not efficient enough to replace cases as quickly as they are naturally removed, then prevalence and incidence are doomed to decline—ultimately to zero (see online appendix for more details).

Ross's model, published in the second edition of his book The Prevention of Malaria (Ross 1910, 1923), became the basis for subsequent modelling of vector-borne infections. He concluded, as also crucially cited in Bacaër (2010):

> As a matter of fact all epidemiology, concerned as it is with the variation of disease from time to time or from place to place, must be considered mathematically, however many variables are implicated, if it is to be considered scientifically at all. To say that a disease depends upon certain factors is not to say much, until we can also form an estimate as to how largely each factor influences the whole result. And the mathematical method of treatment is really nothing but the application of careful reasoning to the problems at issue.

In the late 1920s and 1930s, William Kermack and Anderson McKendrick introduced continuous time epidemic models that incorporated stochastic (random) elements of infection and recovery (McKendrick 1926; Kermack and McKendrick 1927; Harvey 1943; Bacaër 2010). They showed how to include age, or time since infection, as continuous variables, and how to generalize models to deal with time-dependent infectiousness and recovery rates. Together with Hamer, Soper, and Ross (Hamer 1906; Ross 1923; Soper 1929), they laid the early foundations for modern mathematical modelling in epidemiology. Crucially, they realized that epidemics are driven by the rate of contact between susceptible and infectious individuals.

Because the simplest models do not distinguish between individuals, all uninfected people are equally likely to be infected and the probability that a susceptible person is infected is simply proportional to the number (or proportion) of people who are already infected. This so-called principle of mass action, first formulated to model the rates of chemical reactions, now found applications in infectious disease epidemiology. It simply captures the idea of two (sub) populations being so perfectly mixed that the 'contact rate' is proportional to each (sub) population size (or proportion) individually.

When considering the control of transmissible diseases, the single most important concept is the basic reproduction number, conventionally written as R_0 and most commonly thought of as the number of secondary cases produced by a single primary case introduced into an otherwise entirely uninfected population. The roots of this concept can be traced through the work of Alfred Lotka, Ronald Ross, and others, but its first modern application in epidemiology was by George Macdonald (Lotka 1923; Ross 1923; Macdonald 1957), in constructing population models of the spread of malaria. The concept was extended and developed by Dietz (Dietz 1980) and by Anderson and May and their various colleagues and coworkers. The collaboration of Anderson and May culminated in their definitive book *Infectious Diseases of Humans: Dynamics and Control* (Anderson and May 1991).

The number of publications applying mathematical models to answer questions regarding infectious diseases has increased dramatically in the last two or three decades, due largely to the unprecedented growth in computational power. Models are increasingly relied upon to inform public health policy and funding decisions, examples of which will be discussed later in this chapter.

Key concepts for modelling infectious disease are outlined in Box 5.16.3.

Box 5.16.3 Basic concepts crucial to the modelling of infectious agents

(See online appendix for a more detailed treatment)

- **Basic reproduction number**: number of new cases directly produced by one case in a purely susceptible population.
- **Disease duration**: length of time individuals spend in an infected state.
- **(Disease) mortality**: rate of death (usually probability per year).
- **Equilibrium**: a scenario in which the state of the model world is unchanged by time.
- **Herd immunity**: collective ability of a population to resist or extinguish an outbreak, despite containing susceptible individuals.
- **Incidence**: rate of occurrence of infections.
- **Prevalence**: proportion of a population affected by a condition at a point in time.
- **Model scenario**: 'model world' history with fully specified initial conditions and rules.
- **Parameter**: a variable which expresses a rule—a model *input*.
- **State variable**: a variable capturing a model *output*.

Types of model worlds, or mathematical models

In this section we give conceptual definitions of, and brief remarks on, some of the most important model-world/model 'types' encountered in disease modelling. The list is far from a systematic review, and avoids overanalysis of grey areas. The categories are cross cutting. For example, compartmental models may be deterministic or stochastic, and they may be discrete or continuous, in any combination.

Static vs. dynamic models: In *static* models, the rules affecting individuals are fixed. For example, the incidence of infection is given a predetermined value, and no interaction between individuals is considered. In *dynamic* models, the rule defining parameters may explicitly be varied over time, and the time-dependent prevalence of an infection may be allowed to play a role in transmission of the infection. Static models, commonly used in cost-effectiveness analyses, can track individual life trajectories involving costs and benefits of health system interactions, within a context that is declared by fiat. Here *static* doesn't mean that nothing happens in the model world. Rather, the term indicates that we are ignoring the emergent impact (feedback) of factors like health system interventions, or prevalence of disease, on the overall context in which individuals will find themselves. In a *dynamical* model world, such feedback is considered. For example, an intervention may change infectious pressure in some non-trivial way as time passes. This can be crucial to sensibly modelling long-term costs and benefits. A *static* model can estimate long-term costs for individuals who received the vaccine or not, but it cannot account for herd

immunity—i.e. the indirect protection individuals obtain by interacting with immune individuals.

Compartmental vs. individual-based models: In a *compartmental* model, population members move between compartments or health states according to rates determined by parameter values that are defined at an aggregate level. Within each compartment, all members are equal in every detail, and indeed utterly lose their individual identity. When noting *how many* population members flow out of a compartment, during a given period, there is no way we can even sensibly ask such questions as *which* population members they are, or how long they have been in the compartment they are leaving. An *individual-based* model (also known as a *microsimulation* model or *agent-based* model) explicitly keeps track of individuals' particular states, and, in principle, their individually tailored tendency to experience events. Infection, relationship formation or dissolution, death, and so on, are simulated as occurring to *particular* individuals. A key feature of this approach is that it allows for extensive, complex, and not-consciously-predetermined heterogeneity between individuals, whereas compartmental models must explicitly enumerate all possible states in which population members may be found. Both of *compartmental* and *individual-based* types of models can be either *dynamic* or *static*.

Deterministic vs. stochastic models: In *stochastic* models, there is intrinsic randomness in the occurrence of events. The key epidemiological processes, such as transmission of infection, are governed by *probabilities*, or *probabilities per unit time* (or time step). These probabilities are of course not just an expression of vagueness. They capture, or, depending on your point of view, emerge from, specific dynamical rules. In *deterministic* models, the future is entirely *determined* by the choice of a combination of 'initial conditions' and 'dynamical rules'. Usually, when we construct a *deterministic* model world, we are actually thinking of capturing some aspects of real world dynamical processes that are at least in part *stochastic*. Then we usually interpret the outputs of a *deterministic* model to be *average numbers of events* (like infection or death), or *average sizes of population groups* (like people requiring long-term HIV treatment) that are *expected* (rather than strictly *determined*) in a model world where the dynamics are ultimately *stochastic*.

Population-level vs. *in vivo* models. Models of biological processes within an individual, from the life cycles of cells and infectious agents to

pharmacokinetic and pharmacodynamic processes, are often referred to as *in vivo* models. In contrast, population-level models focus on a higher level/scale of aggregation. Usually this point is not explicitly noted, as most epidemiological models disregard complex biological details. However, it is useful to be aware that there are interesting questions at both scales, and the fundamental modelling process is very similar, and even leads to similar structures and processes, like healthy and infected subpopulations, cycles of death and reproduction, transmission of infection status through contacts and environmental exposures, etc.

Discrete models vs. continuous models. In discrete models, features, such as time itself, or other quantities expressing the state of the model world, jump in discrete steps. For example, individual level models always have whole numbers of people. In continuous models features vary continuously, being able, in principle, to take any value in some sensible range. We can run a continuous model until 0.4 people are left alive. It is not at all unusual for a single model to simultaneously conceive of some aspects continuously and others discretely—for example, discrete jumps in population states occurring against a continuous timeline and continuously varying risk factors.

Fundamental models

This section aims to capture the essence of dynamical modelling by demonstrating, in considerable detail, including conceptual digressions, some very simple models. We hope to show that modelling is really not something strange or difficult, accessible only to a closed circle of initiates or specialists.

Useful epidemiological models can be complex, and many models appearing in the literature are, we would argue, unnecessarily complex. ‘Occam’s razor’ is the tried and tested philosophical heuristic that encourages us to cut away unnecessary details and assumptions in our search for explanations, and it applies with great consistency in formal modelling. We noted earlier that one should always *start* with the simplest possible model worlds in which we can see any trace of the issues we are exploring—and we would now add that indeed one should also stay as simple as possible at all times. The examples next are particularly stripped down—so much so that one could be forgiven for wondering whether they are to be taken

seriously, or whether they really illustrate real contemporary disease modelling. In fact, they fully illustrate a conceptual framework for approaching population dynamic models, and, in particular, disease transmission models.

We trust readers can recall some advanced high school mathematics. Familiarity with basic calculus will greatly help. Even readers with no prior modelling experience, but a willingness to take the plunge, notepad in hand, should be able to follow most, if not all, of the details of these examples, and should find here a foundation for a deeper reading of the modelling literature.

‘Exponential’ growth or decline

The notion of ‘exponential growth’ (or decline) is one of the most important recurring themes in all dynamical modelling—from the pure sciences like physics and engineering, to highly applied social sciences like population dynamics. Exponential growth has little to do with the popular notion that something has grown ‘exponentially’, which often simply means that something has grown ‘enormously’ (explosively?).

Technically, exponential growth simply means that some quantity is changing (growing or shrinking) at a rate *proportional to how large it currently is*, with both the size and the growth continuously adjusting to maintain this relationship, at least for a while, if not indefinitely.

Exponential growth can lead to surprisingly rapid growth in some cases, but it can also be very slow. A simple savings account that attracts a measly few percent interest per year, in a society where inflation runs in the double digits, will experience face value exponential growth in the monthly balance statements, but leave us with (exponentially) less and less purchasing power over time.

Let’s consider a question of the most fundamental kind in population dynamics:

> **Question 1:** *Given a population of 1 million individuals, with individual mortality (m) of one percent (0.01) per year, how many will (are expected to) be alive after 10 years?*

Following our noses, we now build a model world in which this question can be explored. It involves nothing more than spelling out what the rules are:

There are neither births nor migrations in or out of the population.
Each member of the population faces a predetermined fixed probability, *m*, of death each year.

And just like that, our model world has been created. The assertion that we begin with 1 million individuals is not so much a dynamical rule, as a statement about *initial conditions*. To go beyond the purely conceptual, so that we can calculate something useful using the tools of mathematics, we now consider whether there is some subtlety in how we think about time.

Thinking of time as passing in discrete steps

One view on time passing is to accept that either the model world itself, or the mathematics we are using to describe it, is somehow vague on exactly when people die, and that it is only the number of deaths in each year, or alternatively, the population size at the beginning/end of each year that is well-defined.

Another way of spelling things out in gory detail is to come up with a quite contrived (even if absurdly comical) ‘dynamical process’:

At the beginning of each year, 1% of people are handed death notices, requiring them to report to the Grim Reaper at the end of the year.
Just after the deaths all occur just at the stroke of midnight on New Year’s Eve, a new round of death notices is handed out to the survivors.

We can even ponder the possible arcane details of how the Grim Reaper’s administration works. For example:

Perhaps the death scheduler in chief does some rounding off, to the nearest whole number, when deciding, explicitly, according to a formula, how many death notices to issue each year, based on how many people are currently alive.
Perhaps people are required to roll a one hundred sided die, which has one side with a 0, and 99 sides with a 1. A death is scheduled for every person who rolls a zero.
Perhaps there really aren't well-defined individuals here—just an annually revised declaration of the number of individuals. And perhaps we don't much care whether this is a whole number. Then we can simply schedule $(0.01)N$ deaths, even if N is not a multiple of 100. For example, if there are 250 people at the beginning of a year, we will let 2.5 people die that year, without worrying that this is not possible in the real world. In fact, one could argue that 2.5 deaths are not even possible in our model world. The fractional population members could be said to be a trick to create a simple mathematical approximation to what happens in our model world.

The last of these three, even though it is in some sense the weirdest, departing even from our real world experience of people being counted in whole numbers, is actually the most simple, analytically. Allowing for non-whole numbers of individuals is an important and widely recurring device. It allows the stated individual-level risk or *probability of death* to be directly manifested, within each time step, as the actually realized population proportion which dies. This is the crux of focusing on 'dynamical principles', which are well-defined, rather than 'actual realizations' of a dynamical rule, which are usually subject to some sort element of randomness in the real world. This hints at a recurring theme. We can often buy a lot of analytical efficiency, not to mention simplicity and clarity, with just a little bit of abstraction, just a bit of appropriate slippage between either:

the real world and a model world, or
(and this is a point of view not often spelled out) a model world and a set of equations which purports to capture aspects of the model world.

The abstraction from real to model world has at least two distinguishable aspects:

In order to get off the ground, we need some simplification and spelling out of the basic rules that embody what the investigation is about, such as births, deaths, and infection.
There is additionally often a need to decide on some very mundane details, which may at first appear irrelevant, or extraneous to the main question. Everything that can happen in the model world needs to be spelled out with sufficient specificity for the mathematics to be able to enter the discussion, and possibly, as a bonus, for the analysis not to be overly complicated.

It is well understood among scientists, engineers, etc. that there is much value in drawing semi schematic diagrams (suggestive of key aspects of reality) to help visualize complex biological systems like cells or organs. These pictures are usually called cartoons, which is by no means a disparaging term. It's just another example of the process building a model world to support a conversation where there are in principle many complications that we cannot possibly consider in detail. Similarly, it is absolutely normal for modellers to make up superficially silly sounding explanations for how things are decided, coordinated, implemented, and tracked, in model worlds. We enthusiastically encourage an unrestrained approach to making up model world rules, and visualizations to support thinking about all the required details.

Having made quite a meal of the construction of this most simple model world, we are ready to assign actual mathematical relations to what is happening. We introduce the state variable, which is just the population size (N), and which we need to evaluate at discrete values of time, namely 1, 2, 3, … (years in our example, but it could be any unit of time such as days, weeks, months, even hours, and so forth). Letting N_t denote the population after t years, we can now finally proceed to build equations.

The flipside of the declaration about the *proportion of deaths*, m, per year, is that $(1–m)$ is the *proportion which survives* each year. Hence we have the following expressions for the population at later times, in terms of the population at the start:

$$N_1 = (1-m)N_0$$

$$N_2 = (1-m)N_1 = (1-m)(1-m)N_0,$$

and so forth. These lead us naturally to identifying a general expression governing the population dynamic. Given the population at an arbitrary time *t*, we can write the population 1 year later:

$$N_{t+1} = (1-m)N_t.$$

This kind of relation is called a *difference equation*. It tells us something about a function, intended to be evaluated/interpreted at particular steps—in this case, *N*, should be a whole number, interpreted as time measured in years. More specifically, this particular difference equation tells us how the function is related to itself when it is evaluated at two values separated by one time step. Even more specifically, increasing the 'independent variable' (time) by one (interpreted as the passing of one step of time, i.e. one year) causes the function *N* to change by a given multiplicative factor—which we happen to express as (1–*m*).

Difference equations are a widely applicable way of capturing a range of idealized (model world) dynamical rules, like:

Each year, the staff complement of a rapidly growing company doubles.
Each month, my bank adds to my account 0.5% of the average value I have maintained in my account during that month.
Each month, the amount of money under my mattress increases by exactly the $100 I put there on payday.
After I reach age 65: every month, my retirement fund will be affected by two components of change: a decrease by a fixed pay out to me (my pension) and an increment of 0.5% of the value of the fund the moment before the pension pay out.

Most people are somewhat familiar with 'algebraic equations'. A 'linear' equation can be solved for a unique value of *x* that satisfied $Ax + B = 0$, for some given values of A and B (the answer is $x = -B/A$). A quadratic equation ($AX^2 + Bx + C = 0$) can be solved for potentially two values, and so on. Difference equations are not used to solve for *particular values* of *x* in this sense. The '*solutions*' of difference equations are fully specified

functions N_t, which, in order to be solutions, have to obey the difference equation for at least a specified *range of values* of the step variable—in our case, all positive whole (natural) numbers. It can be very challenging to solve such equations, but in this case (Equation 1), it is not difficult to figure out that the population after *t* years is given, completely generally, by

$$N_t = (1-m)^t N_0.$$

Even without any substantial mathematical analysis, we should expect that, as the number of years grows indefinitely large (tends to infinity), the population in Question 1 becomes indefinitely small (tends to zero). This is reflected in key features of the model formulae:

the population size after *t* steps (years) is the original population size, multiplied by *t* copies (or *t* th power) of the factor (1–*m*), and
(1–*m*) is positive but less than one.

An often wielded algebraic/analytical concept is that powers of any positive number less than one can be made as close to zero as we want, by taking a high enough power.

The very particular property that in each time step the population:

shrinks by a factor

or, put differently:

sheds a number of members in proportion to how large the population was at the beginning of that time step

is known as the *exponential decline* of the population. Of course, we can also turn it around and define exponential growth as increasing by a certain *factor* in each time step.

According to the discrete time analysis, then, the concrete answer to Question 1 is given, to the nearest whole number, by:

$$N_{10} = (1-0.01)^{10} \cdot 10^6 = 0.9910 \cdot 10^6 = 904,384.$$

Figure 5.16.1 shows the decline of a population over successive years. The annual mortality has been increased substantially, relative to our preceding example, in order to show the effect more starkly in a small number of steps.

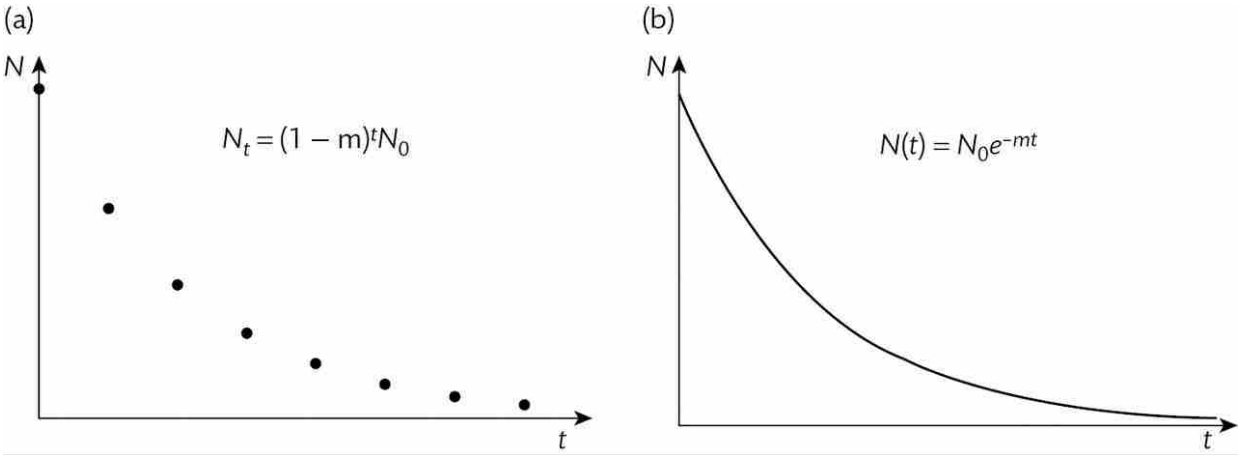


Fig. 5.16.1 Exponential decline (of a model world 'population'), observed (a) in regular time steps, and (b) continuously.

Thinking of time as passing continuously

With time, *t*, now allowed to take any value, not just whole numbers, we want to find a fully specified function of time, *N*(*t*)which expresses, for any (positive) real number *t*, the population size at time *t*. The crucial conceptual leap here is to construct a slightly different kind of model world, still inspired by the same underlying question (the 'Question 1' referred to previously) but a world that fundamentally deals with time continuously. Just as we discovered some details that needed to be spelled out in the discrete time model world, here we are left with choices about the representation of the population itself. Firstly: should the population be allowed to vary continuously, or should the population jump in steps of one, at appropriate times?

As before, we will allow no births or migration to take place in the *continuous time* model world. Before we confront the problem of tracking and updating the population size, we need to adapt the stated mortality dynamic as being a constant *mortality rate m*. In discussing what this means, and how, ultimately, to incorporate the idea into a *mathematical model*, we

will use the language of calculus, but develop the argument intuitively rather than formally.

The recurring elementary task, in specifying the behaviour of a continuous time model world, is to consider what happens to the system when time t advances by a very ('infinitesimally') small increment, usually called *dt* (time intervals which are not necessarily infinitesimal are usually called Δt). The deceptively simple Figure 5.16.2, which does little more than remind us to think about a short time interval, *dt*, is such a useful crutch, that we in all seriousness recommend it be drawn whenever continuous models are being developed.

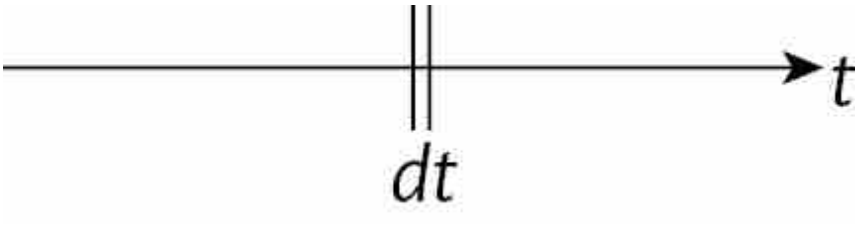


Fig. 5.16.2 The fundamental diagram of dynamical modelling in continuous time. Attention is drawn to what happens during an extremely small ('infinitesimal') interval of time, of duration *dt*. In such a small interval of time, there may be some discernible, and describable, tiny ('infinitesimal') changes in our model world, but they will not, during this interval of infinitesimal duration, feed back into any *changes in how the model world behaves*. This is a fundamental technique of modelling—and thus we are able to derive *differential equations*, whose solution(s), which we obtain by the tools of calculus, will describe how the model world behaves over time.

In the real world, if we were to consider ever smaller time windows, it would become more and more likely that no notable events occur in these time windows. Indeed, the very idea of a time increment, or window, of size *dt*, is of one that is small beyond all measurement and practical meaning. Nevertheless, it turns out to be perfectly sensible to say that the (very small/infinitesimal) probability *dP* of having an event occur, subject to a rate *m*, is just the rate itself, times the (infitesimal) size of the proposed interval, i.e.

$$dP = m^* dt,$$

This could be seen as a fundamental *definition* of a rate, and in the minds of many modellers, it is seen just like that.

While this is a bit abstract at first blush, it has a simplicity, and some analytical advantages, that is not as nicely manifest when accumulating events over real-sized time steps. When we want to know how many people die in a year, all sorts of complications arise. What if different times of year bring intrinsically different hazards? Or what if any of several non-lethal events (like disease acquisition), which can happen in complex ways during the year, then in turn bring with them the risk of complications and death? It can be very difficult to just write down a rule summarizing what all can all happen in a year. In fact—figuring out what can happen in a year is probably the research question, or a crucial non-trivial subquestion. We can hardly expect to get far with a method for solving this problem if that method starts by writing down the solution.

In a sufficiently small time step, however, it becomes much simpler. A key simplifying concept is that there can be no variation in rates during the interval *dt*. This is the nature of infinitesimal intervals—their most important property for modelling purposes. Events may happen within infinitesimal intervals, but these events cannot provide feedback to the rules of the game quickly enough to change the rules during the very interval under consideration.

It is in principle quite possible to construct a model world in which time passes continuously, and individuals die at well-defined points in (values of) time. These points in time can be ‘chosen according to’ (technically, ‘generated by’) a suitably well-defined (but not necessarily constant) *mortality rate*. We will, however, allow the population to be decreased by a continuous ‘flow’ of deaths, which is often much simpler—much like the non-whole number populations sizes we proposed in the discrete time formulation. This is usually done via a rule which forces the model world to track, at all times, the ‘average’ behaviour of a related model world with discrete jumps in population counts. For many investigations, this doesn’t imply any important loss of information or insight, and these models of continuously varying population sizes are indeed a mainstay of scenario models. They allow us to bring to bear, in a very direct and efficient way, powerful ideas and theorems of calculus, which make it possible to uncover detailed and robust analytical insights, and efficiently generate precise and

informative numerical/graphical representations of very complex dynamical processes.

We want to take the step from:

Modelling probabilities of events, to
Modelling a continuous flow that captures the expected accumulation of events.

In a *continuous time continuous-population-size* model world, there is always something happening—populations are flowing into each other smoothly in some way. The way to turn out ideas about probabilities into population flows is as follows:

Retain the idea that every member of some population (potentially itself a subpopulation of some wider group that we are also explicitly tracking) is, by virtue of their presence in the particular population, subject to an *event rate m*.
Hence retain the idea that each population member has an infinitesimal probability, *dP*, of experiencing the relevant event, in a time slice *dt*, with $dP = m * dt$.
Crucially, *declare* that the (infinitesimal) *proportion* of population members which experience the event in the time slice *dt*, is precisely equal to the (infinitesimal) per-individual probability *dP* which we just defined as our dynamical rule in the stochastic view. This is where we build in the rule that what happens at all times is the expected/mean behaviour of a cousin model world where events are truly stochastic and fluctuations are allowed.
Conclude that the number of population members which experience the event in a time slice *dt* is given by $dN = * dP$.

In an infinitesimal time slice *dt*, then, there will be an infinitesimal reduction in the population, due to an infinitesimal number of deaths. All the processes can now be thought of as continuous flows, just as we usually think of the flow of water as proceeding with infinite smoothness, even though we know that in the real world, there are a trillion trillion water molecules tumbling around in a small glass. It’s seldom important to know any details of how those molecules are bumping into each other.

So we have come as far as noting that, between time *t* and time $t + dt$, there will be

$$dP = dPN = mN(t)dt$$

deaths. This is more usually written as:

$$\frac{dN}{dt} = -mN.$$

Which is achieved simply by:

dividing both sides of the equation by *dt*
inserting an explicit minus signs to capture that deaths lead to a *reduction* in the population *N*.

The meaning of *dN/dt*, known as a 'derivative', is, by analogy with the concept of *gradient* or *slope*, an expression for the rate of change of *N*, with the passage of time, *t*. This equation can be solved (technically, 'integrated') by standard methods developed in any calculus course, to yield an expression for the population $N_t$after time *t*:

$$N(t) = N(0)e^{-mt},$$

The symbol *e* denotes the 'base of the natural logarithm', a number (approximately 2.718) which, for analytical convenience, is very much *the* preferred base through which to define powers. Setting

$$m = 0.01, \text{and}$$

$$t = 10,$$

we can calculate the answer to Question 1 given by the continuous model, as follows, rounded to the nearest whole number:

$$N(10) = 10^6 e^{-10 \cdot 0.01} = 904,837.$$

This differs slightly (by about 1 part in 2,000) from the answer given by the discrete model in Equation 2, and the respective solutions will diverge further, as the time passes. This somewhat fine but conceptually interesting point is not an error *per se* in either method, or a sign of some fundamental incompatibility of approaches, and certainly not a reason for concern about the formal manipulations we have performed. In fact, it simply reflects our slightly casual correspondence between the notion of (cumulative) *annual risk of death*, and (instantaneous) *probability per unit time of death* (with time measured in years). This detail is explored a little more in the online appendix to this chapter, along with other fundamental modelling concepts.

The value of m in Equation 3 in the continuous time model is the intrinsic *risk per unit time*. When we change our minds about the unit of time in which we want to work (days versus years, for example) then rates change their values accordingly. A rate defined *per year* does not take the same value as that same rate, defined *per day* (just like a length measured in centimetres does not assume the same value as the same physical length measured in inches). By choosing suitably different units, the value of a rate parameter can take any rate at all—as close to zero as we want, or as large as we want. However, the cumulative probability of an event occurring in a *particular period, T*, like one year for example, is an actual specific number between zero and one. If we now want to work in days, we should also update the question to ask what is the probability of the event occurring in 365 days, and the answer should be the same as the probability in one year. In practice, the cumulative probability of a common epidemiological event of interest actually happening to a particular individual, during an actual year of waiting, is a small number—like a few per cent. In this case, the instantaneously defined probability per year of the event happens to be a very similar number—and so subtle the difference is frequently not apparent.

A theorem of calculus translates *rates* (which are measured in a way that is sensitive to our choice of time units) into well-defined *probabilities* of events occurring in a particular physically meaningful period T (the value of which depends on the choice of units):

$$R(T) = e^{-mT}.$$

Using the corresponding value for $1–R(t)$, the probability of surviving for t years, in place of $(1–m)$ in Equation 1, we can make the *discrete time model*

solution correspond *exactly* to the *continuous time model solution* in Equation 4. All we need to do is slightly adjust the values of *m* to account for the fact that they mean something slightly different in the two cases. Now the only difference between the discrete time *mathematical model* and the continuous time *mathematical model* lies in the fact that we did not intended to evaluate the discrete time model for non-whole-numbered values of *t*—the formulae are in fact identical!

The models we have just constructed also apply to a susceptible population being infected at a constant rate, which might then be called 'force of infection', 'incidence', or 'hazard rate'. Similar problems may be framed in many other contexts, modelled by essentially the same equations: an infected/infectious population disappearing due to a constant rate of recovery, the mass of water remaining in a towel drying on a clothes line, the temperature (relative to room-temperature) of a cooling cup of coffee, the decaying mass of a radioactive substance. Exponential growth and decay are the fundamental recurring theme of population models, as has been pointed out extensively by Lev Ginzburg and Mark Colyvan in their book (Ginzburg and Colyvan 2004).

Know your model world building blocks

It is not necessary here to explore in detail how continuous and discrete time models differ or correspond in more complex cases, but the following remarks may be helpful.

Continuous time models, though they may be daunting to people who are not comfortable with calculus, are in fact usually mathematically much simpler than discrete time models.
The rules which apply over an ‘infinitesimally’ small period of time, *dt*, do not need to be updated over this period of time, because the system itself accrues irrelevantly small (infinitesimal) changes in this interval. This is the key simplification offered by thinking in infinitesimal steps.
Hence, without much technical effort, we can usually solve for the state of the system at the end of an infinitesimal time step, relative to a specified initial state, by simply translating the rules of our model world from words into algebra.
Thus we derive ‘differential equations’.
Then, if we want to know how the system changes over a substantial period of time, we invoke the tools of calculus to ‘solve’ the differential equations.
‘Solutions’ to differential equations are not values of populations at particular times, in the sense that ‘values’ are ‘solutions’ to the more familiar ‘algebraic’ equations. Solutions to differential equations are ‘functions’. This means they are fully specified formal relations between population states and time, which, to be valid solutions to differential equations, provide the values/sizes of population variables at infinitely many points in time, spanning some period over which the rules, captured in the differential equations, apply.
Solutions of differential equations capture internally consistent ‘histories’ of model worlds in particular scenarios.

Discrete time step models may require either extra technical sophistication, or additional assumptions of convenience, in order for the fundamental time-step relations to be written down.

These relations are called ‘difference equations’, and their solutions too are mathematical functions (of time).
For functions to be solutions to difference equations, they are only required to correspond to model world states at certain discrete times.
Sometimes these discrete time steps need to be taken seriously, so that it may not be sensible to simply read off the system state at non-whole-number times.
For instance, it may be possible to build a year-on-year model of an insect population, which accounts for the net result of a brief breeding season with many births and deaths, but does not attempt, nor claim, to describe the intermediate states when many eggs, larvae, and pupae are being formed and eliminated.
Similarly, seasonal models of human disease may not be intended to model the details of brief outbreaks, but merely account for net impact of typical outbreaks in terms of mortality, morbidity, vaccines used, accumulated immunity, and so forth.

Often, indeed usually, neither differential nor difference equations can be neatly solved with explicit formulae. Then it becomes necessary to use computational means to provide numerical approximations, graphical representations, and the like. These numerical methods form a mature and robust field of knowledge in their own right, and can deal routinely with most epidemiologically interesting models, so it is not necessarily a big deal if we cannot analytically solve the central equations in a model-based investigation. Of course, it is always useful to have analytical solutions, if they can be found.

For extensive in-depth investigations, challenges may arise with computational power and run-times required. Exploring many alternative scenarios in an attempt to ‘fit’ models to data, and exploring the sensitivity of outcomes to inputs and assumptions, may involve many millions of complex calculations, and demand quite sophisticated use of data management technology and computer clusters.

To avoid proliferation of analytical threads and details, the remaining examples in this section will be demonstrated with the continuous formulation, chosen for its conceptual clarity and pedagogic value. The development of intuition about the applicability of calculus is a powerful problem solving capacity. Now let’s return to our simple population dynamics.

Question 2. *What will happen to the population of Question 1, with a death rate of 1% per year, if there is immigration of individuals at an average rate of 200 per year?*

We assume a constant per capita mortality rate m (a probability per unit time of death) as before, and now, in addition, add a constant immigration rate b (a number of people coming in per unit time). In other words, in a small time increment *dt*, the change in population is given by:

$$dN = dbt - mNdt.$$

Hence the derivative (rate of change) is given by:

$$\frac{dN}{dt} = b - mN(t).$$

This 'differential equation' can be solved, by relatively elementary methods of calculus, to yield:

$$N(t) = \frac{b}{m} + \left(N(0) - \frac{b}{m}\right)e^{-mt}.$$

Note that the second term of the solution has the previously discussed 'exponential decline' structure. Whatever value is taken for the initial population $N(0)$, this second term fades after a large period of time, so that, as time t tends to infinity, the population size tends to a constant equilibrium value, namely b/m which in our case is 200,000.

The key point is that this is the eventual state of the system *no matter what the (positive) initial value* $N(0)$ *was* (see Fig. 5.16.3). A corollary to this is that, for our particular original population size to be maintained, an average influx of 1,000 individuals per year would be required, since then $N_E = b/m = 1{,}000/0.001 = 1{,}000{,}000$. Indeed, Equation 5 shows directly, without passing through the formalities of obtaining a solution, that the rate of change of the population will be zero when $N = b/m$. We need just ask what it takes to have $dN/dt = 0$, bearing in mind that $dN/dt = b–mN$. This idea of exploring the behaviour of dynamical (difference, or differential) equations in their own right, without directly constructing and manipulating

potentially complicated ‘solutions’, is a normal, and very informative part of the investigation of the properties of model worlds.

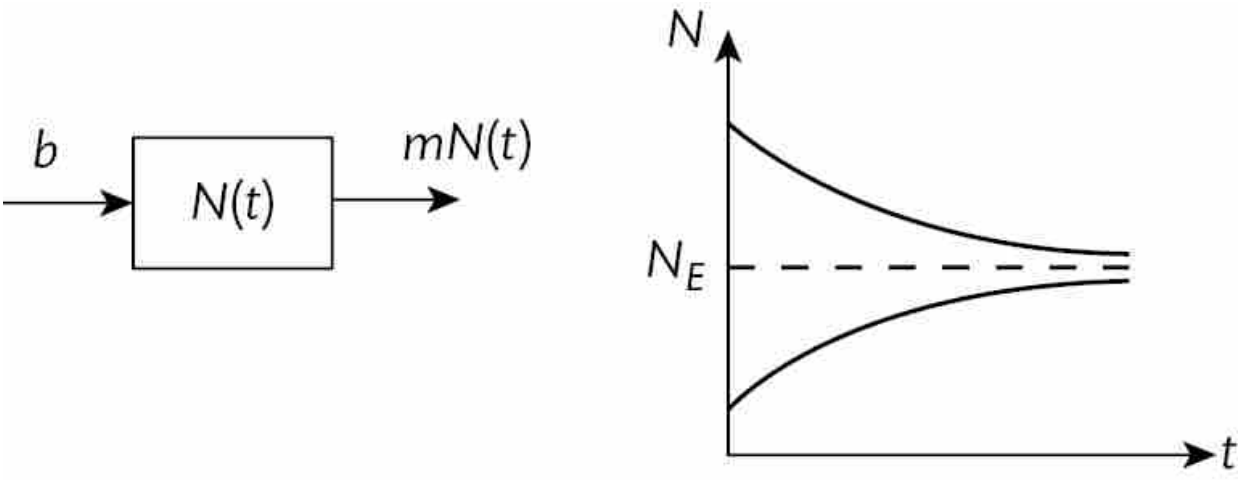


Fig. 5.16.3 Stable equilibrium of a model world population subject to a constant death *rate*, i.e. probability of death per unit time, given by *m*, and being replenished with a steady inflow of *b* new members per unit time. No matter what the initial population, the model world rules eventually drive the population size ever closer to the intrinsic equilibrium value $N_E = b/m$.

The volume of water in a vertical-sided tank can be used as a helpful concrete mental image for this model world. It also makes for a cheap and informative physical demonstration. Consider water in a tank, filled to some level above a small outlet, from which we declare (because we can do that, in our model world) that the *rate of flow* is proportional to the *pressure* at the outlet. Pressure in a simple liquid is just proportional to depth, and the height of the water level above the outlet is proportional to the volume$V(t)$ of water which stands above the outlet (see Fig. 5.16.4).

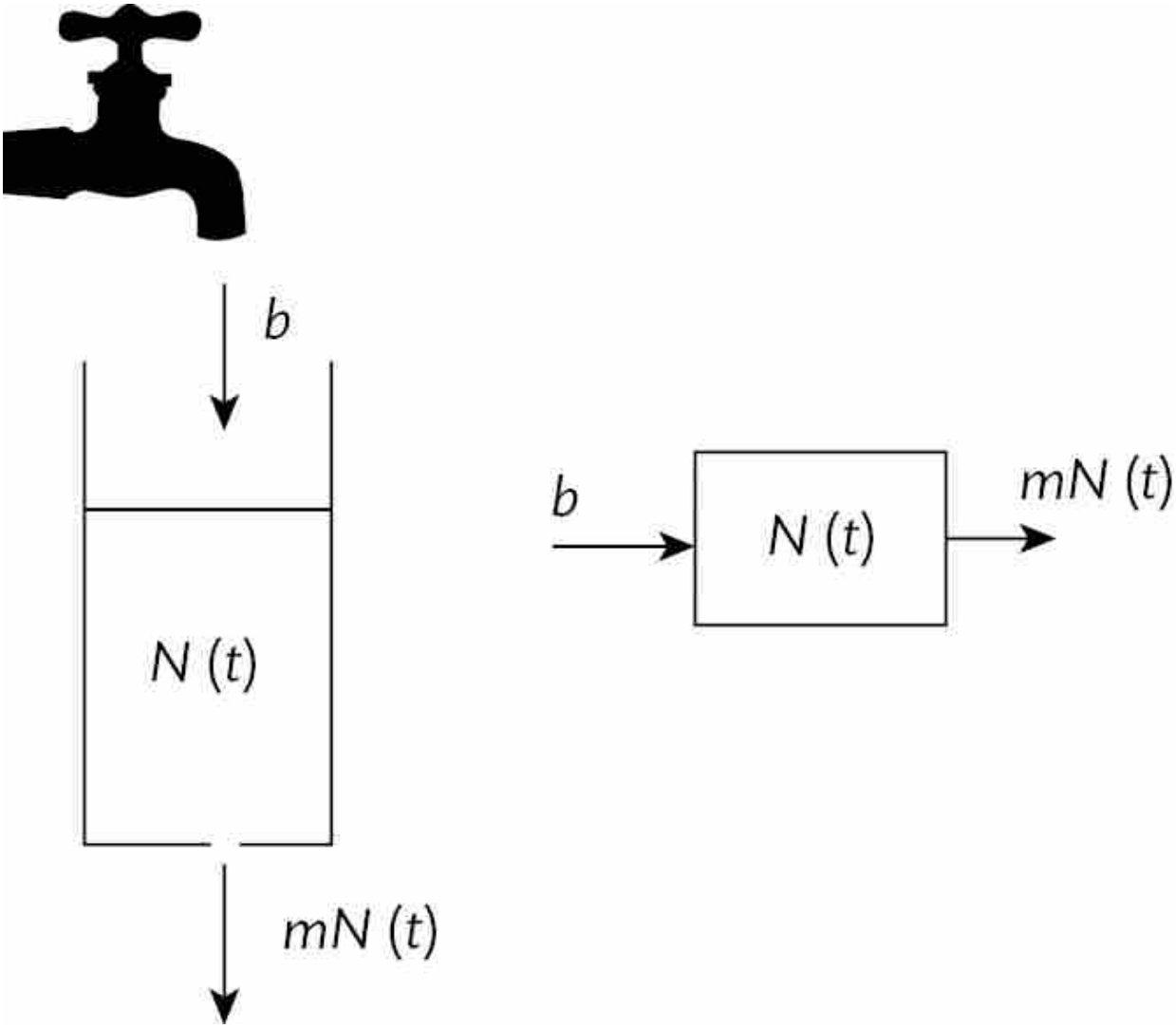


Fig. 5.16.4 Cartoon *Punctured water-tank* and schematic *flow diagram* representation of model world population $N(t)$ with constant inflow b and ('linearly') variable outflow $mN(t)$.

If the tap feeding into the tank is shut off, water will just run out of the tank through the hole on the side, at a rate that is always proportional to the height of the (ever declining) water level above the hole. In other words, the level will *exponentially decay* down to the level of the hole. If the tap produces a steady stream of water, the level in the tank will approach an equilibrium value at which the outflow (which depends on the water level) matches the inflow (which is fixed). In principle, the tank may overflow if the hypothetical equilibrium water level is higher than the top of the tank, but if we want to, we can make our model world tank automatically be big enough.

This inflow–outflow process can be applied to the population of red blood cells in a human body. Bone marrow normally produces a roughly steady supply of about $2 * 10^6$ cells per second, or $1{,}728 * 10^8$ per day, while a nearly constant *proportion* of cells die daily. If we assume the body maintains an equilibrium population of about $2 * 10^{13}$ (20 trillion!) red blood cells at any moment, that mortality rate can easily be calculated from our model. (It's really worth trying it out—the 'real world' answer can be looked up readily enough online)

Question 3. *What will happen to the population of Question 1, with a per capita death rate of 0.01 per year, if there is no immigration, but a constant per capita birth rate of 0.04?*

Once more, we assume a constant per capita mortality rate (probability per unit time of death) *m*. The new part is in the interpretation of fertility. Most naively, we would say that the probability of any population member giving birth to an offspring in any infinitesimal time, *dt*, is given by:

$$dP_b = bdt.$$

This is pretty much all we can mean by constant fertility rate, and it is a rather crude model of fertility. Clearly our model world does not know about the different roles of men and women in reproduction, and there is also no concept of pregnancy or fertility variations within and between individuals. Nevertheless, the fertility rate b can be chosen to produce some reasonable dynamics. Certainly, a nuanced discussion of fertility, and other demographic processes like changing age structure of a country over time, cannot be supported by models lacking explicit age structure, but these kinds of complications are mainly about proliferation of detail, not mainly about new concepts. The conceptually important step is to translate the probability of birth, per population member, into the actually realized (infinitesimal) number of births (*dB*) in the generic infinitesimal time slice *dt*:

$$dB = bNdt.$$

Considering also the mortality, the net change in population, during *dt*, is given by:

$$dN = bNdt - mNdt,$$

hence the derivative (rate of change) is given by:

$$\frac{dN}{dt} = (b-m)N.$$

This can be solved to give:

$$N(t) = N(0)e^{(b-m)t}.$$

In Question 3 we proposed values of m = 0.01, b = 0.04. The model world manifests an exponential population *increase* when $m < b$, a *constant* population when $m = b$, and exponential *decline* when $m > b$ (see Fig. 5.16.5). Exponential growth or decline is also called 'geometric' increase or decline (but less frequently, these days). When there is exponential growth, it is easy to show that there is a fixed 'population doubling time', and in the case of exponential decline there is a fixed population 'halving-time', sometimes called the 'half-life'.

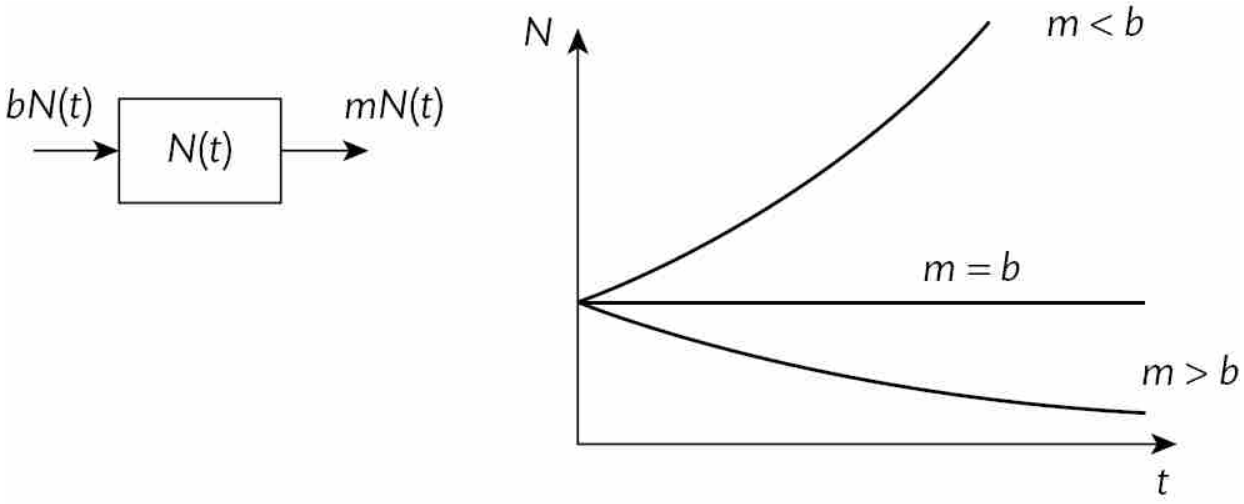


Fig. 5.16.5 Exponential growth (m < b), decline (m > b), and equilibrium/maintenance (m = b) arising from competing replenishment and attrition, both of which proceed in direct ('linear') proportion to the instantaneous population size.

That populations, even under much less simplistic assumptions, tend to increase ‘geometrically’ was first observed and studied by the great mathematician Euler in the mid-eighteenth century (Euler 1748), before being more famously explored by Thomas Malthus (Malthus 1798). Although exact exponential relations such as those presented here arise only under the idealized conditions in model worlds, their conceptual importance and wide applicability can hardly be overstated. With a technically small but far-reaching mathematical twist, exponential functions also form the basis of much modelling in pure science and engineering, including the study of oscillations and waves.

Susceptible–infectious compartmental models

The introduction of multiple health or disease states into our model worlds is naturally crucial to modelling disease transmission and progression. We will also need to expand our options for thinking about rates. Thus far, the populations in our example models have been allowed to:

grow by fixed in- or out-flows (immigration or cell generation), or
experience growth or decline processes scaled in proportion to the instantaneous population size (our fertility and mortality models).

Next, we consider the most basic model of *actual disease transmission*, previously noted as utilizing *the principle of mass action*. Thus, the number of new infections occurring in a time window of duration *dt* is proportional to the number of infection-susceptible individuals, and, importantly, also proportional to the number of infectious individuals. To be proportional to two things at the same time requires being proportional to the product of the two.

This means we need to choose a value for the constant of proportionality, relating the product of infectious and susceptible individuals to the rate of occurrence of new infections. This parameter can be understood to capture both:

the *probability per unit time* of any one member of one group coming into contact with any one member of the other, and
the *ordinary probability* (not *per unit time*) of such a contact leading to transmission of infection.

If the rate of interaction is twice, in model world A, what it is in model world B, but the probability of successful transmission, given an interaction, is half, in world A, what it is in world B, then emergent value for the final model transmission rate is the same in the model equations for A and the model equations for B.

So we have another interesting and useful aspect of introducing figurative daylight between the 'model world' discussion and 'mathematical model' analysis: we can be explicit about details in model worlds which do not directly manifest in the model equations, and we can see how some views of models (and data) cannot distinguish one dynamical rule (model world) from another. This has important consequences for interpretation of, and inferences intended to be based on, models. If we only think about equations, we will necessarily leave aspects of model worlds undetermined. This is not a problem *per se*, but it is good to be able to keep track of one's mental images, assumptions, etc. and to know which rules and principles are manifesting; where, when and how, in our *model worlds* and *mathematical equations*.

Question 4. *How can we describe the prevalence of gonorrhoea in a closed sexually active population?*

Let $S(t)$ and **I(t)** denote the numbers of susceptible and infectious people at time t. We can build a model world based on the following rules:

Individuals mix randomly, any pair of them coming together with the same probability (per unit time) as any other pair.
People become infectious as soon as they are infected.
They recover at a constant per capita rate r per day, at which point they immediately become susceptible again.

In a small period of time, dt, the increments in numbers of susceptible and infectious individuals are:

$$dS = -KI(t)S(t)dt + rI(t)dt$$

$$dI = KI(t)S(t)dt + rI(t)dt,$$

where K is the mass action proportionality constant. Seen from the point of view of an individual, this indicates that the probability of infection, in a small time slice dt around the time t, is given by (see also Fig. 5.16.6).

$$dP = KI(t)dt$$

Which can also be stated as the 'incidence' of infections, among susceptible individuals. This is an important measure of the intrinsic risk associated with an infectious disease epidemic.

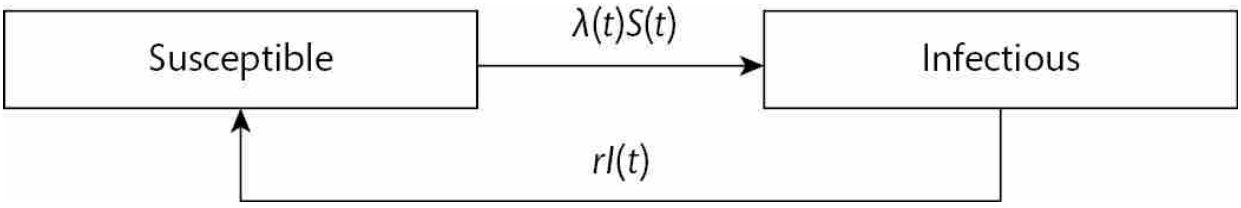


Fig. 5.16.6 Flow diagram of a basic transmission and recovery model. Individuals move from the *susceptible* class to the *infected* (and *infectious*) class with an incidence λ(t) that is explicitly allowed to vary dynamically, in order to capture potentially complex rules for contacts with individuals in the *infectious* class. Infected individuals move back to the susceptible state by recovering with a (constant) probability per unit time r. Equation 6 captures the dynamics for a simple gonorrhoea model by modelling contact according to the principle of mass action (i.e. perfectly random mixing).

Another important measure of infectious risk is what is usually called the basic reproductive number, usually denoted R_0. This is essentially to be understood as the number of new cases that would be produced *by a single infectious person* who is introduced into a large population of people susceptible to infection. The key point is that we are only counting persons *infected by* this artificially introduced infectious individual (sometimes called the *index case*), and not those infected, in turn, by others who acquire the infection over the course of the epidemic. Calculating R_0 for complex model worlds can be a painful exercise in rearranging the model equations, but the important points to understand are:

If the index case is not expected to infect at least one additional individual before either recovering or dying, or somehow ceasing to be infectious, then there is no prospect of sustained spread for the infection.
If the index case is expected to infect numerous individuals, then there will be, in a deterministic model world, at least a period over which the number of infectious cases increases.

Depending on the rules for transmission, there will be some dynamical feedback as the infection spreads, reducing the number of still-susceptible cases. Unless there is a mechanism to replenish the susceptible population, this group will eventually decline to the point that the infected individuals, even if there are many of them, simply don't have enough potential contacts with susceptible individuals to sustain spread. In this case the epidemic 'burns out' without infecting everyone.

Epidemic model equations can be subjected to some considerable instructive detailed analysis beyond our immediate scope. Hethcote and Yorke (Hethcote and Yorke 1984) noted that their model has prevalence declining to zero for plausible estimates of the average partner change rate, the transmission probability per partnership, and the duration of infection.

Indeed, as just noted, model worlds involving infections for which survival or cure confers immunity, or for which infection is incurable, generically manifest epidemics that die out (potentially after rising to a peak) unless there is a sufficient source of new susceptible individuals. In our gonorrhoea model world, the cured individuals recover to be susceptible again, so, in principle, an endemic equilibrium might have been produced, though at apparently infeasible values for certain rates. As it was, the model led to a conundrum in the context of data on gonorrhoea as it could not reasonably explain the low, stable equilibrium prevalence in the United States in the 1980s.

Question 5. *In what kind of model world can we maintain an observed low endemic prevalence of gonorrhoea in a closed sexually active population?*

Hethcote and Yorke (1984) went on to build a further model (Fig. 5.16.7) in which they divided the sexually active population into two groups: low activity (the majority) and high activity—2% of the population with higher partner change rate. They showed that with plausible estimates of the

relevant parameters they could convincingly model the low gonorrhoea prevalence shown by the data.

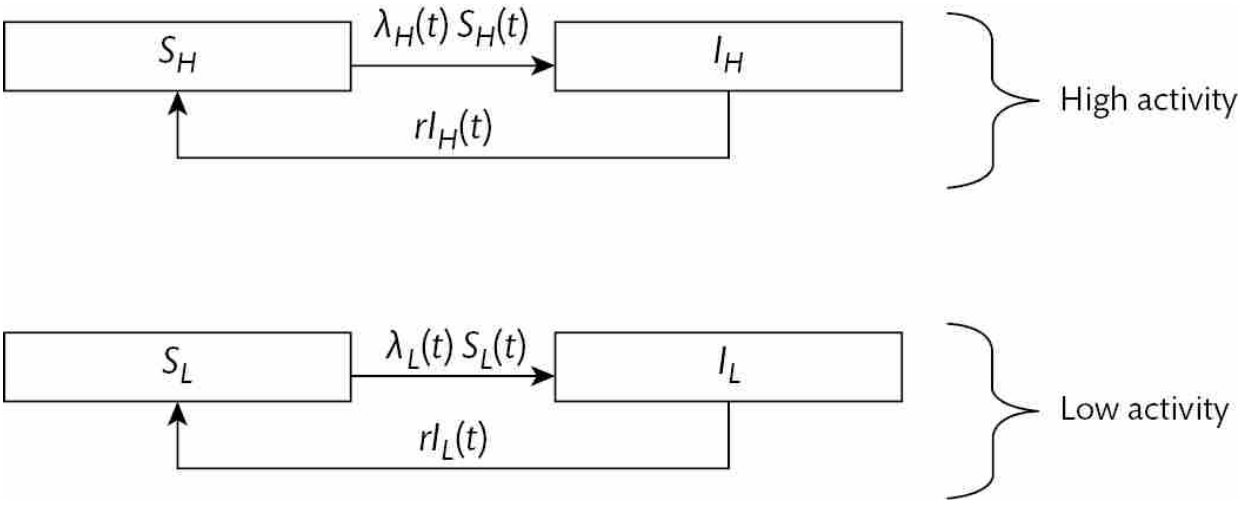


Fig. 5.16.7 Flow diagram of a transmission model with simple contact heterogeneity. The population is divided into a high- and low-activity (contact rate) class. Within each, individuals move from the susceptible class to the infected class with the applicable activity-level-specific incidence λH(t) or λL(t). They move in the opposite direction through a common (and constant) recovery rate *r*.

It is very common that simple first models of systems sensibly capture initial insights and data, but fail to reproduce crucial features which are known to arise in the real system. It is also very typical that one seeks to address this by identifying some important further refinement of population classes. This will require thinking about estimates of subpopulation sizes, and details of contact and mixing rules. The work of Cooke, Hethcote, Nold and Yorke (Cooke and Yorke 1973; Yorke et al. 1978; Hethcote et al. 1982; Hethcote and Yorke 1984) highlighted the importance of incorporating in modelling both risk-heterogeneity, and also of accounting for the class of asymptomatic individuals. The implications of their work helped to inform US control policy for sexually transmitted infections in the 1980s.

Our somewhat dry technical examples have only considered:

Single populations experiencing in- and out-flows.
Simple structured populations with susceptible and infected classes, and high or low contact tendencies.

Other key population structures which may be important to consider are:

Sex/Gender
Age
Location
Relationship status
Finely resolved disease states such as exposed but clinically healthy and uninfectious, or
Detailed viral and immunological markers
Treatment access and utilization

Putting to one side the list of potential complications, we want to return to our opening remarks for this section. Model building and interpretation, while deserving care and attention, and benefiting from technical skill and experience, should not be seen as obscure activities left to the obscure practices of a secret society.

Example applications

Modelling can help us to understand the natural history of the infections or diseases that we are studying. Before we start to build models to explain these different time trends, it is essential to learn more about the biology of the disease and think of other data that is crucial in understanding the unfolding of disease incidence, such as treatment and vaccine availability and coverage.

Historically important infectious diseases

Figure 5.16.8 shows the number of reported cases for each of four diseases in England and Wales between 1930 and 2000 (Public Health England n.d.). Before we can begin to make models to explain these different trends, we need a better understanding of the biology and control of the individual diseases.

Diphtheria is a bacterial upper respiratory tract infection spread through contact with saliva. A vaccine has been available since the 1920s, yet the number of cases in the United Kingdom remained very high at around 60,000 per year between 1930 and 1940. Only when mass childhood

immunization was introduced in 1940 did the number of cases consistently drop to effective elimination in 1950.

Measles is highly contagious and almost everyone who come into contact with aerosols from infected people will get the disease. Survival of an infection ensures lifelong immunity and so does vaccination, which has been available since the late 1960s. In Figure 5.16.8 it is clear that England and Wales, almost without fail, experienced a measles outbreak every second year prior to vaccination. Why would this be? Since measles is an immunizing disease, a large outbreak can only occur once the pool of susceptible people has been sufficiently replenished. This implies that the birth rate in England and Wales was such that measles outbreaks occurred every second year. We can use a mathematical model to show that the interepidemic period would be longer for lower birth rates and vice versa.

Pertussis, or whooping cough, is a highly contagious bacterial upper respiratory tract infection and those infected may cough for up to three months. Outcomes of treatment with antibiotics are poor. Before the introduction of a vaccine in the 1950s, incidence of whooping cough was very high at around 100,000 cases per year. Numbers of cases fell dramatically until unfounded beliefs that the vaccine causes brain damage led to reductions in coverage to a low of 30% in 1978. Major epidemics followed soon after and coverage gradually increased and has remained at around 90% since the 1990s. The whooping cough vaccine is ideally given in three doses to infants and one booster dose in young children due to waning immunity.

Polio is in a race with Guinea worm disease to be the second human disease, after smallpox, to be eradicated from the world. According to the World Health Organization (WHO), only 33 cases of polio are estimated to have occurred worldwide in 2018. This viral disease is highly contagious and spreads through the oral-faecal route. The rapid rise in reported cases in the United Kingdom in 1947 remains unexplained to this day. One theory suggests that improvements in water and sanitation lead to an increase in the age of infection and adolescents are much more likely to develop severe symptoms than infants. However, age distribution of cases did not change from before to after the large outbreak. The second theory suggests that a strain of the polio virus previously not endemic to the United Kingdom caused the large outbreak, but no data are available on the circulating strains prior to the outbreak to confirm this theory (Agerholm 1953). Mass

vaccination with an inactivated vaccine from 1956 and then a live, oral vaccine from 1962 lead to rapid decline in notified polio cases and the last case of wild virus polio infection in the United Kingdom occurred in 1984.

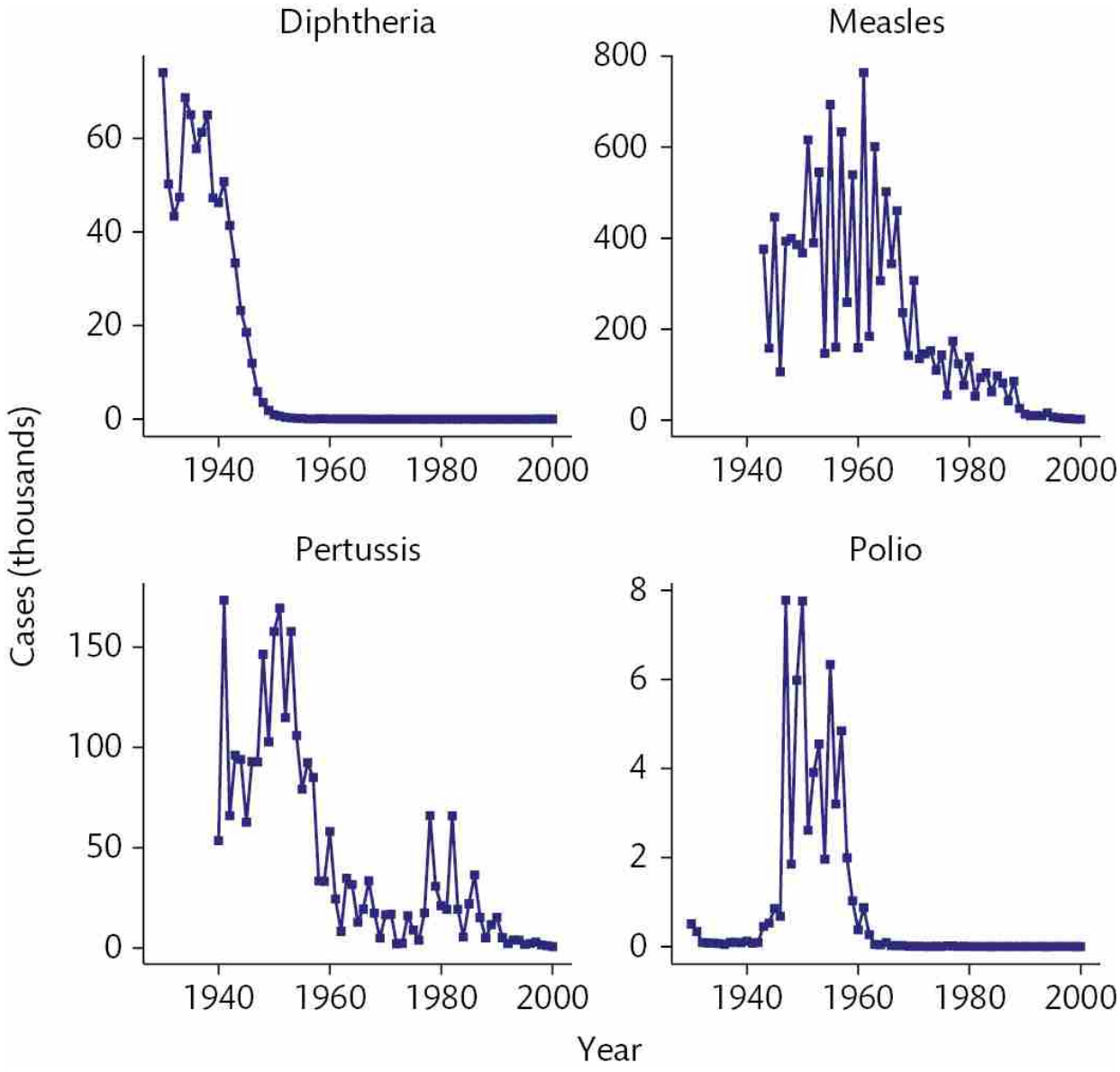


Fig. 5.16.8 Disease incidence trends in England and Wales for diphtheria, measles, pertussis, and polio, from 1930 to 2000. The data points indicate the recorded annual number of cases *in multiples of one thousand*.

Modelling can help us to understand the time trends of these four diseases and help inform decisions that policymakers need to make. All four of these

diseases are preventable through vaccination. We can collect data on the coverage of vaccination over time and develop a mathematical model to see if the increase in coverage (and decrease in the case of whooping cough) sufficiently explains the trends in prevalence over time. Then we can use the model to see whether different vaccination strategies would have had similar impact. For example, would vaccination of targeted age groups have had similar impact on diphtheria and polio prevalence as mass vaccination? What impact would changes in vaccine schedule have on whooping cough prevalence?

Application Laundry list

The following applications demonstrate examples of mathematical modelling in the context of public health:

Encapsulating the quantifiable essence of a complex situation in a simple framework which can then be used to make estimates. Dual infection or co-infection is a good example—infection with human papillomavirus (HPV) is more likely to persist in HIV infected individuals, which may increase incidence of cervical cancer.

Translating descriptions of behaviour at one scale (e.g. individual-level epidemiology of a disease) into descriptions at another scale (e.g. population dynamics). The HPV vaccines used in most public health settings protect against the two oncogenic HPV types that can be found in 70% of cervical cancer cases—will this lead to a 70% reduction in population-level cervical cancer incidence in the long term?

Enabling extrapolation, with quantified uncertainty, from data measured over a relatively short term, to a projection of what is likely to happen in the long-term. HPV vaccines have been available since 2007 and observed efficacy is good—so far. By setting rules in our model world about long-term efficacy, we estimate impact on cervical cancer incidence, a disease that can take decades to develop from initial HPV infection.

Synthesizing data from different fields into a single framework, thus enabling checks on consistency. In modelling the impact of an HPV vaccine on cervical cancer incidence, in a setting with high HIV prevalence, we need clinical information concerning the rates of progression/regression of cervical disease; we need epidemiological data concerning time trends in HIV, HPV, and cervical disease; we need data on cancer screening and vaccination coverage; and we need economic data if we are to assess the cost-effectiveness of new cancer prevention strategies.

Providing clear guidelines for new data collection strategies when trying to answer specific questions where critical features are missing from existing data. This may include identifying what type of data, how much data, and at what temporal and spatial resolution it should be collected, within financial and operational constraints.

Identifying the factors that sustain the spread of infection within a population and estimating the relative importance of various factors/parameters, thus informing control and prevention programmes. In the case of HIV, factors include sexual behaviour, the use of condoms, the extent of ‘intergenerational’ sex, and age at sexual debut. Modellers might seek, for example, to compare the effects on HIV

transmission of sexual heterogeneity, concurrency, and different mixing patterns.

Making predictions over time of the probable dynamics of prevalence, incidence, and infection-related mortality, under different interventions and comparing the impacts and synergies of different combinations of interventions.

Identifying threats to the success of an existing intervention. A model can be used to show that the scale-up of HPV vaccination may have a less than desired effect on cervical cancer incidence, if coverage of screening is not maintained for several more years.

Maximizing cost-effectiveness of interventions, hence enabling optimization of use of resources to combat an infection. Models can assist in evaluating the economics of different control strategies, and can help to decide whether programmes should re-allocate funding in response to the new data or scientific breakthroughs. Providing the new 9-valent HPV vaccine plus HPV DNA testing may be more expensive in the short term, but may reduce cancer rates, and very expensive cancer treatment, more drastically in longer term than vaccination with 2-valent vaccine and cytology based screening.

Concluding remarks

The role of models is to formalize our understanding of, or hypotheses about, data and processes, so that we can tap into powerful tools developed by mathematicians over many years. This allows us to restate our thoughts about systems in an orderly fashion, demonstrating details of their internal workings, in the context of tough political and technical questions about investment in, and management of, public health systems.

We hope readers of this chapter will take away the following key messages:

The conceptual heart of modelling is the construction of interesting 'model worlds'—realms in which we declare how things work, in order to explore dynamics that, in the real world, are too complex to track in detail, or are not amenable do direct experimental control that might teach us what we are looking to learn.

Building model worlds is not a specialist activity to be performed only by technical modellers. Anyone who has a vested interest in a question can contribute to model world construction and exploration.

There are some technical details that must be faced in producing and manipulating mathematical equations of various kinds, which capture the processes we have built in to our model worlds. While this is indeed a specialist activity, there are a number of recurring themes, and even non-specialists can achieve quite a high level of familiarity with important modelling concepts.

ACKNOWLEDGEMENTS

The authors owe a particularly big thanks to Gavin Hitchcock and Brian Williams, who coauthored the previous edition of this chapter together with Alex Welte, and then declared that they were happy for the present authors to adopt, revise, and arbitrarily recycle as they saw fit. GH additionally provided many useful comments on the new version. While we have made deep and extensive changes, relative to the previous edition of the *Oxford Textbook of Global Public Health*, this chapter is in many ways recognizable as a revision of that original work.

FURTHER READING

Anderson, R.M. and May, R.M. (1991). *Infectious Diseases of Humans: Dynamics and Control*. Oxford: Oxford University Press.

Bacaër, N. (2010). *A Short History of Mathematical Population Dynamics*. London: Springer-Verlag.

Brauer, F., Castillo-Chavez, C., and Castillo-Chavez, C. (2001). *Mathematical Models in Population Biology and Epidemiology*. London: Springer.

Keeling, M.J. and Rohani, P. (2011). *Modeling Infectious Diseases in Humans and Animals*. Princeton, NJ: Princeton University Press.

Vynnycky, E. and White, R. (2010). *An Introduction to Infectious Disease Modelling*. Oxford: Oxford University Press.

REFERENCES

Agerholm, M. (1953). Epidemic poliomyelitis and its prevention. *Lancet*, 287–93.

Anderson, R.M. and May, R.M. (1991). *Infectious Diseases of Humans: Dynamics and Control*. Oxford: Oxford University Press.

Bacaër, N. (2010). *A Short History of Mathematical Population Dynamics*. London: Springer-Verlag.

Bernoulli, D. (1760). Essai d'une nouvelle analyse de la mortalité causée par la petite vérole et des avantages de l'inoculation pour la prévenir. In Histoire de l'Académie Royale des Sciences (Paris) avec Mém des Math et Phys and Mém, pp. 1–45. (English translation by Blower. *Reviews in Medical Virology*, 14, 275–88.

Cooke, K. and Yorke, J. (1973). Some equations modelling growth processes and gonorrhea epidemics. *Mathematical Biosciences*, 16, 75–101.

Dietz, K. (1980). Models for vector-borne parasitic diseases. In: Barigozzi, C. (ed.) *Vito Volterra Symposium on Mathematical Models in Biology*. Berlin: Springer.

Dietz, K. and Heesterbeek, J. (2002). Daniel Bernoulli's epidemiological model revisited. *Mathematical Biosciences*, 180, 1–21.

Euler, L. (1748). *Introduction to Analysis of the Infinite*. New York: Springer Science & Business Media.

Ginzburg, L. and Colyvan, M. (2004). *Ecological Orbits: How Planets Move and Populations Grow*. Oxford: Oxford University Press.

Hamer, W.H. (1906). *Epidemic Disease in England: The Evidence of Variability and of Persistency of Type*. London: Bedford Press.

Harvey, W. (1943). A.G. McKendrick 1876–1943. *Edinburgh Medical Journal*, 50, 500–6.

Hethcote, H. and Yorke, J. (1984). *Lecture Notes in Biomathematics: Gonorrhea Transmission Dynamics and Control*. Berlin: Springer-Verlag.

Hethcote, H., Yorke, J., and Nold, A. (1982). Gonorrhea modeling: a comparison of control methods. *Mathematical Biosciences*, 58(1), 93–109.

How to Solve It (n.d.). [Online] Available at: https://en.wikipedia.org/wiki/How_to_Solve_It

Kermack, W. and McKendrick, A.G. (1927). Contributions to the mathematical theory of epidemics I. *Proceedings of the Royal Society of London, Series A*, 115(772), 700–21.

Lotka, A.J. (1923). Contribution to the analysis of malaria epidemiology. I. General part. *American Journal of Hygiene*, 3(Supp.), 1–36.

Macdonald, G. (1957). *The Epidemiology and Control of Malaria*. Oxford: Oxford University Press.

Malthus, T.R. (1798). *An Essay on the Principle of Population*. London: J. Johnson.

McKendrick, A.G. (1926). Applications of mathematics to medical problems. *Proceedings of the Edinburgh Mathematical Society*, 44(1), 98–130.

Polya, G. (1945). *How To Solve It*. Princeton, NJ: Princeton University Press.

Public Health England (n.d.). *Notifiable Diseases: Historic Annual Totals*. [Online] Available at: https://www.gov.uk/government/publications/notifiable-diseases-historic-annual-totals

Ross, R. (1895). Some observations on the crescent-sphere flagella metamorphosis of the malaria parasite within the mosquito. *Translations of the South Indian Branch of the British Medical Association*, 6, 334–50.

Ross, R. (1899). Life-history of the parasites of malaria. *Nature*, 60(1553), 322–4.

Ross, R. (1910). *The Prevention of Malaria* (2nd edn). London: John Murray.

Ross, R. (1923). *Memoirs: With a Full Account of the Great Malaria Problem and its Solution*. London: John Murray.

Soper, H.E. (1929). The interpretation of periodicity in disease prevalence. *Journal of the Royal Statistical Society. JSTOR*, 92(1), 34–73.

Starfield, A.M., Smith, K.A., and Bleloch, A.L. (1990). *How to Model It: Problem Solving for the Computer Age*. New York: McGraw-Hill.
Yorke, J., Hethcote, H., and Nold, A. (1978). Dynamics and control of the transmission of gonorrhoea. *Sexually Transmitted Diseases*, 5(2), 51–6.

Online appendix: key modelling concepts

Here we variously recap and expand on the key modelling concepts of Box 5.16.3, illustrating them by default with reference to HIV. Further elucidations of these concepts may be found in many modelling texts (see, for example, 'Further reading').

Basic reproduction number R_0

This very important concept is somewhat more subtle than usually noted, but there is a valid heuristic: the total number of secondary cases that arise (or are expected to arise, on average) from direct infection by one primary case introduced into a (very large) completely susceptible population.

Disease duration

The length of time for which people have a disease, infection, or medical condition before they are treated, cured, or die. For example, without treatment, people will live with HIV for about 10 years, on average.

(Disease) mortality

The proportion of (infected) people that die per unit of time, for instance within a year. Since HIV-positive people live for about 10 years without treatment, the mortality rate is about 10% per year, averaged over the course of disease.

Epidemic growth rate

The rate at which the prevalence of infection rises at the start of an epidemic. For example, the prevalence of HIV initially grew at a rate of 60% per year

in South Africa.

Equilibrium

A scenario in which the state of the model world is unchanged by the passing of time. This is a subtle concept, and may refer to a scenario in which it is implicit that individuals are entering and leaving various population groups, but in such a way that the total numbers in each group (the ‘state variables’ which define the population) are unchanging. This requirement of ‘detailed balance’ between processes moving individuals into and out of population groups can only be precisely realized in a model world.

It may appear strange or contrived to have finely adjusted all dynamical rules and state variables to achieve such a seemingly delicate balance. The key point to note here is that equilibrium can readily and naturally emerge dynamically from the system over time, even if it was not originally evident. Although real-world systems rarely display anything close to true equilibrium, the concept is very important—and not just for the purposes of formal analysis of mathematical models. Practically, the ideal of equilibrium is a useful concept that enhances insight, for example in understanding where trends, under current conditions, are in principle headed. For example, are we essentially on the way to disease elimination, or apparently doomed to high endemic persistence.

Herd immunity

As indicated throughout this chapter, the heart and soul of dynamical modelling is the consistent propagation, through time, of rates of occurrence of epidemiological events. Crucially, these rates need to be consistently adapted as the state of the population changes from the application of the rules. The feedback loops which this implies are often characterized by critical ‘tipping points’ such as embodied in R_0. One of the most important insights that comes from this kind of analysis is that it is not necessary to protect (e.g. through vaccination) every individual of a population, in order to achieve the elimination of an otherwise endemic infection. Thus ‘herd immunity’ is the collective ability to clear an existing, or resist an invading, epidemic even when some individuals are fully susceptible.

Incidence

This can be thought of as the richly informative cousin of prevalence, and warrants a lengthier discussion because the understanding and estimation of incidence, and its susceptibility to intervention, is ultimately the central technical problem of all modelling of infectious agents. There are several valid ways to define incidence, but ultimately they all refer to the notion of the number of new cases that occur within a given population and period of time. This is sometimes stated as a raw number of cases per time window, such as 1,000 cases of severe influenza needing handling at a major hospital within a month. However, a more 'intrinsic' notion of incidence captures *risk per unit time* experienced in an exposed population. This risk is often expressed as the proportion of a susceptible group who are, or would be expected to become, infected in a period of time. When modellers and statisticians speak of event rates, in general, what they mean is the probability per unit time of an event occurring, clearly stated as referring either to an individual or to some population.

A subtlety is that, even when we express incidence in *probability per year*, for example, this is not entirely the same as the probability of an infection *actually occurring within a year*. Firstly, risk can vary over time, but this changing *instantaneous* risk can still be expressed in units of probability per year. Secondly, even if the risk is constant, any particular rate (probability per unit time) can be expressed in as large, or as small, a numerical value as one wants, by choosing different units of measurement of time. This is just like noting that an object's length can be made as large or small a number as we like, by choosing different units in which to measure length. However, the probability of an event actually occurring saturates at a value of one, which is the limit of certainty of occurrence. This is true no matter how high the rate (probability per unit time) or how long the window of time for which we wait. For hazards of no more than a few per cent, such as the annual probability of HIV acquisition in high-risk settings, the incidence as an instantaneously conceived probability per year, and the probability of infection during the course of a year, are numerically similar enough that they can be regarded as the same thing—especially when we consider the imprecision with which they can be estimated. On the other hand, a flu outbreak which brings an infection risk of 20% per week, sustained for 4 weeks, will affect about 59% (not 80%) of individuals—that is, $1—(0.08)^4$, or one minus the probability of escaping infection for 4

weeks. To press the point: a flu outbreak which brings an infection risk of 30% per week, sustained for 4 weeks, clearly cannot affect 120% of the population! It will affect about 76%.

There is no simple general relationship between prevalence and incidence—a source of much inconvenience and technical pain for epidemiologists. It is simpler at *equilibrium*: if an epidemic of HIV reaches a 'steady state' so that all rates, and the resultant prevalence, are constant, then the number of new infections each year will be equal to the number of deaths. Thus the incidence, expressed in units of events per unit time per total population (as opposed to per susceptible individual) is equal to the ratio of prevalence to duration (with a consistent set of units to measure time). Hence a 'stable' HIV epidemic (without effective treatment—see '(Disease) mortality', earlier) has a value of incidence (expressed as infection risk per year) of about one-tenth of prevalence. In reality, there is no equilibrium, and incidence and prevalence vary in complex ways as functions of both calendar time and age.

Model scenario

Mathematical models, in the basic abstract sense, capture rules about processes in a model world, a world which *by definition* functions according to *rules* captured in certain mathematical statements. This however leaves (wide) open the question of *what actually hap*pens in this model world, as the rules by themselves do not spell out a 'starting point' or 'initial condition' from which the system is then driven by the clearly stated rules. For example, we may have captured rules about the transmission of HIV, but when applied to a population in which there is at some point no infected individual, these rules will not lead to the spontaneous appearance of an outbreak. When populations have structure, such as heterogeneity in age, gender, and various other 'risk factors', these also need to be specified at some initial time, in order for the rules to be applied to a concrete 'realization' of the model world. This fully specified realization of a possible history of a model world, in which rules as well as initial conditions are spelled out, is often called a 'scenario'. A scenario may also involve arbitrary 'midstride' impositions of rule changes, such as the initiation of a vaccination or treatment campaign at some point in time. In epidemiology, such 'interventions' by human agency are usually treated as arbitrary

changes to the model ‘rules’, rather than as events which unfold according to some complex higher level sociological rules—as this level of emergent behaviour is almost impossible to build into a model world. Model world rules are usually best kept as close as possible to rules of nature that can plausibly be modelled.

Parameter

Any variable in a model which captures a *rule* about the model behaviour, such as the rate of recovery from infection, the probability that a ‘contact’ leads to ‘transmission’, and so forth. (See also ‘State variable’).

Prevalence

The proportion of a population or group who are currently suffering from a disease, infection, or medical condition. For example, in South Africa about 10% of adults are infected with HIV.

State variable

Any variable in a model which captures the state of the system, such as the number of living individuals, the number of people in a particular risk category, and so forth. (See also ‘Parameter’).

Note that mathematically there may sometimes be no way to distinguish state variables from parameters in a model equation. The distinction is conceptual, and usually straightforward: parameters are usually model inputs, and state variables are usually model outputs.

5.17

Public health surveillance

Nguyen Tran Hien, James W. Buehler, and Ann Marie Kimball

Introduction to public health surveillance

Public health surveillance provides the epidemiological foundation for modern public health practice. The ongoing monitoring of disease or health trends within populations informs what public health actions are taken and reflects whether those actions are effective. Surveillance may involve monitoring of diseases and other health-related conditions as well as their antecedents, characteristics, and consequences. Surveillance can guide the local response to individual cases of disease or more broadly inform public health programmes and policies. A key function of surveillance is to identify circumstances that merit further public health scrutiny, such as groups or locations that are disproportionately affected or changes in disease occurrence or severity. Person, place, and time are the key epidemiological attributes of disease distribution in populations, and knowledge of these attributes afforded by surveillance allows targeting of public health programmes. General principles that underlie the practice of surveillance are essentially the same for all countries, regardless of economic development. Defining surveillance objectives depends on programme goals and on what information is needed, who needs it, and how it will be used. Desirable but potentially competing attributes of surveillance include completeness, timeliness, representativeness, high predictive value, acceptability, flexibility, simplicity, and low cost. Public health surveillance is conducted in many ways, depending on the nature of the health event under surveillance, the nature of healthcare and information infrastructures, the population involved, resources available, and information needs. The widespread and expanding use of the internet, electronic media, communication technologies, and mobile computing have enabled

innovations in public health surveillance that reach far beyond traditional methods. Surveillance systems should be periodically assessed to determine whether information needs are being fulfilled and whether the resources are being used effectively. Although surveillance methods were originally developed as part of efforts to control infectious diseases, basic concepts of surveillance have been applied to all areas of public health.

In many resource-poor countries, challenges to meeting needs for population health information are heightened and include potential tensions between the perspectives of external funders and in-country officials, between the interests of those responsible for disease-specific 'categorical' programmes and those interested in strengthening general surveillance infrastructures, and between advocates for using limited resources to strengthen information capacities versus those who would focus spending to address urgent service needs. At the same time, recent innovations hold promise for improving surveillance in resource-poor settings, where precedents in developed countries that can impede innovation are absent or less established.

Definition of surveillance

People who are responsible for protecting and promoting the health of populations need ongoing, timely, and reliable information about the health of the populations they serve. The term 'surveillance' encompasses the multiple processes that public health agencies employ to meet this need for information. Surveillance can include information about specific diseases, health risks, or more general indicators of overall population health. As an integral part of public health practice, surveillance is an ongoing, systematic process of information collection, analysis, interpretation, visualization, dissemination, and connection to public health programmes (Thacker and Berkelman 1988). Surveillance systems are organized networks of people and activities dedicated to managing and maintaining surveillance for specific conditions. Surveillance systems can operate at various levels within countries, from local to national, or on a global scale.

In 1963, Langmuir established the modern concept of surveillance as 'the continued watchfulness over the distribution and trends of incidence through the systematic collection, consolidation, and evaluation of morbidity and mortality reports and other relevant data' together with timely and regular

dissemination to those who ‘need to know’ (Langmuir 1963). In 1968, the twenty-first World Health Assembly described surveillance as the systematic collection and use of epidemiological information for the planning, implementation, and assessment of disease control; in short, surveillance implied ‘information for action’ (WHO 1968). Over time, the scope and methods of surveillance have been adapted to meet the information needs of public health programmes that address a broad spectrum of infectious and non-infectious diseases and acute and chronic conditions. For brevity, the term ‘diseases’ or ‘conditions’ will be used as shorthand throughout this chapter for diseases, injuries, risk factors, and other health-related events that are the focus of public health programmes and surveillance systems.

Surveillance can include monitoring of the incidence, prevalence, antecedents (such as behavioural risks or hazardous exposures), characteristics, and outcomes of disease and other health conditions of public health importance. According to the US Centers for Disease Control and Prevention (CDC), a ‘public health surveillance system is useful if it contributes to the prevention and control of adverse health-related events and can be useful in contributing to [public health programme] performance measures, including health indicators that are used in needs assessments and accountability systems’ (CDC 2001). Surveillance should begin when there exists, or is likely to exist, a public health problem that merits attention, and it should be discontinued or revamped when the information it yields is no longer needed or useful. Surveillance systems can operate at various levels within countries, from local to national, or on a global scale.

The International Health Regulations (IHR) of 2005 provide a legal and practical framework among all World Health Organization (WHO) member countries to protect against the international spread of disease. The IHR require that all countries can detect, assess, report, and respond to public health threats that have the potential to spread across borders. So the IHR call for strengthening of national capacity for surveillance and control of public health events of national and international concerns (WHO 2016).

Establishing a surveillance system

Establishing a surveillance system requires a statement of objectives, definition of the disease or condition under surveillance, and implementation of procedures for collecting information (or obtaining information from

existing sources), managing, analysing, visualizing, interpreting, and disseminating the information. In addition, as the number of surveillance systems increases, attention should be given to whether there are multiple systems collecting information within a population and, if so, whether these systems collect common data elements and adhere to shared information standards. This is important to allow for information sharing across systems, to minimize duplication of effort, and to simplify the work of those who are asked to report to these systems. Attention should also be given to whether surveillance for a particular condition can be done similarly in different locations, increasing the likelihood that comparison of disease incidence or prevalence across multiple localities or countries is meaningful.

Surveillance systems are information loops, or cycles, that involve healthcare providers, public health agencies, and the public in a process of information collection, analysis and interpretation, and feedback. The cycle begins when events of public health concern occur and is completed when information about these events is made available and applied in public health programmes. This may involve multiple levels that should be specified in developing a surveillance system, ranging from the anticipated local response to individuals with cases of disease to the use of information aggregated from multiple localities to inform national policies. The delivery of information to constituents and policymakers (Langmuir 1963) is essential to the completion of the surveillance cycle. To the extent that surveillance data can be released without compromising privacy or confidentiality, they should be placed in the public domain to allow full use and benefits, with appropriate documentation.

Weaknesses in any part of the information chain can affect the entire surveillance process, and potential constraints should be anticipated. For example, when surveillance draws upon information arising from the use healthcare services, surveillance will be affected by the availability and quality of those services, including laboratory and other diagnostic services (Berkelman et al. 1994) and by the accuracy with which diagnoses are recorded or coded. No information resource is perfect, and establishing surveillance requires consideration of options, each with strengths and limitations, for how surveillance should be conducted at each step in the surveillance cycle. Having data that are complete and without error is desirable, but the effort required to obtain such data in a timely manner may not be feasible or sustainable.

Objectives of surveillance

Goal of public health surveillance is to provide information that can be used for health action by public health personnel, government leaders, and the public to guide public health policy and programmes (Smith et al. 2013). Surveillance is essential for monitoring events that might jeopardize the health of a population so that appropriate prevention and control measures can be implemented in a timely manner. An effective surveillance system includes the following: collection, reporting, and consolidation of data; routine analysis and interpretation of data; feed-forward of surveillance data to decision makers; feedback of surveillance data to those providing the data and other interested parties; detection, evaluation, and response to unusual patterns in the data and quality assurance (PAHO WHO 2014).

The objectives of surveillance systems will be shaped by the public health programmes they serve. For example, a programme to eradicate an infectious disease requires intensive surveillance in the later stages of the campaign that emphasizes identification of all people with the disease (Hinman and Hopkins 1998).

The purpose of surveillance is to meet the needs of public health programmes for information about the health of the populations they serve. The following are purposes of public health surveillance:

To define public health priorities.
To characterize disease patterns by time, place, person, and disease characteristics.
To detect epidemics, health problems.
To monitor changes, measure trends of diseases, and health-related events.
To suggest hypotheses for further investigation.
To estimate magnitude and scope of health problems.
To guide and evaluate prevention and control programmes, including assessment of effectiveness and/or adverse consequences.

Taken together, population-level measures of time, place, person, and disease characteristics can lead to an understanding of the impact of various diseases and insights into individual and social determinants of their occurrence. Surveillance might lead to hypotheses that can be tested in more definitive investigations and can identify individuals who could be enrolled

in further epidemiological studies. In some instances, surveillance alone can provide compelling evidence of the impact of prevention programmes, such as the sharp decline in the incidence of a vaccine-preventable disease coincident with the widespread use of a new vaccine (Seward et al. 2002). In other situations, surveillance can contribute to evaluating the impact of interventions, but more detailed investigations are necessary to explain trends that are shaped by multiple factors. Low incidence or prevalence does not necessarily mean that certain diseases do not deserve public health attention, including continuation of surveillance, since this might reflect the impact of successful prevention efforts that must be sustained in order to prevent re-emergence. Information on disease trends gained through surveillance, when combined with other information, can also be used to predict future disease trends. For example, with the ageing of the 'baby boom' population in countries such as the United States, projections of disease prevalence in the elderly, such as the anticipated prevalence and cost of dementia (Hurd et al. 2013), are garnering increased attention.

Surveillance for a disease or other health problem should have clear objectives. These objectives should include a clear description of how data that are collected, consolidated, and analysed for surveillance will be used to prevent or control the disease. When establishing the objectives of surveillance, several issues should be considered, including: health-related event and case definition; population under surveillance; data sources and data collection approach; planned uses of the surveillance data; legal authority and frequency of data collection; organization of the surveillance; integration with other surveillance and health information systems; data processing, analysis and management; information dissemination methods and ethical issues (CDC 2012c).

An important use of surveillance is to help set public health priorities. In addition to recent measures of incidence or prevalence, other factors that can shape priority setting include likely future trends, the potential for public health or other interventions to have an impact, severity, disparities and attendant social justice concerns, and public interest. To the extent that surveillance data are used in priority setting, deliberations may give rise to concerns about surveillance methods, how surveillance data are queried and analysed, and the accuracy of estimates of disease impact. For example, even though mortality monitoring is the oldest form of surveillance and remains important in assessing the impact of different diseases (Thacker 2010),

different agencies using mortality data may arrive at different conclusions, depending on the use of overall mortality measures, age-specific measures, or approaches that seek to integrate measures of mortality and wellness. Table 5.17.1 provides two estimates of the number of deaths (in millions) for leading causes of death globally in 2010, one from the Institute for Health Metrics and Evaluation (IHME) and another from the WHO (Liu et al. 2000). Such estimates from different organizations rarely line up, and disparities in estimates can cause consternation among donors and national governments seeking to prioritize interventions.

Table 5.17.1 Estimates of the global number of deaths, by cause, 2010, Institute of Health Metrics and Evaluation (IHME) and World Health Organization (WHO)

	IHME	WHO
Cause of death	Deaths (millions) (95% CI)	Deaths (millions) (95% CI)
Pneumonia/lower respiratory infections	0.847 (0.736–0.981)	1.396 (1.189–1.642)
Diarrhoeal diseases	0.666 (0.544–0.763)	0.801 (0.555–1.182)
Malaria	0.676 (0.445–1.002)	0.564 (0.432–0.709)
HIV/AIDS	0.126 (0.105–0.150)	0.159 (0.131–0.185)
Neonatal sepsis or meningitis	0.514 (0.318–0.841)	0.573 (0.388–0.789)
Total deaths in children <5 years of age	6.841 (6.439–7.302)	7.622 (no CI is given)

CI, confidence interval.

Source data from Institute of Health Metrics and Evaluation, University of Washington, Seattle, WA, USA (unpublished report) and WHO via Child Health Epidemiology Reference Group, Department of International Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA; *The Lancet*, Volume 379, Issue 9832, Liu L. et al., Global, regional, and national causes of child mortality: an updated systematic analysis for 2010 with time trends since 2000, pp. 2151–61, Copyright © 2000.

Efforts to estimate the impact of different diseases can be done periodically as data and the resources to analyse them become available, but this approach does not lend itself to frequent and ongoing monitoring where disease patterns are shifting because: (1) data quality is highly variable from different surveys or settings and often repurposed from its original collection

goal; (2) cross-sectional survey information typically becomes available 1–2 years after surveys are completed; (3) estimates for very remote and information-poor geographic locations are often based on imputed data from neighbouring localities, and little ‘ground truthing’ is done to verify imputed data; and (4) assignment of causes of death, especially when based on verbal autopsies, is plagued with potential misclassification.

Health problems under surveillance

It is necessary to decide exactly what disease or health problem will be under surveillance. Surveillance may frequently be conducted for any of several points along a spectrum, ranging from exposure to an adverse outcome. It is important to consider which manifestations or stages of a disease should be under surveillance. For example, manifestations of ischaemic heart disease include angina pectoris, acute myocardial infarction, and (sudden) death, and precursors may include hypertension or behavioural risk factors such as tobacco use, unhealthy dietary habits, or lack of exercise. Criteria developed for selecting and prioritizing health problems for surveillance are described as following (CDC 2012b):

Public health importance of the problem: incidence, prevalence, severity, sequel, disabilities, mortality caused by the problem, socioeconomic impact, communicability, potential for an outbreak, public perception and concern, and international requirements.
Ability to prevent, control, or treat the health problem: preventability, control measures, and treatment.
Capacity of health system to implement control measures for the health problem: speed of response, economics, availability of resources, and what surveillance of this event requires.

Target population

The population under surveillance may be defined on the basis of where people live, work, attend school, consume water or food, or use healthcare services. Alternatively, the population may be defined on the basis of where health events occur. For example, surveillance of traffic injuries aimed at identifying roadway hazards could include all injuries that occur in a target community, regardless of whether affected people live in that community. In

some cases, target populations for public health surveillance may be animals. For example, animal surveillance systems for West Nile virus and avian influenza have been developed, recognizing that these pathogens of concern to human health also infect animals and may alert public health authorities of potential transmission to humans.

Questions addressed by surveillance can be summarized using epidemiological measures of disease characteristics (what), person (who), place (where), and time (when):

Disease characteristics: what are the characteristics of disease, such as the clinical or laboratory manifestations, severity, and outcomes? For example, understanding the prevalence of the sensitivity or resistance of infectious agents to antimicrobial drugs is critical to informing healthcare and public health responses, and monitoring the prevalence of different microbe strains can be important for understanding the spread of different infectious or for vaccine development. Global influenza surveillance is used to anticipate which strains of influenza virus will be most prevalent during upcoming influenza seasons and to develop annual influenza vaccines (Cox et al. 1994). Differences in the spectrum or severity of disease among affected people might pinpoint differences among groups in genetic susceptibility, the co-occurrence of other illnesses, or access to healthcare.

Person: who is affected? Understanding the characteristics and commonalities of affected people, such as their age, gender, occupation, race, ethnicity, habits, customs, or social or economic status, provides further insights into potential exposures or behaviours that can affect the risk of disease. Combined with information about the characteristics of the population at large, as from census data or population surveys, surveillance can provide critical insights into variations in the rates of disease among different groups. For example, in 1981, shortly after the disease later named acquired immunodeficiency syndrome (AIDS) was recognized and before the aetiological agent (human immunodeficiency virus (HIV)) was identified, by describing the sexual, blood-related, or perinatal exposures of affected people, surveillance combined with early epidemiological investigations helped to determine the infectious nature of the disease, identify probable modes of transmission, inform early prevention recommendations, and calm fears about the potential for widespread transmission through casual contact with affected people (Jaffe et al. 1983).

Place: where is disease occurring or not occurring? Understanding geographic variations in incidence or prevalence can inform where investigations and interventions should be directed and can provide important clues to environmental, occupational, water- or food-borne, behavioural, or other risk factors. Alternatively, geographic variations might reflect differences in detection of disease resulting from variations in the availability, quality, or use of public health or healthcare services.

Time: are disease trends upwards, downwards, or stable? Do downward trends coincide with interventions, such as the introduction of a vaccine,

treatment, or prevention programme? Do upturns coincide with aggravating situations? Changes in trends might reflect the underlying health of populations or they might reflect changes in how surveillance is conducted or changes in medical or public concern about particular conditions. A key purpose of surveillance is to recognize changes in trends that might herald the onset of an outbreak or epidemic, prompting investigations to confirm the observed upturn in disease and, if so, identify the likely cause and guide interventions.

Case definition

The case definition is fundamental to any surveillance system because it is the formal declaration of what is to be monitored. The case definition is primarily a criterion for determining who or what is to be counted but can also serve other objectives, such as providing a guide to clinicians for diagnoses. Where feasible, consistent use of case definitions across geographical areas permits comparison of disease rates. The case definition should be sufficiently inclusive (sensitive) to identify people who require public health attention or to monitor trends reliably but sufficiently exclusive (specific) for surveillance to have adequate predictive value.

Case definitions must be crafted so that they can be applied in a given surveillance context, reflecting the level of expertise of those who report and collect surveillance data and the extent of information that is likely to be available. Surveillance case definitions may also be layered to include those with possible, suspected, presumptive, or confirmed diagnoses, depending on the urgency of need to respond to certain diseases before diagnoses are definitively established. '*Possible*' is classification of cases is based on clinical criteria. '*Probable*' is classification of cases is based on clinical criteria, epidemiological link, and presumptive laboratory criteria. '*Confirmed*' is classification of cases is based on laboratory confirmation. For infectious diseases, confirmation of the diagnosis typically requires a positive laboratory test. But, if laboratory testing is not routinely available, syndromic criteria might be sufficient to guide treatments and to monitor disease trends. For example, in the United States, surveillance criteria for sexually transmitted infections (STIs) such as gonorrhoea or syphilis, incorporate both clinical and laboratory criteria (CDC 2011), reflecting the routine availability of diagnostic testing. However, in many countries,

routine access to such tests is not available, and syndrome-based criteria have been developed to treat and monitor conditions such as urethritis or cervical infection, based on prior assessments that included laboratory testing and informed presumptions regarding likely diagnoses and effective treatments (WHO 2003).

Types of surveillance

There are two approaches to categorize types of surveillance: by the format and by the content of data collection.

Types of surveillance by forms of data collection

Passive and active surveillance: A passive surveillance 'relies on the physicians, laboratory or hospital staff or other relevant sources to take the initiative to report data to the health department'. A key data quality concern of passive surveillance systems is external completeness. *An active surveillance* system is a system that is 'based on the public health officials' initiative to contact the physicians, laboratory or hospital staff or other relevant sources to report data'. Active surveillance typically generates high-quality data (high levels of completeness, validity, and timeliness) through a direct link with data providers. Active surveillance relies on a prompt response from public health professionals, who will usually monitor replies and can respond to low response rates in order to improve external completeness. Completeness for key variables, and thus better representativeness, will be achieved if distribution and composition of data providers are carefully chosen. Data comparability will be improved if data are collected through a common surveillance protocol. The main disadvantages of active surveillance are the substantial resources needed and the subsequent need to focus on high-priority diseases and questions that justify the additional expense.

Compulsory versus voluntary: Surveillance systems rely on processes and individuals. Countries use different approaches to elicit surveillance data from their sources. Some systems make data submission mandatory or quasi mandatory (either by law, professional edict, policy, or guidance), whereas others rely on a voluntary approach. A legal or professional obligation to

report may lead to greater external completeness and may increase timeliness but is unlikely to have an impact on representativeness or validity.

Comprehensive versus sentinel: Comprehensive surveillance systems include reports of cases of infectious diseases that occur within the whole population of the geographical area covered by the surveillance system (national, regional). Sentinel systems rely on notifications from a selected group of physicians, hospitals, laboratories, and other institutions. In the case of sentinel systems, representativeness is often established as part of the system-design process. The choice between sentinel and comprehensive systems is based on surveillance objectives, characteristics of the disease, and the availability of resources (ECDC 2014).

Types of surveillance by content of data collection

Following WHO WPRO, there are two main types of surveillance—event-based and indicator-based (WHO WPRO 2008b). Although the functions differ, these systems are both essential components of a national surveillance system and should complement each other.

Indicator-based surveillance is routine reporting of disease cases including notifiable disease surveillance system, sentinel surveillance, and laboratory surveillance. Reports are from healthcare facilities weekly and monthly. Indicator-based surveillance involves the routine collection of standardized information to establish historical trends and baselines to which the current situation can be compared. Data collection and analysis is very structured and specific to the disease or syndrome being monitored. Public health responses are triggered when pre-defined threshold activity levels are surpassed.

Event-based surveillance is rapid detection, reporting, confirmation, and assessment of public health events, including clusters of disease and rumours of unexpected deaths. Reports are intermediate commonly. Event-based surveillance involves rapidly detecting events that potentially pose a high risk to public health. These systems are designed to detect rare and new events, such as high-impact outbreaks or emerging and unknown diseases, at a very early stage when the event is geographically localized and involves a relatively small number of people, so that rapid assessments and responses can be made to contain or mitigate the situation. In order to accomplish this, event-based systems use broader and more sensitive definitions and flexible

reporting structures than indicator-based systems. Event-based surveillance complements indicator-based surveillance. both systems should be seen as essential components of a single national surveillance system.

Integrated diseases surveillance (IDS): To improve responses to emerging and re-emerging events, African Region has developed an integrated diseases surveillance (IDS) in the advocated for countries to establish community-based surveillance systems since 2001. Community-based surveillance (CBS) is an active process of community participation in detecting, reporting, responding to, and monitoring health events in the community. Community representatives and health workers were to be used as contact persons for the purpose of surveillance. Main activities of IDS are including (WHO Regional Office for Africa 2014):

Use standard case definitions to identify priority diseases, conditions, and events.
Report suspected cases, conditions, and events to the next level of the health system.
Investigate and confirm suspected cases, outbreaks, and events.
Analyse and interpret data collected.
Prepare to respond to outbreaks and other public health events.
Coordinate and mobilize resources to respond appropriately.
Provide feedback within and across levels of the health system.
Monitor and evaluate the system and component activities.
Conduct operational research to inform better practices and processes.

Methods for data collection

A number of methods are used to collect the majority of health-related data for surveillance. These methods can be further characterized by the approach used to obtain information from the sources described previously. Depending on the situation, these methods might be used to obtain information about a sample of a population or events or about all members of the population or all occurrences of a specific event (e.g. birth or death). Careful consideration of the objectives of surveillance and a thorough understanding of the advantages and disadvantages of different sources and methods for gathering data are critical in deciding what data are needed for surveillance and the most appropriate sources and methods for obtaining it.

Sources of data

Typical sources of data for surveillance are as following (CDC 2012c):

Individual persons
Healthcare providers, facilities, and records: physician offices, hospitals, outpatient departments, emergency departments, inpatient settings
Laboratories
Environmental conditions: air, water, animal vectors
Administrative actions
Financial transactions: sales of goods and services, taxation
Legal actions
Laws and regulations

Vital statistics

In countries with well-established vital event registration systems, monitoring trends for various causes of death is a long-standing part of surveillance for life-threatening conditions. Despite the well-recognized limitations in how causes of death are classified on death certificates, monitoring cause-specific mortality trends overall and by age, gender, and geographic location provides a useful, broad-brush picture of the impact and trends for multiple conditions.

There is frequently a lengthy interval between death and collection and analysis of death certificates, which may make such vital statistics less useful for surveillance purposes when more current data are needed. However, summary vital data can be rapidly collected. For example, weekly reporting of deaths from 122 cities to the CDC has been integral to the surveillance of influenza epidemics in the United States (Fig. 5.17.1), providing an indication of the timing and severity of annual seasonal influenza relative to historical norms (Choi and Thacker 1981; CDC 2009).

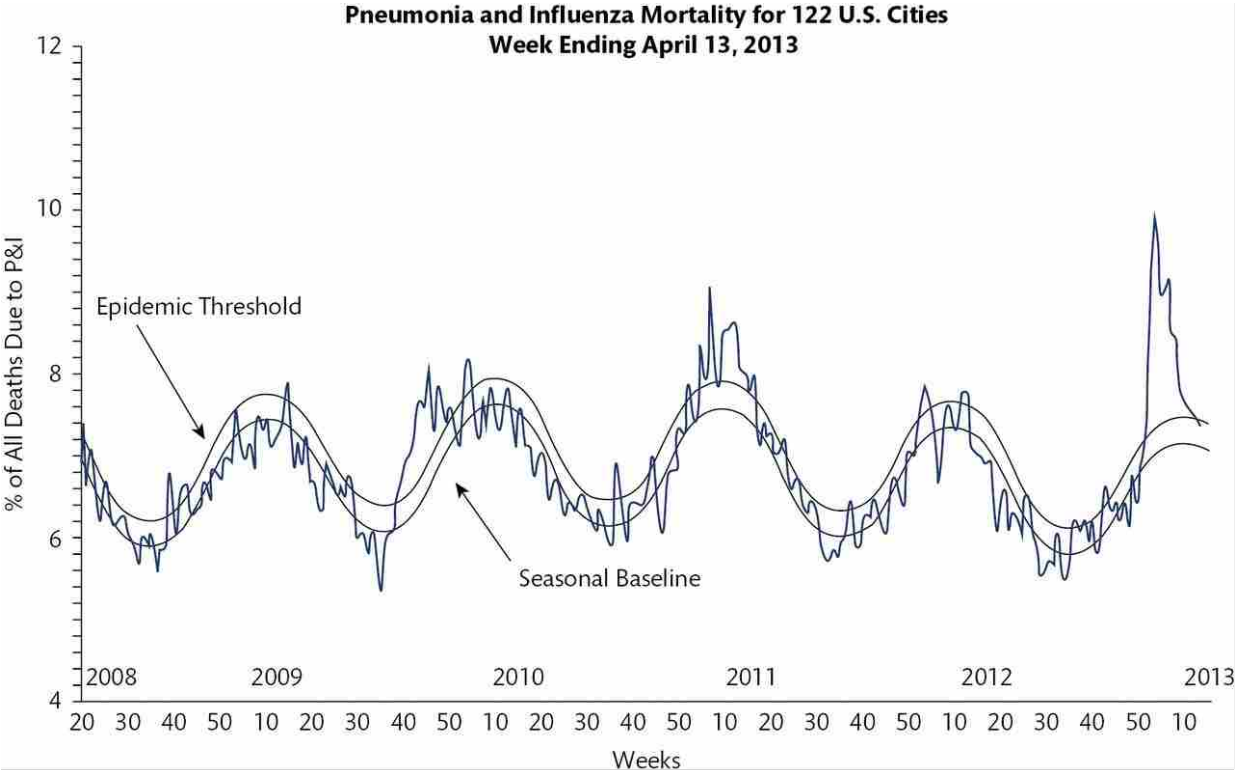


Fig. 5.17.1 Percentage of deaths attributed to pneumonia and influenza (P&I) reported by the 122 Cities Mortality Reporting System, by week and year—United States, 2008–2013.

Reproduced from *FluView, 2013–2014 Influenza Season Week 3 ending January 18, 2014*, Pneumonia and Influenza (P&I) Mortality Surveillance, available from http://www.cdc.gov/flu/weekly/#S2.

Medical examiner and coroner reports

For more detailed descriptions of the circumstances surrounding deaths (including autopsy reports, toxicology studies, and police reports), medical examiner and coroner records may be useful. These reports are typically representative of unexpected or medically unattended deaths or deaths caused by intentional and unintentional injuries and have been used for surveillance of such conditions as heatwave-related mortality and alcohol-related injuries. Systematic necropsy examinations have also been useful in ascertaining the contribution of tuberculosis to mortality of HIV-infected individuals in West Africa. Antibody determinations performed on post-mortem medical examiner samples may be useful for estimating population prevalence of infectious diseases (e.g. dengue) (Rigau-Perez et al. 2006).

Syndromic surveillance

A growing number of public health agencies have been prompted by concerns about the threat of bioterrorism to invest in automated ‘syndromic surveillance’ to detect epidemics more quickly than possible using traditional approaches (CDC 2004; Mandl et al. 2004). This involves monitoring trends in healthcare use for relatively non-specific syndromes (e.g. febrile respiratory illness or gastrointestinal symptoms) or other manifestations of illness (e.g. medication purchases, work, or school absenteeism) that may be apparent and can be monitored before diagnoses are established. Because these systems emphasize timeliness of the entire surveillance process, often enabling tracking of events within a day or less of their occurrence, they generally rely on data from electronic health information systems. These systems use automated methods to harvest electronically stored data, apply statistical tools to detect aberrant trends, and develop visual displays of geospatial disease trends. Despite this emphasis on automation, these systems still require human judgement to interpret reports and determine what level of further assessment or response is warranted when statistical alerts are generated.

Notification

Notifiable disease reporting is the approach traditionally used by public health programmes, based on laws or regulations that require reporting of selected diseases or conditions to public health departments. People or institutions responsible for reporting often include doctors, other healthcare providers, coroners, and medical examiners, laboratories, clinics, and hospitals. With the emergence of large, often commercial, laboratories that serve multiple providers across relatively large geographic areas, public health agencies are increasingly emphasizing laboratory-based reporting. This approach increases the timeliness and completeness of case reporting for diseases that require a laboratory test to diagnose. Because clinical information that accompanies specimens submitted to laboratories is often limited, however, follow-up with the providers who order laboratory tests or with patients themselves is often needed to complete case reporting (Ackers et al. 2000).

The overall completeness of notifiable disease surveillance depends on a succession of events: affected people must be sufficiently ill to seek healthcare, diagnostic tests are performed (which might not occur if patients are treated empirically), and positive results are reported. This approach may

be relatively passive, meaning that public health staff do not reach out to healthcare providers to encourage reporting, or more active, meaning that public health staff may contact or visit providers to prompt reporting or assist in the reporting process. In general, more active approaches to surveillance will result in more complete and timely reporting. For example, in the late 1970s, a relatively large shigellosis outbreak in Washington, District of Columbia, United States, was not recognized by the health department's surveillance system because fewer than a third of outbreak-associated cases had been reported as required. A review of the surveillance system found that surveillance was not only incomplete but also not representative because reporting varied considerably among different facilities. In hindsight, adherence to reporting requirements lagged because the health department had taken a very passive approach and had done little to inform and remind local healthcare providers of their reporting requirements (Kimball et al. 1980).

Completeness of reporting may be affected for multiple other reasons: if patients and doctors seek to conceal the occurrence of diseases that carry a social stigma, if healthcare providers are unaware of regulations, or if interest in particular diseases is limited. Even if reporting is incomplete, temporal trends and geographic patterns can be determined as long as the proportion of cases detected remains consistent over time and across geographical areas. Conversely, lack of representativeness and heterogeneity of reporting further limit the usefulness of incomplete data (Lowndes and Fenton 2004).

Example: notifiable disease surveillance in China

The China Infectious Disease Automated-alert and Response System (CIDARS) is an extension of the nationwide electronic system for notifiable disease reporting that was initiated in 2004 following the severe acute respiratory system (SARS) outbreak. The system focuses on 39 infectious diseases or syndromes that clinicians are required to report to the Ministry of Health. CIDARS added a formal process for analysis and follow-up that specified statistical thresholds for prompting alerts, based on the number of cases relative to historical precedents or temporal–spatial clustering; procedures for notifying responsible local officials via mobile telephone when those thresholds were exceeded within a locality; responsibility for preliminary assessments and, if indicated, more definitive field

investigations; and responsibility for documenting and reporting the results of those assessments and investigations. In an initial evaluation of CIDARS conducted in a sample of ten provinces, for nine relatively rare but potentially severe infections, 308 alerts were identified during a 2-year period in 2008–2010, of which 69 (22%) merited a field investigation, leading to detection of nine cholera outbreaks. For other more common but less severe diseases, there were more than 100,000 statistical alerts, of which 1.4% were deemed to merit further attention, leading to detection of 167 disease outbreaks (Yan et al. 2012).

Healthcare providers networks

Networks of healthcare providers have been organized for several decades, primarily to gather information on selected health events. Most have been organized by practising doctors on a voluntary basis; in many European countries, these networks have formed firm relationships with both public health authorities and academic centres. The strengths of sentinel provider systems include the commitment of the participants, the possibility of collecting longitudinal data, the flexibility to address a changing set of conditions, and the potential to conduct focused studies of the incidence, prevalence, care, or outcomes of various conditions, typically in primary care settings (Fleming et al. 2003). The main limitation of this type of system is that the population served by participating doctors may not be representative of the population.

Disease registries

Registries are comprehensive longitudinal listings of people with particular conditions. They often include detailed information about diagnostic classification, treatment, and outcome. Registries were initially established primarily for epidemiological research on individual diseases or conditions to develop aetiological hypotheses and to identify cases for further research (Weddell 1973). Registries have also been used to ensure the provision of appropriate care and to evaluate changing patterns of medical care; unlike other disease information systems, they cut across the different levels of severity of illness and may provide information over time about individual people. Registries can also be valuable for monitoring disease incidence and distribution. To focus on selected diseases or conditions, registries often

develop a constituency that promotes participation and reporting. Most registries rely on numerous sources of data for case detection including, but not limited to, hospitals, laboratories, and death records; few registries rely primarily on doctor notification.

Example: birth defects. Surveillance for birth defects was first initiated in response to the thalidomide tragedy in the late 1950s and early 1960s; registries were established to provide reliable baseline rates for specific birth defects and to detect increases in the prevalence of birth defects as a means to identifying human teratogens. The CDC has conducted birth defects surveillance in metropolitan Atlanta, Georgia, United States, since 1967 by using multiple sources of ascertainment of all serious birth defects observed in stillborn and liveborn infants and recognized by signs and symptoms apparent in the first year of life (Correa-Villasenor et al. 2003).

The International Clearinghouse for Birth Defects Surveillance and Research consists of 40 registries globally (Botto et al. 2006). The International Clearinghouse conducts a spectrum of surveillance activities that includes monitoring of selected conditions, and the exchange of information about birth defects clusters under investigation, findings of unpublished studies, and rumours about the occurrence of birth defects. Clearinghouse participants work at the intersection of surveillance and research in their assessment of the potential teratogenicity of first-trimester use of medications.

Laboratory surveillance

Laboratory surveillance may be conducted as part of a notifiable diseases system, sentinel surveillance system, or other system. In addition to supporting routine case reporting, laboratory-based surveillance can allow for the use of molecular epidemiological tools, more detailed evaluation of microbe strains or types, and monitoring of drug sensitivity. The use of molecular tools to enhance surveillance of pathogens is growing in many countries to facilitate outbreak detection, especially in situations where disease is associated with widely distributed food products, where the number of cases of illness due to a particular infection is insufficient to prompt attention, but where epidemiological linkage of geographically disparate cases allows detection of distinctive outbreak-associated pathogen strains. In the United States, PulseNet is a network of public health

laboratories that performs DNA ‘fingerprinting’ on bacteria that may be food-borne (Swaminathan et al. 2001). The network permits rapid comparison of these ‘fingerprint’ patterns through an electronic database and has detected multiple outbreaks that likely would have been otherwise unrecognized. For example, in 2008–2009, PulseNet detected and enabled tracking of an outbreak of salmonellosis associated with contaminated peanut butter with over 500 cases of illness detected in 43 states (CDC 2009). Molecular methods are also increasingly applied in surveillance of the HIV epidemic. For example, surveillance systems in Canada (Jayaraman et al. 2006) routinely characterize newly diagnosed HIV infections with respect to the HIV subtype.

Sentinel surveillance

Sentinel surveillance is the most efficient way to collect high-quality data in a timely way. It is typically carried out with indicator-based surveillance and involves systematically collecting data on a routine basis from a limited number of surveillance sites. Data collected in a well-designed sentinel system can be used to signal trends, identify outbreaks, and monitor the burden of disease in a community, providing a rapid, economical alternative to other surveillance methods. Because sentinel surveillance is conducted only in selected locations, however, it may not be as effective for detecting rare diseases or diseases that occur outside the catchment areas of the sentinel sites. Ideally, the sites are chosen to be representative of the population so that the information gathered can be applied to the population as a whole. For example, the spectrum of influenza infections varies widely in terms of severity and various indicator-based systems are used to monitor these syndromes. Specifically, sentinel influenza-like illness (ILI) surveillance is focused on mild manifestations of influenza in the outpatient setting while sentinel severe acute respiratory infection (SARI) surveillance is used to monitor persons with more severe illness who have been admitted to a hospital for their respiratory illness. Selecting sentinel SARI surveillance sites should consider following issues: representativeness of the population of interest (climate zones, demographic, and socioeconomic characteristics); feasibility and sustainability (staff and leadership motivation; mechanisms for collection, storage, and transport of clinical specimens; high-quality diagnostic laboratory, ability to reliably manage and report surveillance data; long-term funding); disease burden (population-

based incidence, healthcare seeking rates); and availability of reliable numerator and denominator data (PAHO 2014). Influenza surveillance is often multifaceted and includes measures of morbidity (e.g. office or hospital emergency department visits for 'influenza-like illness' or hospitalizations among people with laboratory-confirmed influenza), mortality (e.g. deaths attributed to 'pneumonia and influenza'), and laboratory-based indicators (e.g. viral isolates from participating laboratories where virus type, alignment with vaccine strains, or sensitivity to antiviral medications can be tested) (CDC 2012a).

Population health surveys

A survey is an investigation that uses a 'structured and systematic gathering of information' from a sample of 'a population of interest to describe the population in quantitative terms'. The majority of surveys gather information from a representative sample of a population so that the results of the survey can be generalized to the entire population. Behavioural data can be collected from samples of the population at large, from populations at particular risk for the health condition of interest, and from populations affected by the health condition of interest (Davy 2006). When conducted in person, objective measures of health can also be obtained, such as measurements of height, weight, blood pressure, or collection of specimens for laboratory testing. Population-wide household surveys, such as the National Health Interview Survey in the United States (CDC 2013b) or the General Household Survey in England and Wales (Twigg 1999), have provided information on personal health practices such as alcohol use and smoking, disabilities, and doctor encounters.

In countries where most residents have telephones, telephone interviews have the advantage of lower cost per interview, allowing for wider geographic coverage and more local estimation of the prevalence of various health indicators. For example, in the United States, all states conduct the Behavioural Risk Factor Surveillance System (BRFSS) to monitor a spectrum of health-related behaviours at the state, and in many instances local level. The BRFSS, like all telephone-based surveys, has had to adapt to the declining use of landline telephones and the increasing use of mobile telephones, trends that have varied among demographic groups within the population (CDC 2013c).

Inclusion of behavioural measures has been advocated by the WHO for HIV and STI surveillance programmes. Increasingly, demographic and health surveys are conducted to include both behavioural and biological measures, and to use sampling methods that provide representative data on a population level (Kenya Central Bureau of Statistics et al. 2004).

Surveys of populations at special risk may also play an important role in guiding public health response. Such systems may be relatively simple, in the form of episodic surveys of convenience samples (e.g. episodic convenience samples of men who have sex with men attending gay bars in Scotland) (Williamson et al. 2006) or may be formalized as ongoing surveillance systems. Such surveys can provide data on levels of risk behaviours, on adoption of desired preventive behaviours, and on exposure to public health prevention messages.

Healthcare information systems

Surveillance systems often depend on existing health data collection systems. These systems may be either integral to surveillance or serve as an adjunct to surveillance for specific diseases or conditions. Examples include use of hospital discharge databases and insurance claims data. The advantage of this approach is that it involves use of information systems that have been previously established for other purposes and can be tapped for surveillance applications for relatively little additional expense. In addition, these information systems often represent very large numbers of people. The main limitation is that diagnoses may be inaccurately or variably recorded or coded. A growing number of healthcare providers are using electronic health records to foster automated exchange of information from patients' medical records. From the public health perspective, these 'health information exchanges' provide a focal point for obtaining surveillance information within communities and for providing feedback to clinicians (Public Health Informatics Institute 2005).

Second generation surveillance

Second generation surveillance systems do not propose any new methods of data collection. Rather, they focus existing methods on appropriate populations and subpopulations, and combine them in ways that have the greatest explanatory power. For HIV surveillance, it means making best use

of other sources of information, communicable disease surveillance, reproductive health surveys, etc. to increase understanding of the HIV epidemic and the behaviours that spread it. These are most useful in reducing the spread of HIV and in providing care for those affected. Second generation surveillance systems include biological surveillance (sentinel serosurveillance in defined subpopulations, regular HIV screening of subpopulations, HIV screening of specimens); behavioural surveillance (repeat cross-sectional surveys in the general population and in defined subpopulations); and other sources of information (HIV and AIDS case surveillance, death registration, STI, and tuberculosis (TB) surveillance (UNAIDS/WHO 2000).

Internet-based approaches to surveillance

The widespread use of the internet and electronic media has led to innovations in population health monitoring that reach far beyond traditional surveillance methods. In 1994, the International Society for Infectious Diseases launched the Program for Monitoring Emerging Diseases (ProMed), which allowed individuals to post reports and updates on disease events on a public-access internet site, with a moderator assuring the professionalism of the website. Information posted on ProMed includes alerts, reports of epidemics, and local observer reports, with over 30,000 subscribers worldwide (Madoff and Woodall 2005). For example, the earliest public notice of the outbreak later known as SARS was in the form of a query posted on ProMed in early 2003 by a physician who learned from an acquaintance of widespread severe respiratory illness in Guangzhou, China. This report and other updates to ProMed alerted the global public health community to this epidemic and played a key role in charting its course of spread to multiple countries (Yu and Madoff 2004).

The Global Public Health Intelligence Network (GPHIN) is managed by the Public Health Agency of Canada and is geared to providing early warning for potential threats to public health worldwide. GPHIN uses an ongoing, automated process to monitor online media reports in six languages worldwide for early indicators of events such as disease outbreaks, natural disasters, food or water contamination, potential biological or chemical terrorism, and other threats. GPHIN operates 24 hours a day, 7 days a week in ‘near real time’. Access to GPHIN is limited to agencies determined to have a legitimate public health mandate, as determined by the WHO and

Health Canada (Public Health Agency of Canada 2013). The Google.org Flu Trends monitors trends in internet searches using terms that have been observed to correlate with influenza activity. While this approach yields trends that parallel other indicators of ILI, its public health application is yet to be determined (Valdivia et al. 2010).

The growing use of the internet, including use of mobile phones that provide mobile internet access, is enabling a form of public 'participatory epidemiology' or 'crowd sourcing' where interested members of the public or members of online communities interested in specific health problems are encouraged to post reports of illness or disease outbreaks. The application of such utilities and services is an area of ongoing development and evaluation (Freifeld et al. 2010).

Use of big data for surveillance

Today, use of big data has been shown to improve precision in selected disciplines of public health. Big data is an extremely large volume of structured, semi-structured and unstructured data sets that may be analysed computationally to reveal patterns, trends, and associations, especially relating to human behaviour and interactions. Such voluminous data can come from myriad different sources, such as business transaction systems, customer databases, medical records, internet clickstream logs, mobile applications, social networks, the collected results of scientific experiments, machine-generated data, and real-time data sensors used in internet of things environments. Disease surveillance using big data has shown efficacy in air pollution, antibiotic resistance, cholera, dengue, drowning, drug safety, electromagnetic field exposure, influenza A H1N1, Lyme disease, monitoring food intake, whooping. Effective signal detection often leads to attempts to predict future signals. Predicting public health risk leads to a chance to implement preventive interventions (Dolley 2018).

Data collection instruments

The desire to collect detailed information must be tempered by the need to limit data collection to information that can be reliably and consistently collected and used over the long term. Forms, questionnaires, or other data collection instruments that are too extensive will not be welcomed by those on whom the surveillance system must depend. While computerized systems

may permit more detailed data collection, collecting large volumes of data may exceed the capacity of public health agencies to make effective use of those data. Standards for exchanging electronic health information are critical as healthcare information systems and public health surveillance systems become increasingly automated. In the United States, the Health Information Technology for Economic and Clinical Health (HITECH) Act in 2009 provides incentives for healthcare providers to adopt electronic health records and requires that doctors and hospitals use their electronic health record systems to report quality-of-care measures to government-sponsored healthcare insurance programmes for the elderly and poor and immunization service and certain disease data to state and local public health programmes, using national information standards (HealthIT.gov 2013).

Analysis and management of surveillance data

Data analysis

Uses of surveillance data begin with descriptive analyses according to the basic epidemiological parameters of time, place, and person, including the use of standard tests for statistical significance. Typically, public health surveillance data are completed, summarized, and reported over specified time intervals (days, weeks, months, or years) and for specific geopolitical jurisdictions (cities, counties, states, provinces, nations, etc.). Incidence or prevalence rates can be calculated using various population denominators, typically census data, and comparisons of rates over time or across geographic areas should be adjusted using standard methods to account for differences in population characteristics, such as differences in the age distribution of populations. The use of maps or graphics is common to illuminate differences in disease incidence or prevalence by time, place, or personal characteristics such as age or gender.

Analyses of trends are generally designed to determine whether current or recent levels of disease differ from levels that would be expected based on historical precedent, which may be defined over a period of years to days, depending on the availability of historical data and the epidemiology of different diseases or conditions. Detection of epidemics or evaluation of suspected epidemics is a key function of surveillance. Nonetheless, many epidemics are first recognized by astute healthcare providers who observe an

increase in disease occurrence, instead of being detected by surveillance systems, because the cycle of surveillance has not had time to detect an upturn. Even in such instances, however, surveillance can be useful to help assess whether recently observed cases of disease exceed expected levels. Surveillance is most likely to provide the initial recognition of potential epidemics in situations where cases are occurring over a sufficiently large geographic area or time period so that the increase would not be apparent to individual practitioners.

The key to discerning whether observed trends are in excess of expected trends is excluding the possibility of random variations or 'noise' in observed disease levels. One approach to resolving this dilemma is to employ various techniques to 'smooth' the data, especially for graphic displays of temporal trends (Devine and Parrish 1998), including the use of moving averages or the aggregation of data points into longer but still relevant time intervals (by week instead of by day or by month instead of by week). A spectrum of advanced analytic techniques for evaluating trends in surveillance data has evolved to model various types of historical trends, such as seasonal variations or long-term secular trends, to estimate predicted or expected levels of disease based on those trends, and to determine whether observed levels are statistically different from expected levels (Brookmeyer and Stroup 2003). The use of geographic information systems extends the traditional use of maps to display surveillance data by allowing for analyses that allow geospatial assessment of disease and locations of risk exposures or preventive or therapeutic services (Waller and Gotway 2004). Recognition of clusters of disease is common but often difficult to distinguish from random variation, giving rise to various statistical techniques to detect and characterize clusters (Kulldorff 1997). The increasing automation of the surveillance process, especially for systems of syndromic surveillance that track daily trends in disease and seek to detect changes in disease trends quickly, has spawned the development of statistical methods and computer software algorithms for automatic trend aberration detection, including techniques that are derived from cluster detection and industrial quality control methods (Burkom 2007).

Regardless of whether analyses involve straightforward descriptive statistics or advance modelling techniques, any analyses of surveillance data must take into account potential limitations of the data, including variations in their completeness, representativeness, timeliness, or data quality over

time or among different locations. Data quality reflects features and characteristics that ensure data are accurate and complete and that they convey the intended meaning. It refers to individual records and to data system. Data of high-quality improve the intended use in operations, decision-making, and planning. These are some of the data quality issues which might be encountered, including: poorly completed forms; under-reporting, over-reporting, and duplicate reporting; unsystematic data collection and untruthful reporting; inconsistent reporting formats/forms; inconsistent reporting periods, and calculation errors.

Aggregation of data

Surveillance data may be in the form of individual patient records or aggregate counts and tabulations. In some instances, there may be a need at the local level to maintain records on individual people to direct follow-up services but for regional or national reports to be in the form of aggregate counts. Individual-level data permit more flexibility of analysis than aggregate data. This can be useful when routine analyses raise questions that require more detailed analyses.

Linkage of surveillance data to other information sources

Given the complexity of establishing new data and information systems, there is increasing interest in the combination of existing databases for surveillance purposes. This allows the combination of information about individual case-patients from multiple sources. For example, linkage of birth certificates (which include the birth weight and information on other infant or maternal characteristics) and death certificates (which include information on the age and cause of death) allows determination of neonatal or post-neonatal deaths rates for infants in different birth weight or gestational age groups (Buehler et al. 2000). Linkage of HIV and tuberculosis records enables surveillance for the co-occurrence of these two infections (Ahmed et al. 2007).

Techniques involved in data linkage are based on comparing records from different information systems, assessing the likelihood that records from two different sources represent the same individual, and combining the data for those individuals whose records match (Clark 2004). In using linked record systems, assessments should be done to determine what proportion of linked

records do not actually represent the same individual and, conversely, what proportion of records should have been but were not correctly linked.

Data management

Critical to the ongoing nature of surveillance is the management of surveillance data, including the maintenance of historical data and procedures for incorporating new information into surveillance databases. Surveillance data often need to be updated. Information that was initially missing may become available; follow-up investigations may yield supplemental information; people initially classified as meeting or not meeting a case definition may be reclassified; errors may be identified and corrected; and duplicate case reports may be recognized and culled. One approach to handling these and other changes is to maintain both provisional and final records. Measures such as physical and technical protection for keeping surveillance data in confidentiality should be implemented.

Data reporting

The flow of information

All countries have laws or regulations that require the reporting of certain diseases by healthcare providers to a government agency. Regulations specify not only who must report and the list of diseases to report, but also how to report, how promptly, etc. In most jurisdictions, reporting is required from primary care provider, clinic, hospital, and/or sometimes the laboratory to local health agency, usually the district or subdistrict. The local health agency is usually responsible for case investigation and action as necessary. The local health agency forwards the report or numbers up to the up next governmental level (province or state or equivalent intermediate level), who then compile the data and forward the reports or numbers to national level-Ministry of Health (MOH). The MOH reports a few conditions required by IHR to the WHO. In turn, both WHO and most national MOHs produce reports or bulletins that get back to the lower levels.

Surveillance may rely on data acquired explicitly for surveillance purposes or on secondary uses of data obtained from existing information sources. Regardless, there is a flow of information throughout the surveillance process, which should be planned. For surveillance systems that

depend on reporting individual cases of disease or tabulations of disease counts, people responsible for reporting may be all or selected healthcare providers in a defined area, including laboratories. Reporters may also be asked to collect specimens needed by public health agencies for laboratory confirmation or characterization of specific diseases or exposures.

Surveillance systems collect and provide data on a regular basis, ranging from daily to annually or, in some instances, every few years when surveillance depends on information from surveys that are conducted intermittently. This periodicity should be specified. For surveillance of a given disease or condition, acceptable intervals between the occurrence of events of interest and the eventual availability of reports from surveillance systems will depend on the nature of the targeted condition and how data are needed and used to inform public health actions.

For the reporting, the '*zero reporting*' method should be used. It means that a 0 (zero) on the reporting form should be recorded when no cases of an immediately reportable disease have been diagnosed during the week. Submitting a zero for each immediately reportable disease when no cases were detected during the week tells the staff at the next level that a complete report has been filed.

Data transmission

The mode of data transmission will depend on both the need for timeliness and availability of resources. While reliance on postage of paper forms or facsimile transmission continues, there is growing use of telecommunications and the internet to transmit surveillance information. In places where internet access remains limited, the use of hand-held mobile devices and mobile telephones for data collection and transmission is growing. In 2012, the World Bank estimated that three-quarters of the world's population have access to a mobile telephone, providing a relatively inexpensive and widely available technology to improve access to health information and public health surveillance (World Bank 1993, 2012).

Dissemination of surveillance information

Communication of surveillance data is an essential step in the surveillance chain. Appropriate feedback must be given to those providing the data, such

as analyses that allow them to compare their experience to that of the community at large and other reports that enable them to see the value of the information resource to which they contributed. When appropriate, people or institutions providing the data should be credited for their contributions and acknowledged. Public health professionals, policymakers, or others who may be responsible for taking action or setting the direction of public health programmes in response to surveillance data must receive the information that they need from the surveillance system on a timely basis and in an appropriate format for their use. In addition, these data should be available as widely as possible to all who may be interested in analysis and interpretation of these data. Websites are increasingly used to disseminate surveillance data as with the US CDC's *Morbidity and Mortality Weekly Report* (http://www.cdc.gov/MMWR), the WHO's *Weekly Epidemiological Record* (http://www.who.int/wer), and the European *Eurosurveillance* (http://www.eurosurveillance.org/).

The data must be provided on a regular basis, with the frequency of surveillance reports commensurate with the nature of the surveillance system, the characteristics of the disease under surveillance, and information needs. To improve the timeliness of surveillance data, provisional or preliminary data may be released, with appropriate caveats, in advance of completion of more definitive reports.

The format for dissemination varies with the target audience; a creative design will help to make the information stand apart from other documents and receive greater attention. Most policymakers and clinicians would prefer to see the data interpreted using graphics that provide a quick visual image that highlights key observations, accompanied by an abbreviated summary text. The important role that graphs and maps can play in visually decoding large quantities of data has been clearly demonstrated, with graphic displays giving the reader an understanding of large and complex datasets not conveyed easily in other ways (Tukey 1977; Tufte 1983). In contrast, many epidemiologists and other scientists will find access to more detailed data to be most useful.

Evaluation of surveillance systems

Surveillance systems should be periodically evaluated to ensure that important public health problems are under surveillance, that the information

provided is useful, and that resources dedicated to surveillance are being used wisely. Such evaluations may include a mix of quantitative and qualitative assessments (CDC 2001, 2004; Silk and Berkelman 2005; Buehler et al. 2009). Evaluation may also be considered in a broader strategy of continuous quality assurance (Krause 2006). Reflecting the steps that are taken in establishing a surveillance system, an evaluation of a surveillance system should include a review of its objectives, a detailed description of its operation, an assessment of its performance using the surveillance attributes as a guide, and recommendations.

Samuel describes the main the attributes used to evaluate a surveillance system as the following:

Acceptability: Willingness of persons and organizations to participate in the surveillance system.
Cost-effectiveness: Relationship between the expected outcomes and the costs of surveillance to achieve these outcomes. Surveillance system costs include direct costs (personnel and material resources), indirect costs (resulting from preparedness and response to surveillance findings), and prevention benefits or costs from a societal perspective (e.g. effects of the information generated on decision making and population health).
Data quality: Completeness and validity of data in the surveillance system.
Flexibility: Ability of the surveillance system or its processes to adapt with little additional time or resources to changing epidemiology, information needs, technologies, or clinical practices.
Predictive value positive (PVP): Proportion of cases reported to the system that actually have the health condition under surveillance. For event-based surveillance, PVP represents the probability that a detected outbreak is of public health significance and requires response.
Representativeness: Ability of the system to accurately describe the occurrence of a health condition under surveillance over time and its distribution in the population by place and person.
Security: Processes and methodologies to keep surveillance data and information confidential, available, and accurate.
Sensitivity: Sensitivity of a surveillance system to (a) case detection (proportion of individuals who have the condition of interest), (b) outbreak detection (probability that the surveillance system will detect a significant increase in a health condition in time or space), and (c) case definition (ability of the case definition criteria to accurately represent the health condition of interest and classify the cases to which it is applied).
Simplicity: Structure and ease of operation of the system across the surveillance process cycle from data collection to use. Systems should be as simple as possible while still meeting their objectives.
Stability: Ability to collect, manage, and provide data without failure (reliability) and to be operational when needed (availability).
Standards use: Use of data exchange, messaging, or other information technology standards by a surveillance system that enhances the ability of the system and its software applications to communicate, exchange data, and use the information that has been exchanged.

Timeliness: Time between any two steps in the surveillance process. Steps in the surveillance process will vary by system. The relative importance of timeliness of surveillance process intervals varies by surveillance objective and health event under surveillance (Groseclose and Buckeridge 2017).

The attributes of public health surveillance systems are related to one another; strengthening one system attribute could adversely affect another attribute of a higher priority. Some attributes are more important for selected surveillance purposes and health outcomes than others. A recent systematic review of 99 articles examining animal and human health surveillance system evaluations identified 23 different attributes of surveillance systems (Drewe et al. 2012). Almost one-half of the evaluations assessed only one or two attributes; the attributes assessed most frequently were sensitivity, timeliness, and data quality. The investigators noted with some concern that surveillance objectives for the system being evaluated were often not stated, and therefore the rationale for assessing the selected attributes was not always evident.

Recommendations about the present state and future potential of surveillance system should be made based on conclusions. The conclusions should state whether surveillance is meeting its objectives and whether it is operating efficiently. Recommendations should be realistic, feasible, and clearly explained. Feedback to health facilities and stakeholders is important. Prioritizing plans and developing a timetable for surveillance improvements might be helpful (CDC 2012c).

Organizational structure

Surveillance systems require an organizational structure that defines the roles and responsibilities of participants in all stages of the surveillance cycle, including participants at different levels of government (local, state, regional, national, or international). This includes articulation of procedures for alerting those responsible for prevention and control activities when surveillance data identify situations that merit further attention, such as an upturn in observed disease incidence that might herald the onset of an outbreak.

Ethical issues in public health surveillance

As surveillance does not fall under the research, however, it may create some challenges on ethical issues. For surveillance, people have experienced legal, social, physical, and psychological harm after the public disclosure of health data. The collection of population-level health data can lead to mandatory quarantines, isolation of groups and individuals, and seizure of property. There is potential for intrusions of privacy and discrimination. Data that associate high HIV rates with particular communities, for example, can lead to stigmatization of members of those communities.

Surveillance, when conducted ethically, is the foundation for programmes to promote human well-being at the population level. It can contribute to reducing inequalities. WHO has developed guidelines on ethical issues in public health surveillance. This document can assist everyone involved in public health surveillance, including officials in government agencies, health workers, NGOs, and the private sector. The guidelines cover (i) the broad responsibility to undertake surveillance and subject it to ethical scrutiny; (ii) the obligation to ensure appropriate protection and rights; and (iii) considerations in making decisions about how to communicate and share surveillance data (WHO 2017). Ethical guidelines for surveillance address following topics: developing effective mechanisms for ethical surveillance dissemination; ensuring data are secure and collected only for legitimate health purposes; identifying and minimizing potential harms before surveillance; assisting countries' lack of adequate resources to undertake surveillance on their own; respecting values and concerns of communities in surveillance; communicating results of surveillance to relevant target audiences; ensuring that identifiable data are appropriately secured; and sharing data with all parties involved and other nations and international health agencies.

Public health surveillance in under-resourced countries: special considerations

General principles that underlie the practice of surveillance are essentially the same for all countries. These include the steps of system design and the principle of collecting and using information for action. Of interest, the revised IHR, which were adopted by the WHO in May 2005 and came into force in June 2007, provide a powerful mandate to strengthen surveillance

systems globally (Baker and Fidler 2006). However, information and other surveillance resources vary widely among countries.

Tracking disease trends, particularly infectious diseases, has been the main reason surveillance systems have been initiated in under-resourced countries. Surveillance data are also needed to establish rational public health priorities and to evaluate the impact of large-scale prevention and treatment programmes, such as those established for malaria, tuberculosis, and HIV/AIDS. Many infectious diseases and other conditions of public health import occur in settings with only rudimentary healthcare and few laboratory resources. With large-scale treatment programmes underway, the need for an adequate laboratory infrastructure has become even more critical, to assure that the treatment is appropriate, that diagnoses are accurate, that microbes are not becoming resistant to therapies, or that viral or bacterial strains represented in vaccines are not being replaced by other strains. Lack of definitive diagnoses may hinder surveillance and response efforts as well as resulting in inappropriate therapy and adverse consequences, such as increases in drug resistance.

The success of surveillance depends on having sufficient resources, the simplicity of reporting procedures, personal rapport among people in the network, regular feedback, and the visibility of actions taken in response to case reports or analyses of surveillance data. For example, programmes to eradicate polio make extensive use of surveillance to monitor progress and focus efforts towards reaching their goals. A major accomplishment of the polio eradication programme has been to develop an integrated global virological surveillance network to track the genetic profile and transmission patterns of circulating polio viruses (Pallansch and Sandhu 2006). Eradication programmes must rely on such highly targeted surveillance, which becomes more important (and expensive) as eradication of the target disease approaches.

In many under-resourced countries, the absence of an effective vital event registration systems for births and deaths shapes how surveillance is conducted. Use of the verbal autopsy, based on interviews to determine the cause of death, may assist in following mortality patterns in places without routine death registration (Kumar et al. 2006). The sensitivity of this approach in classifying causes of death may be lower for acute febrile conditions such as malaria, than for maternal deaths, injuries, tuberculosis, and AIDS. In addition, different techniques in conducting verbal autopsies

may result in different sensitivities of classifications for specific conditions (Setel et al. 2006). A potentially simple and rapid alternative to such mortality surveillance is to monitor severe morbidity through periodic surveys of hospitals for the number of admissions attributed to particular conditions. Again, the success of such surveys is largely dependent on the accuracy of diagnoses.

Other sources of health information may include UNICEF, WHO, international conferences, non-governmental organizations, and population laboratories (e.g. the International Centre for Diarrhoeal Disease Research, Bangladesh). Although some health problems are similar in many under-resourced settings, relying on data from other countries or districts can often create major problems when there are large geographical differences in the incidence of specific conditions.

The design of surveillance systems must consider the security and safety of surveillance staff, geography and accessibility, population dispersion and mobility, type of health system, and literacy. Problems (more common but not unique to under-resourced countries) may include limited personnel available for public health, coincident operation of multiple surveillance systems within a target population, lack of laboratory capacity, and infrastructure and communications constraints (e.g. lack of equipment, supplies, or electrical power).

Solutions to address the lack of personnel for public health and prevention have included voluntary systems (using community health workers, traditional birth attendants, or village volunteers). More familiar solutions are public health training programmes designed to meet human resource gaps. Concerns about the cost-effectiveness of short-term training have resulted in establishing long-term programmes to build capacity within countries. Applied field epidemiology training programmes have been applauded for boosting the number of trained personnel (Cardenas et al. 2002).

Concurrently, the increase in availability of computers to analyse and transmit surveillance data and the decrease in the cost of such technology offer increased opportunities for surveillance. Epi Info™ is a freely available computer program designed to assist public health data management and analysis (CDC 2013a); this tool has been used successfully in both well-resourced and under-resourced countries, and is available in multiple languages. Epi Info™ and other information systems should be seen as tools

to be used to provide data to policymakers and others to inform decisions that will improve health. The use of mobile phones and hand-held devices is proving useful in many areas. By leveraging the availability of mobile telephone networks and international telecommunication systems, few additional resources are needed to provide for collection of information and for two-way exchange of information needed for response.

Conclusion

Public health surveillance is a continued watchfulness over health-related problems through systematic collection, consolidation, and evaluation of relevant data. Public health surveillance has historically galvanized and guided prevention and control programmes for infectious diseases, ranging from smallpox eradication, immunization campaigns to prevent childhood diseases, programmes to prevent HIV infection and AIDS, and responses to emergent diseases such as SARS or pandemic influenza. Surveillance has also taken on increased visibility and importance in evaluating and guiding prevention and control efforts for a broad spectrum of non-infectious diseases and conditions. Surveillance provides a stimulus to keep prevention and control activities moving in the right direction, guiding the response to individual cases as well as public policy.

Effective public health interventions depend upon a continuing and reliable source of information; resources should be allocated for the maintenance of the surveillance systems and for their regular evaluation. The data must be timely and representative of the population; they must be analysed and interpreted with feedback to the reporters and disseminated as widely as possible, including to those formulating and implementing public health policy. Allocation of resources devoted to surveillance for the purpose of informing prevention and control activities must be balanced with resources needed to implement those activities.

ACKNOWLEDGEMENTS

The authors thank James W. Buehler and Ann Marie Kimball for their contributions as authors of previous editions of this chapter.

REFERENCES

Ackers, M.-L., Puhr, N.D., Tauxe, R.V., and Mintz, E.D. (2000). Laboratory-based surveillance of salmonella serotype typhi infections in the United States: antimicrobial resistance on the rise. *Journal of the American Medical Association*, 283(20), 2668–73.

Ahmed, A.B., et al. (2007). The growing impact of HIV infection on the epidemiology of tuberculosis in England and Wales. *Thorax*, 62, 672–6.

Baker, M.G. and Fidler, D.P. (2006). Global public health surveillance under new international health regulations. *Emerging Infectious Diseases*, 7, 1058–65.

Berkelman, R.L., et al. (1994). Infectious disease surveillance: a crumbling foundation. *Science*, 264, 368–70.

Botto, L.D., et al. (2006). Fostering international collaboration in birth defects research and prevention: a perspective from the International Clearinghouse for Birth Defects Surveillance and Research. *American Journal of Public Health*, 96, 774–80.

Brookmeyer, R. and Stroup, D.F. (2003). *Monitoring the Health of Populations: Statistical Principles and Methods for Public Health Surveillance*. Oxford: Oxford University Press.

Buehler, J.W., Prager, K., and Hogue, C.J. (2000). The role of linked birth and infant death certificates in maternal and child health epidemiology in the United States. *American Journal of Preventive Medicine*, 19(Suppl. 1), 3–11.

Buehler, J.W., Whitney, E.A., Smith, D., Prietula, M.J., Stanton, S.H., and Isakov, A.P. (2009). Situational uses of syndromic surveillance. *Biosecurity and Bioterrorism: Biodefense Strategy, Practice, and Science*, 7(2), 165–77.

Burkom, H. (2007). Alerting algorithms for biosurveillance. In: Lombardo, J.S. and Buckeridge, D.L. (eds.) *Disease Surveillance: A Public Health Informatics Approach*. Hoboken, NJ: John Wiley & Sons, pp. 143–92.

Cardenas, V.M., et al. (2002). Improving global public health leadership through training in epidemiology and public health: the experience of TEPHINET. Training Programs in Epidemiology and Public Health Interventions Network. *American Journal of Public Health*, 92, 196–7.

Centers for Disease Control and Prevention (CDC) (2001). Updated guidelines for evaluating public health surveillance systems. *Morbidity and Mortality Weekly Report*, 50, 1–35.

Centers for Disease Control and Prevention (CDC) (2004). Framework for evaluating public health surveillance systems for early detection of outbreaks. Recommendations from the CDC Working Group. *Morbidity Mortality Weekly Report*, 53(RR05), 1–11.

Centers for Disease Control and Prevention (CDC) (2009). Multistate outbreak of Salmonella infections associated with peanut butter and peanut butter-containing products—United States, 2008–2009. *Morbidity and Mortality Weekly Report*, 58(4), 85–90.

Centers for Disease Control and Prevention (CDC) (2011). *STD Surveillance Case Definition.* [Online] Available at: http://www.cdc.gov/std/stats10/app-casedef.htm

Centers for Disease Control and Prevention (CDC) (2012a). *Overview of Influenza Surveillance in the United States*. [Online] Available at: http://www.cdc.gov/flu/weekly/overview.htm

Centers for Disease Control and Prevention (CDC) (2012b). *Diabetes Complications*. [Online] Available at: http://www.cdc.gov/diabetes/statistics/complications_national.htm

Centers for Disease Control and Prevention (CDC) (2012c). Public health surveillance. In: *Principles of Epidemiology in Public Health Practice: An Introduction to Applied Epidemiology and Biostatistics* (3rd ed.). Washington, DC: CDC.

Centers for Disease Control and Prevention (2013a). *Epi Info™*. [Online] Available at: http://wwwn.cdc.gov/epiinfo/

Centers for Disease Control and Prevention (2013b). *National Health Interview Survey*. [Online] Available at: http://www.cdc.gov/NCHS/NHIS.htm.

Choi, K. and Thacker, S.B. (1981). An evaluation of influenza mortality surveillance, 1962–1979. I. Time series forecasts of expected pneumonia and influenza deaths. *American Journal of Epidemiology*, 113, 215–26.

Clark, D.E. (2004). Practical introduction to record linkage for injury research. *Injury Prevention*, 10, 186–91

Correa-Villasenor, A., et al. (2003). The Metropolitan Atlanta Congenital Defects Program: 35 years of birth defects surveillance at the Centers for Disease Control and Prevention. *Birth Defects Research. Part A, Clinical and Molecular Teratology*, 67, 617–24.

Cox, N., Brammer, T.L., and Regnery, H. (1994). Influenza: global surveillance for epidemic and pandemic variants. *European Journal of Epidemiology*, 10(4), 467–70.

Davy, M. (2006). Time and generational trends in smoking among men and women in Great Britain, 1972–2004/05. *Health Statistics Quarterly*, 32, 35–43.

Devine, O. and Parrish, R.G. (1998). Monitoring the health of a population. In: Stroup, D.S. and Teutsch, S.M. (eds.) *Statistics in Public Health*. New York: Oxford University Press, pp. 59–91.

Dolley, S. (2018). Big data's role in precision public health. *Frontiers of Public Health*, 6, 68.

Drewe, J.A., Hoinville, L.J., Cook, A.J.C., Floyd, T., and Stark, K.D.C. (2012). Evaluation of animal and public health surveillance systems: a systematic review. *Epidemiology & Infection*, 140, 575–90.

European Centre for Disease Prevention and Control (ECDC) (2014). *Data Quality Monitoring and Surveillance System Evaluation—A Handbook of Methods and Applications*. Stockholm: ECDC, pp. 7–13.

Fleming, D.M., Schellevis, F.G., and Paget, W.J. (2003). Health monitoring in sentinel practice networks: the contribution of primary care. *The European Journal of Public Health*, 13(Suppl. 1), 80–4.

Freifeld, C.C., et al. (2010). Participatory epidemiology: use of mobile phones for community-based health reporting. *PLoS Medicine*, 7(12), e1000376.

Groseclose, S.L. and Buckeridge, D.L. (2017). Public health surveillance systems: recent advances in their use and evaluation. *Annual Review of Public Health*, 38, 57–79.

HealthIT.gov (2013). *HITECH Act*. [Online] Available at: http://www.healthit.gov/policy-researchers-implementers/hitech-act-0

Hinman, A.R. and Hopkins, D.R. (1998). Lessons from previous eradication programs. In: Dowdle, W.R. and Hopkins, D.R. (eds.) *The Eradication of Infectious Diseases*. New York: Wiley, pp. 19–32.

Hurd, M.D., Martorell, P., Delavande, A., Mullen, K.J., and Langa, K.M. (2013). Monetary costs of dementia in the United States. *The New England Journal of Medicine*, 368(14), 1326–34.

Jaffe, H.W., et al. (1983). National case control study of Kaposi's sarcoma and *Pneumocystis carinii* pneumonia in homosexual men: epidemiologic results. *Annals of Internal Medicine*, 99, 293–8.

Jayaraman, G.C., et al. (2006). A population-based approach to determine the prevalence of transmitted drug-resistance HIV among recent versus established HIV infections: results from the Canadian HIV strain and drug resistance surveillance program. *Journal of Acquired Immune Deficiency Syndromes*, 42, 86–90.

Kenya Central Bureau of Statistics, Ministry of Health (MOH) and ORC Macro (2004). *Kenya 2003 Demographic and Health Survey: Key Findings. Calverton MD, USA, MOH and ORC Macro*. [Online] Available at: http://www.measuredhs.com/pubs/pdf/SR104/SR104KE03.pdf

Kimball, A.M., Thacker, S.B., and Levy, M.E. (1980). Shigella surveillance in a large metropolitan area: assessment of a passive reporting system. *American Journal of Public Health*, 70(2), 164–6.

Krause, G. (2006). From evaluation to continuous quality assurance of surveillance systems. *Eurosurveillance*, 11(11), 204–5.

Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics—Theory and Methods*, 26(6), 1481–96.
Kumar, R., et al. (2006). Validity of verbal autopsy in determining causes of adult deaths. *Indian Journal of Public Health*, 50, 90–4.
Langmuir, A.D. (1963). The surveillance of communicable diseases of national importance. *The New England Journal of Medicine*, 268, 182–92.
Liu, L., et al. (2000). Global, regional, and national causes of child mortality: an updated systematic analysis for 2010 with time trends since 2000. *The Lancet*, 379(9832), 2151–61.
Lowndes, C.M. and Fenton, K.A. (2004). Surveillance systems for STIs in the European Union: facing a changing epidemiology. *Sexually Transmitted Infections*, 80, 264–71.
Madoff, L.C. and Woodall, J.P. (2005). The Internet and the global monitoring of emerging diseases: lessons from the first 10 years of ProMED-mail. *Archives of Medical Research*, 36, 724–30.
Mandl, K.D., et al. (2004). Implementing syndromic surveillance: a practical guide informed by the early experience. *Journal of the American Medical Informatics Association*, 11, 141–50.
Pallansch, M.A. and Sandhu, H.S. (2006). The eradication of polio—progress and challenges. *The New England Journal of Medicine*, 355, 2508–11.
Pan American Health Organization (PAHO) and WHO Regional Office for America (2014). *Operational Guidelines for Sentinel Severe Acute Respiratory Infection (SARI) Surveillance*, pp. 11–16. [Online] Available at: https://www.paho.org/revelac-i/wp-content/uploads/2015/10/2015-cha-operational-guidelines-sentinel-sari.pdf
Public Health Agency of Canada (2013). *The Global Public Health Intelligence Network (GPHIN)*. [Online] Available at: http://www.phac-aspc.gc.ca/gphin/
Public Health Informatics Institute (2005). *Public Health Opportunities in Health Information Exchange. Topics in Public Health Informatics (2005)*. [Online] Available at: http://www.phii.org/sites/default/files/resource/pdfs/Opportunities_0605.pdf
Rigau-Perez, J.G., et al. (2006). Medical examiner samples: a source for dengue surveillance. *Puerto Rico Health Sciences Journal*, 25, 67–9.
Setel, P.W., et al. (2006). Core verbal autopsy procedure with comparative validation results from two countries. *PLoS Medicine*, 3, e268.
Seward, J.F., et al. (2002). Varicella disease after introduction of varicella vaccine in the United States, 1995–2000. *Journal of the American Medical Association*, 287(5), 606–11.
Silk, B. and Berkelman, R.L. (2005). A review of strategies to enhance completeness of reporting. *Journal of Public Health Management and Practice*, 11, 191–200.
Smith, P.F., et al. (2013). Blueprint version 2.0: updating public health surveillance for the 21st century. *Journal of Public Health Management and Practice*, 19, 231–9.
Swaminathan, B., et al. and the CDC PulseNet Task Force (2001). PulseNet: the molecular subtyping network for foodborne bacterial disease surveillance, United States. *Emerging Infectious Diseases*, 7, 382–9.
Thacker, S.B. (2010). Historical development. In: Lee, L.M., Teutsch, S.M., Thacker, S.B., and Louis, M.S. (eds.) *Principles and Practice of Public Health Surveillance*. Oxford: Oxford University Press, pp. 1–17.
Thacker, S.B. and Berkelman, R.L. (1988). Public health surveillance in the United States. *Epidemiologic Reviews*, 10, 164–90.
Tufte, E.R. (1983). *The Visual Display of Quantitative Information*. Cheshire, CT: Graphics Press.
Tukey, J.W. (1977). *Exploratory Data Analysis*. Reading, MA: Addison-Wesley.
Twigg, L. (1999). Choosing a national survey to investigate smoking behavior: making comparisons between the General Household Survey, the British Household Panel Survey and the Health Survey for England. *Journal of Public Health Medicine*, 21, 14–21.
UNAIDS/WHO (Joint United Nations Programme on HIV/AIDS and World Health Organization) (2000). *Working Group on Global HIV/AIDS and STI Surveillance. Guidelines for Second*

Generation HIV Surveillance. WHO/CDS/CSR/EDC/2000.5 UNAIDS/00.03E. [Online]. Available at: https://www.who.int/hiv/pub/surveillance/en/cds_edc_2000_5.pdf

Valdivia, A., López-Alcalde, J., Vicente, M., Pichiule, M., Ruiz, M., and Ordobas, M. (2010). Monitoring influenza activity in Europe with Google Flu Trends: comparison with the findings of sentinel physician networks—results for 2009–10. *Eurosurveillance*, 15(29), 2–7.

Waller, L.A. and Gotway, C.A. (2004). *Applied Spatial Statistics for Public Health Data*. Hoboken, NJ: John Wiley & Sons.

Weddell, J.M. (1973). Registers and registries: a review. *International Journal of Epidemiology*, 2, 221–28.

Williamson, L.M., et al. (2006). Increases in HIV-related sexual risk behaviour among community samples of gay men in London and Glasgow: how do they compare? *Journal of Acquired Immune Deficiency Syndromes*, 42, 238–41.

World Bank (1993). *World Development Report 1993: Investing in Health*. Oxford: World Bank and Oxford University Press.

World Bank (2012). *2012 Information and Communications for Development: Maximizing Mobile*. Washington, DC: World Bank.

World Health Organization (WHO) (1968). *Report of the Technical Discussions at the 21st World Health Assembly on 'National and Global Surveillance of Communicable Diseases*. Geneva: World Health Organization.

World Health Organization (WHO) (2003). *Guidelines for the Management of Sexually Transmitted Infections*. Geneva: World Health Organization.

World Health Organization (WHO WPRO) (2008b). *A Guide to Establishing Event-Based Surveillance*. [Online]. Available at: https://iris.wpro.who.int/handle/10665.1/10421

World Health Organization Regional Office for Africa (2014). *Integrated Disease Surveillance and Response in the African Region: A Guide for Establishing Community Based Surveillance*. [Online]. Available at: https://www.afro.who.int/publications/integrated-disease-surveillance-and-response-african-region-guide-establishing-0

World Health Organization (WHO) (2016). *International Health Regulations (2005)* (3rd ed.). Geneva: World Health Organization.

World Health Organization (WHO) (2017).*WHO Guidelines on Ethical Issues in Public Health Surveillance*. Geneva: World Health Organization.

Yan, W., Nie, S., Xu, B., Dong, H., Palm, L., and Diwan, V. (2012). Establishing a web-based integrated surveillance system for early detection of infectious disease epidemic in rural China: a field experimental study. *BMC Medical Informatics and Decision Making*, 12(1), 4.

Yu, V.L. and Madoff, L.C. (2004). ProMED-mail: an early warning system for emerging diseases. *Clinical Infectious Diseases*, 39(2), 227–32.

5.18

Life course epidemiology and analysis

Elizabeth Rose Mayeda, Alexandra M. Binder, and Lindsay C. Kobayashi

Introduction

Life course epidemiology primarily aims to understand disease aetiology, with the goal of identifying effective strategies to promote population health and health equity. Successful interventions must be delivered at the right time in a person's life, and life course epidemiology focuses on evaluating how exposures at different periods of life may have different effects on subsequent health outcomes. The integration of a life course approach to epidemiologic research is central for identifying effective policies and population-level interventions to promote population health and health equity. In this chapter, we summarize conceptual models in life course epidemiology, discuss causal inference, biological mechanisms, and areas of innovation in life course epidemiology, and we close with future directions in life course epidemiology.

Conceptual models in life course epidemiology

Life course epidemiologic research is guided by conceptual models developed by Kuh and Ben-Schlomo that describe how the timing of exposure to a risk or protective factor influences health later in the lifespan (Ben-Shlomo and Kuh 2002; Kuh et al. 2003; Ben-Shlomo et al. 2016). We illustrate major conceptual models here with directed acyclic graphs (DAGs). We distinguish between models, all of which assume there is at least some effect of exposure on health: the immediate risk model, the trajectory, or chains of risk model, the critical and sensitive periods models, and the cumulative risk model.

In the *immediate risk model*, the exposure has an immediate effect on health; after the exposure is removed, the risk of the health outcome starts to decline (Fig. 5.18.1a). Examples of the immediate risk model include household firearm ownership and suicide rates among youth and adults (Siegel and Rothman 2016; Knopov et al. 2019), use of benzodiazepines and risk of falls among older adults (Díaz-Gutiérrez et al. 2017), and the effects of shift work and night work on workplace accidents and car accidents following night work shifts (Kecklund and Axelsson 2016).

In the *trajectory model* (also called 'chains of risk model'), the exposure trajectory is 'sticky': exposure at each life course period sets the stage for exposure at the next life course period. The initial insult does not directly harm health later in life, but the first exposure leads to subsequent exposures that directly affect health. Figure 5.18.1b illustrates an immediate risk and trajectory model, where the exposure in childhood and early adulthood do not directly influence midlife health, but exposure in childhood influences exposure in early adulthood, which in turn influences exposure in midlife. An example of the trajectory model would be if health behaviours established in childhood influence adult health only via influencing health behaviour patterns in adulthood.

In the *cumulative risk model*, each exposure period has an independent effect on health (Fig. 5.18.1c). This can occur with or without a trajectory model, where the exposure is 'sticky'. Note that Figure 5.18.1c displays a cumulative risk with trajectory model, as exposure at each life course period has an independent effect on the health outcome as well as exposure at the subsequent life course period. For example, socioeconomic position tends to be 'sticky' across the life course, and indicators of socioeconomic position at each life stage (e.g. early life, mid life, later life) may also have independent effects on health outcomes in later life. Using data from the US Health and Retirement Study, Marden et al. observed results consistent with the cumulative risk model for indicators of socioeconomic position at three life periods in relation to later-life memory function and decline (Marden et al. 2017). Testing the cumulative risk model can pose some challenges to causal inference, as illustrated in their study: confounders of any mid- or later-life exposure that arise from early-life exposure may also mediate effects of early-life exposure on the health outcome (often referred to as 'time-dependent confounder-mediators'), and statistical adjustment for these variables could bias estimated effects of the exposure on the outcome. This

challenge can be addressed by strategies such as inverse probability weighting to block the confounding effect of the confounder-mediator variable on the mid- or later-life exposure, without blocking its mediating effect on the early-life exposure to outcome relationship (Robins et al. 2000; McCulloch 2015).

In practice, the terms ‘sensitive period’ and ‘critical period’ are often used interchangeably, but they are conceptually distinct. We use the same DAG to illustrate both models (Fig. 5.18.1d). In the *sensitive period model*, an exposure during a specific life course period has a particularly strong influence on certain aspects of health. During other periods, the exposure has a weaker impact on health. For example, the Moving to Opportunity (MTO) for Fair Housing Demonstration Program was a randomized trial in the United States in the 1990s that randomly assigned families living in high-poverty public housing projects to receive subsidized housing vouchers to move to lower-poverty neighbourhoods (Leventhal and Brooks-Gunn 2003). Chetty et al. (2016) examined the impact of moving to a lower-poverty neighbourhood on children’s long-term outcomes using tax records and found that moving to a lower-poverty neighbourhood at younger ages, particularly before age 13 years, increased college attendance and earnings in adulthood, and reduced single parenthood rates. These findings suggested that childhood is a sensitive period for the impact of neighbourhood and long-term outcomes. Another randomized controlled trial that was influential in identifying sensitive early childhood periods for cognitive and behavioural development is the Bucharest Early Intervention Project in Bucharest, Romania. Beginning in the year 2000, this study randomized infants and toddlers living in institutions who were abandoned at or soon after birth to move to foster care (treatment) or to remain in institutional care (control) (Zeanah et al. 2003; Nelson et al. 2007; Black et al. 2017). Other examples of evidence for childhood and adolescence as sensitive periods includes studies of age at arrival and later-life health outcomes among immigrants to the United States (Roshania et al. 2008; Osypuk et al. 2015a).

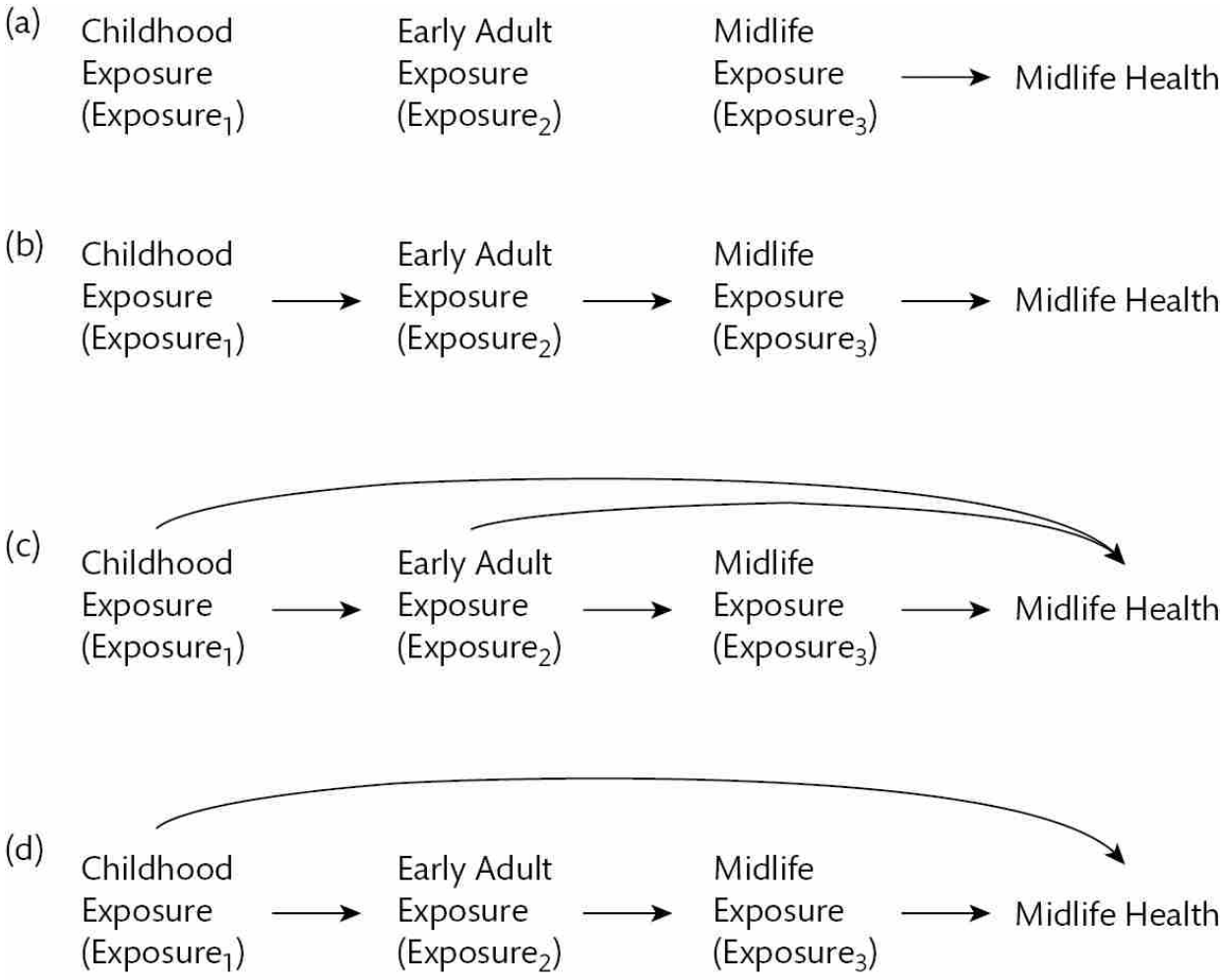


Fig. 5.18.1 Life course models: (a) Immediate risk model; (b) Immediate risk and trajectory model; (c) Cumulative risk and trajectory model; (d) Childhood critical/sensitive period and trajectory model.

An essential aspect to defining a sensitive period is distinguishing it from the induction period. The induction period refers to the time between exposure and the initiation of the disease process. To disentangle these mechanisms, investigators must consider time since exposure separate from the life course milestones that are hypothesized to define the sensitive period. This may be difficult if study recruitment is defined by the postulated sensitive window. It is also important to bear in mind that the induction period may impact the degree to which a sensitive window can be detected in a specific study. As an example, consider a hypothetical investigation of the association between gestational hypertension and cardiovascular disease (CVD) among the subset of parous women in a large established cohort. We will assume the investigators found that gestational hypertension during first

pregnancy was associated with modestly elevated risk of CVD in the 10-years following that pregnancy, but a substantially elevated risk of CVD in the 40-year incidence of CVD. These results would suggest that the induction period for the influence of gestational hypertension on CVD is, on average, longer than 10-years. However, these results do not reveal whether the association between hypertension and CVD is stronger when hypertension occurs during pregnancy versus hypertension during other life course periods. The association between hypertensive disorders of pregnancy, such as preeclampsia, and subsequent CVD has been well-replicated across observational studies (Garovic et al. 2010; Tooher et al. 2017; Ying et al. 2018). One theory behind this relationship is that hypertensive disorders of pregnancy may be an early indicator of a reduced capacity to accommodate cardiometabolic stress (Sattar and Greer 2002). Correspondingly, women with hypertensive disorders of pregnancy have a much higher risk of future hypertension (Tooher et al. 2017; Ying et al. 2018). Regardless of whether gestational hypertension has a stronger association with CVD incidence than hypertension during other periods of the life course, it does characterize a high-risk subgroup of women who could potentially benefit from early CVD prevention and screening efforts.

The *critical period model* is a more stringent form of the sensitive period model that assumes that an exposure has a specific impact on disease aetiology during a discrete life course period. Outside of that timeframe, the exposure has little to no impact on that facet of the disease process. Classically, critical exposure windows have been suggested to overlap with periods of developmental plasticity, such as during gestation, early childhood, puberty, and adolescence. For this reason, this model has often been considered synonymous with the concept of ‘developmental programming’ and the Developmental Origins of Health and Disease hypothesis (Barker 1995; Barker 2012), which is discussed later in the chapter. A recent example of the critical period model is a study in California that documented a harmful association between prenatal exposure to ambient fine particulate matter exposure and risk of preterm birth, results which highlighted specific gestational periods of vulnerability in mid and late gestation (Sheridan et al. 2019).

An important nuance to the critical period model is that a particular exposure can influence distinct aspects of the disease process during different life course periods. This is exemplified in the association between

body size and breast cancer risk, which differs across life course periods. Prior to menopause, women with an adult body mass index (BMI) classified as overweight or obese have a modestly decreased risk of breast cancer incidence compared to those with a BMI in the normal weight range (van den Brandt et al. 2000; Reeves et al. 2007). In contrast, after menopause, women with a BMI in the overweight or obese range have a modestly increased risk of breast cancer incidence relative to those with a BMI in the normal weight range (van den Brandt et al. 2000; Reeves et al. 2007). Distinct from the associations observed earlier, greater body fatness during adolescence is associated with decreased risk of both pre- and postmenopausal breast cancer, independently of adult BMI (Baer et al. 2005; Bardia et al. 2008; Baer et al. 2010).

Identification of the appropriate life course model for a given research question should be grounded in theory and relevant prior literature. If the researcher believes there is theoretical motivation that more than one life course model may be appropriate for their research question, alternative life course models can be evaluated empirically. In practice, it can sometimes be challenging to differentiate between the effects of age of onset of an exposure (reflecting a sensitive period model) and duration of exposure (reflecting a cumulative risk model) unless an analytic strategy is adopted to evaluate the extent to which the data are consistent with these two alternative models. For example, multiple studies have reported a harmful association between hypertension in midlife and later-life cognitive health (Kivipelto et al. 2001; Whitmer et al. 2005; Knopman et al. 2009; Debette et al. 2011). Power et al. (2013) evaluated whether the association between hypertension in midlife and later-life cognitive function reflected age at onset (sensitive period model) or duration of exposure (cumulative risk model) in the Normative Aging Study, a cohort study of men living in the greater Boston area. The authors found that age at onset of hypertension was not associated with cognition independent of duration, providing support for the cumulative risk model for the association between hypertension and later-life cognitive health (Power et al. 2013).

What are the goals of life course epidemiology?

Research questions in analytic epidemiology can be broken down into two broad categories: predictive research questions and causal research

questions. Life course epidemiology can include both research focused on predictive questions (e.g. does birth weight predict attained height in adulthood?) and research focused on causal questions (e.g. does exposure to maternal cigarette smoking *in utero* influence behavioural and cognitive outcomes in children?). Most research questions in life course epidemiology studies aim to draw causal inferences from the data, and throughout the rest of the chapter, we assume causal inference is the goal. The goal of drawing causal inferences in life course research is to identify the determinants of states of disease and health, in order to provide evidence towards identification of effective strategies to intervene and promote population health and health equity. Researchers should keep their research goal in mind when specifying research questions, designing studies and selecting analytic tools, and considering the strengths and limitations of their approach.

For example, a life course perspective can be taken when trying to identify effective strategies to prevent Alzheimer's disease and related dementias, a major public health challenge (Brookmeyer et al. 2007; Hebert et al. 2013). Clinical onset of dementia typically occurs in later life, but mounting evidence suggests that dementia-related brain pathology begins developing many years—likely decades—prior to clinical onset (Alzheimer's Association 2019). This conceptualization of the disease process suggests the optimal time to intervene to prevent development of dementia-related pathology is in early or mid life. Alternatively, if the researcher aims to identify factors that contribute to resilience to cognitive decline in the presence of dementia-related pathology, they may consider how exposures at different points in the life course may help buffer against the functional consequences of dementia-related pathology to prevent or delay cognitive decline and the onset of cognitive impairment.

Causal inference in life course epidemiology

Distinguishing between alternative conceptual models is important because these models have implications for designing interventions and prioritizing resources. Under all but the 'immediate risk' model, intervening to change exposure in childhood would provide the greatest benefit to improve adult health outcomes. Under the cumulative risk model, it is not possible in adulthood to completely remediate the consequences of childhood exposures. Under the critical or sensitive childhood period models,

intervening to improve exposure in adulthood has no effect on adult health. Causal inferences about life course exposures, especially when disentangling alternative conceptual models, are often challenging. More details about inference in epidemiology can be found in Chapter 5.14.

An essential, but frequently overlooked step in conducting meaningful research is to clearly define the research question. The counterfactual contrast of interest and appropriate methods to estimate the counterfactual contrast depends on the research question. This critical first step of causal inference was clearly articulated by Kaufman and Hernán (2012): 'No method is so sophisticated that it can relieve the investigator of the obligation to provide a well-formulated causal question in the first place.' Once the research question is clearly defined, major threats to internal validity in life course epidemiology include confounding, collider-stratification bias, measurement error, and reverse causation.

Bias arising from confounding from unmeasured (or poorly measured) common causes of the exposure and outcome of interest is a major concern in observational epidemiology (Fig. 5.18.2a). Clearly specifying the assumed causal structure relevant to the research question of interest using DAGs can help guide researchers towards identifying and controlling for potential confounders in a given study. Time-dependent confounder-mediators (introduced under the cumulative risk model description earlier in the chapter) are of special relevance in life course epidemiology.

Collider-stratification bias arises from conditioning on a common outcome of the exposure and the outcome of interest and related causal structures (Fig. 5.18.2b) (Hernán et al. 2004). In life course epidemiology, a special case of collider-stratification bias, called 'survival bias', can arise from conditioning on survival when the exposure of interest and the outcome of interest are correlated with survival. Survival bias is potentially relevant across the life course, as it is relevant to studies estimating the effects of prenatal exposures on health after birth (Liew et al. 2015) as well as research estimating effects of life course exposures on later-life health (Mayeda et al. 2018a; Mayeda et al. 2018b).

Reverse causation refers to when an exposure-outcome association arises from the outcome of interest influencing the exposure of interest (instead of vice versa) (Fig. 5.18.2c). The aetiology of many chronic diseases may begin years before onset of clinical disease. Often, the latency period (i.e. subclinical period following exposure) is unknown. As a result, it can be

difficult to rule out reverse causation as a potential explanation for an observed exposure-outcome association or identify the relevant life course period in which to intervene on the exposure to prevent disease. For example, reverse causation is an important concern for research on determinants of Alzheimer's disease and related dementias because of the long latency period (Singh-Manoux et al. 2017; Singh-Manoux et al. 2018).

Measurement error, including error in measurement of exposures, outcomes, confounders, and mediations, present a major challenge in life course epidemiology (Fig. 5.18.2d). Clearly defining the construct of interest is an important first step in identifying an appropriate measure of the construct. Inadequate measurement can result in conceptual inadequacy (i.e. the researcher asks a vaguely defined question and, as a result, the study yields a meaningless answer). Inadequate measurement can also result in missing data because research participants do not know how to answer or will decline to answer the interview items. If interview items are worded in such a way that most participants provide the same response regardless of their true level of the construct being measured, the study measures will have poor variability. Inadequate measurement can also result in poor validity and reliability (i.e. the measurements do not align with the underlying construct the researcher aims to measure or the study measure measures the underlying construct with a high degree of error). Given that life course epidemiology often seeks to estimate effects of exposures or requires control for potential confounders that precede the study period, measurement error is a major concern in life course epidemiology. Validation studies are a useful tool to validate study measures and to provide data to support useful quantitative bias analysis (Osypuk et al. 2015b; Niesner et al. 2013; Lash and Olshan 2016; Hinkle et al. 2017; Luke et al. 2018).

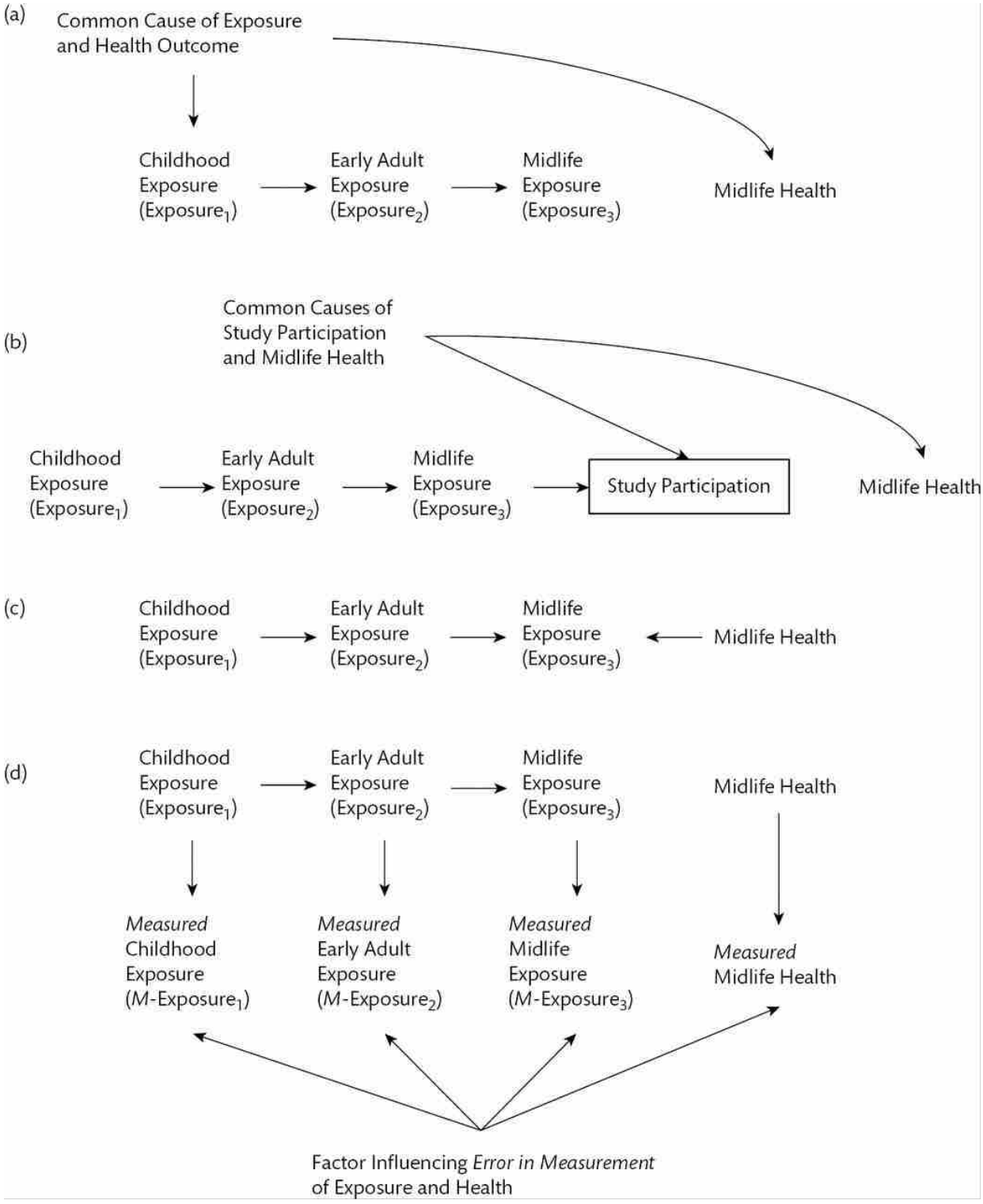


Fig. 5.18.2 Examples of sources of bias in life course epidemiology studies. Note that in all examples, there is no effect of the exposure of interest at any life course period on midlife health, but confounding, collider-stratification bias, reverse causation, or measurement error induce a spurious association between the exposure of interest and outcome of interest: (a) confounding: a common

cause of exposure the health outcome confounds the association between the exposure of interest and the health outcome of interest; (b) collider-stratification bias: the exposure of interest and determinants of the health outcome of interest both influence study participation; (c) reverse causation: the health outcome of interest influences the exposure of interest; (d) measurement error: error in measurement of the exposure of interest and health outcome of interest are correlated.

In addition to threats to internal validity, generalizability (extending study results from the analytic sample to the target population of inference) is also an important challenge, as the target population of inference almost always differs from the study sample in life course research (Lesko et al. 2017; Westreich et al. 2017; Stuart et al. 2018). A researcher may aim to generalize results to a larger target population with different distributions of socioeconomic and health characteristics than their study sample. For example, a researcher may wish to generalize results from a US-based multisite cohort study established in 2010 to the general US population in 2010. It may also be of interest to generalize results across time periods, for example, generalizing results from a national cohort study in the United Kingdom conducted in 1960 to the United Kingdom general population in 2020. For social exposures, there is potential for the correspondence between measures and constructs to differ across time and place. For example, Hayward and Gorman's influential paper 'the long arm of childhood' found that early-life social factors were powerful predictors of mortality in mid- to late-life in US men born in the early twentieth century (Hayward and Gorman 2004). One early-life social factor that was associated with elevated mortality in adulthood was having a 'mother working outside the home', which was likely a marker of early-life socioeconomic adversity in the early twentieth century, rather than representing the 'effect' of having a mother who worked outside the home. Markers of socioeconomic status are not necessarily constant across place and time, which should be considered in life course epidemiology.

Methods for drawing causal inferences using observational data are essential in life course epidemiology, and common tools such as DAGs clearly represent the challenges. DAGs are a widely used tool in epidemiology for visually encoding assumptions about causal relations among the exposure, outcomes, and covariates (Glymour 2017). DAGs can be particularly helpful in distinguishing between covariates that represent potential confounders (i.e. variables that represent common causes of the effect of exposure on the outcome) and covariates represent potential mediators (i.e. variables that are on the causal pathway of interest). Clearly

differentiating between potential confounders and mediators helps researchers plan analytic strategies accordingly. Often, several life course models may seem plausible for a given research question, but different life course models have inconsistent empirical predictions, so empirical research results based on analyses with clearly articulated assumptions can provide evidence on which is most plausible for each outcome in a given target population.

A major study design challenge in life course epidemiology is that the length of follow-up required to collect information across the human lifespan or in the case of research across successive generations, multiple human lifespans. As a result, many life course studies that use data from cohorts established in mid-to-later life but wish to examine early-life exposures must rely on retrospective collection of exposures. For example, a researcher may wish to estimate the effects of childhood exposures on health in late life in a given target population using a cohort where study participants were enrolled at ages 65 years and older. In this case, measurement error in the exposure is a major concern, as the researchers may need to rely on retrospective reporting of early-life experiences that occurred many decades prior to the time of the study. Additionally, researchers would be estimating the effects of early-life exposures on the late-life health outcomes among the people who survived to late-life. If the exposure influences survival, collider-stratification bias may arise from conditioning on survival to late life (Mayeda et al. 2018a; Mayeda et al. 2018b).

Biological mechanisms in life course epidemiology

Understanding the biological mechanisms driving the association between life course exposures and later-life health can inform the development of successful interventions to promote population health and health equity. In this section, we discuss several areas of active research advances on the biological mechanisms in life course epidemiology.

Embodiment

Embodiment refers to the processes through which our bodies physically incorporate aspects of the material and social world in which we live throughout the life course (Krieger 2005). These exposures ultimately manifest as states of health or disease; they ‘get under the skin’ to affect

human physiological systems and influence risks of health outcomes (Marmot and Wilkinson 2001). Embodiment is a core concept of the ecosocial theory of disease distribution, an influential epidemiologic theory that explicitly frames how social experiences become biologically embedded (Krieger 2005; Krieger 2011). Ecosocial theory posits that there are multiple, non-mutually exclusive pathways of embodiment that occur across spatial and temporal scales to shape the distribution of health and disease outcomes in a population (Krieger 2011). The concept of embodiment has explicitly informed some hypotheses in social epidemiologic research, such as the role of historical trauma in physical health outcomes among American Indians and Alaska Natives, the causes of US racial disparities in CVD, and gender differences in autism spectrum disorder (Gravlee 2009; Walters et al. 2011; Cheslack-Postava and Jordan-Young 2012). Many life course and social epidemiologic studies implicitly use the construct of embodiment by assuming that material and social exposures have biological effects on human bodies that result in health and disease outcomes. Embodiment inherently takes a life course perspective, as it refers to how material and social conditions across the life course lead to biological processes that ultimately result in population patterns of health and disease.

Developmental Origins of Disease hypothesis

The Developmental Origins of Disease hypothesis posits that the prenatal and early postnatal environments shape health and disease throughout the life course (Barker 1995; Barker 2012). A key theory underlying this hypothesis is that the developing body adapts to the prenatal and early postnatal environment to maximize use of available resources and minimize harm from exogenous insults. This developmental programming can persist through cell divisions to shape adult phenotypes and how the body responds to the environment across the life course. In a seminal paper in life course epidemiology, Barker reported that foetal undernutrition in middle to late gestation, which leads to disproportionate foetal growth, programmes later coronary heart disease risk (Barker 1995). Findings from research studies following people exposed to the Dutch Hunger Winter of 1944–1945 further highlighted the negative health implications of prenatal nutrient deprivation (Painter et al. 2005). A severe food shortage caused during the German occupation of the Netherlands created a tragic natural experiment of the effect of prenatal exposure to famine and health throughout the life course

(Painter et al. 2005). Among a number of adverse health impacts, offspring that were exposed to the famine during early gestation had normal birth weights, but an increased risk of obesity during adulthood (Ravelli et al. 1999). A more recent extension of the Developmental Origins of Disease hypothesis extends beyond one individual's lifespan to include intergenerational transmission of risk of life course health and disease (Barker 2012).

Epigenetic modifications

Epigenetic modifications are frequently proposed to contribute to the delayed or enduring effects of exposures on chronic disease outcomes (e.g. sensitive or critical periods). Epigenetics refers to stable, mitotically heritable regulation of gene expression that is not due to changes in the DNA sequence (Berger et al. 2009). Epigenetic modifications, such as DNA methylation, non-coding RNA, and histone modifications, can influence gene expression pre-transcriptionally (e.g. by altering DNA accessibility and chromatin structure) and/or post-transcriptionally (e.g. by targeting specific transcripts for degradation). The public health interest in the epigenetic programming of chronic disease is driven by three general characteristics of these modifications: their role in shaping development, their supposed temporal stability, and their responsiveness to environmental cues. Together, these characteristics promote the hypothesis that epigenetic modifications mediate the influence of exposures during critical periods of development on adult health. Studies of families exposed to the Dutch Hunger Winter, described earlier, generated some of the first evidence that prenatal famine exposure could have a lasting influence on epigenetic patterns (Heijmans et al. 2008; Tobi et al. 2009). Compared to unexposed same-sex siblings, individuals exposed to famine periconceptionally had lower DNA methylation within the imprinting control region of the insulin-like growth factor II *(IGF2)* gene in whole blood collected six decades later (Heijmans et al. 2008). IGF2 plays an essential role regulating gestational growth and development, and deviation from normal expression patterns is associated with major health implications (Bergman et al. 2013).

There are several complications to establishing the role of epigenetic patterns in disease aetiology (Michels et al. 2013). Many of the determinants of epigenetic patterns are still being identified, making observation studies integrating epigenetic analysis particularly susceptible to unmeasured

confounding. Like any biomarker study, effect estimates can be greatly impacted by batch effects, which can arise due to a number of factors, including variability in sample collection, processing, storage, and technical artefacts in the biomarker assay. All somatic cells in the body are genetically similar, but the epigenome differs by cell type (Roadmap Epigenomics Consortium et al. 2015), making these studies particularly sensitive to the choice of tissue assayed (Jaffe and Irizarry 2014). Generally, an exposure is assumed to influence disease risk by modifying gene regulation in a particular 'target' tissue. For example, we might expect that ambient air pollution impacts the incidence of respiratory disease by modifying gene regulation in lung tissue. If it is not feasible to assay epigenetic patterns within the target tissue, it may not be possible to capture the contribution of epigenetic modifications to the disease mechanism. In most established cohorts, often the only biospecimen collected is whole blood, which is still a heterogeneous mixture of cell types. An essential preliminary question for researchers seeking to integrate epigenetic analysis into their study questions is whether they will be able to assay the hypothesized target tissue or a surrogate tissue that captures shared variation with the target tissue. If the answer is no, and the goal is to understand the disease aetiology, the study might not be worth pursuing. However, if the intent is to identify a non-causal biomarker of exposure or subclinical disease, the study may have merit.

Epigenetic epidemiology is a young field, and many other challenges in the conduct of these studies are logistical (e.g. cost of assaying genome-wide epigenetic variation, statistical methods for integrating multiple types of epigenetic modifications into the analysis) and will evolve with time. In spite of these challenges, the capacity of epigenetic epidemiology to provide insight into the lasting effects of exposures on health motivates the continued expansion of the field. An area of growing interest and debate is whether exposures during an individual's lifetime can exert intergenerational or transgenerational effects on offspring through epigenetic mechanisms. Transgenerational inheritance refers to the transmission of acquired traits to offspring (Walker et al. 2018). In contrast, intergenerational effects occur when multiple generations are directly exposed during critical windows of susceptibility. For example, maternal grandmother's exposures during pregnancy (F0), could exert intergenerational effects on the traits of the developing foetus (F1; child) and the germline of that foetus (F2;

grandchild). One illustration of these intergenerational associations observed in prospectively collected Swedish nationwide registries is the increased risk of asthma among children whose grandmothers smoked cigarettes during pregnancy, controlling for maternal smoking behaviours (Bråbäck et al., 2018; Lodge et al. 2018). Transgenerational inheritance is difficult to prove because it requires at least three generations away from gestating female exposure, and two from either non-gestating female or male exposure to prove it is not a direct intergenerational effect of the exposure (Heard and Martienssen 2014; van Otterdijk and Michels 2016). To date, the degree to which epigenetic information can be conveyed across generations is still unclear in humans (Heard and Martienssen 2014; Perez and Lehner 2019).

Biological age

Most research in life course epidemiology considers chronological age as a measure of the ageing process, but developing accurate measures of biological age is an active area of research. Biological ageing has been conceptualized as acquired changes to cell function, based on extensive epidemiologic and experimental evidence that gene regulation is sensitive to exposures across the life course. There are several postulated measures of biological ageing that are shown to predict health outcomes independently of chronological age, including telomere length (Müezzinler et al. 2013). More recently, researchers have started to use specific variations in DNA methylation to estimate ‘epigenetic age’, which has been consistently associated with mortality and age-related diseases across cohorts, independent of chronological age (Horvath and Raj 2018). These ‘epigenetic clocks’ have been shown to be responsive to changes in environmental exposures, health-related factors, and social and behavioural factors (Horvath et al. 2014; Quach et al. 2017; Nwanaji-Enwerem et al. 2017). While current biological ageing markers require further validation, they could represent a major advance in life course epidemiology by allowing investigation of life course determinants of discrepancies between chronological age and biological age.

Biomarkers in life course epidemiology

An important consideration in life course epidemiology utilizing biomarkers is whether the life course period represented by biomarkers aligns with the

life course period the researcher aims to measure for their research question. For example, if researchers are interested in estimating health effects of lead exposure in adults and wish to identify an appropriate biomarker of lead exposure, they must first consider the lead measure most representative of their lead exposure of interest. For example, researchers might be interested in the acute health effects of recent lead exposure or they might be more interested in the health effects of cumulative lead exposure over decades. Selection of the optimal biomarker of lead exposure for a given research question should be guided by the life course exposure the authors seek to measure to address their research question. Circulating blood lead levels capture recent external lead exposure and lead that has been mobilized from tissue stores, primarily from bone tissue (Hu et al. 2007). Mobilization of lead from bone stores is particularly relevant during life course periods of high bone resorption (i.e. pregnancy and ageing) (Hu et al. 2007). Repeated measures of blood lead levels over time can be used as a proxy for cumulative exposure to lead over the time period for which blood lead levels were measured, but such repeated measures are rarely available, with the exception of some occupational cohort studies (Hu et al. 2007). Bone lead levels measured in the tibia or patella bones are considered an indicator of cumulative dose of lead over decades (Hu et al. 2007), with lead levels measured in the tibia considered to represent a longer time period of exposure than patella bone lead (Wilker et al. 2011; Power et al. 2014).

Often, researchers aim to evaluate the long-term influence of exposures during life course periods that occurred prior to the study period. An example of researchers taking advantage of sensitive periods to improve measurement is research on the effects of childhood net nutrition on later-life health outcomes. Net nutrition is determined by the net of nutritional intake and diarrheal or other diseases that affect the intake and absorption of nutrients (Perkins et al. 2016). This is an exposure that may have implications for adult health, particularly in low-income settings where poor early-life nutrition is common in the population. However, childhood nutrition cannot be measured directly in a cohort study that begins after adolescence. A solution that life course epidemiologists have taken is to use adult height as a biomarker to investigate the later-life health outcomes associated with cumulative net nutrition during growth periods in childhood. In early life, net nutrition cumulatively influences height during two growth periods that determine adult height: from conception to approximately two

years of age, and during early adolescence (Perkins et al. 2016). A range of studies using adult height as a marker of net nutrition during sensitive growth periods in childhood have shown inverse associations between adult height and health outcomes such as CVD, stroke, later-life cognitive function, and Alzheimer's disease and related dementias, as well as suicide and all-cause mortality (Magnusson et al. 2005; Schnaider Beeri et al. 2005; Lee et al. 2009; Nüesch et al. 2016; Kobayashi et al. 2019). An important methodological point is that, in studies with older populations, measurements of height may need to account for ageing-related shrinkage, although some research has indicated that statistical adjustment for age in regression modelling is often sufficient to account for shrinkage (Fernihough and McGovern 2015). These studies, whether explicitly or implicitly, use the sensitive life course model because they use adult height as a biomarker of cumulative net nutrition during sensitive growth periods in childhood growth periods. This approach allows for the investigation of an early-life exposure that otherwise could not be investigated in studies of adult health.

Areas of methodological innovation in life course epidemiology research

Causal inference methods

Life course epidemiology largely relies on observational studies. To provide the best evidence for understanding life course mechanisms influencing health and disease, causal inference methods provide useful tools for ameliorating potential sources of bias in life course epidemiology. The adoption of causal approaches to mediation analysis and effect decomposition (Vanderweele 2015) are potential boons for life course epidemiology, although uptake in many areas has been slow. As illustrated in the life course epidemiology models, causal mediation analysis is central to hypothesis testing in life course epidemiology.

Natural experiments are a useful causal inference tool to minimize threats to causal inference in traditional observational studies. Natural experiments mimic randomized controlled trials in that the mechanism determining each person's exposure is independent of their set of potential outcomes, but unlike a randomized controlled trial, treatment assignment is outside the control of researchers (Glymour et al. 2017). More detail about natural experiments and associated methods can be found in Chapter 5.10.

An example of ground-breaking work in life course epidemiology using a natural experiment is research on the health effects of exposure to the Dutch Hunger Winter of 1944–1945 (introduced earlier in the chapter). One of the most famous findings from this natural experiment link prenatal nutritional deficiency to risk of schizophrenia: the birth cohort conceived at the height of the famine had higher risk of schizophrenia up to midlife (Susser and Lin 1992; Susser et al. 1996). These findings were replicated in research on the Chinese Great Leap Forward famine of 1959–1961 (St Clair et al. 2005). A more recent natural experiment example includes a natural experiment in California, where researchers compared preterm births among mothers who lived near coal and oil power plants before and after eight power plants in California closed between 2001 and 2011 (Casey et al. 2018). They found that power plant retirements were associated with a decrease in the proportion of preterm births (<37 weeks' gestation) among mothers living nearby. For example, for mothers living within 5 km of the power plants, there was an estimated reduction in preterm birth from 7.0% to 5.1% (Casey et al. 2018).

An underutilized tool in life course epidemiology is the use of negative controls to evaluate potential bias in study findings by reproducing the original study with a 'negative control' exposure (or outcome) which cannot be causally associated with the original outcome (or exposure), but which is likely to involve the same sources of systematic bias (Lipsitch et al. 2010; Flanders et al. 2011; Arnold et al. 2016). A popular application of negative controls has been to assess bias due to unmeasured or uncontrolled confounding, although the method has been applied to assess the potential for recall bias, diagnostic bias, and selection bias (Suissa and Ernst 2005; Zaadstra et al. 2008; Ivers et al. 2015; Freedman et al. 2016; Matthay et al. 2017; Cohen et al. 2019; Liew et al. 2019; Hamer et al. 2019). Negative control studies can use negative controls for either the exposure variable or the outcome variable of interest. Here, we will discuss negative controls with respect to negative control outcome studies, although the same principles will apply to negative control exposure studies. In the case of a negative control outcome, it is assumed that the exposure does not have a causal effect on the negative control outcome. When used for the detection of bias due to unmeasured confounding, a key assumption of the method is that the negative control outcome (N) and the original outcome of interest (Y) share the same confounding structure with respect to the exposure of interest (A).

With confounders collectively referred to as U, this is referred to as the A-Y and A-N associations being ‘U-comparable’. When U-comparability or approximate U-comparability holds and measurement error in N is negligible (Lipsitch et al. 2010; Sanderson et al. 2018), any observed association between A and N indicates the presence of confounding bias in the A-Y association.

An applied example is the use of negative control outcomes by Hamer et al. (2019) to assess the potential of confounding bias in the association between physical activity and CVD mortality. A large body of observational evidence has indicated protective effects of moderate-to-vigorous intensity physical activity (MVPA) on CVD mortality, yet this association is subject to unmeasured confounding and is difficult to test using randomized controlled trials due to the long follow-up required to ascertain mortality as an outcome. Hence, Hamer et al. used data on 104,851 participants followed for an average of 9.4 (standard deviation: 4.5) years from the Health Survey for England and the Scottish Health Survey to determine whether unmeasured confounding could at least partially account for the observed association between MVPA and CVD mortality, using accidental death as a negative control outcome. In this example, the authors hypothesized no causal association between MVPA and accidental death, excluding specific causes of accidental death that could be related to physical activity (Hamer et al. 2019). They empirically confirmed the assumption of U-comparability by examining the confounding structures of the MVPA-accidental death (A-N) and the MVPA-CVD mortality (A-Y) relationships. After reproducing the analysis of the MVPA-CVD mortality relationship using accidental death as the outcome, the authors observed point estimates in the protective direction for both associations, although the MVPA-accidental death association was substantially weaker than the MVPA-CVD mortality association (hazard ratio = 0.85; 95% confidence interval: 0.54 to 1.38, and hazard ratio = 0.58; 95% CI: 0.49 to 0.68, respectively) (Hamer et al. 2019). After sensitivity analyses including the evaluation of dose-response relationships and alternative negative controls, the authors concluded that the effects of MVPA on CVD mortality are likely to be causal given the consistency and biological plausibility of the association and its robustness to negative control outcomes, but that the true effects are likely of smaller magnitude than previously reported in observational studies.

Other applications of the negative control method for other types of systematic bias involve variations on the method described earlier for the detection of uncontrolled confounding. The potential magnitude of recall bias can be empirically assessed using negative control exposure variables, by including 'probe' questions in research questionnaires about exposures that cannot be causally associated with the outcomes under study, but would be subject to the same degree of recall bias as the exposures of interest (Zaadstra et al. 2008; Lipsitch et al. 2010). Family studies that swap exposures or outcomes between family members are a special case of negative control studies to detect unmeasured confounding, because they allow for the specific identification of confounding by shared genetic or familial environmental factors (Cohen et al. 2019). Negative controls can also be used to assess diagnostic bias, such as in the paradoxical inverse and bidirectional association between diagnoses of cancer and Alzheimer's dementia, which has been consistently observed in the epidemiological literature (Driver et al. 2012; Driver 2014; Schmidt et al. 2017; Yarchoan et al. 2017). Using US Medicare claims linked to National Cancer Institute Surveillance, Epidemiology, and End Results (SEER) data, Freedman et al. observed that new cancer cases were less likely than cancer-free controls to have a history of Alzheimer's dementia, and that they were also less likely than controls to have a prior injury in an automobile accident, with almost identical point estimates (Freedman et al. 2016). Since an automobile accident injury is highly unlikely to cause cancer, the authors concluded that diagnostic bias may explain their findings, whereby cancer is detected less often among patients with a history of unrelated medical conditions (Freedman et al. 2016).

Life course epidemiology lends itself well to the use of negative controls because life course study designs often necessitate including long follow-up periods, the use of self-reported information on exposures that occurred long before the study started, and the mostly observational nature of life course studies all introduce strong potential for various biases. However, there are few true life course epidemiologic studies that have incorporated negative control methods to detect bias and improve causal inference. Increased use of negative control methods by investigators would help strengthen the evidence base for key research questions in life course epidemiology.

Finally, it is important to remember that no study is without limitations, and the triangulation of evidence across diverse study designs with different

strengths, limitations, and potential sources of bias is crucial for strengthening causal inference in life course epidemiology (Lawlor et al. 2016; Vandenbroucke et al. 2016). Ultimately, causal thinking and the thoughtful application of methods to strengthen causal inference within studies and across bodies of literature will help identify the life course determinants of health and disease in order to inform development of effective policies and programmes to promote population health and health equity.

Monte Carlo simulation studies in life course epidemiology

Monte Carlo simulation studies can be used to evaluate the plausibility of hypotheses in life course epidemiology that are difficult to investigate using empirical data. Conducting a simulation study can be conceptualized as conducting an experiment on a computer. After clearly defining the research question the simulation study is designed to evaluate, the first step in conducting a simulation study is for the researcher to express the causal structure of the data-generating process they wish to simulate as a DAG. The second step is to specify the data-generating rules corresponding with the DAG. Since the researcher is generating the data, they are generating the causal effect of the exposure on the outcome (β) in the simulated world of the study. The third step is to generate the data according to the specified rules using a statistical software package (e.g. R, Stata, SAS). The fourth step is to estimate the exposure-outcome association in the simulated data (β). The fifth step is to quantify the magnitude of bias by comparing the estimated exposure-outcome association with the causal effect in the simulated world (versus β). The final step is to run multiple (e.g. 1,000) iterations of sample generation to obtain stable estimates of the effect estimate in the simulated world, and then compare the average effect estimate and other metrics of distribution of the effect estimate (Burton et al. 2006; Morris et al. 2019). This process can then be repeated, varying the causal structure and parameterizations.

An example of the application of Monte Carlo simulations to evaluate the plausibility of explanations for empirical observations in life course epidemiology is a simulation study to evaluate the extent to which selective survival can plausibly explain the age attenuation of Black-White inequalities in stroke incidence in the United States: in midlife, Black Americans have substantially higher stroke incidence than White Americans,

but this inequality disappears by age 85 years (Giles et al. 1995; Kissela et al. 2004; Kleindorfer et al. 2006; Howard et al. 2011). Although the age attenuation of racial inequalities in stroke has been well-documented, the reason for this observation is difficult to evaluate in empirical data. From a public health perspective, it is relevant to understand the explanation for the observed age patterning of racial inequalities in stroke incidence. One possible explanation is that the observed age trends reflect true age heterogeneity in racial disparities in stroke incidence, due to the elimination of the social disadvantage that drives racial disparities in stroke incidence in midlife. This explanation would be most consistent with the immediate risk model for the effects of racism on stroke incidence. An alternative explanation is that the age attenuation of racial inequalities in stroke reflects selective survival: there are major racial disparities in life expectancy in the United States, and, as a result, Black Americans who survive to older ages may represent a more selected, healthier population than White Americans who survive to old age, which could result in lower stroke incidence rates in Black versus White Americans who survive to old age. Notably, the selective survival explanation does not imply the elimination of health disparities at older ages. Using Monte Carlo simulations guided by real data for mortality and stroke incidence, Mayeda et al. (2018a) demonstrated that the observed age attenuation of racial inequalities in stroke incidence could plausibly be explained by selective survival. More broadly, the results of this simulation study demonstrated that selective survival has the potential to obscure social disparities in health.

Sequence analysis to characterize complex life course exposure patterns

Sequence analysis is a promising method to simultaneously incorporate the order, timing, and duration of life course exposures into a single measure. Sequence analysis can be applied to multielement categorical data repeated over time, such as employment history data. Sequence analysis was originally developed by biologists to identify homology between DNA sequences. The sequence of DNA elements on a chromosome can be thought of as analogous to the sequence of life history events over time, an observation that led to the adoption of sequence analysis by sociologists in the 1980s (Abbott and Tsay 2000). The method has gone through waves of popularity, criticism, and methodological development in sociology, where it has been used to characterize a range of sequenced data, including the onset

of social welfare policies by countries, historical patterns of lynching rates in the southern United States, and daily time-use data (Abbott and Tsay 2000).

The most popular application of sequence analysis in life course epidemiology has been for employment history data (Sabbath et al. 2015; Kobayashi and Feldman 2019). Employment is a complex state that people transition into and out of over time on multiple dimensions, such as on weekly hours worked, job type, job skill level, and various forms of non-employment. When combined with life transitions such as childbirth, caregiving, illness, or retirement, sequence analysis for employment history data can provide rich insights into prototypical population trajectories of employment and life histories that may have consequences for a range of health outcomes. Sabbath and colleagues (2015) used sequence analysis to characterize exposure to work–family demands in women from ages 16 to 50 in relation to subsequent mortality risk, observing the lowest mortality rates in married women who stayed home briefly with children before re-entering the paid workforce, intermediate mortality rates in single working mothers and married non-working mothers, and the highest mortality rates for single non-working mothers (Sabbath et al. 2015). Within each of these groups, sequence analysis identified prototypical periods and durations of time that were spent with and without children and in and out of work, allowing a comprehensive exposure assessment that would not be readily possible using manual methods of classification.

Sequence analysis is a two-step data reduction technique that consists of an optimal matching analysis followed by a cluster analysis (Aisenbrey and Fasang 2010). The goal of optimal matching is to minimize the 'cost' required to transform the sequences of any two given individuals to match one another. Cost assignments must be informed by theory and prior hypotheses about the relationships under study, and have been the subject of much debate in the sociological literature (Wu 2000; Aisenbrey and Fasang 2010; Halpin 2010). After distances are calculated, the second step of sequence analysis is a cluster analysis using the matrix of optimal matching distances to distil the extremely large number of potential sequences into a handful of clusters of similar sequences.

In a causal framework, sequence data have the potential to be poorly defined exposures because they consist of multiple heterogeneous elements that have been reduced into a single entity. If theory permits that a given life course exposure trajectory identified using sequence analysis is a single

entity that can be intervened upon, rather than a series of stochastically generated events, then the exposure may be well-defined. A further issue is that control for confounding may quickly become problematic, as the possibility arises of within-sequence confounder-mediators that cannot easily be controlled for. Although it has potential limitations, sequence analysis presents a promising and underutilized approach to simultaneously characterize the aspects of timing, dose, duration, and order of complex life history exposures in life course epidemiology.

Future directions in life course epidemiology

To date, the vast majority of life course research has been conducted in the United States and Western Europe. Examining life course determinants of health in low- and middle-income countries is critical for identifying risk and protective factors for health throughout the life course in countries with rapidly growing older adult populations (Tollman et al. 2016). Research in novel contexts has the potential to expand understanding of the life course determinants of health for people in countries around the world. Until recently, there have been very few data sources to support such research. Over the past decade, new International Partner Studies of the US Health and Retirement Study have been established in several low- and middle-income countries including China, Costa Rica, Mexico, Indonesia, India, and South Africa (https://g2aging.org/). While many of these studies are currently limited to baseline data or two waves of data, their harmonized design will allow for cross-national comparisons of the life course drivers of population health outcomes, especially as longitudinal follow-up continues. For example, studies such as the Longitudinal Study of Aging in India (LASI) and the Health and Aging in Africa: A Longitudinal Study of an INDEPTH Community in South Africa (HAALSI) (Lee et al. 2014; Gómez-Olivé et al. 2018) will provide evidence on ageing outcomes in older populations with limited formal education and low literacy, who grew up under different social, political, cultural, and economic conditions than those in the United States and Western Europe, where most life course evidence to date has been from. These data will be crucial for enhancing evidence on the life course drivers of population health from a global perspective, and for allowing the triangulation of evidence across diverse populations.

Birth cohort studies, such as the British birth cohort studies established in 1946, 1958, 1970, and 2000 (Elliott and Shepherd 2006; Power and Elliott 2006; Wadsworth et al. 2006; Connelly and Platt 2014) are valuable resources for life course epidemiology, but take many years to mature. An efficient and creative strategy for life course epidemiology is to develop a follow-up study of a cohort studied earlier in the life course. For example, the Project Talent Aging Study is a follow-up study of Project Talent, a national study that surveyed over 400,000 high school students in 1960 who were followed longitudinally through 1974 (https://www.projecttalent.org/). Project Talent Aging Study researchers are re-contacting Project Talent participants for a study focused on cognitive ageing. This study has the potential to help researchers understand early-life factors that influence risk of Alzheimer's disease and related dementias in later life.

In addition to expanding epidemiologic research with a life course perspective to include people with diverse life experiences, new areas of development include life course research extending beyond one human lifespan to include intergenerational and transgenerational life course research, as well as the application of innovative methods. Including people with diverse life experiences and the thoughtful application of innovative tools in epidemiologic studies with a life course perspective has the potential to expand understanding of life course determinants of health and ultimately guide social, public health, and medical policies and programmes to promote population health and health equity.

REFERENCES

Abbott, A. and Tsay, A. (2000). Sequence analysis and optimal matching methods in sociology. *Sociological Methods & Research*, 29(1), 3–33.

Aisenbrey, S. and Fasang, A.E. (2010). New life for old ideas: the 'second wave' of sequence analysis bringing the 'course' back into the life course. *Sociological Methods & Research*, 38(3), 460–2.

Alzheimer's Association (2019). 2019 Alzheimer's disease facts and figures. *Alzheimers & Dementia*, 15(3), 321–87.

Arnold, B.F., Ercumen, A., Benjamin-Chung, J., and Colford Jr, J.M. (2016). Brief report: negative controls to detect selection bias and measurement bias in epidemiologic studies. *Epidemiology*, 27(5), 637–41.

Baer, H.J., et al. (2005). Body fatness during childhood and adolescence and incidence of breast cancer in premenopausal women: a prospective cohort study. *Breast Cancer Research*, 7(3), R314–25.

Baer, H.J., Tworoger, S.S., Hankinson, S.E., and Willett, W.C. (2010). Body fatness at young ages and risk of breast cancer throughout life. *American Journal of Epidemiology*, 171(11), 1183–94.

Bardia, A., et al. (2008). Relative weight at age 12 and risk of postmenopausal breast cancer. *Cancer Epidemiology, Biomarkers & Prevention*, 17(2), 374–8.
Barker, D.J. (1995). Fetal origins of coronary heart disease. *British Medical Journal*, 311(6998), 171–4.
Barker, D.J. (2012). Developmental origins of chronic disease. *Public Health*, 126(3), 185–9.
Ben-Shlomo, Y. and Kuh, D. (2002). A life course approach to chronic disease epidemiology: conceptual models, empirical challenges and interdisciplinary perspectives. *International Journal of Epidemiology*, 31(2), 285–93.
Ben-Shlomo, Y., Cooper, R., and Kuh, D. (2016). The last two decades of life course epidemiology, and its relevance for research on ageing. *International Journal of Epidemiology*, 45(4), 973–88.
Berger, S.L., Kouzarides, T., Shiekhattar, R., and Shilatifard, A. (2009). An operational definition of epigenetics. *Genes & Development*, 23(7), 781–83.
Bergman, D., Halje, M., Nordin, M., and Engström, W. (2013). Insulin-like growth factor 2 in development and disease: a mini-review. *Gerontology*, 59(3), 240–9.
Black, M.M., et al. (2017). Early childhood development coming of age: science through the life course. *The Lancet*, 389(10064), 77–90.
Bråbäck, L., Lodge, C.J., Lowe, A.J., Dharmage, S.C., Olsson, D., and Forsberg, B. (2018). Childhood asthma and smoking exposures before conception—a three-generational cohort study. *Pediatric Allergy and Immunology*, 29(4), 361–8.
Brookmeyer, R., Johnson, E., Ziegler-Graham, K., and Arrighi, H.M. (2007). Forecasting the global burden of Alzheimer's disease. *Alzheimer's & Dementia*, 3(3), 186–91.
Burton, A., Altman, D.G., Royston, P., and Holder, R.L. (2006). The design of simulation studies in medical statistics. *Statistics in Medicine*, 25(24), 4279–92.
Casey, J.A., et al. (2018). Retirements of coal and oil power plants in California: association with reduced preterm birth among populations nearby. *American Journal of Epidemiology*, 187(8), 1586–94.
Cheslack-Postava, K., and Jordan-Young, R.M. (2012). Autism spectrum disorders: toward a gendered embodiment model. *Social Science & Medicine*, 74(11), 1667–774.
Chetty, R., Hendren, N., and Katz, L.F. (2016). The effects of exposure to better neighborhoods on children: new evidence from the moving to opportunity experiment. *American Economic Review*, 106(4), 855–902.
Cohen, J.M., Wood, M.E., Hernández-Díaz, S., Ystrom, E., and Nordeng, H. (2019). Paternal antidepressant use as a negative control for maternal use: assessing familial confounding on gestational length and anxiety traits in offspring. *International Journal of Epidemiology*, 48(5), 1665–72.
Connelly, R. and Platt, L. (2014). Cohort profile: UK Millennium Cohort Study (MCS). *International Journal of Epidemiology*, 43(6), 1719–25.
Debette, S., et al. (2011). Midlife vascular risk factor exposure accelerates structural brain aging and cognitive decline. *Neurology*, 77(5), 461–8.
Díaz-Gutiérrez, M.J., et al. (2017). Relationship between the use of benzodiazepines and falls in older adults: a systematic review. *Maturitas*, 101, 17–22.
Driver, J.A., et al. (2012). Inverse association between cancer and Alzheimer's disease: results from the Framingham Heart Study. *British Medical Journal*, 344, e1442.
Driver, J.A. (2014). Inverse association between cancer and neurodegenerative disease: review of the epidemiologic and biological evidence. *Biogerontology*, 15(6), 547–57.
Elliott, J. and Shepherd, P. (2006). Cohort profile: 1970 British Birth Cohort (BCS70). *International Journal of Epidemiology*, 35(4), 836–43.
Fernihough, A. and McGovern, M.E. (2015). Physical stature decline and the health status of the elderly population in England. *Economics & Human Biology*, 16, 30–44.
Flanders, W.D., et al. (2011). A method for detection of residual confounding in time-series and other observational studies. *Epidemiology*, 22(1), 59–67.

Freedman, D.M., et al. (2016). Associations between cancer and Alzheimer's disease in a U.S. Medicare population. *Cancer Medicine*, 5(10), 2965–76.

Garovic, V.D., et al. (2010). Hypertension in pregnancy as a risk factor for cardiovascular disease later in life. *Journal of Hypertension*, 28(4), 826–33.

Giles, W.H., Kittner, S.J., Hebel, J.R., Losonczy, K.G., and Sherwin, R.W. (1995). Determinants of black-white differences in the risk of cerebral infarction. The National Health and Nutrition Examination Survey Epidemiologic Follow-up Study. *Archives of Internal Medicine*, 155(12), 1319–24.

Glymour, M.M. (2017). Using causal diagrams to understand common problems in social epidemiology. In: Oakes, M.J. and Kaufman, J.S. (eds.) *Methods in Social Epidemiology* (2nd ed.). San Francisco, CA: Wiley, pp. 458–92.

Glymour, M.M., Walter, S., and Tchetgen, E.J. (2017). Natural experiments and instrumental variable analyses in social epidemiology. In: Oakes, M.J. and Kaufman, J.S. (eds.) *Methods in Social Epidemiology* (2nd ed.). San Francisco, CA: Wiley, pp. 493–538.

Gómez-Olivé, F.X., et al. (2018). Cohort profile: health and ageing in Africa: a longitudinal study of an INDEPTH community in South Africa (HAALSI). *International Journal of Epidemiology*, 47(3), 689–90.

Gravlee, C.C. (2009). How race becomes biology: embodiment of social inequality. *American Journal of Physical Anthropology*, 139(1), 47–57.

Halpin, B. (2010). Optimal matching analysis and life-course data: the importance of duration. *Sociological Methods & Research*, 38(3), 365–88.

Hamer, M., Bauman, A., Bell, J.A., and Stamatakis, E. (2019). Examining associations between physical activity and cardiovascular disease mortality using negative control outcomes. *International Journal of Epidemiology*, 48(4), 1161–6.

Hayward, M.D. and Gorman, B.K. (2004). The long arm of childhood: the influence of early-life social conditions on men's mortality. *Demography*, 41(1), 87–107.

Heard, E. and Martienssen, R.A. (2014). Transgenerational epigenetic inheritance: myths and mechanisms. *Cell*, 157(1), 95–109.

Hebert, L.E., Weuve, J., Scherr, P.A., and Evans, D.A. (2013). Alzheimer disease in the United States (2010–2050) estimated using the 2010 census. *Neurology*, 80(19), 1778–83.

Heijmans, B.T., et al. (2008). Persistent epigenetic differences associated with prenatal exposure to famine in humans. *Proceedings of the National Academy of Sciences of the United States of America*, 105(44), 17046–9.

Hernán, M.A., Hernández-Díaz, S., and Robins, J.M. (2004). A structural approach to selection bias. *Epidemiology*, 15(5), 615–25.

Hinkle, S.N., Rawal, S., Zhu, Y., Grewal, J., Albert, P.S., and Zhang, C. (2017). Validation of self-reported diagnosis of gestational diabetes at 6-weeks postpartum. *Epidemiology*, 28(5), 747–52.

Howard, V.J., et al. (2011). Disparities in stroke incidence contributing to disparities in stroke mortality. *Annals of Neurology*, 69(4), 619–27.

Horvath, S., et al. (2014). Obesity accelerates epigenetic aging of human liver. *Proceedings of the National Academy of Sciences*, 111(43), 15538–43.

Horvath, S. and Raj, K. (2018). DNA methylation-based biomarkers and the epigenetic clock theory of ageing. *Nature Reviews Genetics*, 19(6), 371–84.

Hu, H., Shih, R., Rothenberg, S., and Schwartz, B.S. (2007). The epidemiology of lead toxicity in adults: measuring dose and consideration of other methodologic issues. *Environmental Health Perspectives*, 115(3), 455–62.

Ivers, L.C., et al. (2015). Effectiveness of reactive oral cholera vaccine in rural Haiti: a case-control study and bias-indicator analysis. *Lancet Global Health*, 3(3), e162–8.

Jaffe, A.E. and Irizarry, R.A. (2014). Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biology*, 15, R31.

Kaufman, J.S. and Hernán, M.A. (2012). Commentary: epidemiologic methods are useless: they can only give you answers. *Epidemiology*, 23(6), 785–6.

Kecklund, G. and Axelsson, J. (2016). Health consequences of shift work and insufficient sleep. *British Medical Journal*, 355, i5210.

Kissela, B., et al. (2004). Stroke in a biracial population: the excess burden of stroke among blacks. *Stroke*, 35(2), 426–31.

Kivipelto, M., et al. (2001). Midlife vascular risk factors and Alzheimer's disease in later life: longitudinal, population based study. *British Medical Journal*, 322(7300), 1447–51.

Kleindorfer, D., et al. (2006). The unchanging incidence and case-fatality of stroke in the 1990s: a population-based study. *Stroke*, 37, 2473–8.

Knopman, D.S., Mosley, T.H., Catellier, D.J., and Coker, L.H. (2009). Fourteen-year longitudinal study of vascular risk factors, APOE genotype, and cognition: the ARIC MRI Study. *Alzheimer's & Dementia*, 5(3), 207–14.

Knopov, A., Sherman, R.J., Raifman, J.R., Larson, E., and Siegel, M.B. (2019). Household gun ownership and youth suicide rates at the state level, 2005–2015. *American Journal of Preventive Medicine*, 56(3), 335–42.

Kobayashi, L.C., Berkman, L.F., Wagner, R.G., Kahn, K., Tollman, S.M., and Subramanian, S.V. (2019). Education modifies the relationship between height and cognitive function in a cross-sectional, population-based study of older adults in rural South Africa. *European Journal of Epidemiology*, 34(2), 131–9.

Kobayashi, L.C., and Feldman, J.M. (2019). Employment trajectories in midlife and cognitive performance in later life: longitudinal study of older American men and women. *Journal of Epidemiology and Community Health*, 73(3), 232–8.

Krieger, N. (2005). Embodiment: a conceptual glossary for epidemiology. *Journal of Epidemiology & Community Health*, 59(5), 350–5.

Krieger, N. (2011). Ecosocial Theory of Disease Distribution: embodying societal and ecologic context. In: Krieger, N. (ed.) *Epidemiology and the People's Health: Theory and Context*. Oxford: Oxford University Press, pp. 202–35.

Kuh, D., Ben-Shlomo, Y., Lynch, J., Hallqvist, J., and Power, C. (2003). Life course epidemiology. *Journal of Epidemiology & Community Health*, 57(10), 778–83.

Lash, T.L. and Olshan, A.F. (2016). Epidemiology announces the 'validation study' submission category. *Epidemiology*, 27(5), 613–14.

Lawlor, D.A., Tilling, K., and Davey Smith, G. (2016). Triangulation in aetiological epidemiology. *International Journal of Epidemiology*, 45–6), 1866–86.

Lee, C.M.Y., et al. (2009). Adult height and the risks of cardiovascular disease and major causes of death in the Asia-Pacific region: 21,000 deaths in 510,000 men and women. *International Journal of Epidemiology*, 38(4), 1060–71.

Lee, J., Shih, R., Feeney, K., and Langa, K.M. (2014). Gender disparity in late-life cognitive functioning in India: findings from the longitudinal aging study in India. *Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, 69(4), 603–11.

Lesko, C.R., Buchanan, A.L., Westreich, D., Edwards, J.K., Hudgens, M.G., and Cole, S.R. (2017). Generalizing study results: a potential outcomes perspective. *Epidemiology*, 28(4), 553–61.

Leventhal, T., and Brooks-Gunn, J. (2003). Moving to opportunity: an experimental study of neighborhood effects on mental health. *American Journal of Public Health*, 93(9), 1576–82.

Liew, Z., Olsen, J., Cui, X., Ritz, B., and Arah, O.A. (2015). Bias from conditioning on live birth in pregnancy cohorts: an illustration based on neurodevelopment in children after prenatal exposure to organic pollutants. *International Journal of Epidemiology*, 44(1), 345–54.

Liew, Z., Kioumourtzoglou, M.A., Roberts, A.L., O'Reilly, E.J., Ascherio, A., and Weisskopf, M.G. (2019). Use of negative control exposure analysis to evaluate confounding: an example of acetaminophen exposure and attention-deficit/hyperactivity disorder in Nurses' Health Study II. *American Journal of Epidemiology*, 188(4), 768–75.

Lipsitch, M., Tchetgen, E., and Cohen, T. (2010). Negative controls: a tool for detecting confounding and bias in observational studies. *Epidemiology*, 21(3), 383–8.

Lodge, C.J., Bråbäck, L., Lowe, A.J., Dharmage, S.C., Olsson, D., and Forsberg, B. (2018). Grandmaternal smoking increases asthma risk in grandchildren: a nationwide Swedish cohort. *Clinical & Experimental Allergy*, 48(2), 167–74.

Luke, B., Brown, M.B., Liu, C.L., Diop, H., and Stern, J.E. (2018). Validation of severe maternal morbidity on the US certificate of live birth. *Epidemiology*, 29(4), e31–2.

Magnusson, P.K., Gunnell, D., Tynelius, P., Davey Smith, G., and Rasmussen, F. (2005). Strong inverse association between height and suicide in a large cohort of Swedish men: evidence of early life origins of suicidal behavior? *American Journal of Psychiatry*, 162(7), 1373–5.

Marden, J.R., Tchetgen, E.J., Kawachi, I., and Glymour, M.M. (2017). Contribution of socioeconomic status at 3 life-course periods to late-life memory function and decline: early and late predictors of dementia risk. *American Journal of Epidemiology*, 186(7), 805–14.

Marmot, M., Wilkinson, R.G. (2001). Psychosocial and material pathways in the relation between income and health: a response to Lynch et al. *British Medical Journal*, 322(7296), 1233–6.

Matthay, E.C., Galin, G., Rudolph, K.E., Farkas, K., Wintemute, G.J., and Ahern, J. (2017). In-state and interstate associations between gun shows and firearm deaths and injuries: a quasi-experimental study. *Annals of Internal Medicine*, 167(12), 837–44.

Mayeda, E.R., et al. (2018a). Can survival bias explain the age attenuation of racial inequalities in stroke incidence? A simulation study. *Epidemiology*, 29(4), 525–32.

Mayeda, E.R., Filshtein, T.J., Tripodis, Y., Glymour, M.M., and Gross, A.L. (2018b). Does selective survival before study enrolment attenuate estimated effects of education on rate of cognitive decline in older adults? A simulation approach for quantifying survival bias in life course epidemiology. *International Journal of Epidemiology*, 47(5), 1507–17.

Michels, K.B., et al. (2013). Recommendations for the design and analysis of epigenome-wide association studies. *Nature Methods*, 10(10), 949–55.

McCulloch, C.E. (2015). Editorial: observational studies, time-dependent confounding, and marginal structural models. *Arthritis Rheumatology*, 67(3), 609–11.

Morris, T.P., White, I.R., and Crowther, M.J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, 38(11), 2074–102.

Müezzinler, A., Zaineddin, A.K., and Brenner, H. (2013). A systematic review of leukocyte telomere length and age in adults. *Ageing Research Reviews*, 12(2), 509–19.

Nelson, C.A., Zeanah, C.H., Fox, N.A., Marshall, P.J., Smyke, A.T., and Guthrie, D. (2007). Cognitive recovery in socially deprived young children: the Bucharest Early Intervention Project. *Science*, 318(5858), 1937–40.

Niesner, K., et al. (2013). Validation of VA administrative data algorithms for identifying cardiovascular disease hospitalization. *Epidemiology*, 24(2), 334–5.

Nüesch, E., et al. (2016). Adult height, coronary heart disease and stroke: a multi-locus Mendelian randomization meta-analysis. *International Journal of Epidemiology*, 45(6), 1927–37.

Nwanaji-Enwerem, J.C., et al. (2017). Associations between long-term exposure to PM2. 5 component species and blood DNA methylation age in the elderly: the VA normative aging study. *Environment International*, 102, 57–65.

Osypuk, T.L., Alonso, A., and Bates, L.M. (2015a). Understanding the healthy immigrant effect and cardiovascular disease. *Circulation*, 132(16), 1522–4.

Osypuk, T.L., Kehm, R., and Misra, D.P. (2015b). Where we used to live: validating retrospective measures of childhood neighborhood context for life course epidemiologic studies. *PLoS One*, 10(4), e0124635.

Painter, R.C., Roseboom, T.J., and Bleker, O.P. (2005). Prenatal exposure to the Dutch famine and disease in later life: an overview. *Reproductive Toxicology*, 20(3), 345–52.

Perez, M.F. and Lehner, B. (2019). Intergenerational and transgenerational epigenetic inheritance in animals. *Nature Cell Biology*, 21(2), 143–51.

Perkins, J.M., Subramanian, S.V., Davey Smith, G., and Özaltin, E. (2016). Adult height, nutrition, and population health. *Nutrition Reviews*, 74(3), 149–65.

Power, C. and Elliott, J. (2006). Cohort profile: 1958 British birth cohort (National Child Development Study). *International Journal of Epidemiology*, 35(1), 34–41.

Power, M.C., Tchetgen, E.J., Sparrow, D., Schwartz, J., and Weisskopf, M.G. (2013). Blood pressure and cognition: factors that may account for their inconsistent association. *Epidemiology*, 24(6), 886–93.

Power, M.C., et al. (2014). Lead exposure and rate of change in cognitive function in older women. *Environmental Research*, 129, 69–75.

Quach, A., et al. (2017). Epigenetic clock analysis of diet, exercise, education, and lifestyle factors. *Aging (Albany NY)*, 9(2), 419–37.

Ravelli, A.C., van Der Meulen, J.H., Osmond, C., Barker, D.J., and Bleker, O.P. (1999). Obesity at the age of 50 y in men and women exposed to famine prenatally. *American Journal of Clinical Nutrition*, 70(5), 811–16.

Reeves, G.K., Pirie, K., Beral, V., Green, J., Spencer, E., and Bull, D. (2007). Cancer incidence and mortality in relation to body mass index in the Million Women Study: cohort study. *British Medical Journal*, 335(7630), 1134.

Roadmap Epigenomics Consortium, et al. (2015). Integrative analysis of 111 reference human epigenomes. *Nature*, 518, 317–30.

Robins, J.M., Hernán, M.A., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5), 550–60.

Roshania, R., Narayan, K.M., and Oza-Frank, R. (2008). Age at arrival and risk of obesity among US immigrants. *Obesity*, 16(12), 2669–75.

Sabbath, E.L., Guevara, I.M., Glymour, M.M., and Berkman, L.F. (2015). Use of life course work–family profiles to predict mortality risk among US women. *American Journal of Public Health*, 105(4), e96–e102.

Sanderson, E., Macdonald-Wallis, C., and Davey Smith, G. (2018). Negative control exposure studies in the presence of measurement error: implications for attempted effect estimate calibration. *International Journal of Epidemiology*, 47(2), 587–96.

Sattar, N. and Greer, I.A. (2002). Pregnancy complications and maternal cardiovascular risk: opportunities for intervention and screening? *British Medical Journal*, 325, 157–60.

Schmidt, S.A.J., Ording, A.G., Horváth-Puhó, E., Sørensen, H.T., and Henderson, V.W. (2017). Non-melanoma skin cancer and risk of Alzheimer's disease and all-cause dementia. *PLoS One*, 12(2), e0171527.

Schnaider Beeri, M., Davidson, M., Silverman, J.M., Noy, S., Schmeidler, J., and Goldbourt, U. (2005). Relationship between body height and dementia. *American Journal of Geriatric Psychiatry*, 13(2), 116–23.

Sheridan, P., Ilango, S., Bruckner, T.A., Wang, Q., Basu, R., and Benmarhnia, T. (2019). Ambient fine particulate matter and preterm birth in california: identification of critical exposure windows. *American Journal of Epidemiology*, *188*(9), 1608–15.

Siegel, M. and Rothman, E.F. (2016). Firearm ownership and suicide rates among US men and women, 1981–2013. *American Journal of Public Health*, 106(7), 1316–22.

Singh-Manoux, A., et al. (2017). Trajectories of depressive symptoms before diagnosis of dementia: a 28-year follow-up study. *JAMA Psychiatry*, 74(7), 712–18.

Singh-Manoux, A., et al. (2018). Obesity trajectories and risk of dementia: 28 years of follow-up in the Whitehall II Study. *Alzheimer's & Dementia*, 14(2), 178–86.

St Clair, D., et al. (2005). Rates of adult schizophrenia following prenatal exposure to the Chinese famine of 1959–1961. *The Journal of the American Medical Association*, 294(5), 557–62.

Stuart, E. A., Ackerman, B., and Westreich, D. (2018). Generalizability of randomized trial results to target populations. *Research on Social Work Practice*, 28(5), 532–7.

Suissa, S. and Ernst, P. (2005). Bias in observational study of the effectiveness of corticosteroids in asthma. *Journal of Allergy and Clinical Immunology*, 115(4), 714–19.
Susser, E.S. and Lin, S.P. (1992). Schizophrenia after prenatal exposure to the Dutch Hunger Winter of 1944–1945. *Archives of General Psychiatry*, 49(12), 983–8.
Susser, E., et al. (1996). Schizophrenia after prenatal famine: further evidence. *Archives of General Psychiatry*, 53(1), 25–31.
Tobi, E.W., et al. (2009). DNA methylation differences after exposure to prenatal famine are common and timing- and sex-specific. *Human Molecular Genetics*, 18(21), 4046–53.
Tollman, S.M., Norris, S.A., and Berkman, L.F. (2016). Commentary: the value of life course epidemiology in low- and middle-income countries: an ageing perspective. *International Journal of Epidemiology*, 45(4), 997–9.
Tooher, J., Thornton, C., Makris, A., Ogle, R., Korda, A., and Hennessy, A. (2017). All hypertensive disorders of pregnancy increase the risk of future cardiovascular disease. *Hypertension*, 70(4), 798–803.
Van den Brandt, P.A., et al. (2000). Pooled analysis of prospective cohort studies on height, weight, and breast cancer risk. *American Journal of Epidemiology*, 152(6), 514–27.
van Otterdijk, S.D. and Michels, K.B. (2016). Transgenerational epigenetic inheritance in mammals: how good is the evidence? *The FASEB Journal: official publication of the Federation of American Societies for Experimental Biology*, 30, 2457–65.
Vandenbroucke, J.P., Broadbent, A., and Pearce, N. (2016). Causality and causal inference in epidemiology: the need for a pluralistic approach. *International Journal of Epidemiology*, 45(6), 1776–86.
VanderWeele, T.J. (2015). *Explanation in Causal Inference: Methods for Mediation and Interaction*. Oxford: Oxford University Press.
Wadsworth, M., Kuh, D., Richards, M., and Hardy, R. (2006). Cohort profile: the 1946 National Birth Cohort (MRC National Survey of Health and Development). *International Journal of Epidemiology*, 35(1), 49–54.
Walker, V.R., et al. (2018). Human and animal evidence of potential transgenerational inheritance of health effects: an evidence map and state-of-the-science evaluation. *Environment International*, 115, 48–69.
Walters, K.L., Mohammed, S.A., Evans-Campbell, T., Beltran, R.E., Chae, D.H., and Duran, B. (2011). Bodies don't just tell stories, they tell histories: embodiment of historical trauma among American Indians and Alaska Natives. *Du Bois Review: Social Science Research on Race*, 8(1), 179–89.
Westreich, D., Edwards, J.K., Lesko, C.R., Stuart, E., and Cole, S.R. (2017). Transportability of trial results using inverse odds of sampling weights. *American Journal of Epidemiology*, 186(8), 1010–14.
Whitmer, R., Sidney, S., Selby, J., Johnston, S.C., and Yaffe, K. (2005). Midlife cardiovascular risk factors and risk of dementia in late life. *Neurology*, 64(2), 277–81.
Wilker, E., et al. (2011). Longitudinal changes in bone lead levels: the VA Normative Aging Study. *Journal of Occupational and Environmental Medicine*, 53(8), 850.
Wu, L.L. (2000). Some comments on 'sequence analysis and optimal matching methods in sociology: review and prospect'. *Sociological Methods & Research*, 29(1), 41–64.
Yarchoan, M., et al. (2017). Association of cancer history with Alzheimer's disease dementia and neuropathology. *Journal of Alzheimer's Disease*, 56(2), 699–706.
Ying, W., Catov, J.M., and Ouyang, P. (2018). Hypertensive disorders of pregnancy and future maternal cardiovascular risk. *Journal of the American Heart Association*, 7(17), e009382.
Zaadstra, B.M., Chorus, A.M., van Buuren, S., Kalsbeek, H., and van Noort, J.M. (2008). Selective association of multiple sclerosis with infectious mononucleosis. *Multiple Sclerosis*, 14(3), 307–13.

Zeanah, C.H., et al. (2003). Designing research to study the effects of institutionalization on brain and behavioral development: the Bucharest Early Intervention Project. *Development and Psychopathology*, 15(4), 885–907.

SECTION 6

Social science techniques

6.1 Sociology and psychology in public health
Stella R. Quah

6.2 Sexuality and public health
Richard Parker, Jonathan Garcia, Miguel Muñoz-Laboy, Marni Sommer, and Patrick Wilson

6.3 Demography and public health
Emily Grundy and Michael Murphy

6.4 Health promotion, health education, and the public's health
Fran Baum

6.5 Development and evaluation of complex multicomponent interventions in public health
Rona Campbell and Chris Bonell

6.6 Economic appraisal in public healthcare: assessing efficiency and equity
David Parkin, Stephen Morris, and Nancy Devlin

6.7 Behavioural economics and health
Alison Buttenheim and Harsha Thirumurthy

6.8 Governance and management of public health programmes
Zhiyuan Hou and Na He

6.9 Implementation science and translational public health
Wafaa M. El-Sadr, Judith Wasserheit, Bryan Wiener, Andrea Howard, Catherine Hankins, Patricia J. Culligan, and Katherine Harripersaud

6.1

Sociology and psychology in public health

Stella R. Quah

Introduction to sociology and psychology in public health

The disciplines of sociology and psychology differ in interesting ways but they also have common links both conceptually and methodologically. They differ in overall scope: psychology typically focuses on the individual while sociology examines the individual's social action (agency) and the community's sociocultural, economic, political, and physical context (structure) within which they live, interact, work, and play. One important similarity between sociology and psychology is that both disciplines have a dedicated subfield for the study of health and illness. These subfields are, respectively, health sociology (also known as medical sociology) and health psychology. Thus, in line with the objective of this volume, the focus of this chapter is on the contributions to public health by health sociology and health psychology, the two subfields most relevant to public health.

Just as health sociology and health psychology have common links, today both disciplines share two essential notions with public health. The first shared notion is the global impact of individuals' actions. This impact is demonstrated in many types of health-related behaviour. One of the most evident examples is the case of infectious diseases. An infectious disease may begin with one individual but, if left unchecked, it can affect entire continents. This principle was highlighted by a British physician, Southwood Smith, in the nineteenth century. Smith (1866) was alerting the health authorities on the need to investigate and control infectious diseases. By the twenty-first century, sociology and psychology have accumulated a wealth of research on the impact of individual and collective social behaviour on health to show that Southwood Smith's notion applies also to

people's behaviour and social conditions associated with non-infectious diseases and, indeed, to all health-related phenomena.

The second shared notion is the importance of trust in interpersonal relations and in the nurturing of the community' cooperation for collective action. This notion is actively scrutinized in sociology and psychology (e.g. Stevenson et al. 2014; Poulin and Haase 2015; Pletzer et al. 2018; Do Nascimiento-Bezerra and Souza-Feitosa 2018; Yost-Dubrow and Dunham 2018). Health sociology examines trust critically in interpersonal and community settings. An example at the interpersonal settings is the physician's imperative of earning the patient's trust in clinical encounters. The community setting is illustrated by the public health authorities' duty to gain the population's trust as the basic step towards the active involvement and collaboration of communities in health promotion and disease prevention and treatment, including the prevention and control of epidemics (Quah 2007a, 2007b). Research findings supporting the importance of building trust are gradually permeating public health (e.g. Adams et al. 2017; Benjaminy et al. 2018; Callis et al. 2018; Datta 2018; Ellenberg et al. 2018; Kim 2018; Nuriddin et al. 2018; Raven et al. 2018).

These two notions—the global impact of individuals' actions and the importance of trust—exemplify the current collaboration of public health, health sociology, and health psychology, as I elaborate in the pages that follow. The discussion is presented in four sections: Sociology and public health; psychology and public health; the contribution of both disciplines to public health training; and conclusion. The sections on sociology and psychology cover the same three themes: the highlights of their respective theoretical contributions; the main methodological approaches that are relevant to public health; and a succinct summary of areas of research that illustrate the contributions of health sociology and health psychology to the field of public health. The list of references documents the arguments made in the text and serves as further reading on the historical as well as most current information on what health sociology and health psychology offer to public health researchers and practitioners.

Sociology in public health

The sociological analysis of health and illness can be traced back to the beginnings of sociology as an area of systematic knowledge in the late 1800s. Sociology's pioneers in Europe investigated the influence of cultural, socioeconomic, and political conditions in the community on individuals' social action. Two of the pioneers were Emile Durkheim (1858–1917) in France and Max Weber (1864–1920) in Germany. In the midst of the development of sociological theorizing, while sociology pioneers were examining the nature and dynamics of society and social action, the spread of infectious diseases was exerting a heavy toll in Europe. The damage caused by epidemics alerted some British and European medical authorities to the possible connection of three phenomena hitherto seen as unrelated by medical experts: social conditions in the community, social behaviour, and public health. Among the first physicians to presume that connection existed were Virchow and Billings. A notable German pathologist, Rudolf Virchow (1812–1902), seen as 'a dominant figure in German medicine' (Shryock 1974, p. 202) was perhaps the first medical scientist who considered medicine as a social science (Elling and Sokołoska 1978, p. 14). John Shaw Billings, an American surgeon and the first publisher of the *Index Medicus*, wrote on the link between 'hygiene and sociology' around 1879 (Shyrock 1974, pp. 182–3; Cockerham 2010, p. 2). In a 1894 article on the social factors of illness another physician, Charles McIntyre, used for the first time the term 'medical sociology' (Cockerham 2010, p. 2). Along the same line of thought, public health practitioners recognized very early the importance of sociology and psychology to public health research and practice (Roney 1960).

The initial phase of health sociology was rooted in research by sociologists working in medical academic settings and in public health or biomedical clinical settings (Elling and Sokołoska 1978). Health sociology research and publications paved the way to the inclusion of health sociology courses in the medical curriculum at American and British universities after the Second World War (Elling and Sokołoska 1978; Cockerham 2010, pp. 2–3), as I explain later. Some of the signposts of the formalization of health sociology as a subdiscipline of sociology are the establishment of the Medical Sociology Section of the American Sociological Association in 1959; the Research Committee on Medical Sociology of the International Sociological Association in 1963[1]; and the Medical Sociology Research Network in the British Sociological Association in 1969 (Reid 1976). From

its early years, health sociology has focused on all aspects of health and illness adeptly classified into two general dimensions by Robert Strauss: 'sociology *of* medicine' comprising the analysis of 'the organizational structure, role relationships, value systems, rituals, and functions of medicine as a system of behavior'; and 'sociology *in* medicine' dedicated to the analysis of the disease process, health-related behaviour, and the sociocultural, political and socioeconomic determinants of health and illness in the community (Strauss 1957, p. 203).

Research publications in health sociology include journal articles and books. Among the most important domain-specific international refereed journals are the *Journal of Health and Social Behaviour* (JHSB), the dedicated health sociology journal of the American Sociological Association; *Social Science & Medicine* (SSM); *Sociology of Health and Illness*; and *Health Sociology Review*, the journal of the Health Sociology Section of the Australian Sociological Association. There are several textbooks in health sociology, one of the most established is Cockerham's (2017), already in its 14th edition. Among serial books is the *New Blackwell Companion to Medical Sociology* (with editions published in 2001, 2005, and 2010) and the *Handbook of Medical Sociology* (HMS). The first edition of the HMS was published in 1963 and it discussed 'the *potential* of the sociological enterprise for understanding the etiology, diagnosis, and treatment of diseases; the ways health care is provided and funded; the societal and communal commitment to the support of health activities; and, indeed, the very definitions of health and illness'. Looking back at the first three editions of 1963, 1972, and 1979, the editors of the fourth edition wrote that, 'the promise of medical sociology [is now] replaced by hard evidence as to its utility' (Freeman and Levine 1989, p. 1). These authors correctly pointed out that health sociology uses the full range of sociological theories and methodologies (1989, p. 7).

In concert with the increase in research and publications by the 1980s, medical sociologists were also 'actively involved in public health programs, and … accredited schools of public health in the United States were required to provide formal training in the social and behavioral sciences' (Coreil and Dyer 2017, p. 600) and there were calls to increase health sociology input in the training of public health practitioners not only in the United States (Gordon and McFarlane 1996) but also in Canada (Massé and Moloughney 2011) and elsewhere. I return to this issue later.

Sociological theories and public health

Historically, the twin theoretical roots of sociology are philosophy and science as they developed in the seventeenth and eighteenth centuries in Europe with the writings of thinkers such as Charles Montesquieu (1689–1755) and Jean Jacques Rousseau (1712–1778) on social forces, power, and social facts; Emmanuel Kant (1724–1804) on the systematic analysis of cause and effect; and Henri de Saint-Simon (1760–1825) about industry, the need for social reforms, and the scientific study of society and social life. Philosophers such as René Descartes (1596–1650), Thomas Hobbes (1588–1679), and John Locke (1632–1704) aimed at 'grand, general, and very abstract systems of ideas that made rational sense', but other thinkers were more concerned with testing those ideas against their 'real world' experiences or more specifically, 'to combine empirical research with reason' following the path of science (Ritzer 1992, pp. 6–11). The latter trend 'led to the application of the scientific method to social issues' seeking to identify possible laws governing social behaviour just as Newtonian physics demonstrated that physical laws rule the physical world (Ritzer 1992, pp. 10–11). This embryonic tension and subsequent symbiosis between the two approaches—a purely theoretical outlook vis-à-vis the imperative to test assumptions empirically—remain at the core of sociology as a field of knowledge today.

Sociology grew from those early ideas into a distinct discipline in the context of wide-ranging social change. Two founding fathers of the discipline were born at the time of the Industrial Revolution and the rise of capitalism in the nineteenth century. The first was the French thinker August Comte (1798–1857) who contributed his treatise on a new discipline that he labelled 'Sociology' and envisaged it as the systematic analysis of society following the model of physics. The second pioneer thinker was Comte's compatriot Emile Durkheim (1858–1917), who promoted the symbiosis of theory and empirical research. Durkheim paid particular attention to, on the one hand, the study of disorder, conflict, and its consequences in society; and on the other hand, to the importance of social facts for the study and planning of social reform to reduce social disorder. Durkheim 'legitimized sociology in France' (Ritzer 1992 p. 17). Durkheim's treatise on the systematic analysis of social facts and his study of social factors associated with suicide are seminal works in sociology.

Two of Durkheim's most influential books are *Suicide* (1897/1951) demonstrating the link between individual behaviour and society's socioeconomic conditions; and *The Rules of Sociological Method* (1895/1964). Durkheim's emphasis on both sociological theory and empirical research paved the way for today's evidence-based research. His conceptual contribution on the notions of 'the socially *normal* and *abnormal*' set the path for 'twentieth century medical sociology' along 'the enlightenment inspired *normal/pathological* bifurcation in medicine' (Thomas 2012, p. 216).

Two other pioneers in sociology were German: Georg Simmel (1858–1918) and Max Weber (1864–1920). Simmel and Weber were cofounders of the German Sociological Society. Simmel's key contribution is his notion of reciprocity and uncertainty in social interaction (Gerhardt 1989, p. 17). Max Weber's theory of social action, authority systems, and bureaucratization, are among many other seminal contributions in his large body of work, and form part of the foundation of contemporary sociology (Ritzer 1992, pp. 25–9). Among Weber's contributions is his ground-breaking theory of social action and social systems (1921/1968), a classical theory in sociology. In health sociology, Weber's conceptual contributions are applied to a wide range of studies including, for example, the analysis of the bureaucratization of healthcare systems and organizations (Fox 1989; Germov 2005), the analysis of power and authority of the medical profession (Starr 1982; Freidson 1988); and the impact of life chances and lifestyle on health-related behaviour (Cockerham 2010).

The body of theory in Sociology has grown exponentially over the past century. Sociological theories may be classified into two types: general theories that address all human behaviour and have 'universal' applicability; and 'middle-range' theories that have limited applicability. The most relevant general theories for the sociological analysis of health and illness are: structural-functional theory, symbolic interactionism, conflict theory, and phenomenology/ethnomethodology. Given the space constraints of this chapter, only the general theories are discussed.

The structural-functional theory (now neo-functionalism) is the contribution of American sociologist Talcott Parsons (1937). This theory addresses societies as social systems and proposes that all components of society whether they are individuals, groups, communities, nations, or global organizations, have functions—both manifest and latent—for their

maintenance and growth. The dominant trend in social organizations, groups, and social interactions is the maintenance of balance or equilibrium. Consequently, conflict is considered dysfunctional to the development of organizations and social relationships, and dysfunctions need to be rectified or balanced through normative values and the correction of unanticipated consequences of action (Ritzer 1992). Parsons (1937) approached the analysis of health-related phenomena from the perspective of structural-functionalist systems theory, offering two 'models of illness': the 'deviance' model and the 'sick-role' or 'capacity' model, both of which incorporate the micro- and macro-levels of the social structure. The 'deviance' model presents illness as a state of dependency for the person affected whereby therapy is expected but seen as a social control system and the physician as an agent of social control. The 'capacity' model proposes that illness is a breakdown of normality leading to the individual's total or partial incapacity to perform normal social roles, and brings the person into the 'sick-role'. The sick-role is a social position with different rights and obligations within the social structure that includes the doctor-patient interaction and the doctor's role with specific rights and obligations (Parsons 1937; Gerhardt 1989, pp. 1–67).

The second influential general theory is symbolic interactionism. It proposes that the individual's subjective perception of the situation through meanings and symbols has a significant impact on his/her behaviour. In contrast to structural-functionalism that is predominantly associated with one theorist—Talcott Parsons—symbolic interactionism represents the conceptual contributions of several theorists and offers two models of illness: the 'labelling' model and the 'negotiation' or 'crisis' model. The labelling model proposes that illness may be socially defined or labelled through a process of normative evaluation. Correspondingly, the negotiation model proposes that patients have control over their illness through negotiation: diagnoses and treatments by the medical profession may be negotiated and treatment involves the patient's active participation (Gerhardt 1989, pp. 73–173). Some analysts consider symbolic interactionism as playing an important part in current health sociology through qualitative methodological designs such as participant observation (Cockerham and Scambler 2010, p. 8).

Conflict theory is the third general sociological theory of relevance for public health. One of the earliest perspectives in conflict theory was the

Marxist postulation that social conditions of deprivation and exploitation lead to illness (Gerhardt 1989, pp. 254–5). More abstractly, conflict theory proposes that the potential for conflict is a normal component of social relationships, thus, negotiation and bargaining are key elements of social interaction. When applied to health phenomena, conflict theory offers two models, the 'loss' or 'stress' model and the 'domination-deprivation' model. The 'loss' or 'stress' model proposes that there is a strong connection between stress and illness; the loss of required social support increases the likelihood of illness; and vulnerability to illness is unequally distributed in society. The body of sociological research on the stress-illness connection was initiated in the 1950s and continues growing today (Pearlin 1989; Thois 2010). The 'domination-deprivation' model argues that the medical profession occupies a dominant position in society by virtue of its expertise vis-à-vis the patient's position of dependence given the patient's lack of specialized knowledge. The domination-deprivation model is also applied to the analysis of social inequality and socioeconomic deprivation in society (Gerhardt 1989, pp. 249–348). Besides providing the conceptual basis for critical and feminist perspectives, conflict theory as a school of thought has been particularly influential in the analysis of health policy: 'one of conflict theory's most important assets for medical sociology [is] the capacity to explain the politics associated with health reform' (Cockerham and Scambler 2010, pp. 10).

A fourth theoretical position is that of phenomenology/ethnomethodology which proposes, as Ritzer (1992, p. 388) summarizes it, that 'people are the product of the very society they create'. This phenomenology model regards illness as belonging to 'a broad category of disturbing occurrences … which breach the taken-for-granted peace of everyday routines' (Gerhardt 1989, p. 190). Compared to the other theories, and particularly, to the abundance of empirical research generated by the other three conceptual orientations—structural-functionalism, symbolic interactionism, and conflict theory—the phenomenology/ethnomethodology school of thought has contributed relatively less to empirical research in health sociology (Cockerham and Scambler 2010).

Sociological research relevant to public health

At the most abstract level, there are two distinctive contributions of health sociology to the study of health and illness in general and public health in particular. The first contribution is the systematic examination of the link between individual and community and 'the social organizational factors that shape the behaviour of groups' (Thorlindsson 2011, p. 21). The second main contribution is 'that health care systems are organizations … embedded within larger institutions, which have been shaped by historical precedents and operate within a specific cultural context' (Quadagno 2010). Health sociology research relevant to public health can be found within the framework of these two key contributions. The most significant social aspects that exemplify the link between individuals and groups in the context of public health are cultural and social class influences on health-related behaviour of individuals and communities; lifestyles; and the creation and transformation of healthcare systems. Cultural values rooted in ethnic and religious ethos together with social class differences shape people's beliefs and attitudes on health and illness and their lifestyles. Healthcare systems originate, develop, and change as part of the interplay of socioeconomic, cultural, and political factors. Let us discuss each of these aspects—culture, social class, lifestyle, and health systems—in more detail.

Culture and public health

Culture is a key term that refers to values and beliefs emanating from one's ethnic and/or religious ethos, typically transmitted across generations as part of the community's identity or 'design for living'. The impact of the physical environment on people's health has been historically recognized. Today we know that the socioeconomic environment is also a crucial factor in the risk of infectious and chronic illness (Armelagos and Harper 2010; Charmaz and Rosenfeld 2010; Baum 2016; WHO 2018). The impact of culture on public health is manifold and it is manifested in at least three main ways: health behaviour, that is, what people do regarding health risks; health attitudes and beliefs, that is, how people feel and think about health and illness; and people's perceptions of and responses to health authorities' public health guidelines and recommendations (Fox 1989; Quah 2007a, 2007b, 2010; Brown et al. 2012; WHO 2018). Perhaps the best example of this threefold impact is the influence of cultural and/or religious norms and

traditions on people's food preparation and diet. Irrespective of what public health authorities advise on food consumption (Harvey 2017, p. 204), ethnic, and/or religious groups differ in this regard as they tend to follow their traditionally sanctioned diet, such as vegetarian cuisine, avoidance of pork, the use of animal fat or refined vegetable oils in food preparation, or the avoidance of fried food at all; preferred beverages; and level of sugar consumption, among other aspects (Cockerham 2010b, pp. 29–32; Idler 2010; Seow 2017, pp. 314–15).

Cultural norms and traditions also guide people's preference for some types of health services over others. For example, cultural healing traditions such as traditional Chinese medicine and Ayurveda represent, respectively, essential aspects of Chinese and Indian cultures and thus the population's preference for and use of these therapeutic practices must be taken into consideration in public health policy (Quah 2003, 2008; Etkin et al. 2017).

Yet another contribution of health sociology research on the impact of culture on public health and the wider healthcare system is the analysis of the provision of healthcare from two perspectives: the micro-cosmos of clinician-patient interaction, and its permutation into the macro-cosmos of the interaction of health authorities with communities and groups. Two examples suffice. It has been found that social stereotypes held by practitioners (whether clinicians or public health practitioners) distort their decisions and interactions with their subjects of care. Clinicians' subjective perceptions of cultural or socioeconomic differences might lead them to link 'certain diagnoses to certain patients' (Smith et al. 2012, p. 162). Similarly, social perceptions of the community held by public health officials might unwittingly influence their public health approaches (e.g. Callis et al. 2018; Nuriddin et al. 2018). These examples also illustrate a related phenomenon: the impact of 'medical constructs', that is, when a physical or mental condition is defined by the health authorities 'as a *disease*' while the same condition is considered normal by the community or at least not a condition that requires medical attention. When this clash of the expert versus lay interpretations occur, the official classification 'has consequences for the social legitimacy of the symptoms, access to treatment, and costs to the health care system' (Conrad and Barker 2010; Rosich and Hankin 2010, p. S4). One illustration of this problem from the field of epidemic prevention is the opposition to mandatory culling of poultry or livestock by some farming communities who fail to appreciate

the connection between disease and the destruction of what for them are healthy farm animals and their families' source of livelihood (Quah 2007a).

Lifestyle and health

Health sociology research provides evidence of another significance influence—related to but distinct from culture—on what people do about health risks: their lifestyle. Cockerham (2010, p. 159) defines health lifestyles as 'collective patterns of health-related behavior based on choices from options available to people according to their life chances'. Life chances is a concept proposed by Max Weber (1968) to represent choices, that is, opportunities and constraints imposed on individuals by their relative position in society. People's lifestyle shapes their patterns of leisure and physical activity: some people have no time for leisure; others are sedentary or engaged in extreme sports; some frequent gyms as networking and thus do regular exercise; others are regularly involved in alcohol drinking, cigarette smoking, and/or hard drugs consumption. Likewise, a person's lifestyle is associated with the risk of serious injuries and the likelihood of contracting infectious diseases. Consequently, it is not surprising to find persistent differences in morbidity and mortality across ethnic groups that are basically of social rather than biological or genetic origin (Krieger 1987; Link and Phelan 1995; Williams and Sternthal 2010; Whyte et al. 2017). Considering that lifestyles vary for men and women particularly in Asia, gender differences are clearly discernible in morbidity and mortality figures (Quah 2011). The importance of culture for public health policy and practice is becoming better recognized and has led some analysts to recommend that public health practitioners be equipped with 'cultural competence' (Perez and Luquis 2008; Waters et al. 2017).

A person's lifestyle is shared and reinforced by his/her social network. Health sociology research demonstrates the active link between social relationships—also known as social ties—on the one hand, and health behaviours and outcomes on the other (Umberson and Montez 2010). Social ties and social networks significantly influenced health-related behaviour leading to obesity, cigarette smoking, 'recreational' drug use, and alcohol consumption, among other risk behaviours (Cockerham 2010a; Jones et al. 2011). However, while the most visible influence of social networks on health is negative, membership in social networks plays a positive role in

well-being and healthy longevity (Lund et al. 2000; Nyqvist et al. 2013). Social networks can also be mobilized in public health policy for the promotion of healthy habits and health risk-avoidance in the community (Mechanic and Tanner 2007; Umberson and Montez 2010; Brown et al. 2011).

The impact of lifestyle on health and illness is also analysed from the life course perspective, a conceptual approach that examines the changes experienced by individuals along their process of physical and socio-psychological maturation and experiences. An increasing body of evidence-based sociological research demonstrate important differences in the exposure to health risks across the life course of different ethnic groups and social classes (see, for example, Brown et al. 2012).

Social class, social stratification, and public health

Social class, a fundamental sociological construct, depicts the relative position of an individual in the social hierarchical system of prestige in his/her society (Weber 1968). In sociological research, a person's social class—socioeconomic status or SES—is ascertained by means of a composite index of three or more indicators, the most common of which are personal income level, years of formal education completed, and occupational prestige.[2] The resultant SES scores may be used as a continuous scale (usually standardized scores ranging from 0 to 100) or as a set of discrete categories based on the arithmetic mean and standard deviation. The simplest presentation of SES scores is a three-level description of lower class, middle class, and upper class. However, quantitative and continuous SES scores are preferred in sociological studies of class differences and patterns of stratification in society (Bendix and Lipset 1966; Nam and Powers 1983; Phelan et al. 2010).

The nearly universal phenomenon of social class differences influencing the person's access to and utilization of healthcare services is supported by sociological research findings: the higher a person's social class, the greater his/her access to healthcare services (Lahelma 2010; Rosich and Hankin 2010; Brown, et al. 2011; Quah and Cockerham 2017). A country's socioeconomic conditions lead to 'persistent health inequalities among social groups' (Rosich and Hankin 2010), inequalities which normally are 'deeply rooted in society' (Phelan et al. 2010). Research findings indicate

that social class reinforces the influence of lifestyle on health behaviour. However, under certain conditions, social class—mostly due to the influence of education and occupation—may have a stronger impact than cultural beliefs and norms on health behaviour, on access to health services, and on a person's exposure to occupational hazards. Indeed, workers have differential exposure to dangerous repetitive tasks, risk of industrial accidents, and environmental pollution, among other health risks embedded in the nature of the job. Consequently, sociologists examine the health impact of each SES component separately: occupational prestige, income level, and educational level as well as the strength of SES as a predictor of health behaviour.

The influence of social class on health status, health behaviour, access to, and use of health services, is one of the earliest and most examined social influences on health. Summarizing the advances of health sociology in the study of social class and health in the last 50 years, Phelan et al. (2010, p. S29) report that their data support 'the theory of fundamental causes'. This theory states that social class 'influences multiple disease outcomes … through multiple risk factors'; impacts 'access to resources that can be used to avoid risks or minimize the consequences of disease once it occurs'; and that those influences tend to persist over time. Their research findings support the theory and have implications for public health policy that are different from results in studies applying 'an individually oriented risk-factor approach'. Thus, these authors recommend that social class-induced health inequalities 'can be reduced by instituting health interventions that automatically benefit individuals irrespective of their own resources or behaviors' (Phelan et al. 2010, p. S37). Research on the social determinants of health and illness—scrutinizing social, economic, cultural, and ideological factors—continues in earnest (e.g. Pearlin 1989; Link and Phelan 1995; Kawachi et al. 1999; Lund et al. 2000; Mechanic and Tanner 2007; Phelan et al. 2010; Williams and Sternthal 2010; Doyal 2014; Baum 2016; Quah and Cockerham 2017; Marmot 2018).

Taken together, the related phenomena represented by the person's lifestyle, life course, and social class, point to the need to focus on families in the investigation of health promotion as well as disease risks. Typically, in addition to genetic background, family members share some main lifestyle features such as diet, recreation patterns, and the physical neighbourhood. Moreover, some disease vectors (e.g. fleas, mosquitos,

mice) settle in family homes. Thus, it is not surprising to find crucial differences in health risks exposure between, for example, individuals living in rural and urban communities and between people living in rich and poor neighbourhoods. With this in mind, Valberg and colleagues (2018 p. 1) discuss the use of 'family relative risk' (FRR) as 'a continuous risk model' for the public health assessment of families at risk at the community level.

Two additional contributions of the sociological study of occupations and professions that increase our understanding of health and illness in society are the analysis of occupational hazards and the role of the medical profession. On the one hand, the comparative analysis of health outcomes, disease prevalence and disease incidence across the wide spectrum of occupations in society consistently reveals clear variations derived mainly from the nature of work performed and occupational hazards (e.g. Christ et al. 2012; NASEM 2018). On the other hand, sociologists also examine the biomedical profession and traditional healing occupations in terms of their relative prestige, their training, internal composition and dynamics, and their impact on the health of populations (e.g. Starr 1982; Friedson 1988; Quah 2003, 2008). Research findings indicate that the ideal trust-based and confidential doctor-patient relationship is corporatized today as the medical profession is overwhelmed by 'the healthcare industry' dominated by employers' 'guidelines, pre-authorization, centralization and financial awards for increased output' (Starr 1982; Freidson 1988; Timmermans and Oh 2010; Vanderminden and Potter 2010, pp. 356–7).

Healthcare systems and public health

In addition to the analysis of the impact of culture and social class on public health, another illustration of the contributions of health sociology is the study of the nature, dynamics, and evolution of healthcare systems at the national and transnational level. When sociologists study healthcare systems they include the approaches taken by nation-states towards the creation, organization, distribution, and management of their healthcare resources. Thus, health policy and healthcare reform are integral parts of this type of studies. Some cross-national and comparative research findings on healthcare systems suggest that in many countries today 'too much priority has been given to hospital care at the expense of the development of primary care, community care and public health' (Stevens 2010, p. 435).

One of several aspects of healthcare systems that have received increased attention by medical sociologists in the past decade is the analysis of flaws in health systems that distort the link between cost and utilization of health services, thus the urgency of healthcare reform (Mechanic and McAlpine 2010). One sector of the healthcare system that has a significant input in healthcare costs is the role of stakeholders such as the pharmaceutical industry in the promotion of medicalization, particularly through the use of market opportunities and marketing methods to sell their products (Rosich and Hankin 2010, pp. S4–5).

Another illustration of health sociology analyses of health systems is the area of health services research. There is a vast literature of studies documenting the influence of social class and ethnic differences on unequal access to and utilization of health services. Three types of studies illustrate this research area. The first type of studies show that unequal access to healthcare services is preserved by institutions and policies that constrain and enable 'the actions of health services organizations, health care providers, and consumers' (Wright and Perry 2010, p. S109). Those studies of multiple dysfunctions of health systems challenge international policy makers and are the subject of social policy interventions based on social science research such as the Disease Control Priorities Project jointly conducted by the World Health Organization, the US National Institutes of Health and the World Bank (Laxminarayan et al. 2006). The second type of health systems research studies is the comparative analysis of health financing approaches and systems in terms of their suitability to the socioeconomic needs and demographic characteristics of the population. One example of this type of research is the study of insurance plans by Zaslavsky and Epstein (2005). The third type of study represent the direct input of medical sociologists on public policy; one recent example is the analysis of changes in state regulations on the collection and access to health relevant population data for research purposes (Hynes 2011).

In addition to the relevance of the analysis of culture, lifestyle, social class, and social systems for public health, health sociology also examines the interaction of these and other factors in the analysis of disease incidence and prevalence in the community. For example, Anderson et al. (2011, p. 390) found that 'social, cultural, and biological issues should all be considered in the assessment of differences in breast cancer survival'. Brown et al. (2011, p. 941) discuss the relevance of social factors in the

diagnosis and management of obesity, diabetes, and asthma. A wealth of studies on the social factors of disease is found in health sociology journals mentioned earlier such as *JHSB*, *SSM*, and *Sociology of Health and Illness* and numerous other international refereed journals on health research.

Sociological research methods and public health

Health sociology applies the full range of sociological approaches and methodological tools to the study of socioeconomic, cultural, and demographic characteristics of populations and individuals that impact their health (Rosich and Hankin 2010; Smith et al. 2012; Baum 2016). Indeed, referring to the analysis of epidemics and health behaviour in general, Hedges et al. (2002) argue that 'social science knowledge in theory and methodology has guided public health research for several decades'. One illustration of this influence is the application of sociological approaches to the analysis of income inequality and health in most of the articles compiled in the book on that topic by Kawachi et al. (1999).

The fundamental position of sociology in general, and health sociology in particular, regarding scientific research is well articulated by Vogt (2007, p. 5) when he highlights the critical tasks of collecting and analysing evidence for the purpose of creating knowledge. In this regard, there is clear compatibility between the goals of research in both fields of knowledge, health sociology, and public health. Still, there are three important characteristics of sociological research methodology that must be taken into consideration: (a) the kinds of data used in sociology; (b) the levels of measurement applied; and (c) the types of research design followed. Let us consider each characteristic in turn.

What kinds of data are used in sociology? Two kinds of evidence or data are used: *primary* and *secondary* data. Primary data are data collected first-hand by the investigator, for example, data on people's attitudes and behaviour obtained through personal interviews, observation, experiments, or surveys; or data extracted by means of first-hand analyses of documents, letters, records, or other type of 'raw' data. In contrast, secondary data are data previously collected by others, for example, official statistics, population censuses, or databases with processed or coded data sets from surveys or completed studies. Many research institutions keep databases of secondary data, usually codified data from survey research studies that can

be purchased by other researchers together with the respective codebook for secondary and longitudinal analyses (Miller and Salkind 2002, pp. 346–53; Hauser et al. 2010). Many data archives are now accessible online.

What levels of measurement are used in sociology? Sociological research uses four levels of measurement: nominal, ordinal, interval, and ratio levels. The most rudimentary is the *nominal* level which uses names or labels to depict different categories of the phenomenon under study (for example, the labels 'male' and 'female' to record a subject's gender; the labels 'Muslim', 'Hindu', 'Christian', 'Other religion' and 'no religion', to record the subject's religious affiliation). Each name or category may be assigned a numerical code but the numbers have no quantitative meaning, they are only labels. This conversion of alphabetical data into numerical data is done for digital data analysis using computerized statistical models, although there are now computer programmes designed for the analysis of alpha and alpha-numerical data. *Ordinal* measurement assigns ranks to categories of the observed phenomenon (example: 'slender', 'plump', 'obese' to record body appearance) but 'the distances between ranks in an ordinal scale are not necessarily equal'. *Interval* measurement uses ranks with a standard and measurable distance between categories but 'have no meaningful zero point'. The typical illustration of an interval level of measurement is the Fahrenheit temperature scale. A common sociological example is the range of a person's attitudes towards a problem or situation. Here are two sociological interview questions that illustrate the *interval* scale: 'Would you say you are very confident, only slightly confident, slightly anxious, or very anxious about what needs to be done to handle the patient?'; 'How would you rate your quality of life? 'Very poor', 'Poor', 'Neither poor nor good', 'Good', or 'Very good'?' The interval scale is more accurate than the nominal or ordinal levels of measurement, but less precise than the *ratio* scale that has both 'equal intervals between categories and a true zero point' (Vogt 2007, pp. 9–10). Examples of ratio scales in sociology are the measurement of the subject's age in years; income in dollars earned per month; formal education in number of years of completed schooling; number of children; and years of marriage.

One of the most widely used interval scales in the sociological analysis of health and illness is the 'Quality of Life' index popularized by the World Health Organization as the WHOQOL (WHO 2004; WHO 2017). The long and brief versions of the WHOQOL and other modalities of the scale are

usually included in studies of a wide range of health conditions from stress reduction to autism, hepatitis C, coronary heart disease, cancer, obesity, gastric bypass, bariatric surgery, neuropathic pain, and periodontal diseases, among other health conditions (e.g. Hamoen et al. 2015; Buset et al. 2016; De Vibe et al. 2017; WHO 2017; National Academies 2018).

What types of research designs are used in sociology? Depending on the nature of the phenomena under investigation and on the investigator's preferred level of measurement, sociologists may use qualitative or quantitative designs or a combination of both. Researchers who have a preference for qualitative research tend to steer clear of measurement altogether or use only nominal levels of measurement. Some of the most common qualitative research designs are participant and non-participant observation of social phenomena, and conversations with research subjects individually or in small groups (Schwartz and Jacobs 1979; Chongo et al. 2018) following the 'grounded theory' approach (e.g. Thapa et al. 2018). One modality of the small group method is the focus group approach whereby the researcher, acting as moderator, gathers together a small group of people—preferably not acquainted with each other—to discuss a particular topic. The focus group method is probably the weakest approach and it should be used with extreme caution in health research and perhaps only in exploratory studies, due to its three most serious limitations: (1) considering that the conversation is public, that is, in front of strangers, the researcher cannot guarantee confidentiality to the focus group participants, but confidentiality is an important component of informed consent required when conducting research with human subjects (Citro et al. 2003) and particularly if the study is on their state of health; (2) the validity and reliability of data collected in focus group discussions are low or non-existent; and (3) focus group participants are typically influenced by social pressure given the dynamics of small group interaction. Thus, the focus group approach is more suitable for exploratory studies of group dynamics (Mechlenborg and Grønkjaer 2018).

In contrast to qualitative designs, a quantitative research design may involve all four levels of measurement but mostly interval and ratio levels of measurement of social phenomena. Similarly, a quantitative research design may include one or more methods of data collection: document analysis, observation, personal interviews with structured or unstructured questionnaires, surveys, experiments, or quasi-experiments (Boudon 1993,

pp. 158–71; Vogt 2007, pp. 8–10). Naturally, as it happens in all fields of knowledge, research methods have their limitations and researchers need to consider the suitability of the methods selected and take steps to reduce as much as possible the probability of error. With the objective of error reduction in mind, medical sociologists may use in the same study a combination of methods from the array available: participant and non-participant observation; in-depth unstructured interviews; and structured interviews using questionnaires with a variety of question formats, from open-ended questions to attitudinal scales. Structured questionnaires are usually the main instrument of data collection in survey research. Survey research is a methodological design in sociology created by Paul Lazarsfeld (Boudon 1993). A prominent sociologist, Paul Lazarsfeld (1901–1976) identified three phases in the history of surveys. The earliest phase was before 1930, when surveys were used for 'the collection of factual information which could help to improve social conditions'. The second phase covered the period 1930 to 1948 when surveys were used to collect data on public opinion and consumer preferences. The third phase, from 1949 onwards, started with the publication of the seminal sociological study *The American Soldier* when the survey method became an important sociological tool for the collection of 'a large body of data' that could be 'made coherent and meaningful by careful statistical analysis' (Lazarsfeld 1968, pp. vii–x; Boudon 1993).

Sociologists using quantitative research methods aim to ascertain and reduce the level of error in the calculation of correlations among phenomena and in inferences drawn from representative samples to the total population (inferential Type I and Type II errors). This objective is facilitated by the application of statistical analysis models ranging in complexity and sophistication from simple measures of association such as chi-square, Pearson's coefficient, ANOVA, and path analysis, to multilevel models such as ordinal logistic regression, structural equation modelling, and hierarchical linear modelling, among others. Examples of some of the most recent health sociology studies applying multilevel models are: the impact of alcohol drinking on academic performance of secondary school students (Crosnoe et al. 2012), perceived discrimination and health among adolescents and young adults (Grollman 2012), the influence of social networks in the management of hypertension (Cornwell and Waite 2012),

and factors leading to sleep disturbance among US adults (Ailshire and Burgard 2012).

Another important aspect of sociological research that must be mentioned is the commitment that mainstream sociology shares with all other fields of scientific research: to pay serious attention to data validity and reliability. Reliability means consistency in the measurement: getting the same results every time the phenomenon is measured with the same instrument. The relation between validity and reliability is important: a reliable instrument may have low validity (Vogt 2007, p. 118). Validity has two general dimensions: internal and external. Internal validity focuses on the research question. The most elementary form of internal validity is content validity and it addresses the question: is the instrument measuring what it intends to measure or something else? There are two additional types of internal validity: criterion-related validity and construct validity. The second dimension is external validity that is, 'the potential for a study's findings to be generalized from one setting to others' (Burchett et al. 2011, p. 238). Similarly, qualitative research methodologists have made efforts to strengthen the credibility and transferability of their findings (Baum 2016; Lincoln and Guba 1985).

This brings us to the matter of sampling. From the perspective of public health, perhaps one of the most relevant methodological features of sociology, and in particular, health sociology, is the study of population trends through the analysis of representative samples, also known as probability samples. A representative sample is a sample drawn from the target population following a procedure that gives each individual in the target population the same probability of being selected. Probability samples allow the researcher to make inferences to the total population while non-probability samples do not. The application of inferential statistical tests such as chi-square, t-test and ANOVA indicate the probability that what was found in the sample does not reflect the situation in the population but it is due to chance or sampling error (Neuman 2000, p. 217; Vogt 2007, p. 12).

On the matter of sampling, two caveats are in order. The first caveat is that there are research topics and research situations for which random sampling is impossible to obtain or is unsuitable. This is the case of health conditions socially seen as stigmatizing, or deviant, so that the affected subjects and/or their families refuse participation in the study; or there is no

comprehensive and reliable record of all members of the target population that could be used as the sampling frame to draw a representative random sample. In situations like these, the researcher may opt for a non-probability sample. The 'purposive sample' is one form of non-probability sample whereby subjects are selected based on specific criteria, rather than randomly (Vogt 2007, p. 81). Naturally, non-probability samples preclude any inferences to the total population as only a probability sample has external validity (Vogt 2007, p. 82).

The second caveat is on the size of the sample. Besides the options of representative or purposive sampling of large populations, there is also the case study approach whereby the researcher studies only one or a few 'cases' of the social phenomenon under investigation. This approach is useful in at least two ways. First, in emergency situations, a case study may be conducted when rapid information about a condition is required, for example, the case study of a person or a family affected by a rare or 'new' illness that may be potentially contagious; or a unique event that could seriously affect large numbers of people. Second, in non-emergency situations, a case study approach is commonly used in exploratory research: that is in situations where the social phenomenon is new or never studied before and the researchers wish to familiarize themselves with the nature of the phenomenon in order to plan a more comprehensive and systematic study (Ragin and Becker 1992).

Psychology in public health

It is generally agreed that the discipline of psychology was born in the late nineteenth century (Norcross and Karpiak 2012) as a continuation of the discovery of 'the neurophysiological location of psychological faculties' among other advances in the study of the human body (Greenwood 2009). The period 1890–1910, was particularly fruitful for the discipline: it produced the seminal book *Principles of Psychology* by American William James in 1890, and Sigmund Freud's *Die Traumdeutung* published in 1900, often deemed as the framework of psychoanalysis and of Freud's subsequent studies of the thread linking personality, mind, and illness (Friedman 2002).

Just as the subfield of health sociology plays an important role within the discipline of sociology, the most relevant area of specialization in psychology for public health is health psychology (Luecken and Gallo 2008). Health psychology is also referred to by some analysts as clinical psychology or clinical health psychology (Boyer and Paharia 2008), and as a distinct and growing subfield (Romanow and Marchildon 2003). Health psychology has a relatively recent history (Albery and Munafò 2008). Introducing his *Health Psychology* textbook, Howard S. Freeman wrote that although health psychology has 'its roots in the fields of medical sociology … public health' and other fields, 'until the late 1970s, there were no programs of study in health psychology' (Friedman 2002, p. 23). The formal introduction of health psychology as a subdiscipline of psychology is highlighted by three chronological signposts: 1974 when health psychology was first proposed as a new subject in the curriculum at the University of California–San Francisco School of Medicine; 1977 when Health Psychology was approved as a new division of the American Psychological Association; and 1984 when the 'First International Symposium of Health Psychology' was held (Stone 1991, p. 3). As Friedman reports, the number of health psychologists has been growing rapidly not only in the United States but in many countries (2002, p. 23).

However, until about one decade ago health psychology had traditionally 'maintained … a pronounced unidimensional focus on mental health' (Arnett 2001, p. 38). Arnett argued that the field of health psychology 'would be significantly enhanced by a more comprehensive understanding of the epidemiology, pathophysiology, and natural history of the wide range of health disorders amenable to psychological interventions' (Arnett 2001, p. 43). Among the health areas that Arnett suggested could be studied by health psychology—in addition to mental health—are weight control, and addictions to alcohol, nicotine, and hard drugs (Arnett 2001, p. 41). Health psychology has developed significantly since the 1970s (Kenkel et al. 2005) although at different pace in different countries. Sweden is seen as a pioneer based on the development of 'psychophysiology of stress at the Karolinska Institute in Stockholm' in the early 1970s (Stone 1991, p. 9).

Psychological theories and public health

One of the best known names in the history of Psychology is Sigmund Freud (1859–1939). Freud was trained as a physician. He launched a new 'school' of psychoanalysis in psychology with his colleague Joseph Breuer promoting the idea that neuroses had a psychological, not a physiological basis (Ritzer 1992, p. 32). Freud based his theories on his observation of patients, making general inferences from individual cases.

In its early development, psychology had a similar concern found in sociology: a positivist trend following the goals of science, to identify universal laws 'expressed in theories that allowed prediction and explanation' (Looren de Jong 2010, p. 746). Over the years this quest promoted a range of theories including psychoanalysis, neuropsychology, and various personality theories some of them spanning the disciplines of sociology and psychology, usually clustered into four groups. These clusters are: (1) social cognition models addressed to the explanation of individual's health-related behaviour and comprising the theory of planned behaviour, the Health Belief Model, the transtheoretical model, and the health action process model; (2) self-regulation models focus on risk behaviours and include cognitive adaptation theory, self-regulation model and the model of self-efficacy; (3) the psychophysiological models cluster that 'describes how psychological processes, particularly in response to stressors, can influence physiological functioning', and includes the gate control theory of pain, the cognitive perceptual model, and the general adaptation syndrome, among others; (4) the fourth and final cluster is the environmental models cluster that includes the operant learning theory, conservation of resources theory, and the demand-control model of work stress (Weinman et al. 2007a, pp. xxx–xxxv; Weinman et al. 2007b). Given the space constraints of this chapter, this is only a summary of the main theories but it conveys the idea of the large and growing array of conceptual perspectives underpinning research in psychology today. Marshall H. Becker and his colleagues (Becker et al. 2007) provide an excellent discussion of the conceptual collaboration between sociology and psychology in the analysis of health behaviour—exemplified by the Health Belief Model—and the fruits of that collaboration in the form of conceptual models and abundant and advancing research (e g. Harmon-Jones et al. 2015; Rivard and Butler 2017; Anderberg and Dahlberg 2018; Diaz-Quijano et al. 2018).

Seen from a different and more abstract perspective, psychology—and all its subfields including health psychology—is engaged in a debate between

two contrasting positions: a contemporary version of the positivistic approach termed ‘rational reconstruction and theory building’ and ‘critical deconstructivism’. The latter approach ‘focuses on theory as dialogue and critique’ and on ‘the collective construction of meaning, not the discovery of objective facts about the mind’ (Looren de Jong 2010, pp. 747–8). The two positions are seen as the bases for ‘two research traditions within academic psychology, that is, social constructionism and cognitive neuroscience’ (Praetorius 2003, p. 511).

The debate over theoretical paths that can strengthen psychology’s contribution to science and psychologists’ professional identity continues unabated (Romanow and Marchildon 2003), but some intermediate schools of thought recognize that theory is an important part of mainstream psychological research (Looren de Jong 2010). At the same time, some psychologists like C.R. Cooper and J. Denner have recommended the inclusion of culture in the discipline’s research repertoire acknowledging that culture affects behaviour. Their recommendation springs from the engagement of psychologists in interdisciplinary research and the fact that culture is a long-standing subject of investigation in sociology and other social sciences but has not received sufficient attention in psychology (Cooper and Denner 1998). Another reason is their concern with the problem of external validity in psychology. They wrote, ‘Psychologists continue to generalize beyond their samples; while studying only their own countries … they assume that psychological phenomena are universal enough to make such studies representative’ (Cooper and Denner 1998 p. 579).

Psychological research and services relevant to public health

The general scope of health psychology research was well delineated by Canadian psychologist John L. Arnett (2001, p. 45) thus: ‘the diagnosis and prevention of illness, the promotion of health, and the amelioration of suffering related to the whole range of health conditions … across the whole spectrum of health’.

Clinical/health psychologists provide direct healthcare but their actual clinical practice is determined by government regulations and, more broadly, by the political system and ideology guiding the healthcare system. American psychologists succeeded in obtaining official licensing and

getting federal statutes modified ‘to gain autonomous inclusion in most federal programs’ providing paid healthcare and to practice in managed healthcare organizations. However, for most of the twentieth century, the training and practice of psychology and health psychology did not attend to the profession’s participation in the public policy process or in ‘programs that … address society’s public health priorities and perceived needs’ (De Leon et al. 1995, p. 493). This situation is common in other countries but it has been changing. For example, in most OECD (Organization for Economic Cooperation and Development) countries, people have access to paid psychotherapy services only if provided by physicians specialized in psychiatry. However, while both psychologists and psychiatrists can provide psychotherapy, ‘PhD-level clinical psychologists receive far more extensive training and education in psychotherapy than physician-trained psychiatrists’ (Romanow and Marchildon 2003, p. 286). In the realm of ideology, in countries with ‘centralized planning for health care’ or in community-based healthcare programmes such as Cuba, Costa Rica and India, ‘the concept of health psychology appeared early and its approach tended to be community-oriented’ in contrast to the individual orientation of the field in the European tradition (Stone 1991, pp. 12–13). Community-oriented psychology today goes beyond ideology. In addition to community-based counselling in disaster situations, other illustrations of the community-oriented approach involve the treatment of ‘mass psychogenic illness’ also known as ‘mass hysteria’ (Friedman 2002, pp. 64–5) and interventions designed for members of particular communities such as homeless youth (Trickett 2009).

The practice areas among health psychologists also vary across countries. In the first two decades of the field, while clinical work was the most common area of activity in the United States and the United Kingdom, the main work of health psychologists in Germany, Austria, France, and Argentina was teaching at medical schools (Stone 1991, p. 10). An overview of current health psychology work relevant to public health is suggested in the 2010 study of American health psychologists by Norcross and Karpiak (2012). These researchers found that the most common area of professional work among American health psychologists is psychotherapy practised either on individuals or as group therapy. Indeed, 76% of the professional health psychologists spent about 35% of their professional activity in psychotherapy in 2010, and it has been the most common area of

work since 1986 when comparative data collection started. The four other main professional activities are: diagnosis/assessment (58%); teaching (49%); clinical supervision (47%) and research/writing (47%) (Norcross and Karpiak 2012, pp. 5–6).

Psychologists propose that some of the services they offer do increase the productivity of health services by lowering public health expenditure. One illustration of this type of services is the provision of individual or group psychotherapies for breast cancer patients, and the treatment of depression and addiction-related conditions through cognitive behaviour therapies (Romanow and Marchildon 2003). Another example is 'evidence-based parenting interventions' to assist parents with the effective management of children and adolescents affected by 'behavioural and emotional problems' (Sanders 2010). In his Presidential address to the Canadian Psychological Association, M.R. Sanders suggested that psychologists could offer evidence-based parenting interventions as a 'public health approach' because the aim of his proposal was 'to reduce the prevalence rates of inadequate parenting at …[the]… population level', a drastic departure from the traditional focus of psychology on 'individual clinical interventions' (Sanders 2010, p. 17). Similar intervention programmes have been tried in the United States, England, and several European countries.

One of the relevant areas of health psychology for public health is stress management. An interesting input was borne out of the collaboration between a public health practitioner and a psychologist: preventive stress management programmes for corporations based on the 'theory of preventive stress management (TPSM)' devised by J.C. Quick and J.D. Quick in 1979 and focused on both physiological stress and organizational stress (Hargrove et al. 2011). According to the TPSM, individuals and organizations react to stressors either positively or negatively. The response to a stressor, or distress, 'is the psychophysiological response of the individual as well as … of a group or organization to fight or to flee'; the outcomes of that response may be 'either positive, such as a heightened alertness and enhanced performance, or negative, such as medical, psychological and behavioural distress' (Hargrove et al. 2011, pp. 182–3). As mentioned earlier, research on stress management was one of the earliest contributions in the field of health psychology by Sweden's Karolinska Institute. Swedish researchers analysed the impact of stressors produced by the individual's social and physical environment including the work setting,

home environment, transportation, and retirement among other factors (Stone 1991, p. 9).

Family psychotherapy is yet another area of psychology that is relevant to public health. It involves various modalities two of which are perhaps the best well-known. They are family therapy involving children and adolescents, and family therapy with older adults. Family psychotherapy involving children and youth is typically offered in collaboration with the school system in the community and deals with children's mental illness as well as with behavioural problems including drug abuse, truancy, and academic performance issues among other aspects. Family psychotherapy involving older adults is gaining ground in countries with a growing elderly population. Psychologists involved in family psychotherapy with older adults deal with a wide range of issues from dementia and other mental illness to intergenerational conflict, the significance of home setting, pain management, mobility, and physical impairment, among other problems (Curtis and Dixon 2005). In their review of recent developments in family psychotherapy with older adults in the United Kingdom, Curtis and Dixon found 'a lack of empirical research into family therapy with late life families' that needs to be addressed and an insufficient number of family psychotherapists doing this type of work compared to the larger number involved in family therapy with children and younger adults (2005, pp. 56–7).

Change in health-related behaviour is another area of work of health psychologists that is relevant to public health. Psychologists apply one or more social cognition theories (used in sociology as well) in individual and group therapy interventions with youth and adults to modify risk-taking behaviours such as alcohol drinking and cigarette smoking, and to reinforce preventive health behaviours such as condom use and healthy diet (Thirlaway and Upton 2009).

Psychological research methods relevant to public health

A distinguishing feature of psychology is its permanent state of critical introspection. This situation applies to its research methodology. Discussing the methodological landscape in psychology, Lambdin (2012, p. 68) argues that a dominant feature is the rift between two camps: experimental psychologists who claim their work 'is science because it is experimental'

and the work of their non-experimental colleagues. Both psychology camps are searching for ways to advance the discipline and correct research gaps and bias. But, as Lambdin rightly points out, common sources of bias are found in all scientific research—irrespective of the knowledge area, whether it is psychology, sociology, biomedicine, or any other: 'the results of any study can be preordained by the selection of stimuli, how variables are operationally defined, the construction of the experimental protocol, the level at which the data are aggregated, or the particular analysis the researcher chooses to utilize' (2012, p. 69). Lambdin suggests that researchers should explain their findings 'without reference to inferential statistics'; and advises applying instead 'a lot of hard thinking and critical reasoning' and focusing on 'effect sizes and confidence intervals which … are together far more informative and meaningful' (2012, pp. 70–1).

The use of statistical inference is only one of several aspects under discussion within the field of psychology. Four main fundamental issues that impact research are actively discussed today: the 'scientific versus humanistic' cultures in psychology; two ontological issues, the 'subjective-objective' and the 'individual-collective tensions'; 'the problem of evaluative criteria'; and 'the problem of competing worldviews or value systems' (Goertzen 2008). The beneficial ongoing debate ranges from critiques of conceptual schools of thought in psychology to practice-oriented assessment of diagnostic material, including approaches to improve the *Diagnostic and Statistical Manual of Mental Disorders*, the internationally recognized prime classification of psychopathology published periodically by the American Psychiatric Association (Acton and Zodda 2005).

Most if not all health psychologists are inclined to see conclusions obtained from an experimental design also known as randomized clinical trial, as significantly better than findings from case studies, but unfortunately it is not possible to use random allocation and true experiments in health psychology (Friedman 2002, p. 21). Thus, in addition to other forms of systematic observation, psychologists have traditionally used attitudinal scales and questionnaires to collect systematic responses from subjects in laboratory situations or one-to-one interviews or small group contexts. In addition to this long-established psychological methodology, some psychologists may occasionally borrow methods from sociology and other fields to study psychological phenomena. For example,

in their investigation of people's beliefs about medicine, Kienhues and Bromme (2012) used a survey technique by sending a questionnaire online to 284 higher school students.

The distinction made between quantitative and qualitative methods of research—discussed earlier—is the same as in sociology. Thus, in general, psychologists have a preference for quantitative methods and for the application of advanced statistical models in data analysis. However, Wienman and colleagues (2007b) suggest that qualitative methods are used now more often than before and discuss several methods that are undergoing improvement. Three examples suffice. The first illustration is studies of health-related behaviour where investigators may use 'self-report, observation by video or audio recording'. Subjects' self-reports are the most common method of data collection in psychology 'but it is open to problems of error and bias'; thus this method is under close scrutiny and efforts are underway to improve it with a view to make it more 'psychometrically sound'. The second illustration is the assessment of stress as the occurrence of negative life events and emotional changes using attitudinal scales such as the 'Impact of Event Scale' and the 'Hospital Anxiety and Depression Scale' or HADS. The third illustration is the measurement of 'health and illness cognitions' such as the perception of health risks through the use of questionnaire scales such as the Multi-dimensional Health Locus of Control scale and the Illness Perception Questionnaire (Wienman et al. 2007b). As is the pattern in sociology, psychologists dedicate considerable effort at examining and improving the validity and reliability of their research instruments (Sutton 2007).

Formalizing the contributions of sociology and psychology to public health training

The relevance of sociology and psychology to public health is amply recognized today not only in public health research and practice but also in public health education, particularly graduate training in public health.

Some researchers maintain that the United Kingdom was the first country to formalize the field of public health (Paccaud et al. 2011, p. 68). It is not surprising then, that the first training programme in public health was launched in 1871 in the United Kingdom. That programme was the

Diploma in Public Health comprising only three subjects, ‘epidemiology, population statistics and communicable diseases’, and given ‘a growing interest in how social determinants of health and illness interacted with disease since the 1930’s’, sociology and psychology were ‘incorporated into public health education’ in the United Kingdom in the 1960s (Cole et al. 2011, pp. 89–90).

The incorporation of sociology and psychology in public health curricula has been sluggish in European countries but in the United States, two private universities included both fields as the forerunners of graduate training in public health: Johns Hopkins and Harvard (Paccaud et al. 2011, p. 69). The American framework for public health training at the university level—the origins of which can be traced back to the Welch-Rose Report published in 1915—implemented in schools of public health include social and behavioural sciences as one of five core areas of the curriculum. The five areas are required by the Council on Education for Public Health (CEPH) that regulates Masters and Doctoral Public Health programmes in the United States (Gebbie et al. 2007; Calhoun et al. 2011, pp. 152–3; CEPH 2011a, p. 14; CEPH 2011b, p. 17; Imperato et al. 2011; Rosenstock et al. 2011, pp. 43–6).

The past two decades have seen serious efforts to improve the quality of medical education curriculum worldwide (WHO 2005). Other international organizations such as the World Federation for Medical Education are leading the improvement of medical education standards by incorporating in the graduate public health curriculum ‘basic biomedical, behavioural and social sciences, general clinical skills, clinical decision skills, communication abilities and medical ethics’ (WFME 2003, pp. 5–10). As part of that effort to improve public health training, the experience of the University of North Carolina—Charlotte [UNC] is illustrative of the trend towards inclusion of health behaviour in the curricula of public health graduate and undergraduate programmes accredited by the Council on Education for Public Health. At UNC the expansion of the public health curricula was motivated by the aim to establish ‘a thriving research base that emphasized population health and health behaviour research’ and ‘responsive and progressive health and human service training programmes’ within an interdisciplinary model (Thompson et al. 2009, p. 3). The UNC’s current 45-credit curriculum for the Master of Science in Public Health consists of professional practice and research in

epidemiology, biostatistics, research methods, 'behavioral and environmental determinants of health' as well as programme planning and evaluation; and the core curriculum includes 'Social and Behavioral Foundations of Public Health' (Thomson et al. 2009, p. 6). A case was made as well for the inclusion of health sociology modules in the Columbia University Medical Center's curricula for internal medicine residents to expose them to 'important societal health issues such as health disparities, social determinants of health and health literacy' (Fiebach et al. 2011, p. S268). Besides training, another important input of sociological and psychological research is the analysis of decision-making processes of junior doctors when selecting post-graduate programmes (Gale et al. 2017).

Public health programmes in Europe have followed the same trend of incorporating social and behavioural sciences as core subjects in graduate schools of public health following accreditation requirements set by the European Agency for Accreditation in Public Health Education or APHEA (Otok et al. 2011; Tulchinsky and McKee 2011). Outside the western sphere, however, the incorporation of social and behavioural sciences into the public health curriculum started only recently. Today, health sociology and health psychology are standard components of many graduate public health programmes in some non-Western countries (Bangdiwala et al. 2011; Pino and Solimano 2011).

A final word is in order. The development of multidisciplinary research on human subjects involving biomedicine, sociology, and psychology in various combinations, has led regulatory agencies to support collaboration by creating and expanding ethical guidelines. This effort also reinforces the fact that research ethics is as important in health sociology and health psychology research as it is in biomedical research. Two of the most relevant dedicated guidebooks have been published by the National Academies Press: *Protecting Participants and Facilitating Social and Behavioral Sciences Research* (Citro et al. 2003) and *Conducting Biosocial Surveys* (Houser et al. 2010). Multidisciplinary research on health and illness continues to expand with the increasing recognition of the relevance of social sciences in public health as well as in clinical medicine (Daw et al. 2012; Betts and Hodgson 2018; WHO 2018).

Conclusions

The preceding pages present a succinct description of what two social sciences—sociology and psychology—and their respective subfields, health sociology and health psychology, offer to the discipline of public health. Perhaps the strongest tie binding health sociology and health psychology is their interest in the systematic analysis of health-related behaviour and their attention to applicability of research findings to the solution of health-related problems in the community.

The essential focus of sociology is on the individual's *social* behaviour, that is, the person's actions and motives addressed to the actions and motives of others in primary and formal relations, as well as in collective entities ranging from small groups to institutions, social systems, nations, and global organizations. Therefore, health sociology analyses actions, beliefs, attitudes, and knowledge of individuals on matters of health and illness; health-related activities of groups, organizations and institutions at all levels of society; and the impact of the socioeconomic, cultural, political, and physical environment upon the actions, attitudes, and beliefs of individuals and collective entities of any size. Consequently, health sociology involves the study of both the micro-level and the macro-level spheres of social action: from the trilogy of individual action (preventive health behaviour, illness behaviour, sick-role behaviour), to dyads and primary groups (for example, doctor-patient relations; family's health, impact of small social networks on health risks), to large and complex organizations and social systems (for example, the analysis of healthcare organizations, insurance, and pharmaceutical corporations, the medical profession, government health policies, and impact of ideologies on the distribution of health resources in the population).

Health psychology's original focus was on the analysis of the individual's mental health. However, over the past two decades health psychology has re-oriented its attention to the analysis of all other aspects of health behaviour in addition to and beyond mental illness. This re-alignment of research interests is a positive development for public health because public health practitioners may now benefit not only from the growing body of knowledge produced by health sociology and health psychology, but they also benefit from the conceptual and methodological expertise of both disciplines.

Knowledge production in the twenty-first century is increasingly leaning towards multidisciplinary collaboration across knowledge domains. This

modality presents a unique challenge: multidisciplinary collaboration works well as long as each participating discipline has something distinct to offer to the collaborative enterprise. Each discipline must continue to push the boundaries of knowledge to increase its own expertise. Health sociology and health psychology continue creating knowledge in their respective areas of interest through research and constant critical examination of their own theories and methodological approaches. At the same time, both disciplines collaborate actively with each other, with public health, and with other biomedical fields for the improvement of people's health. The collaboration involves both applied and theoretical knowledge. The increasing pace of collaboration among health sociology, health psychology and public health is already evident not only in the joint authorship of scientific publications in books and refereed journals, but also in the multidisciplinary composition of the faculty in schools of public health around the world. Indeed, high-ranking public health schools tend to have a greater proportion of faculty from different disciplines.

REFERENCES

Acton, G.S. and Zodda, J.J. (2005). Classification of psychopathology. Goals and methods in an empirical approach. *Theory & Psychology*, 15(3), 373–99.

Adams, R.M., Prelip, M.L., Glick, D.C., Donatello, I., Eisenman, M.D. (2017). Facilitating partnership with community- and faith-based organizations for disaster preparedness and response: results of a national survey of public health departments. *Disaster Medicine and Public Health Preparedness*, 12(1), 57–66.

Ailshire, J.A. and Burgard, S.A. (2012). Family relationships and troubled sleep among US adults: examining the influences of contact frequency and relationship quality. *Journal of Health and Social Behavior*, 53(2), 248–62.

Albery, I.P. and Munafò, M. (2008). *Key Concepts in Health Psychology*. London: Sage Publications.

Anderberg, Mats and Dahlberg, Mikael (2018). Gender differences among adolescents with substance abuse problems at Maria clinics in Sweden. *Nordic Studies on Alcohol and Drugs*, 35(1), 24–38.

Anderson, B.O., et al. (2011). Optimization of breast cancer management in low-resource and middle-resource countries: executive summary of the Breast Health Global Initiative consensus, 2010. *Lancet Oncology*, 12, 378–98.

Armelagos, G.J. and Harper, K.N. (2010). Emerging infectious diseases, urbanization, and globalization in the time of global warming. In: Cockerham, W.C. (ed.) *The New Blackwell Companion to Medical Sociology*. New York: Wiley-Blackwell, pp. 291–311.

Arnett, J.L. (2001). Clinical and health psychology: future directions. *Canadian Psychology/Psychologie Canadienne*, 42(1), 38–48.

Bangdiwala, S.I., et al. (2011). Public health education in Indian and China: history, opportunities, and challenges. *Public Health Reviews*, 33(1), 204–24.

Baum, F. (2016). *The New Public Health* (4th ed.). Oxford, UK: Oxford University Press.

Becker, M.H., Haefner, D.P., Kasl, S.V., Kirsh, J.P., Maiman, L.A., and Rosenstock, I.M. (2007). Selected psychosocial models and correlates of individual health-related behaviors. In: Weinman, J., Johnston, M., and Molloy, G. (eds.) *Health Psychology*. Vol 1. *Theoretical Models and Frameworks*. London: Sage Publications, pp. 3–26. [Originally published in 1977 in *Medical Care*, volume XV(5), 27–46]

Bendix, R. and Lipset, S.M. (eds.) (1966). *Class, Status, and Power. Social Stratification in Comparative Perspective* (2nd ed.). London: Routledge & Kegan Paul.

Benjaminy, S., Schepmyer, A., Illes, J. and Traboulsee, A. (2018). Resilience, trust, and civic engagement in the post-CCSVI era. *BMC Health Services Research*, 18, 366.

Betts, K, and Hodgson, A. (2018). *Advances in Causal Understanding for Human Health Risk-Based Decision-Making: Proceedings of a Workshop in Brief*. Washington, DC: National Academy of Sciences.

Boudon, R., (ed.) (1993). *Paul F. Lazarsfeld—On Social Research and Its Language*. Chicago, IL: The University of Chicago Press.

Boyer, B.A. and Paharia, M.I. (eds.) (2008). *Comprehensive Handbook of Clinical Health Psychology*. New York: John Wiley & Sons.

Brown, P., Lyson, M., and Jenkins, T. (2011). From diagnosis to social diagnosis. *Social Science & Medicine*, 73, 939–43.

Brown, T.H., O'Rand, A.M., and Adkins, D.E. (2012). Race-ethnicity and health trajectories: tests of three hypotheses across multiple groups and health outcomes. *Journal of Health and Social Behavior*, 53(3), 359–77.

Burchett, H., Umoquit, M., and Dobrow, M. (2011). How do we know when research from one setting can be useful in another? A review of external validity, applicability and transferability frameworks. *Journal of Health Services Research & Policy*, 16, 238–44.

Buset, S.L., Walter, C., Friedmann, A., Weiger, R., Borgnakke, W.S., and Zitzmann, N.U. (2016). Are periodontal diseases really silent? A systematic review of the effect on quality of life. *Journal of Clinical Periodontology*, 43, 333–44.

Calhoun, J.G., Wrobel, C.A., and Finnegan, J.R. (2011). Current state in the U.S. public health competence-based graduate education. *Public Health Reviews*, 33(1), 148–67.

Callis, A., et al. (2018). Lessons learned in clinical trial communication during an Ebola outbreak: the implementation of STRIVE. *Journal of Infectious Diseases*, 217(S1), S40–7.

Charmaz, K. and Rosenfeld, D. (2010). Chronic illness. In: Cockerham, W.C. (ed.) *The New Blackwell Companion to Medical Sociology*. New York: Wiley-Blackwell, pp. 312–33.

Chongo, M., Chase, R.M., Lavoie, J.G., Harder, H.G., and Mignone, J. (2018). The Life Story Board as a tool for qualitative research: interviews with HIV-positive indigenous males. *International Journal of Qualitative Methods*, 17, 1–10.

Christ, S.L., Fleming, L.E., Lee, D.J., Muntaner, C., Muenning, P.A., and Caban-Martinez, A.J. (2012). The effects of a psychosocial dimension of socioeconomic position on survival: occupational prestige and mortality among US working adults. *Sociology of Health & Illness*, 34(7), 1103–17.

Citro, C.F., Ilgen, D.R., and Marrett, C.B. (eds.) (2003). *Protecting Participants and Facilitating Social and Behavioral Sciences Research*. Washington, DC: National Academies Press.

Cockerham, W.C. (2010). Health lifestyles—bringing structure back. In: Cockerham, W.C. (ed.) *The New Blackwell Companion to Medical Sociology*. Oxford, UK: Wiley-Blackwell, pp. 159–83.

Cockerham, W.C. (2017). *Medical Sociology* (14th ed.). New York: Routledge.

Cockerham, W.C. and Scambler, G. (2010). Medical sociology and sociological theory. In: Cockerham, W.C. (ed.) *The New Blackwell Companion to Medical Sociology*. Oxford, UK: Wiley-Blackwell, pp. 3–26.

Cole, K., Sim, F., and Hogan, H. (2011). The evolution of public health education and training in the United Kingdom. *Public Health Reviews*, 33(1), 87–104.
Conrad, P. and Barker, K.K. (2010). The social construction of Illness: key insights and policy implications. *Journal of Health and Social Behavior*, 51(S), S67–S79.
Cooper, C.R. and Denner, J. (1998). Theories linking culture and psychology: universal and community-specific processes. *Annual Review of Psychology*, 49, 559–84.
Coreil, J. and Dyer K.E. (2017). Social science contributions to public health: overview. In: Quah, S.R. and Cockerham, W.C. (eds.) *International Encyclopedia of Public Health* (2nd ed), (Vol–6). Oxford: Academic Press, pp. 599–616.
Cornwell, E.Y. and Waite, L.J. (2012). Social network resources and management of hypertension. *Journal of Health and Social Behavior*, 53(2), 215–31.
Council on Education for Public Health (CEPH) (2011a). *Accreditation Criteria—Schools of Public Health. Amended June 2011*. Washington, DC: CEPH.
Council on Education for Public Health (CEPH) (2011b). *Accreditation Criteria –Public Health. Programs. Amended June 2011*. Washington, DC: CEPH.
Crosnoe, R., Benner, A.D., and Schneider, B. (2012). Drinking, socioemotional functioning, and academic progress in secondary school. *Journal of Health and Social Behavior*, 53(2), 150–4.
Curtis, E.A. and Dixon, M.S. (2005). Family therapy and systematic practice with older people: where are we now?. *Journal of Family Therapy*, 27, 43–64.
Datta, S. (2018). Emerging dynamics of evidence and trust in online user-to-user engagement: the case of 'unproven' stem cell therapies. *Critical Public Health*, 28(3), 352–62.
Daw, J., Shanahan, M., Mullan Harris, K., Smolen, A., Haberstick, B., and Boardman, J.D. (2012). Genetic sensitivity to peer behaviors: 5HTTLPR, smoking, and alcohol consumption. *Journal of Health and Social Behavior*, 54(1), 92–108.
De Leon, P.H., Frank, R.G., and Wedding, D. (1995). Health psychology and public policy: the political process. *Health Psychology*, 14(6), 493–9.
De Vibe, M., Bjørndal, A., Fattah, S., Dyrdal, G.M., Halland, E., and Tanner-Smith, E.E. (2017). *Mindfulness-Based Stress Reduction (MBSR) for Improving Health, Quality of Life and Social Functioning in Adults: A Systematic Review and Meta-Analysis*. Oslo, Norway: The Campbell Collaboration.
Diaz-Quijano, F.A., Martinez-Vega, R.A., Rodriguez-Morales, A.J., Rojas-Calero, R.A., Luna-Gonzalez, M.L., and Diaz-Quijano, R.G. (2018). Association between the level of education and knowledge, attitudes and practices regarding dengue in the Caribbean region of Colombia. *BMC Public Health*, 18(143), 1–10
Do Nascimiento-Bezerra, Y.R. and Souza-Feitosa, M.Z. (2018). The affection of the community healthcare agent in the territory: a study with affective maps. *Ciencia & Salude Colectiva*, 23(3), 813–22.
Doyal, L. (2014). The art of medicine—a decade of researching the social aspects of HIV and AIDS. *The Lancet*, 384, 2102–3.
Durkheim, E. (1895/1964). *The Rules of Sociological Method*. New York: Free Press.
Durkheim, E. (1897/1951). *Suicide*. New York: Free Press.
Ellenberg, S.S., et al. (2018). Rigorous clinical trial design in public health emergencies is essential. *Clinical Infectious Diseases*, 66(9), 1467–9.
Elling, R.H. and Sokołoska, M. eds. (1978). *Medical Sociologists at Work*. New Brunswick, NJ: Transaction Books.
Etkin, N.L., Baker, J.D. and Busch, J.N. (2017). Cultural factors influencing therapeutic practice. In: Quah, S.R. and Cockerham, W.C. (eds.) *International Encyclopedia of Public Health* (2nd ed.), (Vol–2). Oxford: Academic Press, pp. 198–201.
Fiebach, N.H., Rao, D., and Hamm, M.E. (2011). A curriculum in health systems and public health for internal medicine residents. *American Journal of Preventive Medicine*, 41(4S3),

S264–9.
Fox, R.C. (1989). *The Sociology of Medicine. A Participant Observer's View*. Englewood Cliffs, NJ: Prentice-Hall.
Freeman, H.E. and Levine, S. (1989). *Handbook of Medical Sociology* (4th ed.). Englewood Cliffs, NJ: Prentice-Hall.
Freidson, E. (1988). *Profession of Medicine. A Study of the Sociology of Applied Knowledge*. Chicago, IL: University of Chicago Press.
Friedman, H.S. (2002). *Health Psychology* (2nd ed.). Upper Saddle River, NJ: Prentice-Hall.
Gale, T.C.E., Lambe, P.L., and Roberts, M.J. (2017). Factors associated with junior doctors' decisions to apply for general practice training programs in the UK: secondary analysis of data from the UKMED project. *BMC Medicine*, 15(220), 1–14.
Gebbie, K., et al. (2007). The National Board of Public Health Examiners" Credentialing Public Health Graduates. *Public Health Report*s, 122(4), 435–40.
Gerhardt, U. (1989). *Ideas about Illness. An Intellectual and Political History of Medical Sociology*. New York: New York University Press.
Germov, J. (2005). Managerialism in the Australian public health sector: towards the hyper-rationalisation of professional bureaucracies. *Sociology of Health & Illness*, 27(6), 738–58.
Goertzen, J.R. (2008). On the possibility of unification: the reality and nature of the crisis in psychology. *Theory & Psychology*, 18(6), 829–52.
Gordon, L.J. and McFarlane, D.R. (1996). Public health practitioner incubation plight: following the money trail. *Journal of Public Health Policy*, 17(1), 59–70.
Greenwood, J.D. (2009). Materialism, strong psychological continuity, and American scientific psychology. *Theory & Psychology*, 19(4), 545–64.
Grollman, E.A. (2012). Multiple forms of perceived discrimination and health among adolescents and young adults. *Journal of Health and Social Behavior*, 53(2), 199–214.
Hamoen, E.H.J., De Rooji, M., Witjes, J.A., Barentsz, J.O., and Rovers, M.M. (2015). Measuring health-related quality of life in men with prostate cancer: a systematic review of the most used questionnaires and their validity. *Urologic Oncology*, 33(2), 69e19–e28.
Hargrove, M.B., Quick, J.C., Nelson, D.L., and Quick, J.D. (2011). The theory of preventive stress management: a 33-year review and evaluation. *Stress and Health*, 27, 182–93.
Harmon-Jones, E., Harmon-Jones, C., and Levy, N. (2015). An action-based model of cognitive-dissonance processes. *Current Directions in Psychological Science*, 24, 3, 184–89.
Harvey, T.S. (2017). Cultural issues and linguistic complications in health communication. In: Quah, S.R. and Cockerham, W.C. (eds.) *International Encyclopedia of Public Health* (2nd ed.), (Vol–2). Oxford: Academic Press, pp. 202–7.
Hauser, R.M., Weinstein, M., Pool, R., and Cohen, B. eds. (2010). *Conducting Biosocial Surveys: Collecting, Storing, Accessing, and Protecting Biospecimens and Biodata*. Washington DC: National Academies Press.
Hedges, L.V., Johnson, W.D., Semaan, S., and Sagolow, E. (2002). Theoretical issues in the synthesis of HIV prevention research. *Journal of Acquired Immune Deficiency Syndrome*, 30 (Supplement 1), S8–S14.
Hynes, M.M. (2011). Sociology, health data, and health equity: one state agency as a site of social change. *Current Sociology*, 60(2), 161–77.
Idler, E.L. (2010). Health and religion. In: Cockerham, W.C. (ed.) *The New Blackwell Companion to Medical Sociology*. New York: Wiley-Blackwell, pp. 133–58.
Imperato, P.J., LaRosa, J.H., Kavaler, F., Benker, K., and Schechter, L. (2011). The establishment of the School of Public Health at the State University of New York, Downstate Medical Center: the first nationally accredited School of Public Health in a public university in New York City. *Journal of Community Health*, 36,–1–13.

Jones, I.R., Papacosta, O., Whincup, P.H., Wannamethee, S.G., and Morris, R.W. (2011). Class and lifestyle 'lock-in' among middle-aged and older men: a multiple correspondence analysis of the British Regional Heart Study. *Sociology of Health & Illness*, 33(3), 399–419.

Kawachi, I., Kennedy, B.P., and Wilkinson, R.G. (eds.) (1999). *Income Inequality and Health. The Society and Population Health Reader*. New York: The New Press.

Kenkel, M.B., DeLeon, P.H., Mantell, E.O., and Steep, A.E. (2005). Divided no more: psychology's role in integrated health care. *Canadian Psychology/Psychologie Canadienne*, 46(3), 189–202.

Kienhues, D. and Bromme, R. (2012). Exploring laypeople's epistemic beliefs about medicine—a factor-analytic survey study. *BMC Public Health*, 12, 759–74.

Kim, H.H.S. (2018). Particularized trust, generalized trust, and immigrant self-rated health: cross-national analysis of World Values Survey. *Public Health*, 158, 93–101.

Krieger, N. (1987). Shades of difference: theoretical underpinnings of the medical controversy on Black/White differences in the United States, 1830–1870. *International Journal of Health Services*, 17, 259–78.

Lahelma, E. (2010). Health and social stratification. In: Cockerham, W.C. (ed.) *The New Blackwell Companion to Medical Sociology*. New York: Wiley-Blackwell, pp. 71–96.

Lambdin, C. (2012). Significance tests as sorcery: science is empirical—significance tests are not. *Theory & Psychology*, 22, 67–90.

Laxminarayan, R., et al. (2006). Advancement of global health: key messages from the Disease Control Priorities Project. *The Lancet*, 367, 1193–208.

Lazarsfeld, P.F. (1968). Foreword. In: Rosenberg, M. (ed.) *The Logic of Survey Analysis*. New York: Basic Books, pp. vii–x.

Lincoln, Y.S. and Guba, E.G. (1985). *Naturalistic Inquiry*. Beverly Hills, CA: Sage Publications.

Link, B.G. and Phelan, J. (1995). Social conditions as fundamental causes of disease. *Journal of Health and Social Behavior*, 35, 80–94.

Looren de Jong, H. (2010). From theory construction to deconstruction: the many modalities of theorizing in psychology. *Theory & Psychology*, 20(6),745–63.

Luecken, L.J. and Gallo, L.C. (eds.) (2008). *Handbook of Physiological Research Methods in Health Psychology*. Los Angeles, CA: Sage Publications.

Lund, R., Modvig, J., Due, P., and Holstein, B.E. (2000). Stability and change in structural social relations as predictor of mortality among elderly women and men. *European Journal of Epidemiology*, 16(12),1087–97.

Marmot, M. (2018). Social causes of the slowdown in health improvement. *Journal of Epidemiology and Community Health*, 72(5), 359–60.

Massé, R. and Moloughney, B. (2011). New era for schools and programs of public health in Canada. *Public Health Reviews*, 33(1), 277–88.

Mechanic, D. and Tanner, J. (2007). Vulnerable people, groups, and populations: societal view. *Health Affairs*, 26, 1220–30.

Mechanic, D. and McAlpine, D.D. (2010). Sociology of health care reform: building on research and analysis to improve health care. *Journal of Health and Social Behavior*, 51(S), S147–59.

Mechlenborg K.T. and Grønkjaer, M. (2018). Focus groups as social arenas for the negotiation of normativity. *International Journal of Qualitative Methods*, 17, 1–11.

Miller, D.C. and Salkind, N.J. (2002). *Handbook of Research Design & Social Measurement* (6th ed.). London: Sage Publications.

Nam, C.B. and Powers, M.G. (1982). *The Socioeconomic Approach to Status Measurement (with a Guide to Occupational and Socioeconomic Scales)*. Houston: Cap and Gown Press.

National Academies of Sciences, Engineering and Medicine (NASEM) (2018). *A Smarter National Surveillance System for Occupational Safety and Health in the 21st Century. A Consensus Study Report*. Washington, DC: The National Academies Press

Neuman, W.L. (2000). *Social Research Methods. Qualitative and Quantitative Approaches*. Boston, MA: Allyn and Bacon.
Norcross, J.C. and Karpiak, C.P. (2012). Clinical psychologists in the 2010s: 50 years of the APA Division of Clinical Psychology. *Clinical Psychology: Science and Practice*, 19(1), 1–12.
Nuriddin, A., et al. (2018). Trust, fear, stigma and disruptions: community perceptions and experiences during periods of low but ongoing transmission of Ebola virus disease in Sierra Leone, 2015. *BMJ Global Health*, 3, e000410.
Nyqvist, F., Pape, B., Pellfolk, T., Forsman, A.K., and Wahlbeck, K. (2014). Structural and cognitive aspects of social capital and all-cause mortality: a meta-analysis of cohort studies. *Social Indicators Research*, 116(2), 545–66.
Otok, R., Levin, I., Sitko, S., and Flahault, A. (2011). European accreditation of public health education. *Public Health Reviews*, 33(1), 30–8.
Paccaud, F., Weihofen, A., and Nocera, S. (2011). Public health education in Europe: old and new challenges. *Public Health Reviews*, 33(1), 66–86.
Parsons, T. (1937). *The Structure of Social Action*. Volumes 1 & 2. New York: McGraw-Hill.
Pearlin, L.I. (1989). The sociological study of stress. *Journal of Health and Social Behavior*, 30, 241–56.
Perez, M.A. and Luquis, R.R. (2008). *Cultural Competence in Health Education and Health Promotion*. San Francisco, CA: Jossey-Bass.
Phelan, J.C., Link, B.C., and Tehranifar, P. (2010). Social conditions as fundamental causes of health inequalities: theory, evidence, and policy implications. *Journal of Health and Social Behavior*, 51, S28–S40.
Pino, P. and Solimano, G. (2011). The School of Public Health at the University of Chile: origins, evolution, and perspectives. *Public Health Reviews*, 33(1), 315–22.
Pletzer, J.L., Balliet, D., Joireman, J., Kuhlman, D.M., Voelpel, S.C., and Van Lange, P.A.M. (2018). Social value orientation, expectations, and cooperation in social dilemmas: a meta-analysis. *European Journal of Personality*, 32, 62–83.
Poulin, M.J. and Haase, C.M. (2015). Growing to trust: evidence that trust increases and sustains well-being across the life span. *Social Psychology and Personality Science*, 6(6), 614–621.
Praetorius, N. (2003). Inconsistencies in the assumptions of constructivism and naturalism: an alternative view. *Theory & Psychology*, 13(4), 511–39.
Quadagno, J. (2010). Institutions, interest groups, and ideology: an agenda for the sociology of health care reform. *Journal of Health and Social Behavior*, 51, 125–36.
Quah, S.R. (2003). Traditional healing systems and the ethos of science. *Social Science and Medicine*, 57(10), 1997–2012.
Quah, S. R. (2007a). On trust and health consensus-building in the governance of epidemics. In: Quah, S.R. (ed.) *Crisis Preparedness: Asia and the Global Governance of Epidemics*. Stanford, CA: Shorenstein Asia-Pacific Research Centre, Stanford University, pp. 113–33.
Quah, S.R. (2007b). Public image and governance of epidemics: comparing HIV/AIDS and SARS. *Health Policy*, 80, 253–72.
Quah, S.R. (2008). In pursuit of health: pragmatic acculturation in everyday life. *Health Sociology Review*, 17(4), 419–21.
Quah, S.R. (2010). Health and culture. In: Cockerham, W.C. (ed.) *The New Blackwell Companion to Medical Sociology*. New York: Wiley-Blackwell, pp. 27–46.
Quah, S.R. (2011). Gender and the burden of disease in ten Asian countries: an exploratory analysis. *Asia Europe Journal*, 8, 499–512.
Quah, S.R. and Cockerham, W.C. (eds.) (2017). *International Encyclopedia of Public Health* (2nd ed.), (Vols 1–7). Oxford: Academic Press.
Ragin, C.C. and Becker, H.S. (eds.) (1992). *What is a Case? Exploring the Foundations of Social Inquiry*. Cambridge, UK: Cambridge University Press.

Raven, J., Wurie, H., and Witter, S. (2018). Health workers' experiences of coping with the Ebola epidemic in Sierra Leone's health system: a qualitative study. *BMC Health Services Research*, 18, 215.

Reid, M. (1976). The development of medical sociology in Britain. *Discussion Papers in Social Research No 13*. Glasgow, UK: University of Glasgow.

Ritzer, G. (1992). *Sociological Theory*. 3rd Edition. New York: McGraw-Hill.

Rivard, J.C. and Butler, A.S. (2017). *Lessons Learned from Diverse Efforts to Change Social Norms and Opportunities and Strategies to Promote Behavior Change in Behavioral Health: Proceedings of Two Workshops*. Washington, DC: The National Academy of Sciences.

Romanow, R.J. and Marchildon, G.P. (2003). Psychological services and the future of health care in Canada. *Canadian Psychology/Psychologie Canadienne*, 44(4), 283–95.

Roney, J.G. Jr (1960). Social sciences in the teaching of public health. *Journal of Health and Human Behavior*, 1(1), 47–52.

Rosenstock, L., Heksing, K., and Rimer, B.K. (2011). Public health education in the United States: then and now. *Public Health Reviews*, 33(1), 39–65.

Rosich, K.J. and Hankin, J.R. (2010). What do we know? Key findings from 50 years of medical sociology—executive summary. *Journal of Health and Social Behavior*, 51, S1–S9.

Sanders, M.R. (2010). Adopting a public health approach to the delivery of evidence-based parenting interventions. *Canadian Psychology*, 51(1), 17–23.

Schwartz, H. and Jacobs, J. (1979). *Qualitative Sociology. A Method to the Madness*. New York: The Free Press.

Seow, A. (2017). International dietary guidelines. In: Quah, S.R. and Cockerham, W.C. (eds.) *International Encyclopedia of Public Health* (2nd ed.), (Vol 4). Oxford: Academic Press, pp. 312–15.

Shryock, E.H. (1974). *The Development of Modern Medicine. An Interpretation of the Social and Scientific Factors Involved*. Madison, WI: University of Wisconsin Press.

Smith, S. (1866). *The Common Nature of Epidemics, and their Relation to Climate and Civilization*. Philadelphia, PA: J.B. Lippincott.

Smith, M., Saunders, R., Stuckhardt, L., and McGinnis, J.M. (2012). *Best Care at Lower Cost: The Path to Continuously Learning Health Care in America*. Washington, DC: The National Academies Press.

Starr, P. (1982). *The Social Transformation of American Medicine. The Rise of a Sovereign Profession and the Making of a Vast Industry*. New York: Basic Books.

Stevens, F. (2010). The convergence and divergence of modern health care systems. In: Cockerham, W.C. (ed.) *The New Blackwell Companion to Medical Sociology*. New York: Wiley-Blackwell, pp. 434–54.

Stevenson, C., McNamara, N., and Muldoon, O. (2014). Stigmatised identity and service usage in disadvantaged communities: residents', community workers' and service providers' perspectives. *Journal of Community & Applied Psychology*, 24, 453–66.

Stone, G.C. (1991). An international review of the emergence and development of health psychology. In: Jansen, M.A., and Weinman, J. (eds.) *The International Development of Health Psychology*, Philadelphia, PA: Hardwood Academic Publishers, pp. 3–17.

Strauss, R. (1957). The nature and status of medical sociology. *American Sociological Review*, 22(2), 200–4.

Sutton, S. (2007). Predicting and explaining intentions and behavior: how well are we doing? In: Weinman, J., Johnston, M., and Molloy, G. (eds.) *Health Psychology*. Vol 3. *Methods and Measurement*. London: Sage Publications, pp. 3–21.

Thapa, S., Hannes, K., Bive, A., Bhattarai, S., and Mathei, C. (2018). Theorizing the complexity of HIV disclosure in vulnerable populations: a grounded theory study. *BMC Public Health*, 18(162), 1–13.

Thirlaway, K. and Upton, D. (2009). *The Psychology of Lifestyle: Promoting Healthy Behaviour*. London: Routledge.
Thois, P.A. (2010). Stress and health: major findings and policy implications. *Journal of Health and Social Behavior*, 51(5), 541–53.
Thomas, C. (2012). Theorising disability and chronic illness: where next for perspectives in medical sociology? *Social Theory & Health*, 10(3), 209–28.
Thomson, M.E., Harver, A., and Eure, M. (2009). A model for integrating strategic planning and competence-based curriculum design in establishing a public health programme: the UNC Charlotte experience. *Human Resources for Health*, 71(7), 1–10.
Thorlindsson, T. (2011). Bring in the social context: towards an integrated approach to health promotion and prevention. *Scandinavian Journal of Public Health*, 39(Supplement 6), 19–25.
Timmermans, S. and Oh, H. (2010). The continued social transformation of the medical profession. *Journal of Health and Social Behavior*, 51(S), S94–S106.
Trickett, E.J. (2009). Community psychology: individuals and interventions in community context. *Annual Review of Psychology*, 60, 395–419.
Tulchinsky, T.H. and McKee, M. (2011). Education for a public health workforce in Europe and globally. *Public Health Reviews*, *33*(1), 7–15.
Umberson, D. and Montez, J.K. (2010). Social relationships and health: a flashpoint for health policy. *Journal of Health and Social Behavior*, 51(S), S54–S66.
Valberg, M., Stensrud, M.J., and Aelen, O.O. (2018). The surprising implications of familial association in disease risk. *BMC Public Health*, 18(135), 1–9.
Vanderminden, J. and Potter, S.J. (2010). Challenges to the doctor-patient relationship in the Twenty-First Century. In: Cockerham, W.C. (ed.) *The New Blackwell Companion to Medical Sociology*. New York: Wiley-Blackwell, pp. 355–72.
Vogt, W.P. (2007). *Quantitative Research Methods for Professionals*. Boston, MA: Pearson.
Waters, E., Gibbs, L., Riggs, E., Priest, N., Renzaho, A., and Kulkens, M. (2017). Cultural competence in public health. In: Quah and W.C. Cockerham (eds.) *International Encyclopedia of Public Health* (2nd Ed.), (Vol 2), pp. 181–6. Oxford: Academic Press.
Weber, M. (1921/1968). *Economy and Society.* (3 Volumes). Totowa, NJ: Bedminster Press.
Weinman, J., Johnston, M., and Molloy, G. (eds.) (2007a). *Health Psychology. Vol 1. Theoretical Models and Frameworks*. London: Sage Publications.
Weinman, J., Johnston, M., and Molloy, G. (eds.) (2007b). *Health Psychology. Vol 3. Methods and Measurement*. London: Sage Publications.
WFME (2003). *Basic Medical Education. WFME Global Standards for Quality Improvement*. Copenhagen: WFME.
Whyte, J.D., Anderson, I.P.S., and Crengle, S. (2017). Populations at special health risk: indigenous populations. In: Quah, S.R. and Cockerham, W.C. (eds.), *International Encyclopedia of Public Health* (2nd ed.), (Vol 5). Oxford: Academic Press, pp. 573–82.
Williams, D.R. and Sternthal, M. (2010). Understanding racial-ethnic disparities in health: sociological contributions. *Journal of Health and Social Behavior*, 51(S), S15–S27.
World Health Organization (WHO) (2004). WHOQOL-BREF. Geneva: WHO.
World Health Organization (WHO) (2005). *Accreditation of Medical Education Institutions. Report of a Technical Meeting*. Schæffergården, Copenhagen: WHO-WFMF [World Health Organization-World Federation for Medical Education] Task Force on Accreditation.
World Health Organization (WHO) (2017). *Model Disability Survey—Survey Manual*. Geneva: WHO.
World Health Organization (WHO) (2018). *Integrating Social Science Interventions in Epidemic, Pandemic and Health Emergencies Response: Report of the Informal Consultation*. Geneva: WHO.

Wright, E.R. and Perry, B.L. (2010). Medical sociology and health services research: past accomplishments and future policy challenges. *Journal of Health and Social Behavior*, 51(S), S107–19.

Yost-Dubrow, R. and Dunham, Y. (2018). Evidence for a relationship between trait gratitude and prosocial behaviour. *Cognition and Emotion*, 32, 397–403.

Zaslavsky, A.M. and Epstein, A.M. (2005). How patients' sociodemographic characteristics affect comparisons of competing health plans in California on HEDIS® quality measures. *International Journal of Quality in Health Care*, 17(1), 67–74.

---

[1] The name of the ISA Research Committee was changed from 'Medical Sociology' to 'Sociology of Health' in 1986 but, the essential nature of the discipline did not change, and the work and objectives of the ISA Research Committee remained the same.

[2] Sociological research over the past six decades around the world has demonstrated with systematic empirical evidence a universal phenomenon first hinted by French sociology pioneer Henri de Saint-Simon in 1819: different occupations are allocated different levels of prestige or deference and rewarded accordingly financially (from low paid, unskilled occupations such street cleaners or bricklayers, to high salary, highly skilled occupations such as neuroscientists). Occupational prestige is measured by means of population surveys whereby each person from a representative sample of the total population is asked to rate in a five-point scale (5 'excellent standing'; 4 'good standing'; 3 'average standing'; 2 'below average standing'; and 1 'poor standing') a comprehensive list of occupations in their society.

6.2

Sexuality and public health

Richard Parker, Jonathan Garcia, Miguel Muñoz-Laboy, Marni Sommer, and Patrick Wilson

Introduction

Over the course of the past 25 to 30 years, there has been an explosion of public health research and programmes focusing on sexuality. Never before in the history of public health have sex and sexual issues been seen as more worthy of attention. In communities around the world, researchers, practitioners, and policymakers have sought to develop new knowledge, to design new programmes and policies focusing on sexual health or sexuality-related health issues.

This recent public health interest in sexuality contrasts significantly with the way in which sexuality was addressed in earlier times. In the late nineteenth and early twentieth centuries, sexuality was considered at best to be a marginal concern to the health of individuals, with attention focused primarily on seeking to control what at the time were called venereal diseases (VDs) (Brandt 1987). Such diseases were characterized as indicators of moral failures, and the primary concern seemed to be with protecting mainstream society from the moral threat posed by such diseases.

Over the course of the mid-twentieth century, this situation gradually changed, as growing interest began to focus not so much on sexually transmitted infections, but on the wider range of what might be thought of as ‘normal’ sexual behaviours of both men and women—less with the goal of protecting them from disease than to seek to promote their health and well-being. This focus reached its zenith in the late 1940s and the early 1950s with the publication of the Kinsey studies in the United States (Kinsey et al. 1948; Kinsey et al. 1953). Yet the significant controversy that surrounded the publication of the Kinsey reports, especially in the cold war political climate of the 1950s quickly led to an almost total curtailment of

work on (and financial support for) the topic of sexuality during the ensuing decades (Corrêa et al. 2008).

It was only in the wake of the emerging HIV/AIDS epidemic in the early 1980s, the consequences of this long-term neglect of public health work focusing on sexuality became apparent. As the devastating proportions of HIV/AIDS—a global epidemic transmitted primarily through sexual contact—began to be perceived, the urgent need for greater understanding of sexual behaviour and for innovative new approaches to sexual health promotion quickly became evident. Thus, a new wave of public health research, intervention, and sexual health promotion began and continued to gain strength over the course of the 1990s and the 2000s.

This chapter seeks to provide an overview of this rapidly growing body of work in public health. It describes the initial public health response to sexuality in the context of HIV and AIDS, as well as the ways in which that response has been gradually broadened over time in order to provide a more comprehensive approach to sexual health and well-being. It also focuses on both the local and the global dimensions of this work, in both developed and developing countries, and as much in the work of local communities struggling to respond to the needs of their own populations, as well as on the part of a range of international agencies that are increasingly seeking to address a range of challenges to sexual health.

The impact of AIDS

The emergence of the HIV and AIDS epidemic in the early 1980s led to a major transformation in the ways in which public health has addressed issues of sexuality. During the early years of the HIV epidemic, one of the most frequently repeated laments of public health officials was the lack of data on sexual behaviour and the need to rely on findings from studies, such as the work of Kinsey and his colleagues in the United States, that were nearly 50 years out of date. The urgent need to respond to the epidemic with more contemporary knowledge on sexual behaviour provided the justification for a massive increase in research funding for work on sexuality among diverse population groups, and made the development of public health interventions to reduce the risk of sexual transmission of HIV an urgent priority (Turner et al. 1989).

Driven by epidemiological concerns, research on sexuality in relation to HIV and AIDS was initially developed in medical institutions, and was shaped by Western biomedical assumptions. Much of the early research carried out in response to the epidemic focused on surveys of risk-related sexual behaviours and on the knowledge, attitudes, and practices linked to risk of HIV infection. Most studies collected quantitative data on numbers of sexual partners, specific sexual practices, sexually transmitted infections, and similar topics related to transmission of HIV. Behavioural research sought to build a scientific foundation for public health policies and intervention programmes that would more effectively prevent HIV infection (Parker et al. 1991).

Very early in the epidemic, however, critical public health and social science research also highlighted the extent to which HIV and AIDS needed to be understood as what Paula Treichler described as ‘an epidemic of signification’ (Treichler 1988). Such work emphasized the ways in which complex cultural meanings shaped social understandings of the epidemic and surrounded it in deeply rooted forms of stigma, discrimination and denial (Parker and Aggleton 2003). It highlighted the importance of critically examining epidemiological and biomedical categories as well as popular understandings related to the epidemic, and emphasized the need to investigate local sexual cultures and the social organization of sexual practices in different social and cultural contexts (Parker 2001).

In early work on HIV among gay and bisexual men, for example, it became clear that many men who have sex with other men may not identify themselves as homosexual on the basis of this behaviour, and that Western biomedical categories such as ‘homosexuality’, ‘bisexuality’, or ‘heterosexuality’ might not adequately describe the lived sexual experience of many individuals. Increasingly, descriptive categories such as ‘men who have sex with men’ (or ‘MSM’) were adopted both in research and in public health practice in order to avoid imposing culturally inaccurate or inappropriate labels and classifications. This led, in turn, to important work rethinking the relationship between sexual orientation, sexual identity, and sexual behaviour. Based on this work, it became evident that assumptions about the relationship between sexual desire, sexual behaviour, and sexual identity—in particular, the assumption that feeling desire necessarily led to enacting behaviour, and that enacted behaviour necessarily led to forming a

distinct sexual identity based on that behaviour—needed to be re-examined and investigated empirically (Dowsett 2003).

Work on homosexuality and MSM opened the way for a broader critique of other key categories that had been used in seeking to understand the epidemiology of HIV infection. Research on 'prostitution', for example, highlighted extensive cross-cultural diversity in the ways in which sex is exchanged for money, gifts, or favours, and emphasized the importance of sex work and sex workers, on the one hand, and other forms of transaction that might characterize many sexual relationships without constituting sex work or identifying those involved as sex workers or clients (de Zalduondo 1991). Similarly, while the female sexual partners of male injecting drug users (or the male sexual partners of female injecting drug users) might constitute meaningful epidemiological categories, they have little or no sociological significance, and members of groups identified through such categories have little or no sense of self-identity or group identification based on these circumstances, making them largely useless for purposes of health promotion and HIV prevention (Kane and Mason 1992). As a result of HIV and AIDS, then, the ways in which public health has approached issues of sexuality have changed significantly, and new understanding of sexual health promotion have been developed that have important consequences for other areas of sexual health, independent of HIV and AIDS.

Public health research on marginalized populations such as female sex workers and male and female injecting drug users and their sexual partners also focused growing attention on the broader synergy that exists in many societies between poverty, racial or ethnic discrimination, gender power inequality, and vulnerability to HIV infection (Farmer et al. 1996). It became increasingly clear that sexual behaviour is not somehow isolated or free-floating, separate from other forms of social experience and social structure. On the contrary, sexuality exists embedded in social hierarchies, and sexual behaviour can only be understood within the context of social processes. One of the lasting lessons of the HIV/AIDS epidemic for understanding sexuality in relation to public health more broadly was therefore that sexual risk behaviour must be understood not just as 'behavioural risk' but as 'social risk'—and, indeed, that the concept of 'risk' should perhaps be abandoned altogether in favour of a greater

emphasis on ‘social vulnerability’ that would be more useful in seeking to understand the social drivers of the epidemic (Parker et al. 2000).

This emphasis on the social determinants of the epidemic also called attention to the synergies that exist between issues and other social problems. For example, two of the major threats to the sexual health of populations in many countries across the globe are substance abuse and imprisonment. The connections between substance abuse and sexuality are complex. On the one hand, substance abuse and binge use of a substance such as alcohol have been associated with condomless sex and coercive or unwanted sexual encounters (O’Hara and Cooper 2015). On the other hand, alcohol, marijuana, opioids, and recreational drugs have been used by different populations to cope with the stressors associated with sexualities and gender expressions that do not conform to heteronormative gender and sexual norms (Slater et al. 2017). Furthermore, for injecting drug users, blood diseases such as HIV can be acquired through use of contaminated syringes or through sexual transmission; which is particularly concerning since sex in exchange for drugs or money or shelter have been associated with people with severe opioid and cocaine use disorders, including injecting drug users (Decker et al. 2012). Thus, there is a complex bidirectional association between substance abuse and the severity of substance use disorders, and populations’ sexual expressions and behaviours, and sexually transmitted infections. While high-risk substance consumption and substance abuse disorders are cause for concern in and of themselves, recent evidence suggests that substance abuse might be associated with the likelihood of sexual risk behaviour, HIV/sexually transmitted infection (STI) infections, late AIDS diagnosis, poor adherence to HIV/STI and substance abuse treatments (Strathdee et al. 2010).

Just as an emphasis on the social dimensions of risk and the multiple social factors that contribute to vulnerability have opened up important insights in our understanding of sexuality in relation to HIV, recent advances in the use of new biomedical approaches for preventing the sexual transmission of HIV infection have opened up new areas for research and intervention in relation to sexuality and public health. The emergence of pre-exposure prophylaxis (PrEP) as an HIV prevention tool for vulnerable communities globally has spurred an ongoing debate within public health regarding the best way to reduce new HIV infections. Efforts have been made to highlight the potential impact that biomedicine (including PrEP

and highly effective anti-HIV treatments) has to eradicate HIV. However, on the ground efforts to disseminate PrEP within marginalized and vulnerable communities have stalled. For example, one study of Black MSM, a group highly disproportionately impacted by HIV in the United States, found that less than one in four were aware of PrEP as a prevention tool (Eaton et al. 2015). Studies have suggested that for many MSM in the United States and globally, stigma related to HIV, lack of access to basic necessities and healthcare insurance, and negative perceptions of PrEP contribute to lack of uptake (Arreola et al. 2015; Garcia et al. 2016b). Additionally, even in cases where access is unhindered, many populations have difficulty adhering to anti-HIV drugs, reducing their potential for prevention and treatment (Wilson et al. 2016). Scholars in the social sciences have provided some context to the problems with placing too much faith in biomedicine to prevent HIV. Sexual minority communities have not always had a voice or choice in HIV prevention, which has led to failures in how public health researchers and practitioners have viewed effective intervention approaches. Understanding the complex personal, social, and political dimensions of using anti-HIV medication for prevention or treatment is critical to next steps in addressing the epidemic (Aggleton and Parker 2015; Auerbach and Hoppe 2015).

The various insights derived from work on sexuality in the context of HIV and AIDS have continued to shape public health approaches to sexuality over the course of the past two decades. They have been especially important in seeking to develop both research and programmes aimed at promoting sexual health among young people, heterosexual women and men, as well as sexual communities and populations outside the mainstream of society that often face special forms of vulnerability due to emerging forms social marginalization and exclusion. Work on sexuality in relation to population health has also heightened awareness of the complex ways in which sexual experience is socially regulated, with consequences that can be highly diverse. A broader focus not just on the absence of disease, but on a fuller realization of true health and well-being, in relation to sexuality as in relation to other aspects of health, requires a broader focus on the relationship between sexual health and sexual rights (Parker 2007). Notably, the work of public health has been shaped, through political agenda setting, cultural framings, and outright displays of indignation with the status quo stemming from AIDS social movements throughout the

world (Altman and Buse 2012). The AIDS social movement has been significant in addressing cultural, structural, and social determinants of health, including sexuality and sexual health (Parker 2011). In the following sections of this chapter, we will seek to address the development of work on these various dimensions of sexuality and public health, explore ways in which sexuality intersects with public health issues such as mental health and violence, and to highlight the ways in which this rapidly growing field of work is developing in the early twenty-first century.

Youth, sexuality, and gendered experience

Young people today represent the largest population across the globe, at 1.8 billion and growing (Patton et al. 2016). This suggests a critical need for attention to their developmental experiences, particularly in relation to sexuality and health. More recently the global health community has advocated for an expanded definition of 'adolescence' (10–24 years old) given the earlier onset of puberty in populations around the world, and the ways in which young people's developmental experience through the life course now expands into their 20s (Sawyer et al. 2018). The definition of 'youth' and the timing of a young person's first experience of sexuality are two important considerations for public health researchers and practitioners designing sexual health interventions. The biologically defined developmental stages of youth tend to place an important emphasis on puberty as the beginning stage of human sexuality Patton and Viner 2007). The onset of puberty brings with it numerous physiological changes and new social pressures for young people transitioning into young adulthood within any given society. Along with adjusting to the new biological changes that are experienced, including the development of secondary sexual characteristics, young people must adjust to the social responses of peers, family, and communities to their obvious physical and emotional sexual maturation. As societies shift from rural to urban, as the advent of formal schooling incorporates greater populations of young people, as influences from globalization and new technologies shape the experience of childhood and young adulthood, and as gender norms continue to evolve, the sexual socialization of young people is being altered and transformed (Blum 2007).

Girls and boys experience puberty differently biologically and socio-culturally. Puberty is not only experienced at different ages according to sex but may be marked by differing rites of passage (broadly defined) that depend on contextual and structural factors. Sexual differentiation moulded by social norms can shape the transformation of gendered expectations and sexual opportunities for girls and boys. The socializing norms influencing young peoples' development, perceived societal roles, and ultimately their restraint or engagement in sexual risk-taking behaviours are closely linked to local femininity and masculinity norms. In some cultures, girls are de-sexualized at a young age through procedures such as female genital excision, or what has been widely framed by the feminist human rights movement as female genital mutilation (Boyle 2005). Emphasis is generally placed on fertility as the representation of femininity. Boys also undergo sexual rites of passage, but these are often more 'empowering', signifying the acquisition of manhood and emphasizing the importance of virility. Whereas women are sometimes denied sexual pleasure through these procedures, boys that experience male circumcision are ascribed a role in their communities that associates their sexuality with power and responsibility (WHO 2000).

In recent decades, attention has shifted to the social construction of sexuality for understanding the ways in which young people's sexuality and gendered experience are influenced by the societal pressures and traditions around them (Freud 1994). This differentiation based on the social construction of sexuality has concrete implications for gendered disparities in human development related to both education and health. A gendered division of labour presents significant barriers to closing the gap in illiteracy and access to essential health services in resource-poor settings (Blum and Mmari 2006). In addition, girls are often married-off to older men, and boys are sometimes forced to migrate from rural to urban settings to find work. Both intergenerational sex and migration are factors that contribute to sexual risk.

As young people transition into young adulthood, sexual and reproductive health becomes a priority health issue. This is particularly true for young women in low-income countries, for whom adolescent pregnancy presents particular risks of maternal mortality and morbidity (such as obstetric fistula), infection with HIV and AIDS, and STIs (Hindin and Fatusi 2009). Adolescent pregnancy makes girls vulnerable to dropping out

of school and to mental health problems. Campaigns to prevent STIs and HIV are only beginning to target youth facing structural barriers. The risk of sexual violence is an additional challenge, particularly in societies where gendered power dynamics place young women in positions of little negotiation around sexual relations, and where masculinity norms promote aspects of virility and strength for young men.

In recent years, there has been growing attention to the onset of menstruation (menarche) as a window of opportunity for engaging girls prior to the onset of sexual pressures and relationships (Sommer et al. 2015). Although much of the focus on menstruation has emerged from the water and sanitation community, focused on the need to improve water and toilet facilities for menstruating girls in schools (Sommer and Sahin 2013), the sexual and reproductive health community increasingly recognizes the benefit of engaging girls and their families at menarche as an entry point for talking about sexual and reproductive health as a less taboo topic than contraception and family planning (Chandra-Mouli and Patel 2017). The Global Early Adolescent Study (GEAS), launched four years ago, is seeking to build the evidence on this critical developmental phase in 16 study sites (low-, middle-, and high-income countries) around the world (Saewyc 2017). In particular the focus is on understanding the gendered transition from childhood into adolescence, and the factors that promote healthy sexuality or risky sexual behaviours in adolescence.

However, it is also important to question the assumption that sexuality actually begins at puberty. Some studies have shown that sexual experimentation and initiation might begin before this biological landmark (Bankole et al. 2007). Masturbation is a major part of sexual experimentation, and access to pornography on the Internet is widespread. The international community has paid increasing attention to the effects of globalization on trafficking in children and sex work as major public health problems in the contemporary world. This protection of the rights of children is the responsibility of families, communities, and nation-states.

Much of sex education for youth is regulated through kinship networks and schools (Barker 2007). The family and ‘traditional values’ play an important role in sex education for boys and girls. The age of sexual debut continues to change around the world with influences of modernization and delays in the age of marriage secondary to the advent of formal schooling. The quality of communication between parents and youth influences the

ways in which youth are sexually socialized. Religious values that condemn condom use are particularly influential in shaping the types of messages youth receive from their families and in schools. International debates about what constitutes sex education are shaped by progressive advocates for comprehensive approaches (which might include the distribution of condoms to youth for STI and HIV prevention), and by conservatives who believe that sex education should be limited to teaching sexual abstinence until marriage.

More recently, debates over sex education have begun to discuss sexual orientation because at a very young age children begin to experience bullying, especially in schools and on the internet. School and cyberbullying can lead to increased sexual risk, mental health problems, and substance use among youth.

LGBT youth are among the most vulnerable populations globally. Although much of the literature that documents a higher risk of suicide, mental health problems, HIV/STIs, and violence for LGBT youth focuses on the US context, these patterns are documented worldwide (Grangeiro et al. 2012; Roa 2013). LGBT youth are prone to experiencing homelessness due to community and family rejection across cultural contexts (Blondeel et al. 2018). Many studies identify transactional sex as the main source of survival and vulnerability for LGBT youth that experience social rejection (Baral et al. 2015). Although international activists have managed to include protections for young women, it has been more difficult to address the intersectional needs of young transgender women and other sexual minority youth in transnational documents (Barroso 2014).

Structural interventions that target school policies are crucial for the formulation and implementation of comprehensive sex education to prevent bullying based on sexuality (Jones et al. 2009). In terms of the advancement of comprehensive sexuality education (CSE) globally, the UN conducted a review in 2015 which indicated that over 80% of countries now have policies in place to include such content in schools (UNESCO 2015). Such efforts are the result of growing evidence indicating the benefit to adolescent sexual and reproductive health from the inclusion of such content in schools. In 2013, 20 governments in eastern and southern African (ESA) committed to scaling up CSE in their educational systems (UNESCO 2018a), and in 2018, UNESCO updated their global technical guidance on the provision of CSE (UNESCO 2018b). However there continues to be

push back from many governments on the actual implementation of CSE, with growing conservative and religious influences restricting the delivery of such content to youth.

Heterosexual relations

Public health research on heterosexuality has been relatively limited when compared to the focus on non-heterosexuality. Kinsey's studies began to uncover the gradients of sexuality, as opposed to a dichotomous divide between heterosexuality and homosexuality—yet, the curiosity for deconstructing the accepted norms was in some ways aimed at showing the non-existence of an authentic heterosexuality. Historically, the field of sexuality research is laden with bias towards studying 'deviations' from the sexual norm rather than at seeking to more fully understand the ways in which dominant normative patterns of experience are produced and organized (Wilkinson and Kitzinger 1994). The concept of 'heteronormativity' has been used in critical public health to describe the ways in which men and women form both casual and steady partnerships, which are moulded by social and cultural norms (Kitzinger 2005). This concept has theoretical traction for understanding patterns of sexual behaviour with important implications for public health outcomes.

The fact that heterosexuality is employed as a form of normative domination at the same time that it has largely escaped the gaze of scientific inquiry calls our attention to how public health can benefit from highlighting recent explorations of heterosexuality. Reproduction, gendered relations of power, masculine/feminine ideologies, marriage, and parenting have more recently begun to be explored at the intersection between sexuality and public health. In large part as a result of the HIV and AIDS epidemic, there has also been growing attention to what can be described as the political economy of heterosexual relationships: the ways in which economic exchange relations shape heterosexual interactions far beyond the context of what is traditionally understood as prostitution or sex work (Hunter 2002). The examination of the transactional nature of heterosexual relations has received special attention in sub-Saharan Africa as a result of research seeking to respond to the need for a fuller understanding of vulnerability in relation to the heterosexual transmission of HIV (Leclerc-

Madlala 2003), but is clearly an important, albeit highly variable, part of the organization of partner relations between women and men across widely varying social and cultural contexts that requires increased attention in seeking develop the field of sexuality in public health more broadly.

Another key area of research that focuses on heterosexual relations has been related to human procreation. According to the World Health Organization, reproductive health relates to having a 'responsible, satisfying and safe sex life and … the capability to reproduce and the freedom to decide if, when and how often to do so', safely, effectively, and affordably, and through access to health services (WHO 2011b). This discourse comes after decades of public health policies in countries throughout the world, such as those that supported the mass sterilization of women, with a focus on population control as a means to reduce poverty and increase economic growth. These 'structural' determinants of health severely limited women's sexual agency and sexual decision-making.

Moreover, in heterosexual partnerships these asymmetric gendered power dynamics occur not only at a structural level, but they also have effects on individual health outcomes for both women and men. Ideologies of masculinity have vast implications for public health outcomes globally. Preserving a sense of heterosexual manhood and invincibility has been a barrier to seeking health services (Gough 2013). This has been particularly challenging for public health movements advocating for prostate cancer screening, and heterosexual men are often more averse to anally invasive prostate exams (Gray et al. 2002). The concept of 'machismo' or male chauvinism has often been used to analyse the effects of Latin American heterosexual masculinity and its implications for health. Although studies on machismo have associated this ideology with sexual risk behaviours and intimate partner violence, the concept of machismo has been criticized for stereotyping Latino men (Torres et al. 2002). A more complex understanding of Latin American and Latino male sexuality shows that sexual behaviours are dependent on social and interpersonal context. Same-sex socialization among heterosexual men (Kandel and Davies 1991) has been linked to higher levels of sexual risk taking with women, drug use, and sexual violence.

At the same time, femininity has also shown to have effects on the health of women. Forms of the ideal 'sexy' woman certainly vary across the globe, among cultures, and often depend on social factors such as race and

ethnicity. However, there is an overwhelming prevalence of this context-dependent sexualized ideal, which even in its multiform instantiations has impacts on how women perceive themselves. The epidemics of eating disorders are of the clearest examples of how heteronormative femininity affects women's health. Public health research has shown that heterosexual women are affected by eating disorders such as anorexia nervosa and bulimia more so than non-heterosexual women (Meyer et al. 2001). Heteronormative femininity is a 'social determinant' of women's sexual health according to the Healthy People 2020's definition, which lists the factors that limit the 'availability of resources to meet daily needs, such as educational and job opportunities, living wages, or healthful foods', such as violence and discrimination (Healthy People 2020).

Furthermore, heterosexual relations, procreation, and partnerships often result in men and women taking on the responsibility of parenthood. Although parenting is not solely a heteronomative phenomenon in the current world order, public health studies have focused on the negative effects of single motherhood and absent fathers on the sexual and mental well-being of younger generations (Ishida 2010). For many, maintaining the institution of *heteronomative* marriage as both religiously and socially stable is a solution to potential health problems associated with heterosexual relations (Mahoney et al. 2008). However, more critical public health research has associated ideologies associated heteronormative marriage with infidelity, intimate partner violence, and sexual risk (Hirsch et al. 2007).

Sexual diversity and alternative sexualities

Framing sexual diversity in public health is methodologically and epistemologically complicated, particularly because of how sexual desires, behaviours, and identities converge and diverge. Recent cutting-edge global health research has shown that practices of categorization and classification across the axes of sexuality, gender identity, gender performance, and sexual role have been obscured by simplified models presented by social epidemiologists (Garcia et al. 2016a; Parker et al. 2016). Addressing sexual *identities* might be easier in social and behavioural research because there is more reliability in measures when the respondent identifies either to

themselves or openly to society about being a lesbian, gay, bisexual, transgender, or queer (LGBTQ) individual. Alternative sexualities go beyond identity, however, and a person who practices same-sex sexual behaviours or sex work might not identify with a particular group culture. It becomes even more complicated when speaking of desire that does not necessarily translate into behaviour or identity formation, as we mentioned earlier. Research questions vary depending on these dimensions of sexual diversity and alternative sexualities, and public health has often collaborated closely with social movements to set research agendas.

The broader framework of sexual health has been employed to study health problems related to populations characterized by sexual diversity (Parker et al. 2004). Much of the public health research has focused on behaviour change towards the practice of safer sex (e.g. through condom use and reduction in sexual partners), especially considering the disproportionately high impact of HIV and AIDS on gay and bisexual communities (Chow et al. 2011). Scholars have also focused on the intersections of mental and sexual health, using theories such as ‘minority stress’ to examine factors that affect depression, anxiety, and healthcare utilization (Meyer 2003). Fear of disclosure of having same-sex desires or of practising alternative sexualities, for instance, can have important implications on the ways people interact with their healthcare providers (Klitzman and Greenberg 2002). Sexual diversity interacts with other social factors, such as race and poverty, which lead to differential access to healthcare and exposure to disease (Díaz et al. 2001).

In the sexual health literature, there have been disputes over how to create epidemiological categories that capture sexual diversity. The categories of MSM, men who have sex with men and women (MSMW), and women who have sex with women (WSW), behaviourally bisexual, bisexual, gay, and non-heterosexual are all used in the field of public health depending on the research question involved. Some of these categories, such as MSM, attempt to capture self-identified gays and bisexuals as well as those who practice same-sex sexualities but that do not self-identify as gay or bisexual. The category MSM might strip away the social and political implications of public health work on alternative sexualities, which has been a prominent claim among public health activists. Categorizing populations such as MSM and WSW might be understood as an attempt to objectively perform science. Those advocating for LGBTQ health are part

of a public health social movement that understands non-heterosexuality as having vast variation in sexual performance central to a notion of well-being that hinges on social rights. The epidemiological use of categories such as MSM and WSW, among others, does not fully take into consideration that 'community, networks, and relationships in which same-gender pairings mean more than merely sexual behaviour', and these categories have been criticized by the public social movement as heterosexist (Young and Meyer 2005).

There has been less attention to the particular health disparities that affect lesbians and WSW in the public health literature (Bauer and Jairam 2008). This gap is probably due to the lesser impact that HIV and AIDS has had on these populations, when compared to other 'risk' groups. This has led to gaps in research for women who are vulnerable to HIV, although other sexually transmitted infections such as HPV (which causes cervical cancer in women) have been addressed in non-heterosexual women populations (Charlton et al. 2011). Bisexual women and women who have sex with women and men (WSWM) have had some attention particularly regarding their sexual health. There are important concerns about the unmet needs of lesbians, bisexual women, and WSWM in healthcare due to stigma and discrimination when seeking health (Stevens and Hall 1988). This is especially the case when they fear disclosing their sexualities. Lesbians and bisexuals have been identified as having a higher rate of cigarette smoking than heterosexuals in national studies (Ryan et al. 2001). The health concerns of lesbians are more proximal to those of heterosexual women (relative to comparing gay men with heterosexual men). However, non-heterosexual women encounter barriers in sexual and reproductive health, as well as in making important life decisions, such as childbearing that make them particularly vulnerable to stigma and discrimination (Dahl et al. 2013).

The mental and physical health of transgender men and women has drawn special attention in public health, especially in discussion about sex modification procedures, the formal classification of transgender individuals in the DSM and the ICD, and their access to health more generally (Lombardi 2001; Schwend et al. 2018). Illegal hormone and silicone injection has been of particular concern due to the high cost of gender transformation (Wiessing et al. 1999). Moreover, stigma and discrimination are probably most prevalent for transgender individuals,

limiting their employment and education opportunities. These are social aspects of their gender identities that directly affect their risk for HIV and AIDS and mental health problems, especially because many transgender women resort to commercial sex work because of their limited opportunities (Abdullah et al 2012). In some countries, such as Brazil and Cuba, public health policy debates about such government funding for gender modification are being couched in human rights frameworks and public health safety (Gorry 2010).

Moreover, male and female sex workers are among the most vulnerable populations for HIV infection, sexual violence, and drug use (Mayhew et al. 2009). The public health implications of sex work (and the shift to applying a human rights framework) have been explored through a variety of methods, including qualitative and ethnographic techniques. Innovative public health research and interventions have studied sexual risk and mental health risks throughout the world (Kempadoo 2003). Sex work, intergenerational sex, and gift-exchange for sex has been documented throughout the world, and it has been linked to factors such as food insecurity (Weiser et al. 2007). Progressive public health has shown that sex workers can no longer be thought to be merely victims, but have come to take on the role of the human rights advocates throughout the world pointing to the elements of structural stigma and discrimination that led to their condition.

It is a challenge to study sexual desire and behaviour in populations where assuming an identity puts the person at sociocultural, political economic, and even legal disadvantages. Because of the stigma and discrimination associated with non-heterosexual sexualities, non-politicized alternative sexual communities and/or individuals are often ‘hidden’ or hard-to-reach (Heckathorn 1997).

Public health studies and interventions on hidden population are among the most methodologically innovative. Snowball sampling or chain-referral sampling are common techniques in achieving a theoretically defined number of participants in qualitative research in fields such as anthropology and sociology. Respondent-driven sampling is a type of chain-referral sampling in which study participants refer others to the study, beginning with key informants or seeds that can be randomly selected. This form chain-referral provides a dual incentive for the participant and weights the sample. Respondent-driven sampling lends itself to social network analysis

if referral systems and patterns are tracked. In fact, this can be particularly useful in public health to study sexual networks, networks of drug users or sex workers (Ramirez-Valles et al. 2005). More researchers are beginning to bridge the gap between qualitative and quantitative approaches, as well as forging interdisciplinary perspectives, precisely because they acknowledge the complexities of studying sexual diversity in public health. It is now more common for studies with large probability samples to explore multilevel research questions that require a complimentary approach of ethnography and survey methods.

Social regulation of sexuality

The social regulation of sexuality impacts how public health studies are conceptualized, designed, and executed. A range of formal and informal institutions and community structures define socially accepted expressions of sexuality, delineating cultural and social psychological norms. Social regulation might occur to enforce moral 'tradition', especially when relating sexuality with heteronormative reproductive practices.

Legal frameworks are regulations with implications on sexual practices as well as on gender roles. Throughout the world, laws determine the ways in which sexuality affect economic and political issues, such as through the institution of marriage, which marginalizes non-heterosexuals. Laws that ban sodomy or that deny adoption privileges based on sexual orientation target non-heterosexuals and punish the unmarried. Legal systems also dictate who has control over the female body, what women are allowed to wear, how women are passed down as inheritance through generations of men in some countries, the kinds of sexual pleasure women should be granted, and in some cases laws sanction gender-based violence.

Social networks might regulate sexuality through 'soft' mechanisms such as gossip or 'harder' mechanisms such as ostracism from communities if non-conforming sexualities or gender roles are expressed. Families and schools are among the most important institutions in directly providing sex education for children and adults throughout the world. While families are often thought of as spaces for social support, conservative family values that condemn homosexuality, for example, might create structural vulnerabilities to mental health problems, including depression, substance

use, and thoughts of suicide among LGBT youth. Schools and their political agendas on sexuality vary, having a major impact on the sexual formation of the individual and her/his interaction with peers. Policy interventions, such as those that ban bullying in schools, might have vast implications public health outcomes.

Community-based groups and civil society organization such as faith-based organization have varied on their position on sexuality and public health. Some religious traditions, such as Pentecostal evangelicals—which have been proliferating in numbers throughout the world in the twentieth and twenty-first centuries, have been the most adamant proponents of public health campaigns that promote the reduction of sex partners, faithful relationships, waiting until marriage, among other more conservative approaches to regulating sexuality at the expense of excluding non-heterosexuals and gender non-conforming individuals from their messages and flocks. In research on Afro-descendent religion in countries such as Brazil, some faith-based organizations have shown to embrace alternative sexualities and sexual diversity, as well as have a strong platform for women's equality and rights (Garcia and Parker 2011).

Also noteworthy, the Internet is among the most influential social/virtual institutions in contemporary life that shape and regulate sexuality and sexual norms. The pornography industry is ever growing and few public health studies have systematically analysed the association between exposure to pornography and sexual behaviour, some finding that increased exposure to pornography is associated with higher sexual risk taking (Eaton et al. 2012). Social networks and smart phone applications (Apps) such as Grindr (which uses advanced GPS systems to locate proximal partners in the vicinity of the individual essentially to meet up and have sex) are often used to meet new sexual partners and fulfil sexual fantasies that were once only imagined. Even though the Internet has been more liberating than limiting in its forms of social regulation, conservative civil society groups are ever present on the worldwide web with moralizing, normalizing, and moderating tendencies that often make people afraid of having excessive sexual pleasure.

The public health implications of the social regulation of sexuality are extensive. Regulatory groups and mechanisms include not only legal frameworks, but also the family, schools, religious groups/institutions, formal community leadership structures, informal structures that generate

social (and virtual) capital, just to name a few, which have been considered important in public health studies and interventions related to sexuality. When considering the possibility of targeting sexual behaviours through community and ecological interventions and studies, we call for new methodologies that analyse both heterosexual and non-heterosexual practices and pleasures. For example, cellular phone technologies and innovative forms of ethnographic mapping that examines trends on virtual social networks on the Internet could amount to a greater understanding of how sexual desires are enacted. These can also be used to target vulnerable populations that might be excluded from participation in civil society organizations through self-selection or through social sanctioning.

Sexual health and sexual rights

The notion of sexual rights is the counterpart to social regulation. The emphasis on sexual rights has helped to constitute a fairly new to the field of public health and this framework draws on interdisciplinary and intersectoral approaches to improving collective health outcomes in what can be considered a public health social movement. Even though international mobilization has dominated the use of the term ‘sexual rights’ as a rallying framework in global activism, sexual rights are often only regulated effectively as they are interpreted and understood on the local level in families, at schoolyards, by community-based groups and non-governmental organizations (Garcia and Parker 2006).

The language of sexual rights in public health stems greatly from the application of a human rights framework to addressing not only individual, but also social and structural vulnerabilities. Human rights are general understood as universal and fundamental rights deserved equally by all people. The International Conference on Population and Development (ICPD) held in Cairo in 1994 was a watershed moment for the transnational dissemination of discourses of sexual rights, although the official document of the ICPD did not include in it a clear mention of the term sexual rights after much heated debate among conservatives and progressives (Petchesky 2000).

Although public health leaders have partnered with activists throughout the world to push for sexual equality to be considered part of our

fundamental rights, it has been difficult to include sexual rights as part of framing documents at the level of the United Nations because of their contentious nature. International discourses about reproductive and sexual rights have been rooted primarily in Western individualism, even though notions of bodily integrity and personhood have a long history in non-Western contexts, in which they are more intertwined with social and cultural norms and collective rights (Corrêa and Petchesky 1994). To address vulnerabilities that are contingent on context, culture, and structures, it is paramount to include access to resources and the community-level conditions that might enable or restrict sexuality and health.

Moreover, sexual rights emerged under the umbrella of reproductive rights (which focused primarily on gender inequality, violence against women, population control, among other heterocentric issues), but health rights activists and LGBTQ groups have pushed the envelope to develop a concept that goes beyond human reproduction and addresses non-heterosexual health and rights-based issues (Miller 2000). Although sexual health is often confounded with reproductive health, there has been an attempt to differentiate the two by the World Health Organization. These include the right of all persons, free of coercion, discrimination, and violence, to: ‘a state of physical, mental and social well-being in relation to sexuality. It requires a positive and respectful approach to sexuality and sexual relationships, as well as the possibility of having pleasurable and safe sexual experiences’ (WHO 2011a).

While sexuality can include all of these dimensions, not all of them are always experienced or expressed. The ‘responsible’ exercise of human rights requires that all persons respect the rights of others. Within the framework elaborated by the World Health Organization (WHO), sexual and reproductive health are central aspects of life at all ages and encompasses sex, gender identities and roles, sexual orientation, pleasure, and reproduction. Sexuality is experienced and expressed in thoughts, fantasies, desires, beliefs, attitudes, values, behaviours, practices, roles, and relationships. Sexuality-related vulnerabilities highlight why these issues ought to be framed as inextricably tied to other human rights, political and civil liberties, and economic rights (Parker et al. 2014).

The protection from harm, be it for heterosexual or non-heterosexual populations, has been the main focus of sexual rights movement (Parker et

al. 2013). Nonetheless, sexual pleasure has also been conceptualized as an integral part of sexual rights by activists and also by progressive public health campaigns. One example of harnessing the power of the sensual in social marketing is the eroticization of the condom. Rather than showing the condom in isolation as a prophylactic barrier to natural sex, campaigns throughout the world have used sex symbols to entice safer sex—almost as a form of harm reduction in which people are assumed to be having sex in any case, so these campaigns make 'safer', sexy. Sexual rights frameworks that are emerging throughout the world are giving birth to new debates about gender and sexual equality that goes beyond reproduction (not to downplay the utmost importance of debates about new reproductive technologies that are more inclusive of non-heterosexuals).

Conclusion

During the closing decades of the twentieth century and the beginning of the twenty-first, the field of sexuality and public health has taken shape as one of the fastest growing areas of research and intervention in relation to population health. It is a field that has been characterized historically as highly polemical, and frequently marked by controversy and disagreement based on conflicts in relation to fundamental values and ethical positions. But at least since the emergence of the HIV and AIDS epidemic in the early 1980s, we have also realized that ignoring issues of sexuality, sweeping them under the carpet or leaving them hidden away in the closet, does nothing to promote the realization of true sexual health and well-being. This is especially worrisome when sexual norms, as we have pointed out, lead to grave disparities in health. On the contrary, it is only by bringing sexuality out into the open, by freeing it of stigma, prejudice, and discrimination, and by building a culture of sexual rights, freedoms, and protections, that we can truly advance the cause of sexual health and well-being. In this sense, it is clear that public health work on sexuality in the early twenty-first century both grows out of and is aligned with the long tradition of public health as itself a social movement—a movement based on principles of social justice and human rights, capable of promoting health for all based on a foundational set of ethical and political principles. Far from a peripheral

concern related to a taboo topic, few areas of work are so central to a public health agenda for the present.

REFERENCES

Abdullah, M.A., et al. (2012). Is social exclusion pushing the Pakistani Hijras (Transgenders) towards commercial sex work? A qualitative study. *BMC International Health and Human Rights*, 12(1), 32.

Aggleton, P. and Parker, R. (2015). Moving beyond biomedicalization in the HIV response: implications for community involvement and community leadership among men who have sex with men and transgender people. *American Journal of Public Health*, 105(8), 1552–8.

Altman, D. and Buse, K. (2012). Thinking politically about HIV: political analysis and action in response to AIDS. *Contemporary Politics*, 18(2), 127–40.

Arreola, S., et al. (2015). Sexual stigma, criminalization, investment, and access to HIV services among men who have sex with men worldwide. *AIDS and Behavior*, 19(2), 227–34.

Auerbach, J.D. and Hoppe, T.A. (2015). Beyond 'getting drugs into bodies': social science perspectives on pre-exposure prophylaxis for HIV. *Journal of the International AIDS Society*, 18(Suppl. 3), 19983.

Bankole, A., Biddlecom, A., Guiella, G., Singh, S., and Zulu, E. (2007). Sexual behavior, knowledge and information sources of very young adolescents in four sub-Saharan African countries. *African Journal of Reproductive Health*, 11(3), 28–43.

Baral, S.D., et al. (2015). Male sex workers: practices, contexts, and vulnerabilities for HIV acquisition and transmission. *The Lancet*, 385(9964), 260–73.

Barker, G. (2007). *Adolescents, Social Support and Help-Seeking Behavior: An International Literature Review and Programme Consultation with Recommendations for Action*. Geneva: WHO.

Barroso, C. (2014). Beyond Cairo: sexual and reproductive rights of young people in the new development agenda. *Global Public Health*, 9(6), 639–46.

Bauer, G.R. and Jairam, J.A. (2008). Are lesbians really women who have sex with women (WSW)? Methodological concerns in measuring sexual orientation in health research. *Women & Health*, 48(4), 383–408.

Blondeel, K., de Vasconcelos, S., García-Moreno, C., Stephenson, R., Temmerman, M., and Toskin, I. (2018). Violence motivated by perception of sexual orientation and gender identity: a systematic review. *Bulletin of the World Health Organization*, 96(1), 29–41L.

Blum, R. and Mmari, K. (2006). *Risk Protective Factors Affecting Adolescent Reproductive Health in Developing Countries: An Analysis of Adolescent Sexual and Reproductive Health Literature from Around the World*. Geneva: WHO.

Blum, R.W. (2007). Youth in sub-Saharan Africa. *Journal of Adolescent Health*, 41(3), 230–8.

Boyle, E.H. (2005). *Female Genital Cutting: Cultural Conflict in the Global Community*. Baltimore, MD: Johns Hopkins University Press.

Brandt, A. (1987). *No Magic Bullet: A Social History of Venereal Disease in the United States Since 1880*. New York: Oxford University Press.

Chandra-Mouli, V. and Patel, S.V. (2017). Mapping the knowledge and understanding of menarche, menstrual hygiene and menstrual health among adolescent girls in low-and middle-income countries. *Reproductive Health*, 14(1), 30.

Charlton, B.M., et al. (2011). Reproductive health screening disparities and sexual orientation in a cohort study of U.S. adolescent and young adult females. *Journal of Adolescent Health*, 49(5), 505–10.

Chow, E.P., Wilson, D.P., and Zhang, L. (2011). What is the potential for bisexual men in China to act as a bridge of HIV transmission to the female population? Behavioural evidence from a systematic review and meta-analysis. *BMC Infectious Disease*, 11(242), 1–17.

Corrêa, S. and Petchesky, R. (1994). Reproductive and sexual rights: a feminist perspective. In: Sen, G., Germaine, A., and Chen, L. (eds.). *Population Policies Reconsidered: Health, Empowerment, and Rights*. Boston, MA; New York, NY: Harvard University Press, pp. 107–22.

Corrêa, S., Petchesky, R., and Parker, R. (2008). *Sexuality, Health and Human Rights*. London and New York: Routledge;.

Dahl, B., Margrethe Fylkesnes, A., Sørlie, V., and Malterud, K. (2013). Lesbian women's experiences with healthcare providers in the birthing context: a meta-ethnography. *Midwifery*, 29(6), 674–81.

de Zalduondo, B.O. (1991). Prostitution viewed cross-culturally: toward recontextualizing sex work in AIDS intervention research. *Journal of Sex Research*, 22(2), 223–48.

Decker, D. et al. (2012). Injection drug use, sexual risk, violence and STI/HIV among Moscow female sex workers. *Sexually Transmitted Infections*, 88(4), 278–83.

Díaz, R.M., Ayala, G., Bein, E., Henne, J., and Marin, B.V. (2001). The impact of homophobia, poverty, and racism on the mental health of gay and bisexual Latino men: findings from 3 US cities. *American Journal of Public Health*, 91(6), 927–32.

Dowsett, G.W. (2003). Some considerations on sexuality and gender in the context of AIDS. *Reproductive Health Matters*, 11(22), 21–9.

Eaton, L.A., Cain, D.N., Pope, H., Garcia, J., and Cherry, C. (2012). The relationship between pornography use and sexual behaviours among at-risk HIV-negative men who have sex with men. *Sexual Health*, 9(2), 166–70.

Eaton, L.A., Driffin, D.D., Bauermeister, J., Smith, H., and Conway-Washington, C. (2015). Minimal awareness and stalled uptake of pre-exposure prophylaxis (PrEP) among at risk, HIV-negative, black men who have sex with men. *AIDS Patient Care and STDs*, 29(8), 423–29.

Farmer, P., Connors, M., and Simmons, J. (eds.) (1996). *Women, Poverty and AIDS: Sex, Drugs and Structural Violence*. Monroe, Maine: Common Courage Press;.

Freud, S. (1994). The social construction of gender. *Journal of Adult Development*, 1(1), 37–45.

Garcia, J. and Parker, R. (2006). From global discourse to local action: the makings of a sexual rights movement? *Horizontes Antropológicos*, 12(26), 13–41.

Garcia, J. and Parker, R.G. (2011). Resource mobilization for health advocacy: Afro-Brazilian religious organizations and HIV prevention and control. *Social Science & Medicine*, 72(12), 1930–8.

Garcia, J., Parker, R.G., Parker, C., Wilson, P.A., Philbin, M., and Hirsch, J.S. (2016a). The limitations of 'Black MSM' as a category: why gender, sexuality, and desire still matter for social and biomedical HIV prevention methods. *Global Public Health*, 11(7–8), 1026–48.

Garcia, J., Parker, C., Parker, R.G., Wilson, P.A., Philbin, M., and Hirsch, J.S. (2016b). Psychosocial implications of homophobia and HIV stigma in social support networks: insights for high-impact HIV prevention among black men who have sex with men. *Health Education & Behavior*, 43(2), 217–25.

Gorry, C. (2010). Transgender health in Cuba: evolving policy to impact practice. *MEDICC Review*, 12(4), 5–9.

Gough, B. (2013). The psychology of men's health: maximizing masculine capital. *Health Psychology*, 32(1), 1–4.

Grangeiro, A., Holcman, M.M., Onaga, E.T., Alencar, H.D., Placco, A.L., Teixeira, P.R. (2012). Prevalência e vulnerabilidade à infecção pelo HIV de moradores de rua em São Paulo, SP. *Revista de Saúde Pública*, 46(4), 674–84.

Gray, R.E., Fitch, M.I., Fergus, K.D., Mykhalovskiy, E., and Church, K. (2002). Hegemonic masculinity and the experience of prostate cancer: a narrative approach. *Journal of Aging and Identity*, 7(1), 43–62.

Healthy People (2020). Determinants of health. [Online] Available at: http://www.healthypeople.gov/2020/about/DOHAbout.aspx

Heckathorn, D.D. (1997). Respondent-driven sampling: a new approach to the study of hidden populations. *Social Problems*, 44(2) 174–99.

Hindin, A. and Fatusi, A. (2009). Adolescent sexual and reproductive health in developing countries: an overview of trends and interventions. *International Perspectives on Sexual and Reproductive Health*, 35(2), 58–62.

Hirsch, J.S., Meneses, S., Thompson, B., Negroni, M., Pelcastre, B., and Del Rio, C. (2007). The inevitability of infidelity: sexual reputation, social geographies, and marital HIV risk in rural Mexico. *American Journal of Public Health*, 97(6), 986–96.

Hunter, M. (2002). The materiality of everyday sex: thinking beyond 'prostitution'. *African Studies*, 61, 99–120.

Ishida, K. (2010). The role of ethnicity in father absence and children's school enrollment in Guatemala. *Population Research and Policy Review*, 29(4), 569–91.

Jones, N., Moore, K., Villar-Marquez, E., and Broadbent, E. (2009). *Painful Lessons: The Politics of Preventing Sexual Violence and Bullying at School*. London: Overseas Development Institute.

Kandel, D. and Davies, M. (1991). Friendship networks, intimacy, and illicit drug use in young adulthood: a comparison of two competing theories. *Criminology*, 29(3), 441–69.

Kane, S. and Mason, T. (1992). 'IV drug users' and 'sex partners': the limits of epidemiological categories and the ethnography of risk. In: Herdt, G. and Lindenbaum, S. (eds.) *In the Time of AIDS: Social Analysis, Theory and Method*. Newbury Park: Sage Publications, pp. 199–222.

Kempadoo, K. (2003). Globalizing sex workers' rights. *Canadian Woman Studies*, 22(3), 143–50.

Kinsey, A., Martin, C.E., Gebhard, R.H., and Pomeroy, W. (1953). *Sexual Behavior in the Human Female*. Philadelphia, PA: Saunders.

Kinsey, A., Martin, C.E., and Pomeroy, W. (1948). *Sexual Behavior in the Human Male*. Philadelphia, PA: Saunders.

Kitzinger, C. (2005). Heteronormativity in action: reproducing the heterosexual nuclear family in after-hours medical calls. *Social Problems*, 52(4), 477–98.

Klitzman, R.L. and Greenberg, J.D. (2002). Patterns of communication between gay and lesbian patients and their health care providers. *Journal of Homosexuality*, 42(4), 65–75.

Leclerc-Madlala, S. (2003). Transactional sex and the pursuit of modernity. *Social Dynamics*, 29, 213–33.

Lombardi, E. (2001). Enhancing transgender health care. *American Journal of Public Health*, 91(6), 869–72.

Mahoney, A., Pargament, K.I., Tarakeshwar, N., and Swank, A.B. (2008). Religion in the home in the 1980s and 1990s: a meta-analytic review and conceptual analysis of links between religion, marriage, and parenting. *Psychology of Religion and Spirituality*, S(1), 63–101.

Mayhew S, et al. (2009). Protecting the unprotected: mixed-method research on drug use, sex work and rights in Pakistan's fight against HIV/AIDS. *Sexually Transmitted Infections*, 85(Suppl 2), ii31–6.

Meyer, C., Blissett, J., and Oldfield, C. (2001). Sexual orientation and eating psychopathology: the role of masculinity and femininity. *International Journal of Eating Disorders*, 29(3), 314–18.

Meyer, I.H. (2003). Prejudice, social stress, and mental health in lesbian, gay, and bisexual populations: conceptual issues and research evidence. *Psychological Bulletin*, 129(5), 674.

Miller, A.M. (2000). Sexual but not reproductive: exploring the junction and disjunction of sexual and reproductive rights. *Health and Human Rights*, 4(2), 68–109.

O'Hara, R.E. and Cooper, M.L. (2015). Bidirectional associations between alcohol use and sexual risk-taking behavior from adolescence into young adulthood. *Archives of Sexual Behavior*, 44(4), 857–71.

Parker, R., Aggleton, P., and Perez-Brumer, A.G. (2016). The trouble with 'Categories': rethinking men who have sex with men, transgender and their equivalents in HIV prevention and health promotion. *Global Public Health*, 11(7–8), 819–23.

Parker, R., Di Mauro, D., Filiano, B., Garcia, J., Muñoz-Laboy, M., and Sember, R. (2004). Global transformations and intimate relations in the 21st century: social science research on sexuality and the emergence of sexual health and sexual rights frameworks. *Annual Review of Sex Research*, 15, 362–98.

Parker, R., García, J., and Buffington, R.M. (2014). Sexuality and the contemporary world: globalization and sexual rights. In: Buffington, R.M., Luibhéid, E., and Guy, D. (eds.) *A Global History of Sexuality: The Modern Era*. Malden, MA: Wiley-Blackwell, pp. 221–60.

Parker, R., Garcia, J., and Muñoz-Laboy, M. (2013). Sexual social movements and communities. In: Tolman, D.L. and Diamond, L.M. (eds.), *The American Psychological Association's Handbook of Sexuality and Psychology*. Washington, DC: American Psychological Association, pp. 229–50.

Parker, R. and Aggleton, P. (2003). HIV and AIDS-related stigma and discrimination: a conceptual framework and implications for action. *Social Science and Medicine*, 7(1), 13–24.

Parker, R., Easton, D., and Klein, C. (2000). Structural barriers and facilitators in HIV prevention: a review of international research. *AIDS*, 14(suppl. 1), S22–S32.

Parker, R. (2001). Sexuality, culture and power in HIV/AIDS research. *Annual Review of Anthropology*, 30, 163–79.

Parker, R. (2007). Sexuality, health, and human rights. *American Journal of Public Health*, 97(6), 972–3.

Parker, R. (2011). Grassroots activism, civil society mobilization and the politics of the global HIV/AIDS epidemic. *Brown Journal of World Affairs*, 17(2), 21–37.

Parker, R.G., Herdt, G., and Carballo, M. (1991). Sexual culture, HIV transmission, and AIDS research. *The Journal of Sex Research*, 28, 77–98.

Patton, G. and Viner, R. (2007). Pubertal transitions in health. *The Lancet*, 369, 1130–9.

Patton, G.C., et al. (2016). Our future: a *Lancet* commission on adolescent health and wellbeing. *The Lancet*, 387(10036), 2423–78.

Petchesky, R.P. (2000). Sexual rights: inventing a concept, mapping an international practice. In: Parker, R.G., Barbosa, R.M., and Aggleton, P. (eds.) *Framing the Sexual Subject*. Berkeley, CA: University of California Press, pp. 81–103.

Ramirez-Valles, J., Heckathorn, D.D., Vázquez, R., Diaz, R.M., and Campbell, R.T. (2005). From networks to populations: the development and application of respondent-driven sampling among IDUs and Latino gay men. *AIDS and Behavior*, 9(4), 387–402.

Roa, C.A.P. (2013). Factores asociados con riesgo de suicidio de adolescentes y jóvenes autoidentificados como lesbianas, gays y bisexuales: estado actual de la literatura. *Revista Colombiana de Psiquiatría*, 42(4), 333–49.

Ryan, H., Wortley, P.M., Easton, A., Pederson, L., and Greenwood, G. (2001). Smoking among lesbians, gays, and bisexuals: a review of the literature. *American Journal of Preventive Medecine*, 21(2), 142–9.

Saewyc, E. (ed.). (2017). Special issue: the global early adolescent study: an exploration of the factors that shape adolescence. *Journal of Adolescent Health*, 61(4), S1–S54.

Sawyer, S.M., Azzopardi, P.S., Wickremarathne, D., and Patton, G.C. (2018). The age of adolescence. *The Lancet Child & Adolescent Health*, 2(3), 223–8.

Schwend, A.S., Winter, S., Chiam, Z., Smiley, A., and Grinspan, M.C. (2018). Depathologising gender diversity in childhood in the process of ICD revision and reform. *Global Public Health*, 13, 1585–98.
Slater, M.E., Godette, D., Huang, B., Ruan, W.J., and Kerridge, B.T. (2017). Sexual orientation-based discrimination, excessive alcohol use, and substance use disorders among sexual minority adults. *LGBT Health*, 4(5), 337–44.
Sommer, M. and Sahin, M. (2013). Overcoming the taboo: advancing the global agenda for menstrual hygiene management for schoolgirls. *American Journal of Public Health*, 103(9), 1556–9.
Sommer, M., Sutherland, C., and Chandra-Mouli, V. (2015). Putting menarche and girls into the global population health agenda. *Reproductive Health*, 12(1), 24.
Stevens, P.E. and Hall, J.M. (1988). Stigma, health beliefs and experiences with health care in lesbian women. *Image Journal of Nursing Scholarship*, 20(2), 69–73.
Strathdee, S.A., et al. (2010). HIV and risk environment for injecting drug users: the past, present, and future. *The Lancet*, 376(9737), 268–84.
Torres, J.B., Solberg, V.S.H., and Carlstrom, A.H. (2002). The myth of sameness among Latino men and their machismo. *American Journal of Orthopsychiatry*, 72(2), 163–81.
Treichler, P.A. (1988). AIDS, homophobia and biomedical discourse: an epidemic of signification. In: Crimp, D. (ed.) *AIDS: Cultural Analysis and Cultural Activism*. Cambridge MA: The MIT Press.
Turner, C.F., Miller, H.G., and Moses, L.E. (eds.) (1989). *AIDS: Sexual Behavior and Intravenous Drug Use*. Washington, DC: National Academy Press.
UNESCO (2018a). Education for health and well-being. [Online] Available at: https://en.unesco.org/themes/health-education
UNESCO (2015). *Emerging Evidence, Lessons and Practice in Comprehensive Sexuality Education—A Global Review 2015*. Paris: United Nations Educational, Scientific and Cultural Organization.
UNESCO (2018b). *International Technical Guidance on Sexuality Education: An Evidence-Informed Approach* (Revised edition). Paris: United Nations Educational, Scientific and Cultural Organization.
Weiser, S.D., et al. (2007). Food insufficiency is associated with high-risk sexual behavior among women in Botswana and Swaziland. *PLoS Medicine*, 4(10), 1589–98.
Wiessing, L.G., Van Roosmalen, M.S., Koedijk, P., Bieleman, B., and Houweling, H. (1999). Silicones, hormones and HIV in transgender street prostitutes. *AIDS*, 13(16), 2315–16.
Wilkinson, W. and Kitzinger, C. (1994). The social construction of heterosexuality. *Journal of Gender Studies*, 3(3), 307–17.
Wilson, P.A., et al. (2016). Sexual risk behavior among virologically detectable HIV-positive young men who have sex with men. *JAMA Pediatrics*, 170(2), 125–31.
World Health Organization (WHO) (2011a). Reproductive health. [Online]. Available at: http://www.who.int/topics/reproductive_health/en/
World Health Organization (WHO). (2011b). Sexual health. [Online]. Available at: http://www.who.int/topics/sexual_health/en/
World Health Organization (WHO) (2000). *What About Boys? A Literature Review on the Health and Development of Adolescent Boys*. Geneva: WHO.
Young, R.M. and Meyer, I.H. (2005). The trouble with 'MSM' and 'WSW': erasure of the sexual-minority person in public health discourse. *American Journal of Public Health*, 95(7), 1144–9.

6.3

Demography and public health

Emily Grundy and Michael Murphy

Introduction

The health and healthcare needs of a population cannot be measured or met without knowledge of its size and characteristics. Demography is the scientific study of population and is concerned both with the measurement, or estimation, of population size and structure and with population dynamics—the interplay between fertility, mortality, and migration which determines population change. These are pre-requisites for making the forecasts about future population size and structure which largely determine the health profile of a population and should underpin public health planning. Understanding the current structure of a population, and looking forward to its future, requires a review of its past. The number of very old people in a population, for example, depends on the number of births eight or nine decades earlier and risks of death at successive ages throughout the intervening period. The *proportion* of very old people depends partly on this numerator but more importantly on the denominator, the size of the population as a whole. The number of births in a population depends on current patterns of family building, and also on the number of women 'at risk' of reproduction—itself a function of past trends in fertility and mortality. Similarly, the number and causes of deaths are strongly influenced by age structure.

Demography is largely concerned with answering questions about *how* populations change and their measurement. The broader field of population studies embraces questions of *why* these changes occur, and with what consequences.

This chapter presents information on demographic methods and data sources, their application to health and population issues, information on

demographic trends and their implications, and the major theories about demographic change. The aim is to illustrate and elucidate the complex inter-relationship between population change and human health.

Global issues

Figure 6.3.1 shows that the world's population has grown rapidly since the middle of the last century and was estimated to be 7.79 billion at mid-2020 (UN 2019). While it took an estimated 123 years (from 1804 to 1927) for the world to increase its population from 1 to 2 billion, the increase from 7 to 8 billion is expected to take less than a tenth of that time (2011–22). The United Nation's medium projection suggests a further increase of some 1.9 billion between 2020 and 2050 (UN 2019). Population growth is thus expected to continue for the rest of this century, with the global population reaching just under 11 billion by the end of it. Even if there is a reasonable chance that global population growth may cease by around 2070 (Lutz et al., 2017), there will by then be nearly three times as many people as a hundred years earlier. This prospect of global population stability also masks huge differences between regions and between richer and poorer countries.

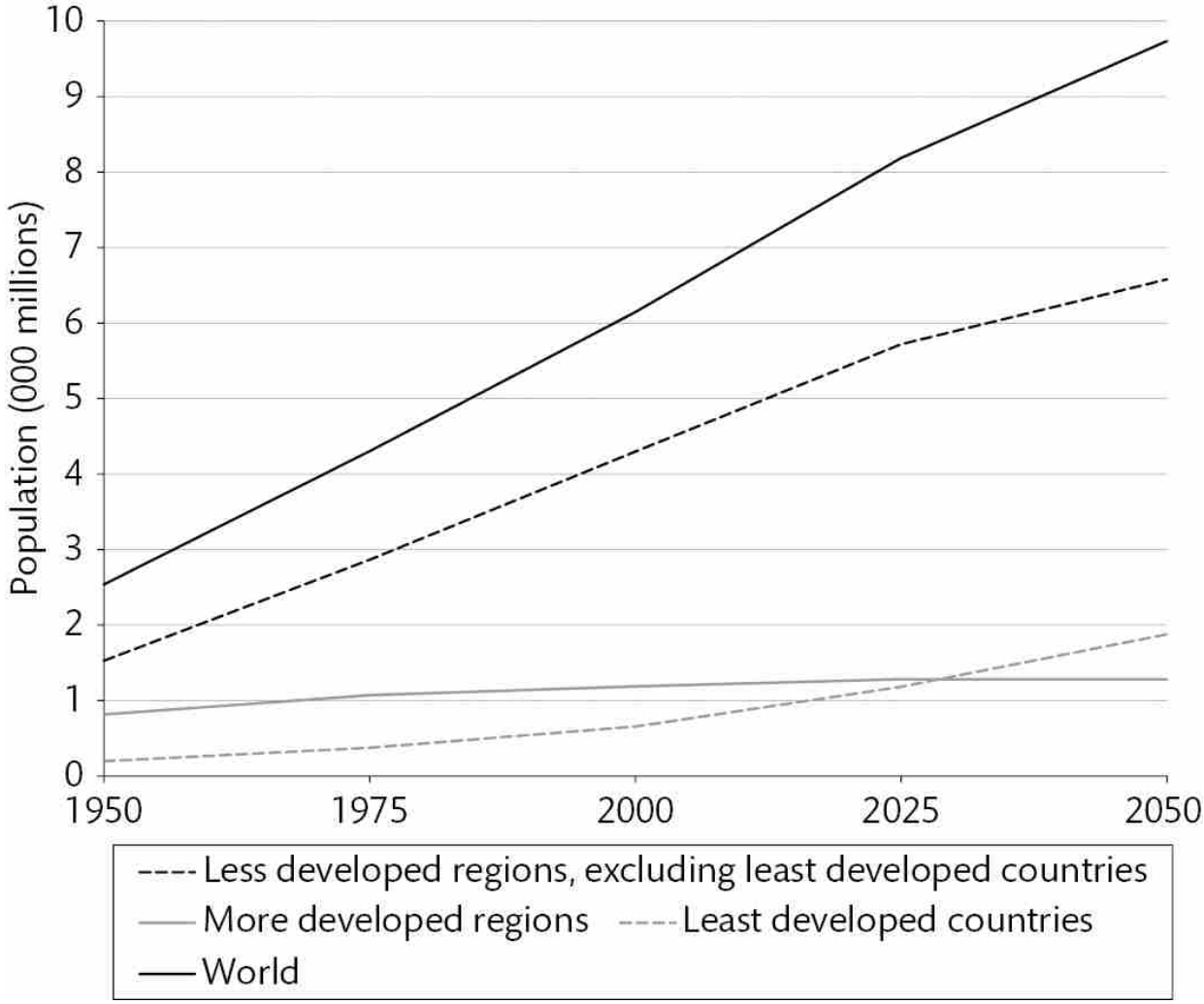


Fig. 6.3.1 Population and projected population of the world by development status.
Source: data from United Nations 2019.

Between 1950 and 2000, 77% of world population growth occurred in countries currently designated by the United Nations as less developed (excluding the least developed, see Box 6.3.1 for definitions); 13% in least developed countries and 10% in more developed regions. Between 2000 and 2050, medium term projections suggest that population growth in more developed regions will account for only 3% of the total with 63% occurring in less developed countries and 34% in the least developed. These projections imply that by 2050 the share of the world's population living in currently more developed regions will account for only 13% of total world population—compared with 32% a century earlier—while the representation

of those in the poorest countries will have increased from 8% of the total in 1950 to 20% in 2050.

Box 6.3.1 Country and regional classifications by level of development

The United Nations classifies countries into 'more' and 'less' developed and also identifies a group of 'least developed' countries. The more developed category includes all of Europe, North America, Australia, New Zealand, and Japan. The least developed countries are mostly in sub-Saharan Africa but also include Afghanistan, Bangladesh, Cambodia, and Myanmar. The classification has some anomalies in that some wealthy Asian countries are counted as less developed (e.g. Republic of Korea, Singapore, Qatar) whereas some poorer former Eastern bloc countries are treated as more developed (e.g. Albania, Belarus, Moldova), although in some more detailed schema these are identified as 'countries in transition'.

The World Bank employs a classification based on gross national income per capita which divides countries into high-, middle-, and low-income groups, with a subdivision of the middle into upper and lower. Some of the countries (principally from Eastern Europe) classified by the UN as developed fall into middle-income categories, while some of the UN less developed group are classified by the World Bank as middle-income (principally Latin American) or high income (e.g. Republic of Korea, Singapore).

Membership of the Organisation for Economic Co-operation and Development (OECD) is also sometimes used as an indicator of developed country status; members include Russia and Mexico, both of which are classified by the World Bank as middle rather than high-income countries.

The Human Development Index compiled by the United Nations Development Programme takes into account factors other than income, such as school enrolment, literacy, and levels of mortality.

Regional groupings employed by different international agencies also vary slightly.

Further details of the current classifications are available on the relevant organizations' web sites.

While some regions grapple with the needs of rapidly growing populations, such as large increases in requirements for maternal and child health services and schools, others face challenges of population ageing and, in some cases, population decline. By 2025, it is expected that over one person in five in Western Europe will be aged 65 or more and, in some countries such as Japan (where the population is already starting to decline), South Korea, Spain and Italy, projections suggest that over a third of the population will be aged 65 and over by 2050 (UN 2019). The proportions—and absolute numbers—of older people are also increasing substantially in many low- and middle-income countries, including some of the most populous such as China, India, Indonesia, and Brazil. Some low income countries with young populations will also experience very large increases in the number (but not the proportion) of older people, including the very old. For example, the number of people aged 65 and over in Kenya is expected to

increase by 460% between 2015 and 2050, compared with an increase of 70% in high-income countries. In 2015 half (53%) of the world's older old people aged 80 and over live in less developed countries, by 2050 that proportion is expected to exceed 70% (UN 2015). These variations have enormous implications for health and healthcare priorities in, and beyond, the populations concerned. Divergence in population growth between regions of the world is also fuelling mass migration, which itself has implications for global population health (Fernandes and Pereira Miguel 2009).

These hugely differing rates of growth arise from differences in vital rates and associated large variations in age structures. The global Total Fertility Rate (see Box 6.3.2 for derivation of total fertility rate (TFR)), currently about 2.5 children per woman, is half of the value in 1950 and over half the world's population now live in countries with 'below-replacement rate' fertility. In many currently developed countries the shift to lower fertility started in the late nineteenth or early twentieth centuries; what has changed in the more recent past has been fertility decline in many regions currently designated less developed and further falls in fertility in many richer countries. In a number of European countries, including Poland, Portugal and Spain, and some Asian countries, such as the Republic of Korea, Japan, Singapore, and Taiwan, women now have on average only 1.4 children or fewer, and people aged 65 and over outnumber children under 15. In sub-Saharan Africa, women on average have 4.5 children each, 40% or more of the population is aged 15 or under and only 3% aged 65 or more (UN 2019).

Many more developed countries have had below replacement level fertility since the 1970s, but their populations have nevertheless continued to grow in the past half century. The reasons why populations continue to increase in such circumstances are, firstly, that the number of births depends both on the number of women of childbearing age and how many children they each have. If the former is increasing so too may the number of births, even if women individually have fewer children each than their predecessors. In Iran, for example, the Total Fertility Rate (births per woman) declined from 6.24 in 1970–75 to 1.75 in 2010–15 but there were nevertheless more births in the latter than the former period and the population still has a high rate of natural increase (more births than deaths) (UN 2019). This reflects the large increase in the number of women of childbearing age determined by the earlier high level of fertility. One

consequence of this large built-in momentum is that it takes many decades before low fertility results in low population growth. Additionally, national population numbers may be ‘topped-up’ by in-migration, often by young adults who add to population numbers both by their own presence and because of the children they have; in many European countries about a quarter of births are to foreign-born mothers.

Levels of mortality, and associated differences in age and cause distribution of death, also vary markedly around the world. In general mortality had been declining, and the late twentieth century crises associated with the HIV/AIDs pandemic and the collapse of the Soviet Union had receded 20) before the COVID-19 pandemic emerged as a new challenge to world population health. In sub-Saharan Africa, the number of HIV/AIDS deaths in 2017, 940 thousand, was half the peak figure in 2004, although HIV/AIDS still represents a substantial and long-term public health challenge and progress towards meeting targets for reducing infections and deaths has been disrupted by the COVID-19 pandemic (UNAIDS 2020). Prior to COVID-19, there were also concerns about an apparent stalling in rates of mortality improvement in the United States and some other high-income countries attributed variously to ‘deaths of despair’ from suicide and drug and alcohol use (Case and Deaton 2015), rising obesity (Preston et al. 2018) and, in the United Kingdom, Government austerity policies (Green et al. 2017). Nevertheless, in the United States life expectancy at birth is close to 80 (and above 80 in some other high-income countries), while in some sub-Saharan countries it is below 60. Differences in the level of mortality, and in population age structure, are associated with huge variations in the distribution of deaths by age and cause. As illustration (Fig. 6.3.2), in Sierra Leone, 37% of all deaths in a year were of infants and children aged under 5, compared with 0.3% in Japan. Conversely, 68% of all deaths in Japan and 48% in Chile, a middle-income country, were of people aged 75 and over; equivalent proportions for Egypt and Sierra Leone were 30% and 8%, respectively.

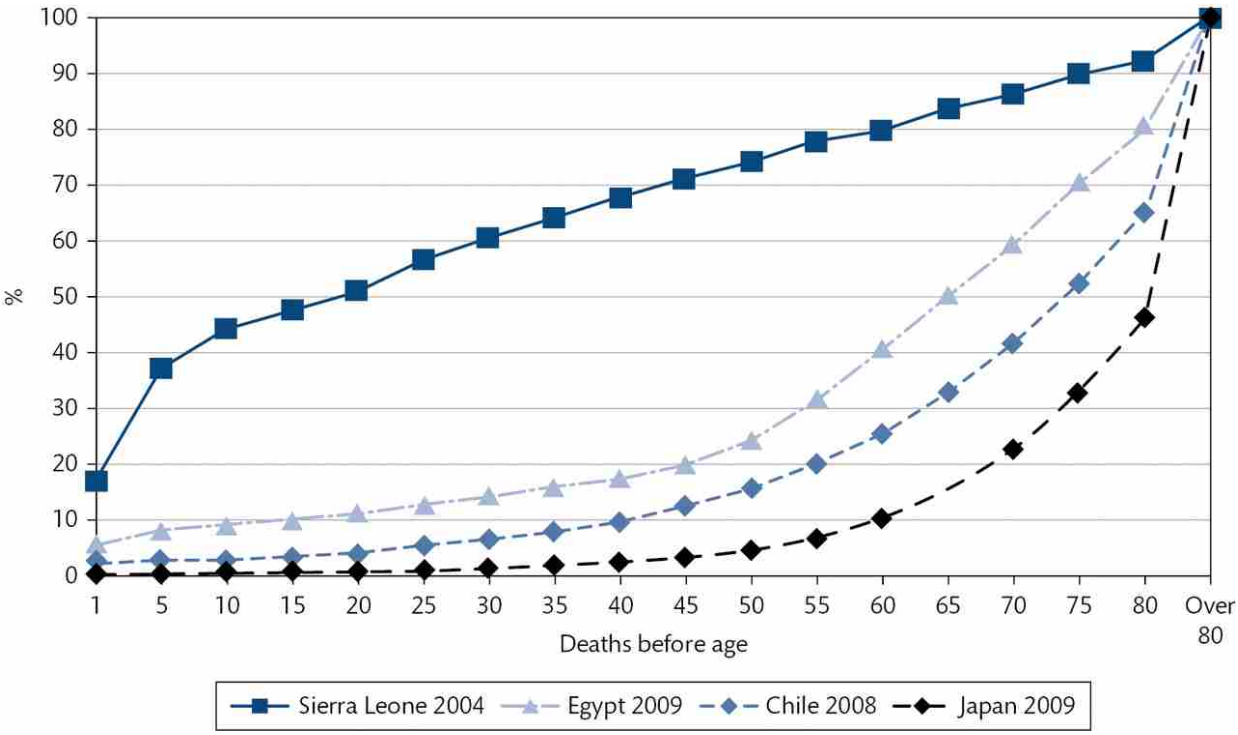


Fig. 6.3.2 Cumulative distribution of deaths by age; Sierra Leone 2004, Egypt 2009, Chile 2008, Japan 2009.

Source: data from United Nations Department of Economic and Social Affairs, Demographic Yearbook 2011, ST/ESA/STAT/SER.R/41, United Nations, New York, USA, Copyright © United Nations 2012, available from http://unstats.un.org/unsd/demographic/products/dyb/dybsets/2011.pdf.

Closely related to variations in the distribution of deaths by age are differences in the cause structure of death. In parts of the world, communicable diseases and reproductive and child health still present the most pressing public health problems. In high-income countries prior to the COVID-19 pandemic, these causes accounted for only 7% of deaths and over 70% were the result of non-communicable diseases (WHO 2018). Some countries face a ‘double burden’ with still high rates of communicable disease and poor maternal and child health but also rising rates of chronic non-communicable disease driven both by age structure changes and changes in lifestyles.

The watershed that separates populations with high fertility, relatively high mortality, young age structures and rapid growth, from those with low vital rates, older age structures, and slow or no growth, is conceptualized as the demographic transition. Identifying, and explaining, this and associated profound changes in health has been a central pre-occupation of modern demography (Lee 2003). Before turning to these issues, the basic methods

and materials of demographic analysis must be considered and the issue of population dynamics—*how* populations change—addressed.

Demographic data and methods of analysis

In the seventeenth century, John Graunt, a London merchant, used data from the London Bills of Mortality to devise an early life table, leading to him being dubbed the 'father of modern demography'. However, while Graunt had information on *numbers* of deaths, he lacked data on the population at risk and could not compute death *rates*. Essentially all demographic analysis requires data both on the population 'stock' and on 'flows' in and out—births, deaths, and migration. The traditional sources of information on the former are population censuses and, for the latter, vital registration systems.

Population censuses

The first 'modern' censuses were undertaken in Scandinavia in the eighteenth century. Censuses spread throughout Europe during the nineteenth century and most of the rest of the world in the twentieth. As well as basic questions about age, sex, marital status, and place of residence, data on other characteristics such as employment, education, and housing are often collected. The United Nations recommends that censuses be conducted at least decennially in years ending in 0 or 1.

Censuses have many strengths and are often the only source of data for small areas or population subgroups. Although primarily a tool for collecting data on population 'stock', censuses have also been used to find out about vital events. Many countries use censuses to provide data on recent internal migration (through questions on place of residence one or more years earlier) and immigration (through questions on country of birth and/or date of entry for those born elsewhere). Indirect estimation techniques developed by William Brass and others mean that questions on number of children born and number who have died, on widowhood and orphanhood, are widely used to assess mortality levels and trends using both censuses and surveys in countries with deficient vital registration systems (Preston et al. 2001).

However, censuses involve huge costs, the challenge of ensuring acceptable data quality and the problem of timeliness, as they are normally conducted only every ten years. Approaches to reducing cost (and improving quality) include use of sample censuses, either for the census as a whole, as

in China, or for more detailed questions, as in Canada; rolling censuses where parts of the population are covered each year, an approach pioneered in France, and increasing encouragement of online responses, as in the United Kingdom. Census taking requires not only a reasonable administrative infrastructure, but also the co-operation of the population to be enumerated. Some countries, including Germany and the Netherlands, have given up taking censuses because the latter is lacking and now rely on large scale surveys or 'virtual censuses' and population registration data. The twenty-first century is seeing more countries adopting alternatives as the information required by governments becomes more complex and the difficulties of mass data collection escalate. In 2010, only 11 of the 27 current EU members conducted traditional censuses (Valente 2010).

When censuses are taken, difficulties arising from errors and omissions are common, even in countries with a long history of census taking. Young, geographically mobile adults; recent (especially undocumented) immigrants; members of minority ethnic groups; infants and the very old are those most likely to be under-enumerated—some of the very groups that policy makers and health professionals may be most keen to know about.

Groups such as seasonal migrants, military personnel, people temporarily away from home and those with more than one residence also present problems. Not only are they more likely to be missed, but a decision has to be made about whether they should be assigned to their place of usual or legal residence (assuming it can be determined) or counted as belonging to the place of enumeration. The former system is termed *de jure*, the latter *de facto*. The issue of assigning people to some place of usual residence is important as often resources are allocated based on population size and characteristics. Moreover, it is essential to try and ensure that demographic events recorded in one system (vital registration) are attributed to the population actually 'at risk' of experiencing them. In richer countries, for example, most deaths occur in hospitals which may draw patients from a wide area. If these decedents are not assigned to the locality where they lived prior to hospital admission, areas including large hospitals will appear to have very high mortality rates while in others, recorded mortality will be artificially low.

Under-enumeration is usually assessed through census validation surveys (surveys of a sample of census addresses in which intensive efforts are made to contact non-respondents and check information supplied by respondents)

and comparisons with population estimates from other sources. Beyond, ensuring near-complete enumeration, the quality of the data collected is also a major concern.

In many populations, people may not always know their exact age and some approximation is reported or made by an enumerator. 'Heaping' on ages ending in 0 or 5 is a common result. Heaping can be detected by looking at the age distribution and applying various tests of consistency and such data are normally adjusted before publication. More serious problems arise when reported age is based on other characteristics, such as marital status, number of children, or grandparent status, as clearly any analysis of, for example, age at first marriage, will be biased if people's report of their age is influenced by their marital status.

Vital registration

Data on demographic events, as well as on population characteristics, are needed. In richer countries these are drawn from vital registration. Compulsory registration of births and deaths was established in most European countries during the nineteenth century. In England and Wales, for example, civil registration was introduced in 1837. Subsequent improvements to the system included those following the 1874 *Births and Deaths Registration Act* which made parents legally responsible for registering births and required attending physicians to supply information on cause of death. Other revisions have since been made, for example, the inclusion of first mother's and later father's age and in 2012, recording of all children previously borne, rather than just legitimate ones, on the confidential section of birth certificates. Most high-income countries have well-established registration systems with complete, or very near complete, coverage. In poorer parts of the world, however, vital registration systems are frequently seriously incomplete or non-existent, although there are exceptions and some countries, including India and China, have sample registration systems for selected areas. Some assessments suggest that only about a third of deaths estimated to occur globally are registered and reported to the World Health Organization, although if the sample registration systems in India and China are considered as sufficiently representative of their national populations, this proportion rises to 72% (Mathers et al. 2005, 2009). A review (Mikkelsen et al. 2015) concluded that there were modest improvements in civil registration and vital statistics

systems between 2005 and 2012 with registration of close to 40% of deaths, and 65% of births of children under 5 years.

The quality of the information supplied and coded is of course very important. No-one registers their own death and the information obtained from proxy informants may be inaccurate. Differential reporting of age, occupation, marital status, or other characteristics in the census and in other sources, such as death certificates, presents a further difficulty. Numerator-denominator discrepancies may introduce serious bias into the analysis of mortality at advanced ages, or by characteristics such as occupationally-defined social class, marital status, or ethnicity (Williams et al. 2006). The Nordic and some other European countries avoid these problems by maintaining well-developed register-based systems that link vital registration data to population, occupational, and educational registers. One study of marital status differences in mortality at older ages found marked differences in results for countries using traditional and register-based systems (Murphy et al. 2007).

Cause of death

Death certificates are the major source of information on cause of death. In richer countries, cause of death is generally certified by a physician and coded according to the International Classification of Diseases (ICD) which originated from work undertaken by the 19th century British medical statistician, William Farr. The tenth revision of the ICD came into use from 1994 and included nearly twice as many codes as ICD-9. ICD-11 was released in mid-2018 to allow member states to prepare for implementation in 2022. National preferences, as well as ICD revisions, may influence assignment of cause of death, as illustrated in a number of classic papers in which case studies of deaths were distributed to physicians in different countries. Growing awareness of particular conditions may also influence coding practices. For example, deaths attributed to Alzheimer's disease and other dementias doubled between 2000 and 2016 making this the fifth largest cause of death globally (compared to twenty-sixth in 2000) (WHO 2018). This increase reflects increases in the proportion of deaths occurring in the oldest age groups and changes associated with the introduction of ICD-10, but also effects of increased awareness and some specific

campaigns to increase recognition of these conditions (Moschetti et al. 2012).

Older people, now the vast majority of decedents in low mortality populations, are more likely to suffer multiple pathologies and the number of conditions recorded on death certificates has been increasing. Choice of one over another as the 'true' underlying cause of death is bound to be partially arbitrary. Multiple coding of death certificates and analyses by all mentions of a condition may be more informative but such data are available in only a few countries (Anderson 2011). Variations in death certificate coding reflecting differences in medical knowledge and diagnosis; in the extent to which autopsies are used; in classification systems and the quality of registration systems are a major factor complicating analyses of trends over time or between countries.

Deaths assigned to symptoms, signs, and ill-defined conditions such as 'old age' or 'senility' or other causes lacking diagnostic meaning, sometimes referred to as 'garbage codes', present a particular problem. Mathers et al. (2005) in an investigation of coverage and quality of cause of death coding in 2003, found that in some countries over 40% of deaths were assigned to these 'causes'. Only 23 countries met their definition of high quality of data with coverage of at least 90% and fewer than 10% of deaths assigned to ill-defined codes.

Variations in coding practices and use of ill-defined codes complicate comparisons over time, as well as between countries. In a classic text, Preston (1976) argued that there was an inverse association between the proportions of deaths assigned to circulatory diseases and to ill-defined causes, and that part of the apparent twentieth century epidemic in heart disease mortality in richer countries may have been an artefactual consequence of improvements in death certification. In England and Wales, for example, 41% of deaths of people aged 80 and over in 1911–15 were assigned to ill-defined causes compared with only 4% a century later.

In countries which lack adequate certification and registration systems, data on deaths by cause are seriously limited. Attempts have been made to develop verbal autopsies, protocols for collecting information from lay informants which can be reviewed by physicians and used to assign cause of death (Wang et al. 2007, 2012). This approach has been useful in a number of small investigations and is being employed on a larger scale, for example in India (Gajalakshmi and Peto 2004). However, one study which tested

physician-certified verbal autopsies in six sites in four poorer countries against gold standard assessment found a concordance of less than 50%, with substantial variability by cause and physician (Lozano et al. 2011). The development of extensive test data sets with which to make comparisons of verbal autopsy findings and calibration against computer algorithms, rather than physician assessments, is an expanding area of enquiry with some promising results (Murray et al. 2011). In the absence of routinely recorded data, estimates may be obtained by modelling. The most ambitious exercise is the Global Burden of Disease programme which has used a large array of sources and methods, including expert knowledge, vital registration, field surveys, surveillance and police and mortuary data to derive estimates of cause-specific mortality by age and sex for 235 separate causes for every country in the world (GBD 2016).

Other data sources

Many countries have a range of surveys which collect more detailed information on, for example, health-related behaviours, family-building strategies, or reasons for migration that would be impossible to collect in a census. In poorer countries, where other data sources are scarcer, surveys often present the best source of information on basic demographic parameters. Data quality is potentially better in a survey than a census, as it is more likely that well-trained interviewers can be used. The World Fertility Survey (WFS), an international population research programme launched in 1972 to determine fertility levels throughout the world, and its successor, the Demographic and Health Survey Programme (DHS), have been particularly valuable in providing demographic and health data for a range of poorer countries. Other approaches include multiround surveys, in which respondents are asked about events since last contact, and dual-record systems which involve two independent data collection systems (one often a multiround survey), the results of which are then combined. This method allows some estimation of missed events but is expensive. These approaches are described in more detail in most demographic textbooks (Preston et al. 2001; Rowland 2003; Siegel and Swanson 2004).

The raw materials of demography relate to individuals' most personal experiences—sexual activity, family formation, birth control, reproduction, marital breakdown, illness, and disability. All of these occur in a social framework which attaches value to some of these behaviours and stigmatizes

others. Not surprisingly, respondents in censuses and surveys may be reluctant to disclose non-marital pregnancies, illegal abortions, undocumented migration, or deaths of relatives from HIV/AIDS. Concealment has also been the policy of some national governments which have treated demographic data as official secrets. Additionally, the enormous potential complications arising from people's uncertainties about age or other 'basic' characteristics; uncertain recollections of prior events and the vast scope for administrative errors have to be considered. The demographer's traditional obsession with data quality is hence understandable. Differences in perceptions and reporting of health status are also problematic and have bedevilled attempts to make international comparisons of health status as, even if questions are harmonized, the ways people respond to them are not (Ritu et al. 2002). Increasingly health-related surveys collect information on observer measured indicators, including biomarkers, as well as self-reported information on health and functioning which allows this issue to be addressed, although raising other problems related to intraindividual variability, observer variation, and differences in laboratory practices.

The statistics produced in data sets produced by the UN, WHO, Organisation for Economic Co-operation and Development (OECD), and other national and international bodies series have their origins in what is or has been *done* by millions of people, mediated by what is *said* about these events and experiences, further filtered by how this is *recorded*, processed, and analysed. Some assessment of data quality is generally given but sometimes users may pay insufficient attention to this.

The analysis of demographic data

A standard array of techniques and measures forms the basis of much demographic analysis; the most common of these are described briefly next. Further detail is supplied in a number of textbooks (Preston et al. 2001; Rowland 2003; Siegel and Swanson 2004). Analysis involves not just the application of a particular technique, but decisions about what units of analysis to use and how to group them. A major distinction is between *period* and *cohort* analysis. Period analysis deals with events of a particular time period (for example, mortality rates 2010–2015) while cohort analyses follow the experience of individuals through time. Cohorts in this sense are

defined as groups of people who have experienced the same significant event at the same time. Thus, birth cohorts comprise people born in a particular year or group of years and marriage cohorts those marrying at a particular time. Cohort and life course approaches to analysing mortality and other indicators of population health have an intuitive appeal and are increasingly used, both fuelling and fuelled by a growing number of longitudinal studies. Cohort analysis of time series data may be used in the absence of specially collected longitudinal data.

Cohort and period are two of the dimensions which 'place' persons in time; the third is age. Duration effects (such as duration of marriage, proximity to death, or length of exposure to a particular pathogen) may also be important. Cohort effects may be substantial and, unless allowed for, may mask relationships between age and various risks. Differences in the smoking behaviour of cohorts, for example, have a major effect on the relationships between age and smoking-related disease observed at different periods (Preston and Wang 2006).

Decisions about whether to use individuals, families, households, or geographic areas as units of analysis are often constrained by data availability. Until relatively recently, most census data were only available as aggregate tabulations, but individual-level information is increasingly available. Other innovations include the development of samples including linked census, vital registration and in some cases health service data such as the longitudinal studies available for countries of the United Kingdom and a number of others (Young et al. 2010). In these data sets, individuals' census records are linked with their vital registration records so numerator-denominator biases in, for example, the analysis of mortality are avoided. In Nordic and some other countries, the whole population has been assigned personal identification numbers facilitating linkage of information from a range of registers. Linkage to use of health and care services is also available in some countries, such as Finland.

These advances have greatly extended the material available for analyses of variations in demographic behaviour, and their consequences. They have also raised complex security and confidentiality issues fuelling debate over appropriate restrictions on access to data.

The measurement of fertility

Fertility means the childbearing performance of a woman, couple, or population. Generally, only live births are included. The term fecundity, by contrast, is used to refer to the physiological capability of producing a live-born child. A rough idea of fertility may be gained from using census or survey data to calculate child-woman ratios: the ratio of 0–4 year olds to women aged 15–49. However the survival of infants (and their mothers) and the age structure of the female population affects these ratios, so they are generally only used if no other data are available.

The simplest measure of fertility commonly used is the crude birth rate—the number of births in a particular year per 1,000 population. As the denominator of this includes those not 'at risk' of giving birth (women outside reproductive age groups and men), it is really a ratio rather than a rate. Crude birth rates are influenced by the age structure of the population, but less seriously so than crude death rates. In 2010–15 crude birth rates ranged from less than 10 per 1,000 in parts of Europe to nearly 50 per 1,000 in the highest fertility countries of sub-Saharan Africa (UN 2019).

Slightly more informative is the general fertility ratio—births per 1,000 women of reproductive age (generally defined as 15–49 or 15–44). Where data allow, age-specific fertility rates (births per 1,000 women of a particular age or age group) are preferred. These are frequently summarized using the TFR. Where, as is usually the case, period data are used to calculate this, it indicates how many children women in a hypothetical cohort would have if they experienced current age-specific fertility rates throughout their reproductive life. This measure is sometimes explicitly denoted TPFR (total period fertility rate). In low mortality populations, a TFR of 2.1 is taken to indicate *replacement level* fertility as, under this regime, a cohort of women would be succeeded by a cohort of daughters of the same size (after some allowance for mortality and the fact that 105–106 boys are born for every 100 girls).

One limitation of the TFR is that it is affected by changes in the 'tempo' rather than the 'quantum' of childbearing. If women start delaying their fertility but 'catch up' later, there will be a divergence between cohort and period measures, as the latter will be based partly on the behaviour of earlier cohorts whose timing of births was different. Similarly, if women have children earlier, TFRs will rise, even if eventual family sizes remain unchanged. This means that period measures are much more volatile than

cohort ones. For these reasons, many statistical offices use cohort, rather than period, measures of fertility as the basis for projections.

More sophisticated measures of fertility include parity progression ratios. These indicate the probability of proceeding from one birth to another (for example, what proportion of mothers with two children progress to having a third). Parity progression ratios are normally calculated for cohorts who have completed, or nearly completed, their childbearing but it is also possible to use data on births by birth order to derive period progression ratios (Bongaarts and Feeney 1998).

In the past, demographers often preferred to calculate age-specific *marital* fertility rates (and TFRs and other measures) on the grounds that the unmarried population is not 'at risk' (or at reduced risk) of childbearing. Changes in marital fertility indicative of deliberate attempts to limit family size are regarded as one of the defining features of the fertility 'transition' (see next) and so distinguishing these from changes due to variations in the 'at risk' (married population) has been particularly emphasized. However, rises in non-marital childbearing, which now account for over 40% of births in OECD countries and over 60% in France, mean that restricting analyses to marital fertility is generally no longer appropriate.

Reproduction rates

In the absence of migration and with fixed mortality, populations will grow if mothers replace themselves with more than one (surviving) daughter and decline if they have fewer than one. Theoretically, it would also be possible to measure the replacement of fathers by sons, but in practice the difficulties involved in obtaining paternity data make this infeasible. Reproduction rates thus relate only to female fertility—births of daughters. The gross reproduction rate (GRR) is derived in the same way as the TFR except that age-specific birth rates based only on births of daughters are used in the calculation. The net reproduction rate (NRR) makes an allowance for mortality; specifically, the chance that a daughter will herself survive to childbearing age. The NRR cannot be calculated unless both age-specific fertility and mortality data are available (although it can be approximated using the GRR and appropriate life table survival data). Changes in either fertility or mortality (or both) will mean a divergence between period

measures (based on the experience of a hypothetical cohort) and the experiences of real cohorts.

Summary information on measures of fertility and reproduction is shown in Boxes 6.3.2 and 6.3.3.

Box 6.3.2 Fertility measures

Definitions:

Fertility: the childbearing performance of individuals, couples, or populations.
Fecundity: the physiological capability of producing a live birth.
Parity: the number of children previously born alive (or sometimes number of previous confinements) to a woman or couple. Nulliparous women are those who have borne no children.

Measures:

Crude birth rate: The ratio of births in a year (other specified period) to the average population in the same year/period (mid-year population), expressed per 1,000:

$$\text{CBR} = \frac{\text{number of births}}{\text{mid-year population}} \times 1{,}000.$$

General fertility rate: births to women aged 15–44/49 in a year/period per 1,000 women aged 15–44/49 in the same period:

$$\text{GFR} = \frac{\text{number of births to women aged 15–44/49}}{\text{mid-year population of women aged 15–44/49}} \times 1{,}000.$$

Age-specific fertility rate (ASFR): number of births to women aged x (or x to x + n) per 1,000 women aged × (or × to × + n). 'n' refers to the length of an age interval.

ASFRs are frequently calculated for five year age groups from 15–19 to 40–44 or 45–49:

$$\text{ASFR} = \frac{\text{births to women aged x}}{\text{mid-year population of women aged x}} \times 1{,}000.$$

Total (period) fertility rate (TFR/TPFR): the sum of the age-specific fertility rates for all reproductive age groups for a particular period (usually a year), conventionally expressed per woman. The TFR indicates how many children a woman would have if throughout her reproductive life, she had children at the age-specific rates prevalent in the specified year or period.

x = 49:

$$TFR = \sum_{x=15}^{x=49} f_x$$

where 'f_x' is the age-specific fertility rate at age x. If rates for age groups, rather than single years, are used then the sum of the age-specific rates must be multiplied by the number of single ages included in the group (usually five).

x = 45–49:

$$TFR = 5 \sum_{x=15}^{x=45-} f_x$$

x = 15–19.

Parity progression ratio: the probability of a women of parity x progressing to parity x + 1.

Box 6.3.3 Reproduction ratios

Measures:

Gross reproduction rate (GRR): The sum of the age-specific female fertility rates (births of daughters), for all reproductive age groups for a particular period (usually a year) conventionally expressed per woman. The GRR indicates how many *daughters* a woman would have if, throughout her reproductive life, she had children at the age-specific rates prevalent in the specified year of period. The GRR can be calculated either by summing female age-specific fertility rates (relating to births of daughters rather than all births) or using the formula:

$$GRR = TFR \times \text{Proportion of female births}$$

The proportion of female births can be taken as 0.488 (100/205) in the absence of more detailed information.

Net reproduction rate (NRR): the average number of daughters that would be borne, according to specified rates of mortality and of bearing daughters, by a woman subject through life to these rates. The NRR employs the same fertility data as the GRR, but also takes into account the effects of mortality. An NRR of 1.0 indicates that a population's fertility and mortality levels would result in exact replacement of mothers by daughters.

The measurement of mortality

As for fertility, the simplest measure of mortality is the crude mortality rate, deaths per 1,000 population. This is strongly influenced by age structure. For example, although life expectancy at birth in high-income countries in 2015–

20 is estimated to be some 18 years longer than in low income countries (81 years and 63 years, respectively), the crude death rate—deaths per 1,000 population of all ages—was *higher* in the high-income country group (8.8 compared with 8.1 in low income countries) (UN 2019). This apparent paradox reflects the fact that the proportions of older people with higher risks of death are much greater in high-income countries. Age (and sex) specific rates, or measures based on them (such as life expectancy at birth), are therefore much to be preferred if data are available to calculate them. Both direct and indirect standardization are sometimes used to make comparisons between populations with different age and sex structures. Standardized mortality ratios (SMRs) are calculated using indirect standardization. This involves selecting a set of 'standard' age-(and often sex) specific mortality rates, for example those for a national population, and applying these to the numbers of people in the relevant age groups in the subpopulation of interest—for example, the population of a particular region. This yields an 'expected' number of deaths—the number of deaths there would be in the subpopulation if age-specific death rates were the same as those in the standard population. The ratio of observed to expected deaths gives the SMR. Thus an SMR of 1.24 indicates that mortality in the subpopulation is 24% higher than in the standard population, after allowing for age differences. SMRs are useful summary measures of differences in mortality but give no indication of the *level* of mortality. In direct standardization the adjustment is carried out the other way and a set population structure is used to weight mortality data to produce age-standardized rates. EUROSTAT, for example, produces a European standard population to facilitate comparisons across populations. Comparisons are of course only valid if the same standard population is used.

Age-specific death rates are calculated using the numbers of deaths at age x (or between ages x and x + n) in a particular year as the numerator and the mid-year population of the same age as the denominator. The rate is conventionally expressed per 1,000 or per 100,000 population. The mid-year population is used as a measure of the average population at risk on the assumption that deaths are evenly distributed throughout the year. For some age groups, notably infants, this assumption is invalid. In low mortality populations, deaths in the first 3 days of life may account for half or more of all deaths in the first year of life. Moreover, information on the size of population aged less than 1 year normally comes from birth data (as in 9 out

of 10 years relevant census data will not be available). For these reasons live births in a particular year are conventionally used as the denominator of the infant mortality rate while deaths to infants aged less than 1 constitute the numerator. Some infants dying in a given year will have been born in the previous year and some born in the year in question will die the following year. This can cause distortions if there are large annual fluctuations in numbers of births (or infant deaths) and often three-year averages are preferred. Deaths at very old ages are also not evenly distributed throughout the year and an adjustment is often made to allow for this.

Infant mortality rates (IMRs) were very high in some parts of historical Europe—with 300 or even 400 deaths per 1,000 live births in regions of Russia and Germany at the end of the nineteenth century (van de Walle 1986). In England and Wales at the start of the twentieth century, there were some 140 infant deaths per 1,000 live births. Infant mortality in high-income countries is now extremely low—fewer than two infant deaths per 1,000 live births in some European countries. There have also been huge falls in infant mortality in many poorer countries; in China infant mortality fell from 42 deaths per 1,000 births in 1990 to 7 in 2019. In India the equivalent fall was from 88 to 28. However, rates remain high in some of the very poorest countries—up to 8 deaths per 1,000 live births (WHO 2017).

Variations on this scale have substantial demographic impacts. Infant mortality has also attracted particular interest because of links with fertility behaviour and as an indicator of public health conditions. Particularly in this latter context, perinatal, early, and late neonatal and post-neonatal mortality rates are often distinguished where data allow (see Box 6.3.4).

Box 6.3.4 Mortality measures

Crude death rate: The ratio of deaths in a year (other specified period) to average population in the same year/period (mid-year population), expressed per 1,000:

$$\text{CDR} = \frac{\text{number of deaths}}{\text{mid-year population}} \times 1{,}000$$

Age-specific mortality rate (ASMR): number of deaths to persons aged x (or x to x + n) per 1,000 persons aged x (or to x + n):

$$\text{ASMR} = \frac{\text{deaths to persons aged x}}{\text{mid-year population of persons aged x}} \times 1{,}000$$

Standardized mortality ratio (SMR): the ratio of observed to expected deaths in a study population. Expected deaths are calculated by applying a set of standard age-specific mortality rates to the age distribution of the study population. Standardized ratios are only useful for comparisons. They have no intrinsic meaning.

$$\text{Infant mortality rate IMR} = \frac{\text{deaths to infants aged} < 1 \text{ year in year x}}{\text{number of live births in year x}} \times 1{,}000.$$

sometimes decomposed into *neonatal mortality rates* (deaths of live-born infants during the first 4 weeks) and *post-neonatal* mortality (from 4 to 52 weeks)

The **perinatal mortality rate** measures late foetal deaths (stillbirths) and early neonatal deaths relative to live births.

$$\text{Perinatal mortality rate} = \frac{\text{still births + deaths under 1 week}}{\text{still + live births}} \times 1{,}000.$$

Still births used to refer to deaths of foetuses of 28 or more weeks' gestation, however an earlier threshold of 24 weeks is now more generally used.

Life tables

Life table analysis is a core demographic technique and life tables provide one of the most powerful tools for analysing mortality and other demographic processes. Life tables show the probability of dying (and surviving) between specified ages. They also allow the calculation of various other indicators, including expectation of life. If complete data on the mortality of a birth cohort are available, then a cohort life table may be constructed. However, the use of cohort life tables is obviously only possible retrospectively. More commonly, period life tables, based on mortality rates at a particular time, are calculated. These life tables show death (and survival) probabilities for a hypothetical cohort with an arbitrary radix (number of babies at the beginning) usually set to 10,000, 100,000, or some other multiple of 100.

Specific notation, summarized in Box 6.3.5, is used in life table analysis. The basis of the table is a set of probabilities of dying—$_nq_x$—which are calculated from age-specific death rates. x here refers to age at the start of an interval whose length is specified by n. Thus $_5q_{50}$ refers to the probability of someone alive at 50 dying between age 50 and age 55. The complement of

${}_nq_x$—the probability of surviving—is denoted ${}_np_x$. The (hypothetical) number of survivors at each age is given by l_x; thus l_o equals the radix (of 100,000) and l_{75} the number of survivors at age 75. The number of person years lived in an interval (${}_nL_x$) and the total number of person years lived after a particular age (T_x) are often not shown in published tables but are steps on the way to the calculation of e_x—life expectancy at age x.

Box 6.3.5 Life table measures and notation

x = age attained last birthday
l_0 = the radix of the life table (hypothetical number of babies), usually 100,000
l_x = number of survivors at age x, so l_{65} is the number of persons alive at age 65 in the hypothetical life table population
${}_nq_x$ = probability of dying between age x and x + n, so ${}_4q_1$ is the probability of dying between age 1 and 5 for a person aged 1
${}_np_x$ = probability of surviving between ages x and x + n, so ${}_{20}p_{65}$ is the probability of surviving from age 65 to age 85 for a person aged 65
${}_nd_x$ = number of deaths between age x and x + n
${}_nL_x$ = number of person years lived between age x and x + n
T_x = total number of person years lived after age x
e_x = expectation of life at age x, so e_0 is expectation of life at birth

This measure provides an indicator of mortality which is very largely independent of the age structure of the population since it depends only on age-specific mortality rates. This makes it more useful than either an SMR (which gives no indication of level and, like directly standardized rates, depends on the choice of standard) or a crude death rate (which is strongly influenced by age structure). Life expectancy either at birth (e_0) or further life expectancy at a particular age, say 65 (e_{65}), is calculated by dividing the total person years lived after age 0 or 65 (T_0 or T_{65}) by the number of survivors aged 0 (l_0) or 65 (l_{65}).

Values of life expectancy at birth are sometimes (mis)interpreted as indicators of usual age at death in a particular population. In very low mortality populations where most deaths occur within a relatively small range of ages (see the example of Japan in Fig. 6.3.2), there will be a close

correspondence between median and modal ages of death and life expectancy at birth (which is a mean value). However, in populations such as Sierra Leone where many deaths occur in infancy, there will be a wide divergence. There is also sometimes confusion about the interpretation of values of further life expectancy at a particular age. This is derived from information about the probabilities of death and survival at *subsequent* ages, and so is not influenced by deaths at earlier ages and it is erroneous to think that, for example, the further life expectancy of someone aged 65 will equal life expectancy at birth minus 65. The higher are mortality rates at young ages the greater will be this divergence. For example, in the United States in 1900–01—when mortality rates among infants and children were high—women aged 65 had a further life expectancy of 12 years, even though female life expectancy at birth was only 49 years. In 2016 there was a much closer correspondence between female life expectancy at birth (81 years) and female further life expectancy at age 65 (21 years).

Model life tables

Patterns of age-specific death rates show similarities whatever the level of mortality. Death rates tend to be higher in infancy than later childhood and rise with age from around the age of puberty, although in the oldest age groups rates of increase tend to flatten out. Because of the tendency for death rates at one age to be associated with death rates at other ages in a given population, it is possible to derive hypothetical schedules, called *model life tables*, describing variations in mortality by age and sex, normally in terms of a limited number of parameters which allow for particular features of the mortality pattern of the population considered. Model life tables are derived from empirical data from countries where these are available. They are extremely useful aids for the estimation of mortality by age in populations with defective data. They are also used (in conjunction with fertility data) to show the outcomes of particular fertility and mortality regimes on, for example, population age structure and for making population projections. All demographic texts give further details of their derivation and application.

Other applications of life table analysis

Life tables are widely used to analyse probabilities associated with events other than death, such as risks of divorce or contraceptive use failure and discontinuation rates, and in estimates of disability-free or healthy life expectancy. Many chronic conditions associated with ageing, such as musculoskeletal and sensory impairments, may have serious implications for health status but are not directly life-threatening. Life table methods are used to decompose total life expectancy into 'healthy' and 'unhealthy' or 'disability-free' and 'disabled' components. This can be done using cross-sectional data on morbidity or disability prevalence in conjunction with mortality data, although this has some limitations. More sophisticated (and data demanding) multistate approaches and simulation methods which allow for transitions both to and from disabled states have also been developed (Manton et al. 2006; Wolf and Gill 2009). Despite these technical advances, there is still controversy about *trends* in health and disability, particularly in older age groups. To a large extent this debate arises from measurement problems and the difficulties involved in making comparisons between health indicators derived in different ways, a further reminder of the importance of data quality and measurement.

Multiple decrement life tables allow 'decrements' from more than one event—for example different causes of death. Cause elimination life tables are also used to identify the 'pure' severity of a particular cause of death, although strong assumptions are required. Multistate models allow analysis of a range of transitions, particularly those where re-entries into a particular state, such as being married or living in a certain region, are possible. These more sophisticated applications of course require more detailed data.

The measurement of migration

In 2018, 55% of the world's population lived in urban areas compared with 29% in 1950 (UN 2018). This huge shift, which has important public health implications, has been largely driven by rural to urban migration, especially in Asia and Latin America. Measuring migration represents particular difficulties. The classical definition of internal migration is a permanent or semi-permanent move across an administrative boundary, which means that the extent of migration recorded depends partly on the size of administrative areas. For example, in a country divided into many small areas, a move over 10 kilometres will count as migration, which would not be the case for

countries divided into larger ones. Hence, international comparison of internal migration rates is potentially misleading. Even the distinction between international and internal migration may be problematic if boundaries are unrecognized, contested, or changing. The temporal dimension to migration presents further difficulties; what constitutes permanent or semi-permanent and how should groups such as seasonal migrants be treated?

The reason for defining migration as a move over a boundary is largely pragmatic. Often only moves of this kind are recorded; moreover this is the information required by local administrations. Some countries have registration systems in which changes of address are recorded. More commonly, censuses are used to find out about migration. Questions on usual address one or five years ago allow the proportion of *movers i*n the population to be measured (except for those aged less than one or five). These data also allow inflows *and* outflows between pairs of areas to be measured. *Moves,* as opposed to *movers,* are not directly measured as someone moving several times in the reference period cannot be distinguished from someone moving only once. Those leaving an address and later returning to it cannot be identified either. This means that the length of the reference period used is important; the proportion of movers in the five years preceding a census will *not* equal five times the proportion moving in one year before the same census.

In the absence of direct census data, estimates of migration can be made indirectly using the ‘balancing equation’ referred to next. Differences in the size of a population at two points in time not accounted for by natural change (i.e. births minus deaths) must be due to migration (or data errors). If vital registration data are available, then both births and deaths can be taken into account. If they are lacking, then the survival of groups enumerated in the first of a pair of censuses must be estimated from a life table and the number of expected survivors compared with the number enumerated in the second census (obviously ageing must be allowed for, so the number of 20–29 year olds in the first census will be compared with 30–39 year olds 10 years later). These methods only allow estimation of *net* migration (balance between in-migration and out-migration). Their major weakness lies in the fact that the residual population balance assumed to be due to migration may in fact reflect differences in the quality of the two censuses considered or errors in the estimates of survival used.

Survey data are also used to measure migration and potentially provide illuminating information on the reasons for, and consequences of, migration. However, as migration over long distances is a relatively rare event, even large general population samples may yield relatively few migrants. A similar problem besets samples of international travellers, such as the UK International Passenger Survey, designed to estimate flows of international migrants through port or border surveys. Tourists and business travellers comprise the vast bulk of people entering or leaving so surveys are an inefficient way of identifying immigrants and emigrants. Compiling advance passenger information electronically, and logging both entries and exits, presents a possible alternative method of capturing movements, although one that presents various legal and practical problems. The UK e-borders initiative, for example, was judged some 10 years after its inception to have met some targets, but not the one of improving estimates of in- and out-migration (Vine 2013). Moreover this, and similar initiatives, are designed to gather information on people travelling on a recognized route who are prepared, or compelled, to provide information to secure passage or entry. It is thus of little use as a method for measuring movements of refugees and asylum seekers pursuing other transit modes.

Unfortunately, other data are often lacking as legal and administrative record systems are frequently concerned with citizenship and right of abode rather than international migration *per se* (and virtually never with emigration). Estimates of the size of immigrant populations depend on whether the measure used is based on place of birth or on nationality—the size of the latter is influenced by policies on acquisition of citizenship.

Population dynamics

Any population comprises those who have made an entry and not yet exited. When whole populations of defined geographic areas are considered, the only means of entry are birth or immigration and the only means of exit death or emigration.

The most basic method of demographic analysis is the decomposition of overall population change (P_t—P_0) into its components (B, D, I, E):

$$P_t - P_0 = B - D + I - E$$

where P_t = population at an end of period; P_0 = population at the beginning of a period; and B, D, I, E represent respectively births, deaths, immigrations and emigrations during the same period (B-D is referred to as natural increase and I—E as net migration). Population subgroups may be similarly defined in terms of entries and exits. In the absence of migration, entry to the population aged 75 to 84 is through ageing (passage from 74 to 75) exit is through further ageing (84 to 85) or death. This simple accounting equation is an important one, both methodologically and as a formal reminder of the need to consider *past* as well as current events.

Of the three demographic determinants of population size, structure, and growth, fertility has historically been of much greater importance than either mortality or migration. Every birth represents not just an addition to the current generation of children, but also potentially an exponentially increasing augmentation in the size of future generations. Death carries no such promise of future return. Births increase the population only at age zero, so making it younger, whereas deaths are spread across the whole age range and so have much less impact on age structure. The third determinant —migration—is generally not of significant magnitude to have a major impact on most national populations, although there are exceptions especially when natural increase is close to zero and a few populations, such as some of the Gulf states, are heavily dependent on migrant labour. Increased levels of immigration to many European countries since the 1990s have had an effect on population size and age structure in some. The population of Spain for example, increased by 10.2% between 1999 and 2006 and over 90% of this increase was due to migration (Sobotka 2008). It is also anticipated that the importance of migration as a driver of population change will increase in the next few decades as a result of lower fertility in developed regions (UN 2017).

For social and biological reasons fertility, mortality, and migration have interactive effects. Decreases in mortality among women with reproductive potential, for example, influence not just the size of the age group affected at the time, but also the size of succeeding generations. Declines in male mortality, particularly in populations where large age differences between spouses are common and remarriage of widows is rare, will similarly tend to increase fertility by effectively increasing the proportion of women of reproductive age who are still married. Conversely, reductions in fertility clearly reduce the risk of maternal mortality and in low income settings may

have further beneficial effects on the survival of mothers, infants, or both. Age at motherhood also influences rates of population growth. The average age of mothers at the birth of their daughters is termed the mean length of a generation and is generally around 29 years. A shorter interval will mean more rapid generational succession (and faster population growth).

Migration affects other demographic parameters because migrants differ from the general population. International migrants are generally young and in good health and often move from relatively high- to low-fertility populations. Consequently, immigrants may serve to (temporarily) 'rejuvenate' the host population and, at least initially, often have higher fertility and lower mortality. In England and Wales, for example, in 2011 the TFR for foreign-born women was 2.2. compared with 1.8 for the UK born (Office for National Statistics 2017). Despite the disadvantages they may face, mortality of immigrant groups is often lower than that of host populations because of the differential selection of immigrants—the healthy immigrant effect (Kennedy et al. 2015). The degree of selection tends to vary according to difficulties and distance to be overcome in making an international move. For all these reasons, the demographic characteristics of population subgroups largely comprising immigrants and their immediate descendants may vary substantially from those of the population as a whole.

Population projections

Population projections represent one of the most widely used outputs of demographic analysis. Strictly speaking a projection simply represents the outcome of applying various assumptions about future fertility, mortality, and migration and so differs from a *forecast*, which implies prediction. However, projections are often treated as forecasts and the degree of uncertainty inherent in them is not always sufficiently recognized, although the production of probabilistic forecasts in recent UN projections, makes this more explicit. The most common method of projection is the component method, based on the balancing equation ($P_t = P_0 + B—D + I—E$). Assumptions are made about the three components of change—births, deaths, and migration—and applied to age and sex groups within the initial population to give a projection of future size and structure. To a large extent, assumptions are based on recent trends together with other information, for example, survey data on fertility intentions or (sometimes) models of change in particular causes of death. Forecasting fertility has generally been

regarded as the most problematic area of projection but recently greater attention has been paid to the errors that have been made in forecasting mortality in developed countries. This has little effect on age groups in which survival is high but can have quite substantial impacts on forecasts of the number of older people. Migration may be an important element and may be underestimated, if projections do not take into account feedback loops whereby migrant populations tend to generate further migration (Bongaarts and Bulatao 2000). International migration is also difficult to project as it is affected by events outside the country, is often a sensitive political issue and may be volatile. Partly for these reasons, immigration levels have been consistently underprojected in forecasts in many European countries (Alders et al. 2007).

Population growth

Changes in the size of a population produced by the surplus (or deficit) of births over deaths are termed natural increase (or decrease). A common indicator of growth is the crude rate of natural increase—the difference between the crude birth rate (annual births per 1,000 population) and the crude death rate (annual deaths per 1,000 population). If net migration is zero, this will be the same as the growth rate of the population—the overall annual change in the population divided by the population size—(conventionally expressed as a percentage). In Europe as a whole and Japan, deaths outnumbered births in 2015–20. In other countries, births still outnumber deaths even though fertility rates have been below the level required for *long-term* replacement for 50 years or so. As already noted, this apparent paradox largely reflects the fact that the number of births is a function of the number of potential mothers, to which immigration may also contribute, as well as of their fertility patterns. If the former is increasing so too may the numbers of births, even if women have fewer children each.

The young age structures of many populations in the developing world mean that these populations have a huge built-in potential for growth. Population *momentum* is the measure which gives the ratio of the ultimate size a given population would achieve to current population size if fertility were to immediately fall to replacement level. The population of Africa is projected to increase from 1.2 billion to 2.5 billion between 2015 and 2050 (UN 2019), a consequence of both population momentum and high levels of fertility. In some low fertility countries there are now concerns about

‘negative momentum’—the prospect of decline in population even if fertility rates increase somewhat because of successively smaller cohorts of women in childbearing age groups.

Intrinsic rate of natural increase: stable population theory

Early in the twentieth century Lotka (1907) demonstrated mathematically that a population closed to migration and subject to unchanging age-specific fertility and mortality rates for a long period would eventually have a fixed age structure (in which the proportion in each age group remained unchanged) and would grow at a constant rate. This type of population is called a *stable* population. The fixed age structure of a stable population is independent of the initial age structure—two very different populations subject to the same unchanging rates for a long period would eventually assume the same structure. A particular case of a stable population is a *stationary* population—one in which birth and death rates are constant and in balance and so population growth is zero. The L_x column of the life table is an example of a stationary population. The number of births is fixed (the radix) and the age distribution is also fixed. In non-stationary stable populations, the age structure is also fixed but the size of every age group is growing at the same constant rate as the overall population and the number of births. This is called the *intrinsic rate of natural increase* and is a function of the NRR and the mean length of a generation (approximated by the mean age of childbearing).

Although stable and stationary populations are theoretical constructs, real populations at various times have met the model requirements closely enough to allow stable population theory to be used to develop methods for indirectly estimating fertility and mortality in populations lacking adequate directly derived data. Stable population models are also widely used for insurance, pension, and personnel planning. One of the important results of the work of Lotka and his successors was to show theoretically the important influence of fertility on age structure during the demographic transition.

Age structure

Population pyramids graphically illustrate the current structure of populations and in so doing, also provide insights into both the future and

the past of the population. High fertility populations have a pyramid shape with each successive cohort being larger than its predecessor. The population pyramid for Bangladesh (Fig. 6.3.3a) shows a typical pattern for a population with a history of high fertility but a recent downturn. Each successive cohort is larger than the preceding one, with the exception of the youngest. 'Old' populations, such as that of the United Kingdom (Fig. 6.3.3b), are more rectangular with a gradual tapering at the top. Bulges in population pyramids due to high numbers of births have 'echo' effects when members of large cohorts themselves have children. Thus the baby boom experienced in many populations in the post-Second World War period (precise timing varied between countries) had an echo effect in the1980s.

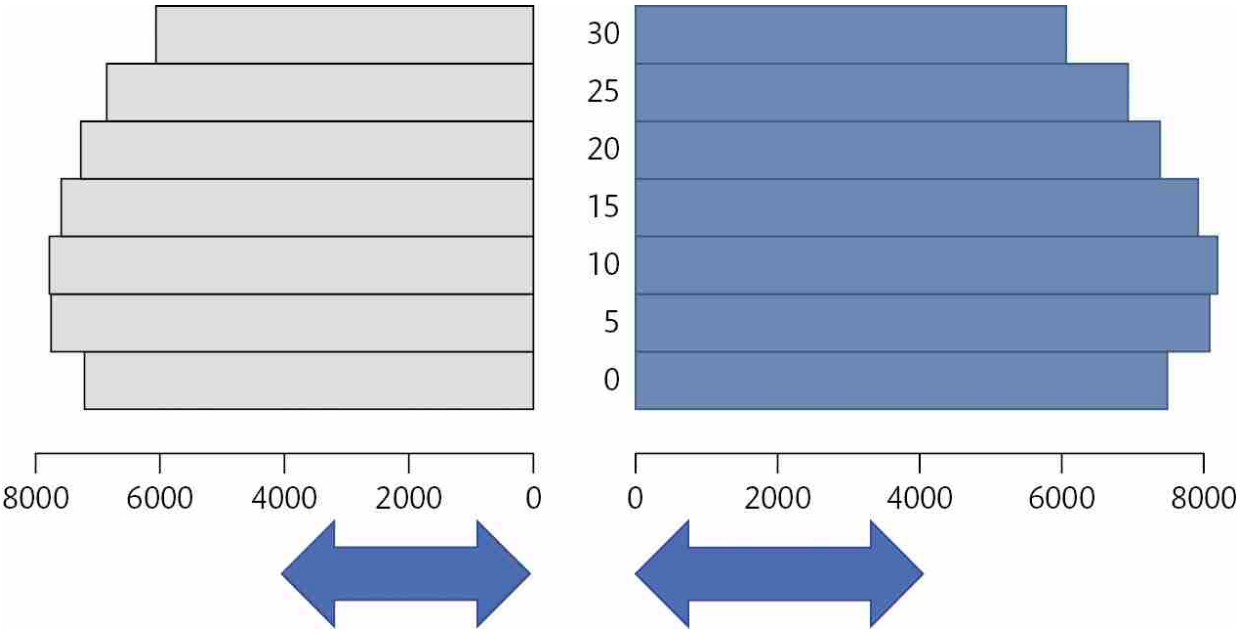


Fig. 6.3.3 Distribution of the population of (a) Bangladesh and (b) the United Kingdom by age and sex, 2010.

Source: data from United Nations, World Population Prospects: The 2010 Revision, United Nations, New York, USA, Copyright © 2012, available from http://esa.un.org/wpp/documentation/pdf/WPP2010_Volume-I_Comprehensive-Tables.pdf.

Historically, and apparently paradoxically, improvements in mortality in those European populations which now have high proportions of old people in fact served to *offset* the trend towards population ageing, as they chiefly benefited the young—and led to increases in the proportions surviving to

have children themselves. However, although fertility has the greatest *potential i*mpact on age structure and population growth, in some circumstances mortality (or migration) may become a more important influence. Many populations in richer countries now have had fertility at or below replacement level for an extended period, life expectancies at birth close to 80 and near universal survival to the end of the (female) reproductive span. In these conditions, further improvements in mortality have the greatest impact at old ages and further population ageing occurs from the apex, rather than, or in addition to, the base of the population pyramid. Mortality changes are now the main motor of the further ageing of a number of populations with already old age structures (Preston and Stokes 2012). Population age structures and associated rates of growth or decline, changes in age structures such as population ageing, and the speed and stage of age structural change all have important economic and health implications which have attracted considerable debate and controversy.

Many economists have pointed out that population growth has often provided a spur to human ingenuity and economic growth. Less positively, the countries which now have the youngest age structures and most rapid rates of population growth are already suffering from land degradation and in many cases, constrained agricultural potential (Alexandratos 2005). Large and growing child populations also hamper efforts at improving human capital through education or improved health (Casterline 2010). Many women in poorer countries, particularly in sub-Saharan Africa, still have unmet needs for family planning with consequent potential negative impacts on their own health and that of their children (Cleland et al. 2014).

Reduced fertility initially produces a ‘demographic dividend’ or ‘window’ when the ratio of children to adults falls and those in prime productive ages, the survivors of larger birth cohorts, account for a higher proportion of the population. It has been argued that this dividend of lower child dependency and higher representation of adults in the prime working age groups played an important part in the rapid economic development of the ‘East Asian Tigers’, like South Korea and also China (Bloom et al. 2000). The next phase, involving high and increasing representation of older people may, it has been suggested, also bring some economic benefit in the form of increased savings (by the large number of older people) and so greater capital available for investment even though gross domestic product (GDP) and national income growth are likely to slow down (Lee and Mason 2017).

The major concerns arising from increases in the proportions of older people relate to effects on productivity and increased need and cost for support systems of various kinds, including pensions, healthcare, and long-term care. In OECD countries, healthcare expenditure is typically three to five times as high for those aged 65 and over as for those aged under 65. However, there is considerable international variation in the proportion of GDP devoted to healthcare spending for older people which bears little obvious relationship to the proportion of older people in the population concerned. (Rechel et al. 2013). Even in rich countries much of the support needed by older people with disabilities is provided by family members, and in many poorer countries families are also important providers of economic support. Population ageing thus has implications for intergenerational transfers at both the population and the family level. Adaptation of policies, including public health initiatives to promote healthy ageing and encouragement of increased economic activity at older ages provide a means of addressing some of these concerns (Rechel et al. 2013; Bloom et al. 2015; WHO 2015).

Some countries have also responded with explicit policy changes designed to increase fertility. The Republic of Korea modified its population policy in 2006 and introduced a range of ‘family friendly’ measures designed to increase fertility; in 2012 Iran shifted from promoting ‘two children are enough’ to ‘at least two children are ideal’, accompanied by large cuts in family planning budgets, and China moved from a ‘one child’ to a ‘two child’ policy in 2015.

In populations which have more recently moved to low fertility and low mortality regimes, the pace of demographic change has been much faster than occurred historically in Europe. The proportion of the Japanese population aged 65 and over doubled, from 7% to over 14%, between 1970 and 1996. In France a similar increase took 130 years to achieve and in Sweden 85 (Kinsella and Phillips 2005). In countries such as China, Thailand, and Brazil an equivalent shift has occurred, or will occur, in around 30 years (UN 2015). More recently ageing populations have thus had a much shorter period in which to adapt to new public health priorities. The origins of these age structure changes lie in the demographic transition.

The demographic transition

Towards the end of the nineteenth century (earlier in France) birth and death rates started falling in a number of European countries. Between 1871–75 and 1911–15 the TFR for England and Wales, for example, dropped from 4.8 to 2.8; by the early 1930s it was below replacement level, a development which was viewed with alarm and led to the first Royal Commission on Population. Although modern methods of contraception were lacking, it was clear that this huge drop in fertility was the result of the deliberate limitation of family size. Half of couples married in the 1870s had six or more children compared with 12% of couples married in 1911–15 (Coleman and Salt 1992). Expectation of life at birth, meanwhile, increased by some 15 years between the end of the nineteenth century and the early 1930s.

Researchers attempting to relate such shifts in demographic regimes to economic and social changes, originated the theory of the demographic transition. The 'classical' view propounded by Notestein (1945) and others was that in 'traditional' societies fertility and mortality are both high and roughly in balance. Change is driven by economic advance which results in lower mortality. Fertility initially remains high, resulting in a rapid period of population growth. After this lag, however, fertility also falls in response to falling mortality and the erosion of 'traditional' pro-natalist values.

This classical view has since been considerably modified (Lee 2003; Bongaarts 2010). Coale and his collaborators in an ambitious project to track the transition in historical Europe suggested that no economic 'threshold' for fertility decline could be identified and that the pattern of decline seemed to follow regional groupings, suggesting a cultural rather than a socioeconomic dimension (Coale and Watkins 1986). Caldwell et al. (1992, 2007) additionally argued that in non-European countries, it was not so much socioeconomic modernization but 'Westernization' involving increased emphasis on the nuclear family and a change in intergenerational wealth flows (resulting in the costs of children outweighing their potential benefits) that was the important trigger of fertility transition.

Research on the historical demographic transition in Europe may seem of limited relevance to contemporary problems. However, this research was fuelled by post-war fears about population growth. By 1950, significant mortality declines had been achieved or initiated throughout the world. In China, for example, expectation of life at birth increased from 43 in 1960 to over 75 by 2010. In this context it seemed imperative to discover the causes of fertility decline and use this knowledge to accelerate fertility 'adjustment'

to falling mortality. Was ‘development the best contraceptive’ as concluded at the stormy 1974 World Population Conference; could change be achieved through intensive family planning programmes, as the experience in Taiwan and South Korea seemed to suggest; or was some combination of these and other factors the key to fertility transition? Studies of societies in which the fertility transition had occurred seemed to offer the best prospect of an answer to these questions. While simple answers to complex questions are rarely forthcoming, Coale (1973) identified three factors which he considered pre-requisites for fertility decline in contemporary populations. These were: that potential parents must think it acceptable to balance the advantages and disadvantages of another child; that some advantage must be gained from reduced fertility and that effective techniques of fertility control must be available. This paradigm, subsequently elaborated into the idea of ‘Ready, Willing and Able’ (to deliberately control fertility), has dominated research on fertility motivations and behaviours ever since (Lesthaeghe and Vanderhoeft 2001).

Trends in fertility in the second half of the last century and early part of this have shown considerable divergence and shed some light on these debates. TFRs are highest in sub-Saharan Africa; with Niger, Democratic Republic of the Congo, Somalia, and Mali, having TFRs above, close to, or above 6 in 2015–20 (UN 2019). In all other world regions there was clear evidence of onset of a transition to lower fertility by the 1970s or 1980s; Japan was the first non-Western country to experience a fertility transition starting after the Second World War. Subsequently, substantial fertility declines have occurred in a number of countries which at that time had only a limited amount of development, such as Sri Lanka, Thailand, China, and more recently Bangladesh. As already noted, one of the largest and fastest recent fertility declines took place in the Islamic Republic of Iran (Abbasi-Shavazi et al. 2009) defying assumptions that socially and religiously conservative Muslim societies with low rates of female employment would be resistant to fertility change.

Recent interpretations of fertility change have reverted to considering development—in its broadest sense rather than restricted to consideration of average incomes—a stronger influence in contemporary populations than ideational changes, although the relevance and role of cultural and policy related factors is recognized (Bryant 2007). Common factors identified in poor rural populations where fertility has fallen significantly are well-

established education systems, improvements in healthcare and in child survival, some form of extrafamilial welfare, well-organized local government, and an organized family planning programme. Potential benefits of investing more resources in fewer children, as a consequence of increasing opportunities in urban or industrial livelihoods, also seem to be important (McNicoll 2006; Bryant 2007). The education of women has been identified as a particularly significant influence on both falling fertility and improved infant survival—and better health throughout the life course (Samir and Lutz 2017).

As noted earlier, fertility rates remain high in much of sub-Saharan Africa. Continuing expressed preferences for large families and other family and social characteristics of African populations have led some to conclude that the model of change seen in Asia and Latin America may not be applicable to Africa, or at least not to all regions of Africa (Caldwell et al. 1992; Caldwell 2007). There is nevertheless evidence of unmet demand for family planning and a strong case can be made, on the grounds of improving maternal and child health, for renewed investment in family programmes which in some places faltered as the policy focus and funding shifted to the HIV/AIDs crisis (Cleland et al. 2014).

Recent trends have also made it clear that the endpoint of fertility transition is not necessarily a fertility level around the ‘replacement rate’. In some countries, such as Guatemala, Peru, Egypt, Turkey, and Ghana, fertility declines ‘stalled’ (sometimes temporarily) at a level higher than two children per woman (Bongaarts 2010). In others they have continued downward, spiralling a new set of concerns about ‘lowest low’ fertility.

The second demographic transition

In contrast to the scenario of high fertility and rapid population growth in some of the poorest countries, a growing number of high-income countries now have concerns about the implications of low fertility. Many of these countries experienced a post-Second World War ‘baby boom’ during the 1950s and early 1960s followed by a 1970s ‘baby bust’ when fertility declined to very low levels. In Scandinavia, France, the United Kingdom, the United States, and other English-speaking countries fertility rates have since fluctuated at levels between 1.6 and replacement level with something of an increase in the first decade of this century—a pattern also seen in many

other industrialized countries. These populations have also experienced a range of family related behavioural changes, included marked increases in cohabitation, non-marital childbearing, divorce, postponed childbearing, and increased levels of childlessness, described by some as a 'Second Demographic Transition' (Lesthaeghe and Niedert 2006). Such behaviours, which have become more usual among Northern European cohorts (and a number of countries outside Europe) born from around the 1950s onwards, are much less prominent in Southern and Eastern Europe and the richer countries of South and East Asia. However, it is mainly in these latter countries that 'lowest low' fertility—rates below 1.3– has been prevalent with 21 countries from these regions having TFRs below 1.3 in 2003. The recent increase in fertility in developed societies meant that only three of these countries, South Korea, Slovakia, and Singapore, had average values below 1.3 in the period 2010–15 (Goldstein et al. 2009; UN 2019). However, the TFR remained below 1.5 in a large number of other countries including Italy, Japan, Poland, Spain, and Ukraine.

An important determinant of very low fertility in many of these societies has been a precipitous fall in marriage rates (Billari and Kohler 2004). Japan, for example, has been transformed from a society with near universal marriage in the early and mid-twentieth century to one in which a fifth of the population will remain never-married at age 45 (Retherford et al. 2001). In this case the erosion of arranged marriages after the Second World War has played a part; changing gender roles may also be important.

The proximate determinants of fertility

One of the contributions of research into the fertility transition has been improved understanding of biosocial influences on reproduction. A huge range of social, economic, cultural, and psychological factors may influence decisions about family building strategies and family size. However, these can only have effect if they are translated into patterns of behaviour or physiological characteristics that influence the risks of conception or delivery.

Conversely, other patterns of behaviour with potentially important influences on fertility may be adopted with little or no thought to these consequences. In a classic paper, Davis and Blake (1956) distinguished a series of 'intermediate fertility variables'; factors influencing exposure to

risk of pregnancy (marriage and coital frequency); risk of pregnancy (such as use of contraception) and pregnancy outcome (spontaneous and induced abortion). The most influential refinement of this work is the Bongaarts decomposition model (Bongaarts 1978), which identified four elements chiefly responsible for observed fertility variations:

the proportion of women married (exposed to risk);
contraceptive use;
induced abortion;
post-partum non-susceptibility to conception (largely determined by breastfeeding practice).

The TFR is dependent on the interactive effect of these variables and hypothetical maximum fertility. In modern 'post-transition' populations, fertility decisions are normally couple- (or woman-) based and are implemented through contraception and abortion. In non-contracepting populations, biosocial factors, notably marriage patterns, breastfeeding practices, sexual frequency and, in some populations, the prevalence of infertility, are of major importance.

The social reproductive span, from entry to end of marriage or any sexual union, nearly always comes later, often much later, than menarche. Fecundity—the potential for bearing children—decreases after the third decade, more sharply after the age of 35 and in most non-contracepting populations the average age at last birth is around 40. Social factors, as well as biological ones, are important influences. Sexual activity may cease before menopause because of widowhood or separation and in some populations, childbearing after becoming a grandmother is disapproved of. In some Western countries fertility rates after age 40 have recently risen, after a long period in which they were falling, reflecting an increase in the postponement of motherhood in order to avoid clashes with educational and career advancement. For some women this may be facilitated by the availability of medically assisted reproduction. In England and Wales, for example, in 2017 fertility rates for women aged 40 and over were slightly higher than for teenage women (ONS 2018).

For those within the effective reproductive span—biologically capable of childbearing and in a sexual union—overall fertility in non-contracepting populations is largely a function of the length of intervals between births. This is determined to a considerable extent by breastfeeding patterns as the

duration of post-partum amenorrhea is much longer for mothers with protracted periods of intensive breastfeeding. Longer birth intervals and increased breastfeeding also have positive effects on infant and child health. Overall deaths before the age of 5 might be reduced by as much as 30% in some countries if closely spaced births were delayed (Cleland et al. 2014).

The epidemiological transition

Transitions from relatively high to low mortality regimes have in all populations been associated with transformations in the age, cause, and sex structure of death. Omran (1971) coined the phrase ‘epidemiological transition’ to describe this process. Changes in the response of societies to health and disease processes also need consideration. The term ‘health transition’ has been proposed as one which embraces both these phenomena.

Substantial falls in death rates from infectious and parasitic diseases and maternal mortality are hallmarks of the epidemiological transition. In England and Wales, over half the gain in life expectancy at birth between 1871 and 1911 was due to reduced infectious disease mortality. Some 20% of the total gain was due to reduced death rates from tuberculosis (Caselli 1991). Declines in these causes of death were greater among the young, the epidemiological transition has also been consistently associated with larger falls in mortality among women than men. Changes in the intrahousehold allocation of resources, declines in causes of death primarily affecting women (such as maternal mortality and respiratory tuberculosis), gender differences in health-related behaviour and in exposure to occupational hazards, and the possibly greater susceptibility of men to stresses associated with socioeconomic changes, may all be underlying factors.

The relative contribution of various eighteenth and nineteenth century developments in promoting the historical epidemiological transition in the West remains a matter of debate. Improved nutrition, better housing and living conditions, public sanitation schemes, and specific public health initiatives, such as smallpox inoculation all have their adherents. In the early twentieth century, improved personal hygiene practices and better infant care were also important. A common thread linking most of these factors is their relationship to overall social and economic development and improvements in standards of living. During the twentieth century, however, developments in medical technology and vector control offered the potential for

'exogenous' mortality decline less dependent on a particular country's level of income and development. One consequence was that the relationship between *per capita* income and life expectancy has shifted to the right (Preston 2007). In 1901, for example, life expectancy in the United States was 49 and income *per capita* was about US$7,300 (2005 purchasing power parity). In 2009 income *per capita* in China was almost identical ($7,400) but life expectancy was 73 years (http://www.gapminder.org/).

Many poor countries have been able to achieve remarkable falls in mortality, especially child mortality, through behavioural change, improved education of women, and introduction of relatively cheap treatments and interventions, such as vaccination and antibiotics (Cutler et al. 2006; WHO 2013). Between 1975–80 and 2010–15, life expectancy at birth increased from 52 to 71 in Bangladesh and from 58 to 69 in Indonesia (UN 2019) and there are now a number of poor or middle-income countries with life expectancies at birth as high as in the United States.

The process of the epidemiological transition (or at least the initial phases) is now complete or underway in much of the world and non-communicable causes of death predominate in all regions, except sub-Saharan Africa.

Recent demographic trends and public health

Population size, growth and age structure are all outcomes of variations in demographic behaviours and all have implications for population health and well-being. Population ageing will almost certainly be the predominant demographic issue of the twenty-first century in nearly all richer and a growing number of poorer countries. The strong association between age and risks of health impairment and disability imply growing needs for support services, even if levels of disability fall. In those countries which are growing old before they grow rich, changes in family support systems for older people may pose additional challenges. A range of strategies for responding to these challenges have been proposed including encouraging longer working lives through review of retirement and pensions policies, better organization of acute and long-term care services and initiatives to promote healthy ageing (Rechel et al. 2013; WHO 2015).

Changes in marriage and family patterns also have other public health implications. In North America and North West Europe (and also Latin America and the Caribbean) high rates of divorce and non-marital

childbearing mean that increasing proportions of children are spending at least part of their childhood in lone parent families. Although causal pathways are difficult to elucidate because of various selection effects, there is evidence indicating poorer health among lone mothers and their children, and among unmarried (especially divorced) people more generally, so these trends have some negative implications. Moreover, in the United States and a number of other populations there are indications of a widening socioeconomic (and ethnic) divergence in family patterns (Cherlin et al. 2016) which may contribute to, or compound, the increases in health disparities observed in many populations.

Continuing improvement in child and adult mortality is projected for the poorer world. In the poorest countries the interaction of rapid population growth, environmental degradation and conflict pose continuing health problems and some demographers have argued that population growth has already reached the point at which disastrous environmental and economic implications are inevitable (Dyson 2010). In other middle and low income countries, patterns of tobacco use are likely to have a substantial effect on health trends in coming decades (West 2006). Rapid increases in childhood obesity pose another challenge (Wang and Lobstein 2006).

Issues such as international migration, economic and cultural globalization, and climate change all have substantial health implications for the rest of the twenty-first century; all interact with demographic patterns and processes. Measuring these trends and assessing their effect on health and demand for healthcare requires an understanding of population dynamics and population-based measures, and suitable demographic data. Demography is thus an essential component of public health.

REFERENCES

Abbasi-Shavazi, M., McDonald, P., and Hosseini-Chavoshi, M. (2009). *The Fertility Transition in Iran: Revolution and Reproduction*. Dordrecht: Springer.

Alders, M., Keilman, N., and Cruijsen, H. (2007). Assumptions for long-term stochastic population forecasts in 18 European countries. *European Journal of Population*, 23(1), 33–69.

Alexandratos, N. (2005). Countries with rapid population growth and resource constraints: issues of food, agriculture and development. *Population and Development Review*, 31, 237–58.

Anderson, R.N. (2011). Coding and classifying causes of death. In: Rogers, R.G. and Crimmins, E.M. (eds.) *International Handbook of Adult Mortality.* Dordrecht: Springer, pp. 47–489.

Billari, F, and Kohler, H.P. (2004). Patterns of low and lowest-low fertility in Europe. *Population Studies*, 58, 161–76.

Bloom, D.E., Canning, D., and Malaney, P. (2000). Population dynamics and economic growth in Asia. *Population and Development Review*, 26(Supplement), 257– 90.

Bloom, D.E. et al. (2015). Macroeconomic implications of population ageing and selected policy responses. *The Lancet*, 385(9968), 649–57.

Bongaarts, J. (1978). A framework for analysing the proximate determinants of fertility. *Population and Development Review*, 4, 105–32.

Bongaarts, J. (2010). Human population growth and the demographic transition. *Philosophical Transactions of the Royal Society of London, Series B, Biological Sciences*, 364(1532), 2985–90.

Bongaarts, J. and Bulatao, R.A. (eds.) (2000). *Beyond Six Billion: Forecasting the World's Population*. Washington, DC: National Academy Press.

Bongaarts, J. and Feeney, G. (1998). On the quantum and tempo of fertility. *Population and Development Review*, 24, 271–91.

Bryant, J. (2007). Theories of fertility decline and the evidence from development indicators. *Population and Development Review*, 33, 101–27.

Caldwell, J.C., Orubuloye, I.O., and Caldwell, P. (1992). Fertility decline in Africa: a new type of transition? *Population and Development Review*, 18, 211–42.

Caldwell, J.C. (2007). *Demographic Transition Theory*. Dordrecht: Springer.

Case, A. and Deaton, A. (2015). Rising morbidity and mortality in midlife among white non-Hispanic Americans in the 21st century. *Proceedings of the National Academy of Sciences*, 112(49), 15078–83.

Caselli, G. (1991). Health transition and cause specific mortality. In: Schofield, R., Reher, D., and Bideau, A. (eds.) *The Decline of Mortality in Europe*. Oxford: Clarendon Press, pp. 68–96.

Casterline, J.B. (2010). *Determinants and Consequences of High Fertility: A Synopsis of the Evidence*. Washington, DC: World Bank.

Cherlin, A.J., Ribar, D., and Yasutake, S. (2016). Nonmarital first births, marriage, and income Inequality. *American Sociological Review*, 81, 771–9.

Cleland, J., Harbison, S., and Shah, I.H. (2014). Unmet need for contraception: issues and challenges. *Studies in Family Planning*, 45, 105–22.

Coale, A. (1973). The demographic transition. In: International Population Conference 1973. Liège: International Union for the Scientific Study of Population.

Coale, A. and Watkins, S.C. (eds.) (1986). *The Decline of Fertility in Europe*. Princeton, NJ: Princeton University Press.

Coleman, D. and Salt, J. (1992). *The British Population: Patterns, Trends and Processes*. Oxford: Oxford University Press.

Cutler, D., Deaton, A., and Lleras-Muney, A. (2006). The determinants of mortality. *Journal of Economic Perspectives*, 20(4), 97–120.

Davis, K. and Blake, J. (1956). Social structure and fertility: an analytic framework. *Economic Development and Cultural Change*, 4, 211–35.

Dyson, T. (2010). *Population and Development: The Demographic Transition*. London: Zed Books.

Fernandes A, and Pereira Miguel J. (eds.) (2009). *Health and Migration in the European Union: Better Health for All in an Inclusive Society.* Lisbon: Instituo Nacional de Saude Doutor Ricardo Jorge.

Gajalakshmi, V. and Peto, R. (2004). Verbal autopsy of 80,000 adult deaths in Tamilnadu, South India. *BMC Public Health*, 4, 47.

GBD Mortality and Causes of Death Collaborators (2016). Global, regional, and national life expectancy, all-cause mortality, and cause specific mortality for 249 causes of death, 1980–2015: a systematic analysis for the Global Burden of Disease Study. *The Lancet*, 388, 1459–544.

Goldstein, J.R., Sobotka, T., and Jasilioniene, A. (2009). The end of 'Lowest-Low' Fertility? *Population and Development Review*, 35(4), 663–99.

Green, M., Dorling, D., and Minton, J. (2017). The geography of a rapid rise in elderly mortality in England and Wales, 2014–15. *Health & Place*, 44, 77–85.
Kennedy, S., et al. (2015). The healthy immigrant effect: patterns and evidence from four countries *International Migration & Integration*, 16, 317.
Kinsella, K. and Phillips, D. (2005). Global aging: the challenge of success. *Population Bulletin*, 60(1), 3–40.
Lee, R. (2003). The demographic transition: three centuries of fundamental change. *Journal of Economic Perspectives*, 17, 167–90.
Lee, R. and Mason, A. (2017). Cost of ageing. *IMF Finance and Development*, 54(1), 7–9.
Lesthaeghe, R. and Vanderhoeft, C. (2001). Ready, willing and able: a conceptualization of transitions to new behavioural forms. In: Casterline, J.B. (ed.) *Diffusion Processes and Fertility Transition. Selected Perspectives*. Washington, DC: National Academy Press, pp. 240–64.
Lesthaeghe, R.J. and Neidert, L. (2006). The second demographic transition in the United States: exception or textbook example. *Population and Development Review*, 32, 669–98.
Lotka, A. (1907). Relation between birth and death rates. *Science*, 26, 21–2.
Lozano, R., Lopez, A.D., Atkinson, C., Naghavi, M., Flaxman, A.D., and Murray, C.J.L. (2011). Performance of physician-certified verbal autopsies: multisite validation study using clinical diagnostic gold standards. *Population Health Metrics*, 9, 32.
Lutz, W., Butz, W.P., and Samir, K.C. (eds.) (2010). World Population & Human Capital in the Twenty-First Century. Oxford: Oxford University Press.
Manton, K.G., Gu, X., and Lamb, V.L. (2006). Long-term trends in life expectancy and active life expectancy in the United States. *Population and Development Review*, 32, 81–105.
Mathers, C.D., Ma, F.D., Inoue, M., Rao, C., and Lopez, A.D. (2005). Counting the dead and what they died from: an assessment of the global status of cause of death data. *Bulletin of the World Health Organization*, 83, 171–7.
Mathers, C.D., Boerma, T., and Ma Fat, D. (2009). Global and regional causes of death. *British Medical Bulletin*, 92, 7–32.
McNicoll, G. (2006). Policy lessons of the east Asian demographic transition. *Population and Development Review*, 32, 1–25.
Mikkelsen, L., et al. (2015). A global assessment of civil registration and vital statistics systems: monitoring data quality and progress. *The Lancet*, 386(10001), 1395–406.
Moschetti, K., Cummings, P.L., Sorvillo, F., and Kuo, T. (2012). Burden of Alzheimer's disease-related mortality in the United States, 1999–2008. *Journal of the American Geriatrics Society*, 60(8), 1509–14.
Murray, C.J.L., Lozano, R., Flaxman, A.D., Vahdatpour, A., and Lopez, A.D. (2011). Robust metrics for assessing the performance of different verbal autopsy cause assignment methods in validation studies. *Popular Health Metrics*, 9, 28.
Murphy, M., Grundy, E., and Kalogirou, S. (2007). The increase in marital status differences in mortality up to the oldest age in seven European countries, 1990–99. *Population Studies*, 61(3), 287–98.
Notestein, F.W. (1945). Population: the long view. In: Schulz, T.W. (ed.) *Food for the World*, pp. 36–57. Chicago, IL: University of Chicago Press.
Office for National Statistics (2017). Births by parents' country of birth 2016. [Online] Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/livebirths/bulletins/parentscountryofbirthenglandandwales/2016
Office for National Statistics (ONS) (2018). Statistical bulletin: births in England and Wales: 2017. [Online] Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/livebirths/bulletins/birthsummarytablesenglandandwales/2017

Omran, A.R. (1971). The epidemiologic transition: a theory of the epidemiology of population change. *Millbank Memorial Fund Quarterly*, 49, 509–38.
Preston, S.H. (1976). *Mortality Patterns in National Populations*. New York: Academic Press.
Preston, S.H. (2007). Response: on the changing relation between mortality and level of economic development. *International Journal of Epidemiology*, 36(3), 502–3.
Preston, S.H., Heuveline, P., and Guillot, M. (2001). *Demography: Measuring and Modelling Population Processes*. Oxford: Blackwell Publishers.
Preston, S.H. and Stokes, A. (2012). Sources of population aging in more and less developed countries. *Population and Development Review*, 38(2), 221–36.
Preston, S.H., Vierboom, Y.C., and Stokes, A. (2018). The role of obesity in exceptionally slow US mortality improvement. *Proceedings of the National Academy of Science*, 115, 957–61.
Preston, S. and Wang, H. (2006). Sex mortality differences in the United States: the role of cohort smoking patterns. *Demography*, 43, 631–46.
Rechel, B., et al. (2013). Ageing in Europe. *The Lancet*, 381(9874), 1312–22.
Retherford, R.D., Ogawa, N., and Matsukura, R. (2001). Late marriage and less marriage in Japan. *Population and Development Review*, 27, 65–102.
Ritu, S.R., Mathers, C.D., Lopez, A.D., Murray, C.L.J., and Iburg, K.M. (2002). Comparative analyses of more than 50 household surveys on health status. Chapter 8.1. In: Murray, C.J.L., Salomon, J.A., Mathers, C.D., and Lopez, A.D. (eds.) *Summary Measures of Population Health Concepts, Ethics, Measurement and Applications*. Geneva: WHO.
Rowland, D.T. (2003). *Demographic Methods and Concepts*. Oxford: Oxford University Press.
Samir, K.C. and Lutz, W. (2017). The human core of the shared socioeconomic pathways: population scenarios by age, sex and level of education for all countries to 2100. *Global Environmental Change*, 42 (2017), 181–92.
Siegel, J.S. and Swanson, D.A. (eds.) (2004). *The Methods and Materials of Demography* (2nd ed.). San Diego, CA: Academic Press.
Sobotka, T. (2008). The rising importance of migrants for fertility in Europe. *Demographic Research*, 19, 225–48
UNAIDS (2020). *Seizing the moment: tackling entrenched inequalities to end epidemics.* New York: Joint United Nations Programme on HIV/AIDS.
United Nations, Department of Economic and Social Affairs, Population Division (2015). *World Population Ageing 2015*. New York: United Nations.
United Nations, Department of Economic and Social Affairs, Population Division (2017). *International Migration Report 2017: Highlights* (ST/ESA/SER.A/404). New York: United Nations.
United Nations (2018). *World Urbanization Prospects: The 2018 Revision*. New York: United Nations.
United Nations, Department of Economic and Social Affairs, Population Division (2019). *World Population Prospects: The 2019 Revision*, custom data acquired via website. New York: United Nations.
Valente, P. (2010). Census taking in Europe: how are populations counted in 2010? *Population and Societies*, 467. [Online]. Available at: https://www.ined.fr/fichier/s_rubrique/19135/pesa467.en.pdf
van de Walle, F. (1986). Infant mortality and the European demographic transition. In: Coale, A.J. and Watkins, S.C. (eds.) *The Decline of Fertility in Europe*. Princeton, NJ: Princeton University Press, pp. 201–33.
Vine, J. (2013). *Exporting the Border'? An Inspection of e-Borders*. London: Independent Chief Inspector of Borders and Immigration.
Wang, H., et al. (2012). Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010. *The Lancet*, 380(9859), 2053–4.

Wang, Y.F. and Lobstein, T. (2006). Worldwide trends in childhood overweight and obesity. *International Journal of Pediatric Obesity*, 1, 11–25.
Wang, L., et al. (2007). Evaluation of the quality of cause of death statistics in rural China using verbal autopsies. *Journal of Epidemiology and Community Health*, 61, 519–26.
West, R. (2006). Tobacco control: present and future. *British Medical Bulletin*, 77–78, 123–36.
Williams, G.M., Najman, J.M., and Clavarino, A. (2006). Correcting for numerator/denominator bias when assessing changing inequalities in occupational class mortality, Australia 1981–2002. *Bulletin of the World Health Organization*, 84(3), 198–203.
Wolf, D.A. and Gill, T.M. (2009). Modeling transition rates using panel current-status data: how serious is the bias? *Demography*, 46(2), 371–86.
World Health Organization (WHO) (2013). *Global Health Observatory Data Repository Mortality and Burden of Disease: Child Mortality.* [Online] Available at: http://apps.who.int/gho/data/#
World Health Organization (WHO) (2015). *World Report on Ageing and Health*. Geneva: World Health Organization.
World Health Organization (WHO) (2017). *Global Health Observatory Data Repository: Mortality and Burden of Disease: Child Mortality.* [Online] Available at: http://apps.who.int/gho/data/#
World Health Organization (WHO) (2018). *Global Health Estimates 2016: Deaths by Cause, Age, Sex, by Country and by Region*. Geneva: World Health Organization.
Young, H., Grundy, E., O'Reilly, D., and Boyle, P. (2010). Self-rated health and mortality in the UK: results from the first comparative analysis of the England and Wales, Scotland, and Northern Ireland longitudinal studies. *Population Trends*, 139, 11–36.

6.4

Health promotion, health education, and the public's health

Fran Baum

Introduction

Health promotion is a complex, ambiguous concept and set of practices. While many have linked it, primarily, to a revolution in health education, its roots go much deeper into the history of public health. It had its contemporary beginnings in the throes of the backlash against bureaucratic and professional dominance exemplified by the new social movements of the 1970s and 1980s. At its heart, health promotion is centred on the values and principles of equity, participation, and empowerment. These concepts are embedded in health promotion's founding document, the Ottawa Charter for Health Promotion. However, exactly how these values are articulated is often ambiguous. In this chapter, the authors contend that health promoters must intensify their reflection on these core values and principles; particularly in the light of the tendency to slip back into a comfortable paternalism, which reinforces existing power imbalances. We are specifically concerned with the precise interpretation of health equity in health promotion. In order to pursue a deeper level of reflection on the meaning of equity, it is argued that health promotion must engage more deeply with recent developments in political philosophy, political economy, and social theory. Following up from the work of the World Health Organization (WHO) Commission on the Social Determinants of Health, one of health promotion's emerging priorities is a reinvigorated push for the global development of 'healthy public policy' as seen in the Health in All Policies (HiAP) movement. This chapter, also addresses the need for health promotion to engage more directly with core social-theoretical concepts related to the generation of social inequalities that underlie health

inequities; finally, we explore the relevance of complexity science and systems thinking for health promotion research.

Health promotion, health equity, and action on the determinants of health: an introduction

Previous attempts (Green and Raeburn 1988; Tones and Tilford 1994; Tones 2002) to situate health promotion within the broad field of public health have often used ‘health education’ as the starting point. This is an entirely sensible approach which we will discuss but it tends to underemphasize the *radical* departure health promotion aims to make from traditional public health approaches in general. Without claiming health promotion means everything all at once, thereby leaving its lofty rhetoric in the realm of the aspirational yet ineffectual, we aim to place health promotion more centrally in the ongoing saga of an increasingly globally aware public health. As one of the acknowledged founders and innovators of the modern health promotion movement has unceasingly argued, health promotion, at its most persistent and radical, heralds a ‘new public health’, not merely a more fine-tuned and effective tool-box for a less paternalistic health education (Kickbusch 1989, 2007).

We will begin with a clear definition of health promotion. Then, by unpacking the dense and sometimes opaque wording that define the elements of health promotion as a concept, we intend to open up some of its central, yet often hidden, connections to much broader themes in contemporary social and political movements and ideas.

Next, we situate health education and its internal critique as an important part of the history of health promotion, while providing more context concerning the specific historical/national trajectories that made the genesis of the modern health promotion movement a mixture of different influences, of which health education is only one.

We will consider how health promotion manages its ambiguous relationship with the history and ideological background of public health, and how it sees itself in relation to the past, present, and future of public health.

We suggest there is a new opportunity for health promotion to reconnect with the avant-garde in public health, which can be examined along two

broad dimensions, namely how to achieve health equity, and the question of health in a global political-economic context. Particular attention will be paid to the recent work following up the WHO Commission on the Social Determinants of Health, led by Sir Michael Marmot, and the renewal of the *Ottawa Charter* action area of 'healthy public policy' in the *Health in All Policies* global movement, spearheaded by the WHO.

In this chapter we argue that there is insufficient engagement with social theory within the health promotion field. We offer some suggestions and map out potential pathways towards more serious reflection on this missing social-theoretical base. Finally, we consider the emerging role of complexity theory and systems thinking in health promotion research.

In general, this chapter on health promotion is *critical* in the positive sense of the word. Previous surveys of the concept and practice of health promotion that provide excellent guidance to the field are referenced. However, we consider that practitioners and researchers in public health can benefit from a reflexive inquiry into the rich ambiguities and tensions that are embedded in the discourse and practice of health promotion. This is particularly the case if, as we argue, the development of health promotion *is* the development of a 'new public health' (Baum 2016).

'Health promotion': a definition and conceptual critique

In this chapter, we will follow the *Ottawa Charter for Health Promotion* (WHO 1986) definition of health promotion as 'the process of enabling people to increase control over, and to improve their health'. However, we will also draw upon the expanded definition in the updated *Health Promotion Glossary*:

> Health promotion represents a comprehensive social and political process, it not only embraces actions directed at strengthening the skills and capabilities of individuals, but also action directed towards changing social, environmental and economic conditions so as to alleviate their impact on public and individual health. Health promotion is the process of enabling people to increase control over *the determinants of health* and thereby improve their health. (WHO 1998, our emphasis)

Here, the 'how' and 'why' *ideology* is linked to the 'what' of the *determinants of health* (WHO 1998). This is crucial, because if health promotion is about anything, it is about *action* taken across the broad

spectrum of health determinants, particularly directed towards the social, environmental, and economic conditions that support health (WHO 1984).

The *Glossary* also emphasizes that 'participation is essential to sustain health promotion action', and identifies the three *Ottawa Charter* strategies for health promotion: '*Advocacy* for health to create the essential conditions for health indicated above; *enabling* all people to achieve their full health potential; and *mediating* between the different interests in society in the pursuit of health.'

These strategies are supported by five priority action areas as outlined in the *Ottawa Charter*:

Build healthy public policy
Create supportive environments for health
Strengthen community action for health
Develop personal skills, and
Re-orient health services (WHO 1986)

As one can see from these definitions, we already have started the 'unpacking'. A more nuanced analysis of some of the key elements in this definition is described in the following subsections: the 'process' of health promotion; 'enabling' and 'empowering'; and for what, the outcome, 'improved health'. This will be followed by an analysis of the *Ottawa Charter strategies* and its *priority action areas* (often called 'action strategies').

As soon as we begin, we find ourselves in murky waters, though not without some guidance. Ironically, although health promotion is, as Tones and Green (2004) noted, an 'essentially contested concept', there has been a remarkable degree of effort, and consensus, concerning its ostensive definition. Few, if any, health promoters dispute the *Ottawa Charter*'s now canonical phrasing (WHO 2005).

The real ambiguity that surrounds the concept of health promotion is embedded in the elision of the concrete meaning of the elements that make up its agreed upon definition.

Health promotion as a process

The emphasis on *process* is important, if only because it warns against reducing health promotion to merely a technical function of public health. It connotes the wider meaning of the concept by signalling that the radical departure and critique of traditional public health lies in its advocacy for changing the *way* we do public health, just as much as *what* we change and *why* we change.

Health promotion, whether it be generated from an internal critique of health education or from other dissatisfactions with the way public health was being practised, is fundamentally concerned with change and, specifically, with the failure of traditional, paternalistic, and professionally dominated public health processes to bring about positive changes in health, particularly for those groups that suffer disproportionately negative health outcomes and the consequent disadvantages. What is substituted is a call for health promoters to create a dynamic, participatory engagement with individuals and communities, to help or 'enable' them to take control over the determinants of their own health.

The first ambiguity we meet when trying to analyse this 'process' turns on whether one interprets the health promotion process as, primarily: a revamped tool-box of health education techniques and social marketing devices, with a rhetorically efficient participatory gloss; or, as a values-based process of communicative interaction that has as its central premise the ethical foundation of respect for human dignity and autonomy. These are certainly polar extremes and there is no doubt that there is room for both aspects in a broad, ecumenical attitude to a diverse field of practical action. Yet, because health promotion is a process often dominated by professionals, in a context where its supposed beneficiaries are often those in a position of relative powerlessness, the tendency for professionals to retreat to an insulated cocoon of technical expertise is strong. The essence of the health promotion process is a focused shift of power from professionals to the community and to individuals within their communities who historically have had less power. To do this, it is crucial that the 'process' we focus on is the one that involves negotiating values, principles, ethics, and power, not the less complicated one of transferring a packet of new skills and technical tools to a community that is presumed to lack capacity. In order to achieve this shift in power, health promoters need to begin by examining their own values and assumptions that inform their actions.

Beliefs and assumptions underlying health promotion

Health promotion practice is influenced by the beliefs and assumptions we hold. While a detailed discussion of this topic is beyond the scope of this chapter, we outline how certain beliefs and assumptions influence how we act in health promoting ways in given situations.

Beliefs are learned through life experiences; they are what we hold as 'true'. They are convictions that influence the way we think, feel, and act. Health promotion practice relies on a set of underlying assumptions that guide those who work in the field. Hartrick et al. (1994) contend health promotion is a 'way of being' that requires certain convictions in order to act in health promoting ways. These include:

All people have strengths and are capable of determining their own needs, finding their own answers, and solving their own problems.
Every person and family lives within a social-historical context that helps shape their identity and social relationships.
Diversity is positively valued.
People without power have as much capacity as the powerful to assess their own needs (people are their own experts).
Relationships between people and groups need to be organized to provide an equal balance of power (this includes professional/client relationships).
The power of defining health problems and needs belongs to those experiencing the problem.
The people disadvantaged by the way that society is currently structured must play the primary role in developing the strategies by which they gain increased control over valued resources.
Empowerment is not something that occurs purely from within (only I can empower myself), nor is it something that can be done to others (we need to empower the group). Rather, empowerment describes our intentional efforts to create more equitable relationships where there is greater equality in resources, status, and authority.
Shared power relations do not deny health professionals their specialized expertise and skills. Rather, professional expertise and skills are used in new ways that result in greater power equity in interpersonal and social relations (Hartrick et al. 1994, p. 87).

So, for example, if we consider the first assumption in the list, believing that people are able to find their own answers and solve their own problems leads one to act in empowering ways because of the belief that people have this capacity to figure things out. On the other hand, if one believes that people need to be told what to do, or that they are not able to figure out issues on their own, it is more difficult to create conditions that are enabling. For some, it might even feel irresponsible to put people in these circumstances or to allow them to have control over these types of decisions.

Enabling and empowerment

At the time of the *Ottawa Charter*, the word ‘enabling’ was favoured, although later this tended to be replaced with the more direct and comprehensive concept of ‘empowerment’. Essentially, this meant that a prerequisite for the new approach was that individuals and communities were to directly participate in the planning and implementation of health promotion activities. The assumption was based on the notion that only by genuinely participating in the health promotion process would people be ‘enabled’ or ‘empowered’ to take control of what determined their health. However, the concept of ‘enabling’ also referred to the more general process of changing the social, economic, and environmental conditions that made it difficult for people to become empowered. There is a deep ambiguity here: it is not clear, for some commentators, whether more macro-scale action, at a policy level, also requires active participation of local communities. There is some room for an interpretation that tends to retain a traditional paternalism when it comes to healthy public policy, leaving the ‘participatory’ aspect of health promotion to the realm of ‘community action’. As will be argued throughout this chapter, health promotion is constantly at risk of sliding back into this paternalistic approach, leaving the more ‘complex’ and high-level ‘technical’ decisions to the experts. Yet, if there is a direct link between human dignity, autonomy, and equity, then *all* aspects of health promotion must integrate the fundamental perspective of participation. In fact, it is argued that the rhetoric of ‘empowerment’ often masks a continuing bureaucratic and

professional dominance of the process of improving public health (Baum 2007).

A key aspect of this ambiguity can be seen when we consider the link to health inequity. Without a genuinely participatory, empowering process, it is those worst off who are left further behind as they suffer, not only a failure to affect those conditions most important to their health, but also a direct assault on their human dignity (Sennett 2003). Those who tend to manage any gains from processes that lack true participation are usually segments of the population that already have access to positions of status and the resources and capacities to take advantage of the interventions on offer. The distinction here is between a situation where already disadvantaged people are assumed to be too ignorant or incapable of participating and thus have solutions imposed on them, and a situation where people of a privileged status delegate, as equals, to professional experts. This is not to say that a participatory process is not better for everyone, regardless of class position or status; rather, it is to emphasize that non-participatory processes have a disproportionate adverse effect on disadvantaged groups.

On the positive side, an empowering health promotion process leaves the ownership and control of a health promotion activity or programme in the hands of the community itself. This is particularly important in communities that have suffered historical social injustices and have thus been actively ‘disempowered’ (an ugly but accurate term). Allowing people to participate in a genuine way in determining not only *what* they want but *how* they want to get it is demonstrably the most effective strategy for change. It is also the *only* strategy for sustaining progress in improving health and shifting control back to the community and away from a negative dependence on bureaucratic and professional power. In this model, professionals are not demons; they are just transformed from arrogant experts into supportive servants of the will of the community.

There is in these simple terms (‘enable’ or ‘empower’) the entire, complex, and ambiguous story of health promotion. All the themes that will be touched upon in this chapter can be traced back to just what is at stake in the ostensive goal of ‘empowering’ people to take control over what determines their health.

What is the ‘health’ in health promotion?

Understanding how we conceptualize health is a key reflective step in health promotion. How we think about health largely determines the types of action we take to promote health. We see next how different historical conceptions of health still shape the contemporary health landscape and continue to sustain ambiguities in how people approach health promotion itself. In the twentieth century, due to the relative success of the sanitation approach to public health and the emergent hegemony of the germ theory of disease, an implicit biomedical definition of health as *the absence of disease* dominated, and along with it, a narrow, individual treatment focus, centred on the healthcare system, was the preferred solution. The WHO had, in 1946, introduced the *positive* definition of health as ‘a state of complete physical, mental, and social well-being and not merely the absence of disease or infirmity’ (WHO 1946). Nevertheless, this definition had little concrete impact on actual health systems, leaving the absence of disease approach to health as the default option when governments turned their attention to the public’s health.

The Lalonde Report, an official document produced by the Canadian Department of Health and Welfare (Lalonde 1974), marked a significant change in thinking about health. Although the report is recognized internationally as the first government document to suggest that health promotion could be a key strategy for improving health, its other, more significant contribution, was to redefine how we view health. Lalonde’s report argued that the healthcare system plays only a small part in determining health, and suggested that health was determined by the interplay between human biology, healthcare organization, environment, and lifestyle. This view of health became known as the ‘lifestyle or behavioural approach to health’, partly because the ‘environmental’ dimension was either ignored or treated narrowly.

With the publication of the discussion paper on concepts and principles of health promotion (WHO 1984) and the endorsement of the *Ottawa Charter for Health Promotion* (1986), a third view of health arose: the socioecological approach. This approach defined health as ‘a resource for everyday life, not the objective of living’. ‘Health is a positive concept emphasizing social and personal resources as well as physical capacities’ (1986, p. 1). In order to reach this state of physical, mental, and social well-

being, people must be able to identify and realize their aspirations, to satisfy their needs and to change, or cope, with their environment. This inextricable link between people and their environment provides the conceptual basis for this socioecological perspective on health and it forms the conceptual base for health promotion practice.

At first glance, these different views of health may appear to be developmental or historical. However, Labonté and others (Labonté 1993; Rootman and Raeburn 1994; Raeburn and Rootman 2007) argue that, in fact, all three views of health (along with many other definitions) continue to be endorsed by different people in the field of health promotion and, furthermore, that the view of health one holds influences one's health promotion practice. Table 6.4.1 illustrates this connection between how we think about health, our view of health, and our actions (health promotion practice). For example, if we hold a view that health is the absence of disease, we are likely to talk about disease processes and risk factors and to manage the problem professionally by prescribing a treatment. If we hold a socioecological view of health, we are more likely to focus on the conditions in which the person is living, the factors that are influencing their ability to meet their needs, and to use enabling strategies to assist the person to have more control over their health.

Table 6.4.1 From concept to action: different approaches to health

	Medical approach	Behavioural approach	Socioecological approach
Health concept	Biomedical; absence of disease or disability	Individualized; physical-functional ability, physical well-being	Positive state connectedness; ability to do things that are important or have meaning; psychological well-being
Health determinant	Disease categories, physiological risk factors (e.g. hypertension)	Behavioural risk factors (e.g. unsafe sex)	Psychological risk factors (e.g. isolation) and socioenvironmental risk conditions (e.g. poverty)
Principal strategy	Surgery, drugs, therapy, illness care, medically managed behavioural change	Advocacy for healthy lifestyle choices	Personal empowerment, small group development, community organization, coalition advocacy, political action
Programme development	Professionally managed	Negotiated with communities and professionals	Managed by community in critical dialogue with supporting professionals and agencies

The *Ottawa Charter* strategies

The three strategies mentioned in the *Ottawa Charter* are: *to advocate*, *to enable*, and *to mediate*. We have already reflected upon the second strategy, as it is part of the definition of health promotion. However, a few words need to be said about the other two strategies.

The concept of *advocacy* receives very little elaboration in the *Ottawa Charter*. In the *Health Promotion Glossary*, it is stated that advocacy 'can take many forms including the use of the mass media and multimedia, direct political lobbying, and community mobilization through, for example, coalitions of interest around defined issues. Health professionals have a major responsibility to act as advocates for health at all levels in society' (WHO 1998, p. 6). This raises one of the many thorny issues that come up when professionals are caught between highly mobilized and often highly critical communities and a state bureaucracy that is extremely reticent about providing funding that sanctions and supports the capacity for critical attention to its policies and programmes. Even at the level of

independent professional organizations, the participation and funding provided by government bodies creates a tension around the organization taking strong critical perspectives. Another aspect of this strategy, as defined in the *Glossary* and glossed over, is the potential contradiction between activities such as 'political lobbying' and 'community mobilization'. Often, the same organization or individual will be less effective as a political lobbyist to the extent they are perceived to be directly associated with community mobilization efforts that the powerful are either indifferent to, or actively disfavour.

Mediation is an even more delicate strategy for health professionals. Its original *Glossary* definition in relation to health promotion was: 'A process through which the different interests (personal, social, economic) of individuals and communities and different sectors (public and private) are reconciled in ways that promote and protect health' (WHO 1986). In the expanded definition of 1998, more explicit emphasis is given to the potential conflicts that often arise between the competing interests mentioned in the original definition. However, the goal of mediation as 'reconciliation' is left unchanged. While there is nothing inherently wrong with the idea of reconciliation, professionals should be extremely self-critical and reflexive when operating with this strategy. Two dangers are apparent with the strategy. First, in striving for 'reconciliation', one may simply paper over a conflict for the purposes of short-term peace, while leaving the principal reasons behind the conflict intact, thereby creating the potential for longer-term embitterment and strategic action by all parties, which ultimately undermines the appearance of agreement. Second, a very real threat to equity can arise when professionals reconcile a conflict between the powerful and the powerless and end up re-enforcing the powerful at the expense of the powerless. This tendency is very strong given the fact that professionals have little to gain personally from any radical re-structuring of power relations. Despite these important caveats, mediation has become even more critical for the success of health promotion in the future, especially in relation to the new global, multilayered context of health governance that the *Bangkok Charter* (WHO 2005) has set out to address and which will be discussed later.

The *Ottawa Charter* action areas

The priority action areas of the *Ottawa Charter* were identified as those areas that were seen at the time of the charter (and still to this day) as critical arenas for health promotion's strategic activities. We will not try and survey the myriad accomplishments of health promotion activity; rather, consistent with our general approach, we will offer a few critical comments on each action area:

Building healthy public policy

There are three elements of healthy public policy emphasized in the *Ottawa Charter*:

If, as the Lalonde Report (1974) argued, the determinants of health lay mainly outside healthcare itself, then policy action must come from policy sectors other than health. The health sector would still play an important, but not exclusive, role in public policy action to support health.
Healthy public policy requires the coordinated use of all policy levers available, including 'legislation, fiscal measures, taxation, and organizational change' (WHO 1986).
Healthy public policy requires the identification and removal of obstacles to the adoption of such policies in non-health sectors.

Without going into a long list of efforts and results in this area, the progress made can be summed up as substantial and encouraging in regard to changes in the rhetoric and discourse around health, in both developed nations and in many of the global institutions responsible for improving health and development worldwide. Conversely, one can equally characterize progress as ephemeral and demoralizing when it comes to the concrete goal of 'coordinated action that leads to health, income and social policies that foster greater equity' (WHO 1986). For a variety of reasons discussed in the later subsection on the political economy of health promotion, given the increasingly urgent crisis of widening inequities in health both between and within societies, the collective policy response of the most powerful countries on earth has been miserly and despicable. To call it 'inadequate' is a gross understatement and an unconscionable euphemism.

When *action* finally starts to catch up with some of the lofty rhetoric behind the calls for 'health in all policies' (Ståhl et al. 2006), health promotion can begin to find some satisfaction in the area of building healthy public policy. The COVID-19 pandemic which started in 2020 has seen many governments in high-income countries adopt social and economic policies which would fit under the rubric of Health in All Policies including increased income support, free childcare, and support to workers to maintain their employment. The WHO (2010) is supportive of and is promoting the use of the Health in All Policies approach. We will discuss some of the latter recent developments in relation to the *Health in All Policies* movement in a later section.

Creating supportive environments

This area forms the basis for what is called the *socioecological* approach to health. Here it is asserted that both the natural and built environments are inextricably linked with people's health. It is crucial to understand that the conceptualization of supportive environments given here is consistent with the expanded, positive understanding of health as a 'resource for everyday life'. It is not merely about threats to physical health, but involves creating conditions that allow people to have 'living and working conditions that are safe, stimulating, satisfying, and enjoyable'. This entails the complex relationships between rapidly changing technologies, working conditions, resource use, climate change, urbanization, and health (among others). At the 2019 global meeting of the International Union for Health Promotion and Education was on the theme *Promoting Planetary Health and Sustainable Development for All* and demonstrated the extent to which the international health promotion community has embraced the need to ensure planetary health. The conference statement defined planetary health in a holistic way:

> Planetary health is the health of humanity and the natural systems of which we are part. 1 It builds on Indigenous peoples' principles of holism and interconnectedness, strengthening public health and health promotion action on ecological and social determinants of health. It puts the wellbeing of people and the planet at the heart of decision-making, recognising that the economy, as a social construct, must be a supportive tool fit for this purpose in the 21st century. (IUHPE 2019)

In considering progress, past endeavours, and future prospects in this area, one must take into account the lofty ambition (and some would say naïvety) of this programme of action. As a project of knowledge development, its referral to the ‘complex interrelatedness’ of contemporary societies is but a cipher for the entire corpus of theoretical and empirical dispute and debate within the social sciences over how to characterize what are now acknowledged to be multiple, interrelated, global, national, regional, and local processes of socioeconomic and cultural change (Held et al. 1999). Later we consider some recent moves towards adopting complexity science and systems thinking to more adequately address this complex interrelatedness.

In the real world, we are not able to coordinate all the best knowledge sources available and neatly calculate what is best for health. Instead, we are left with tools like ‘health impact assessment’ (HIA) and health equity impact assessment (Povall et al. 2014). The action area of creating supportive environments can be seen as a great boon to academic productivity, both theoretical and empirical; yet, before the final judgements of the academy can be handed down, actions must be taken and decisions must be made.

Communities and developers, politicians, and bureaucrats must decide whether to build this or that highway, licence this or that mining operation, enact this or that employment regulation, and build this or that oil pipeline. We are thus forced, by the necessity to decide and act, into an inevitable reduction of complexity. The question is not whether this is a good or bad thing; it is *how*, by what *process*, is complexity reduced? Whither participation and empowerment in a field dominated by professional expertise and the cloistered secrecy of executive and administrative decision-making in both the public and private sectors?

We argue *against* the implication that instruments like HIA inevitably involve participatory processes (Kemm 2006). The natural tendency is always to define ahead of time, objectively, what elements of the built or natural environment are most important for enhancing people’s health. From a utilitarian perspective, locally defined needs and wishes may even legitimately be ignored in the name of some greater good for a larger population. However, health promotion should always err primarily on the side of the fundamental value of the autonomy and the dignity of people and their communities. In this mode, participation is foundational, even in

what are prima facie obvious areas for the guidance of refined professional expertise. In enquiring into the best way to protect and enhance the built and natural environment for health, the first step is to find out what people actually identify as the things that would make life ‘safe, stimulating, satisfying, and enjoyable’. From the professional perspective this route has one incontestable drawback: people are inevitably confused, ignorant, inconsistent, contradictory, and even just ‘wrong’. ‘People’ will disagree with each other; will get annoyed or, even worse, angry; will disrespect experts, politicians, lawyers, and any number of people who ‘actually know’ about the issue. What is feared here is what is fondly called deliberative politics; in other words, the foundation of democratic civil society.

We are, as is universally acknowledged in the health promotion community, a long way from creating supportive environments for health, especially for those suffering gross inequities in social conditions and in consequent health outcomes. What is less often acknowledged is that part of the reason this is so difficult is that we consistently exclude the very people we are meant to be helping from determining the goals, and strategies necessary to move from here to there. Once again, we are led to believe that the ends can justify the means; we can have non-participatory processes as long as we intend to make changes to enhance the lives of the less fortunate. That we end up in a place we did not intend, is inextricably linked to the fact that, at crucial junctures, when inevitable changes of directions and compromises are made (local development processes are a prime example), the people who have an inherent interest in speaking up for the powerless (the powerless themselves) are nowhere to be seen, or are barely heard.

Strengthening community action

This action area is at the very heart of health promotion; in fact, it can be argued that this action area is the one where the basic principles of health promotion lie. You can imagine (wrongly) participation, equity, and empowerment to be contingent add-ons to the other action areas; with strengthening community action the essential unity of all the values of health promotion are embedded as necessary features of its realization. Indeed, what we see in this area is the place where the true spirit of health promotion is anchored in community development as a process. In fact, in

the *Ottawa Charter* itself, there is a strange ellipsis where the term community development is introduced in the section on strengthening community action. It is as if one missed something: there is no linking phraseology relating community development to strengthening community action. This, we surmise, is no error: strengthening community action quite simply *is* community development.

Consider the definition of community development as agreed upon at the International Association for Community Development at a meeting in Budapest in 2004. Community development is a:

> way of strengthening civil society by prioritising the actions of communities, and their perspectives in the development of social, economic and environmental policy. It seeks the empowerment of local communities, taken to mean both geographical communities, communities of interest or identity and communities organizing around specific themes or policy initiatives. (Craig 2005, p. 3)

For a more sustained treatment of the need to recognize the central place processes of community development and empowerment should play in health promotion, Raeburn and Rootman's *People-Centred Health Promotion* is an essential reference (Raeburn and Rootman 1998). Raeburn and Rootman draw heavily, in their chapter on community development, on Meridith Minkler's important piece, 'Improving health through community organization' (1990). In this seminal piece, Minkler outlines the five principles she sees as foundational to community organization or development:

Empowerment
Community competence
Participation
Issue selection
Creating 'critical consciousness'.

As is now obvious, we have run into some of these principles already, the key to which is that the community itself has collective control over the process of identifying issues and planning how to address them. In addition, the important notion of 'critical consciousness' is raised. This refers to the need for critical dialogue in the Freirian sense (Freire 1972); this is particularly important when working with historically oppressed or disadvantaged communities.

Developing personal skills

The new wave of health promotion has often downgraded attention to this critical aspect of its mandate. The *Ottawa Charter* tells us that health promotion 'supports personal and social development through providing information, education for health, and enhancing life skills'. However, since the *Ottawa Charter*, health promoters, with some exceptions, have tended to either ignore or aim strong criticism at the developing personal skills area. This has come about for three reasons. First, as part of the critique of health education, it was argued that individually focused education approaches were generally ineffective in bringing about health promoting behavioural change; instead, a switch to an emphasis on the *social* factors that influence health was necessary to overcome the limitations of traditional counselling and other interventions circumscribed by the discipline of psychology. Second, the emphasis on developing personal skills was associated with the 'victim-blaming' element that many health promoters saw as the consequence of an interpretation of the Lalonde Report and other government documents (particularly, the approach taken in the United Kingdom under the Thatcher governments) in the context of the neo-liberal rolling back of the state's commitment to a strong social safety net. Finally, although developing personal skills is a central mechanism for empowering individuals to take control over their own health, many worried that, in this narrow approach, the *collective* strengthening of communities was adversely effected by too much emphasis on individual empowerment.

All of these concerns are legitimate, though in each of them there is a high risk that we will miss important opportunities by mistaking what are contingent tendencies for essential features.

As has recently been argued, the ignorance of, or even hostility to, work in this area may seriously damage health promotion's potential impact (Godin 2007). First, while many health education and behaviour change approaches are limited in their effectiveness, there are some demonstrably effective interventions that should not be ignored (Kok et al. 1997). Furthermore, we can learn and are learning about why some approaches in this area have not been effective. Second, the fact that this area of health promotion can be enlisted as part of a more general 'victim-blaming' culture of health promotion, does not mean it must be enlisted; in fact, to

the extent that genuinely empowering health education is taken seriously, the resulting improvements in self-esteem should work against victim blaming. Third, individual and community empowerment should not be a zero sum trade-off. It is only when an exclusive focus on individuals is emphasized that we will have the phenomenon of rescuing survivors from a sinking ship. We also note that social media is offering more and more opportunities for effective health promotion. The pervasive use and influence of the Internet and social media in nearly all parts of the world, presents new opportunities and challenges for health promotion and health education (see Chapter 4.3 for extended discussion of new communication technologies). A key feature that differentiates social media from more traditional communication processes is its interactive nature, where communications are not a one-way process, and where users also play an active role. This allows the formation of new online communities, which can enable virtual participation and collaboration among its members and lead to effective mobilization of social movements and the development of personal skills.

In summary, while we must be vigilant against the temptation and limitations of an individually focused, skills development approach, we must also re-engage with the most advanced and progressive elements in this area of work. If we fail to do this, we will jeopardize a key aspect of health promotion.

Reorienting health services

While we have made some important gains in the previous four areas, reorienting health services has proved more difficult. In general, throughout the world, health services remain medically dominated, cure and treatment focused, and individualistic. The *Ottawa Charter* states ‘the role of the health sector must move increasingly in a health promotion direction, beyond its responsibility for providing clinical and curative services’. However, ‘across the world, there appears to have been a stubborn resistance to systematic change in healthcare services and only limited examples of effective and sustainable health services reorientation’ (Wise and Nutbeam 2007). Health services need to embrace an expanded mandate which is sensitive and respects cultural needs. This mandate should support the needs of individuals and communities for a healthier life, and open

channels between the health sector and broader social, political, and physical environmental components.

Health services are broad and far-reaching, with the most complex service for health promotion being the acute care hospital setting. There has been some advancement in this area with the creation of the healthy hospital settings movement, with some research evidence that it is possible to practice from a health promotion perspective even within this particularly medically dominated environment, for nurses at least (Hills 1998). However, we want to focus in this chapter on the area where health promotion should flourish but has not as yet—primary healthcare. Many who work in health promotion would argue that the *Alma-Ata Declaration* (WHO 1978) was the precursor for the *Ottawa Charter*. These two documents share the same values, principles, and basic tenets; the *Alma-Ata Declaration* addresses health systems more particularly while the *Ottawa Charter* has a broader mandate. But it is their relationship that provides the key to reorienting health services. That is, primary healthcare is a place for health promotion to focus its energy in terms of reorienting health services. In fact, the more that health promotion disassociates itself from primary healthcare, the more we give the impression that it is in the domain of medicine, not health. 'The more health promotion becomes distinct from the world of curative care, the more the latter is allowed to continue to be seen as the real work of medicine' (MacDonald 1992).

We want to be clear that when we are talking about primary healthcare, we are not talking about primary care. These terms are often confused or used interchangeably. Primary healthcare refers to the philosophy and principles articulated in the *Alma-Ata Declaration* (WHO 1978). It calls for universal access to health services (universality) and the removal of geographic, social, economic, or cultural barriers to access (accessibility); it demands community participation in planning, operation, and evaluation of health services (participation); it requires integration across health and other sectors such as housing, education, and employment; it recognizes the power of multidisciplinary teams working as equal partners for the health of the community; it focuses on a range of services, determined by the community, that include health promotion, primary prevention, rehabilitative, and curative (essentiality); and, it demands a commitment to equity concerning issues of power and resources (equity and access). Therefore, primary healthcare resists the conceptual and operational

separation of treatment and prevention which fits the engineering model of healthcare, with prestige and often scarce resources going to clinical medicine to the neglect of prevention, promotion, and rehabilitation (MacDonald 1992). The importance of primary healthcare was reinforced by WHO in its Astana Declaration (WHO 2018).

People who work in health promotion and understand its philosophy and principles must be involved in the development and implementation of primary healthcare. Many people working in health promotion are of the opinion that health promotion and healthcare are distinct and separate entities. They are critical of health promoters who talk about healthcare or health service delivery at the same time that they are talking about health promotion. We have a different opinion: it is necessary, not only to talk but to act as health promoters to facilitate primary healthcare reform.

Besides the *Ottawa Charter* outlining our responsibility to take up this challenge, there are two other reasons why it must be people who work in health promotion who participate in the reorientation of health systems to primary healthcare.

First, the health system is controlled by the powerful. There is a hegemony that supports a predominant treatment/cure paradigm. Power resides in these structures and with the health professionals who work in those systems. So, as advocates for equity and social justice, health promoters have a responsibility to take up this challenge. If we continue to work only in the community where we are comfortable, we will avoid confronting one of the greatest challenges of our times: creating a health system that is based on the principles of health promotion. We are not neutral in this process. As Paulo Freire (1972) said, ‘washing one’s hands of the conflict between the powerful and the powerless, means to side with the powerful, not to be neutral’. Kickbusch (1989) confirms this concern. She states, ‘herein lies the great historical opportunity and challenge. Maybe health promotion can break the deadlock of the health policy debate that is basically about medical care and provider dominance’ (p. 14). As she suggests, we are well beyond the burden of proof needed to claim that health promotion is successful—we have demonstrated this through our change in attitudes towards smoking, drinking, and nutrition—even if these are mainly concerns of the middle class and of high-income countries. ‘Accountability and the burden of proof should now lie with the medical system’ (Kickbusch 1989, p. 14).

The second reason that health promoters must take up this challenge can be summarized in one question: if health promoters do not advocate for primary healthcare based on the principles of the *Alma-Ata Declaration* and the *Ottawa Charter*, what model of primary healthcare will dominate our countries?

Health promotion: history and influences

Health education

Health education plays a profound role in the history of health promotion. While it is true that health promotion is an 'essentially contested concept' claimed by a variety of different interests and actors (Green and Raeburn 1988), many of its most prolific commentators, particularly in the area of health promotion research and knowledge development, have been from the field of health education (Green and Kreuter 2005).

These writers have often been concerned with the failure of traditional health education approaches to help motivate individuals to act on health information. Following Tones, we adopt his definition of health education as: 'Any intentional activity which is designed to achieve health or illness-related learning, that is, some permanent change in an individual's capability or disposition' (Tones and Green 2004, p. 7). This refers to what knowledge, attitudes, or skills can be acquired by individuals through a variety of health education processes. In relation to health promotion, key health educators have radically restructured the traditional approaches to influencing health behaviour. Most of this work has revolved around challenging what are now seen to be simplistic and mechanical models of health belief and health decision-making. At the centre of this change has been an adoption of the concept of empowerment and an advocacy for using participatory learning processes that break down the power imbalances between health professionals and lay members of society. Crucially, the relationship between devolving control, developing self-esteem, and bridging the gap between knowledge, attitude, and behaviour is highlighted.

In relation to the overall theme of health equity, the move away from traditional health education models has been critical. Without the concept of empowerment, and the development of capacities and self-esteem, the

traditional ‘health action model’ of raising awareness and changing attitudes to health behaviours tended to exacerbate health inequalities, as the ‘prepared’ middle classes quickly adopted the new healthy practices of more exercise, less smoking, and a healthy diet. The efforts to help population groups that had both the worst health outcomes and the most intransigent health-related social conditions have not been nearly as successful.

A detailed account of these changes can be found in Tones and Green (2004). One very important aspect of his account is how health promoters can learn, as professionals, to overcome the social gap in both power and understanding between them and the groups and individuals they aim to enable and empower. As Tones notes, the ‘holy trinity’ of counselling (respect, empathy, and genuineness) is pertinent here. Often the more socioecological accounts of the health promotion process gloss over this crucial interactive aspect of health promotion. Whether it is with individuals or with groups, health promoters cannot act effectively without using highly developed skills of empathic understanding and facilitation. Particularly with group interactions, where often highly charged community issues are discussed, the professionals striving for neutrality and objectivity will find themselves unable to cope with the anger and resentment felt by people who perceive a history of grave social injustice behind their ‘health’ problems. It is important that these basic counselling skills are imparted to health promoters-in-training before they go out into communities and work with them on issues relevant to their health.

This shift in health education from an information-giving, pamphlet distribution approach to an empowerment liberatory approach has brought renewed interest in Freire’s emancipatory education paradigm (1972). Health promoters in several countries, most notably those in Brazil, have reclaimed and embraced his basic premises and have employed his dialogical problem-posing teaching strategies that help make health education more consistent with the principles and values of health promotion and the ‘new’ public health. Freire’s (1972) model of empowerment education describes a three-stage methodology consisting of listening, participatory dialogue, and action. Freire proposes that the main strategy of empowerment education, critical dialogue, requires us to engage in a process of problem-posing rather problem-solving. Problem-posing is different from problem-solving because it does not seek immediate

solutions to problems. Rather, generative themes arising from the listening phase are ‘codified’ and posed as problematics to raise group consciousness about specific issues. Wallerstein and Hammes (1991) contend that this process recognizes the complexity and the time needed to create effective solutions to societal issues. ‘An effective code shows a problematic situation that is many sided, familiar to participants and open-ended without solutions’ (Wallerstein and Bernstein 1988, p. 383). Freire describes these as ‘generative’ themes because they generate energy and motivate people to act. Freire contends that, through a process of dialogue that reflects on the generative themes raised through listening, people become masters of their own thinking in interaction with others (1972, p. 95). As Wallerstein and Bernstein explain: ‘The goal of group dialogue is critical thinking by posing problems in such a way as to have participants uncover root causes of their place in society—the socioeconomic, political, cultural, and historical contexts of personal lives’ (1988, p. 382).

Freire cautions that ‘the liberating educator has to be very aware that transformation is not just a question of methods and techniques’ (1972, p. 35). If that were the case, we could simply substitute one set of methods for another. ‘The question is in a different relationship to knowledge and to society’ (1972, p. 35).

Public health

Many health promotion researchers preface their scholarly remarks on the birth and development of health promotion with a discussion of its relationship to its older, more developed discipline, public health (Kickbusch 1986; Terris 1992). How this history is understood is perhaps the most telling aspect of how health promotion and its progress are viewed as a contemporary phenomenon. The argument developed next is that there has been a tendency within health promotion to tell a story of public health as a ‘fall from grace’—a fall from its original reforming, perhaps even zealous, focus on the social and environmental causes of ill health, to a more restrictive, preventive biomedical era, and finally, to a broader scale but narrower scope in the ‘lifestyles’ approach focused on individual risk factors and behavioural change (Kickbusch 1986). Health promotion steps into the story to herald the era of a ‘new public health’, as a sort of re-emergence of the spirit of the nineteenth-century socioenvironmental

model, with a modern gloss on the more subtle socioeconomic determinants of health. The purpose of this critical analysis is to challenge this tendency to nostalgia, and to explicate some of its continuing consequences for health promotion's rather schizophrenic relationship to public health. The 'golden age' of public health was influenced by a particular philosophical and political outlook that still finds its expression today in its most modern and rigorous proponents.

The history of health promotion conventionally begins with the publication of the Lalonde Report, entitled *A New Perspective on the Health of Canadians* (Lalonde 1974). The report was the first high-level national government document in the world to advocate for health promotion as a basic strategy for improving population health. It was influential internationally and set the stage for future debate with its concept of the *health field* as the articulation of the argument that the medically dominated healthcare system was only one and perhaps the least significant determinant of health, alongside biology, the physical and social environment, and individual lifestyles. The Lalonde Report relied explicitly for its argumentation on such critiques of the healthcare system found in 'social medicine' as those comprehensively outlined by (McKeown 1976), but which had their roots in the classic public health tradition of William Petty, Johann Frank, Rudolf Virchow, and William Farr (White 1991). The report contains not only the notable tension between an emphasis on individual lifestyles and the subsequently neglected socioenvironmental factors, but also an equal tension, given its own chapter heading, of 'Science versus Health Promotion'. Here it is made very clear that the 'science base' of the health field concept is epidemiology and, in this context, health promotion is seen as that type of action that must be taken even though the pertinent scientific questions have yet to be definitively answered. In some ways, this attitude allowed some initial breathing space for health promotion to prosper; however, by setting up this dichotomy, it ensured that, eventually, when 'science' made its accounting, health promotion would have its day of reckoning with epidemiology.

Meanwhile, many in the health promotion community, especially in Europe and Canada, were starting to develop an independent conceptual basis for their work, based on a rigorous reflection on the type of actions necessary to most effectively promote the health of individuals and communities. Much of this work evolved out of a complex internal critique

of the failure of traditional health education approaches and a more sophisticated understanding of behavioural change (Kickbusch 1986; Tones 1993). Yet, in some countries, such as Canada, the absence of a strong health education tradition contributed strongly to a more socioecological approach to promoting health (with some of its most influential leaders being sociologists and nurses, rather than psychologists and health educators). Furthermore, for a variety of complex reasons (including, again, individual leadership), much of the discourse of contemporary social movements (new leftist, feminist, gay/lesbian, environmentalist) found its way onto the official agenda of major institutions such as the WHO and Health and Welfare Canada (Labonté 1994). As has been recognized by one of the leaders in health promotion internationally, Canada provided a hybrid and fertile mixture of traditional welfare state values and innovative community activism that seemed to provide the perfect ground for a push for the new socioenvironmental approach to health promotion (Kickbusch 1994). Out of this productive interaction between European 'health promotion tourists' (Kickbusch 1986, 1994) and many able Canadian activist/public health practitioners, grew the idea and finally the accomplishment of the *Ottawa Charter for Health Promotion* (1986).

To fully understand the impact that the *Ottawa Charter* had and continues to have, it is important to see that there was a crucial transformation from the epidemiological and bureaucratic dominance of the Lalonde Report to the emerging 'more pluralistic (and messier) social-science paradigm of human and social relations' embedded in the *Ottawa Charter* (Labonté 1994, p. 86). This shift is key to understanding the constant tension between a 'scientific' approach to health promotion and the 'values' underlying the *Ottawa Charter* that is renewed whenever health promoters are asked to more rigorously account for their activities. It raises the uncomfortable question for those who, correctly, see the importance of reconnecting health promotion to the new public health, of just how 'new' the new public health is willing to be, when it comes to its underlying philosophical commitments.

This same tension underlies some of the confusion within the health promotion research community about how to relate to the more recent (Evans and Stoddart 1990) emphasis on 'population health' (Labonté 1997; Poland et al. 1998; Raphael and Bryant 2002). On the one hand, there is a justified admiration for the advocates of population health for their

influential arguments about socioeconomic determinants of health, even so far as to single out progressive population health researchers such as John Frank (Raphael and Bryant 2002). On the other hand, there is the well-articulated angst about the lack of health promotion principles within the population health perspective. The critiques of the population health perspective for its lack of emphasis on values, its weak or non-existent orientation to action, and its somewhat imperious attitude to what is to count as proper 'knowledge', are all cogent and well-aimed. The question is: why should this be a surprise? It is not enough to point out the baleful influence of a replacement ideology for health promotion. Where did it come from, and why is it so influential? Furthermore, is this newly emergent approach (Evans and Stoddart 1990) really so new? How far is it simply a modern, sophisticated renaissance of that very same 'golden era' of public health that health promoters so often return to as their intellectual and moral heritage?

Although health promoters themselves (especially Canadians, who were the ones facing this challenge directly) reacted strongly and were able to defend the rationale for keeping a health promotion focus, the more incisive critiques were often too 'reactionary' and came off rhetorically as overly defensive. A more accommodating response came from within Health Canada itself with Hamilton and Bhatti (1996) introducing the concept of *population health promotion*. This was clearly an attempt to marry these two potentially adversarial positions and to cement the term health promotion as an integral component of population health that could not be ignored.

Why is it that health promotion often seems 'behind the game' in the science debate? Partly, this is due to the fact that, as Labonté says and a *Companion to Social Theory* attests (Turner 2000), the social science world is 'messy'. However, it may also be partially true that, for too long, health promotion has neglected its need to develop an independent 'science base' having unconsciously bought into that original Lalonde dichotomy. This is becoming increasingly clearer as many of the leading proponents in the field are pushing for more intensive theoretical development, a pressure that has become especially acute as the need for demonstrating effectiveness has increased (McQueen 2001; McQueen and Jones 2007; McQueen et al. 2007). To understand why health promotion has such a complex and ambiguous relationship to public health, it is necessary to dig more deeply

into the foundations of modern public health and to unpack its driving philosophy and world view.

Politics and philosophy in public health and epidemiology

It is crucial to understand that the roots of the modern public health epidemiologists' focus on individual risk factors and randomized controlled trials (RCTs) is not in contradiction with or a deviation from the Edwin Chadwicks, the John Simons, and the John Snows of the classical public health. Rather, the full flowering of a utilitarian calculus, an uncompromising economism, and an obdurate scepticism of anything but positivistic scientific knowledge, can be seen as the late fruit of more than three centuries of development in public health and epidemiology.

There is a dilemma and prima facie paradox that health promotion faces when confronting its genealogy in the history of public health. In terms of lives saved and healthy years lived, the early public health interventions to combat the spread of deadly and debilitating communicable diseases cannot be underestimated. However, it is no accident that once the environmental risk factors of the major communicable diseases were effectively neutralized, a shift in focus took place to providing preventive, immunization measures. As Kerr White (1991) so convincingly puts it, the history of public health and epidemiology can be read as successive and iterative 'redefinings of the unacceptable'. Public health has always been concerned with an economistic and utilitarian approach to the health of the population; when things start to 'cost too much', the unacceptable becomes miraculously 'visible'. To understand this history, one has to ignore the facile disciplinary chasm between public health and economics, which has only recently and tentatively been bridged (Evans et al. 1994). While economics became progressively theoretical and mathematical, public health continued the original classical liberal tradition of reformist, practical utilitarianism, most powerfully apparent in the Benthamite tradition's attempt to rationalize government and public services. There is a great irony that the humanitarian idealism (an idealism at the core of health promotion's values base) of the British 'public health doctors', such as Haygarth, Heysham, Thackrah, Baker, and Millar, was never the driving force behind concentrated public health action (Fraser 1973).

As we will discuss in the subsection on health promotion and social justice, public health shares with economics a default, often merely implicit, utilitarian ethics. This shared history is seldom acknowledged, but it is a history that health promotion must confront explicitly. Fortunately, and ironically, recent developments in public health have brought into question the utilitarian approach, finding it inadequate, particularly in relation to the question of health inequity (Anand et al. 2004). The argument fleshed out in the following is that health promotion must forcefully engage in helping public health move in the direction advocated for by Amartya Sen and others (Anand et al. 2004).

Social movements

In this section, we briefly review one of the constitutive ambiguities at the heart of health promotion—many public health practitioners in the 1970s and early 1980s were increasingly cognizant of the lack of participatory involvement of the 'public' in public health programming.

A strong feature of the so-called 'new social movements' in the post-1968 period was a trenchant critique of bureaucratic structures and an increasingly administered society alongside the traditional leftist critique of capitalism. Some public health institutions, particularly urban public health units, decided to transform local public health practice by integrating a participatory model of programming that was heavily influenced by this antibureaucratic critique (Labonté 1994). However, as Dupéré et al. (2007) argue, despite its ambitions, health promotion is still not accurately described as a 'social movement', but rather a 'professional movement that had successfully advanced a discourse about health and the production of health'. Yet, despite this acknowledged status, health promoters have recognized that much of their effectiveness depends on very high levels of social engagement. The more recent emphasis on health in the context of globalization (Labonté et al. 2011) makes the necessity for health promotion to engage with larger social movements, particularly on the global development agenda, even more apparent.

Nevertheless, we find health promotion once again suspended between its constitutive desire to become one with the 'community' and its real position as a mediating professional fraction, often acting on behalf of formal public institutions. In the future, health promotion will have to

sacrifice some of its cherished professional neutrality to choose sides, especially in its responsibility to advocate for health. While this new form of activist engagement must be balanced with the legitimacy attained from professional status, the balance must shift quite radically, given the growing threats to health represented by the inequity of contemporary societies in a globalized world. An example of a social movement which involves many health professionals is the People's Health Movement which operates as a global network with country circles around the world. It also 'watches' international organizations such as the WHO and World Bank and comments on their policies and proposals (Baum et al. 2020).

Health promotion, health inequities, and social justice

We have, throughout this chapter, alluded to the commitment health promotion has to the principle of health equity. This is a fundamental and central value for health promoters and is often the touchstone for deciding why, where, and how to enact health promoting practice and policy. Yet, despite this nearly constant refrain, there is still confusion within health promotion about the theoretical and conceptual basis for its concern with health equity. While most, if not all health promoters, would see health equity as a basic goal of health promotion, seldom is the specific normative dimension that underlies this commitment fleshed out. The basic understanding is that health inequities are undesirable and should be eliminated because they are a set of systematic inequalities in health outcomes that are based on *unjust* inequalities of access to resources that provide for health. The *Health Promotion Glossary* describes what equity in health entails: 'That all people have an equal opportunity to develop and maintain their health, through fair and just access to resources for health.'

However, this definition begs many key questions, such as: what is 'fair' and 'just' access? And, what are 'resources for health'? While health promoters have often reflected deeply on health inequity, much of the appeal has been to an intuitive basis for supporting the elimination of inequity. We argue that this stance is no longer good enough. Health promotion must fully engage with recent work in political philosophy, particularly in the arguments surrounding the concept of social justice that have been developing since the publication of John Rawls' *A Theory of*

Justice (1971). Since Rawls' ground-breaking work, an ongoing debate has taken place concerning what is the proper approach to justice for whole societies (Kymlicka 1990; Aveneri and de-Shalit 1992). More recently this debate has been expanded to consider how we are to think of justice in the global context (Nussbaum 2006). Health promoters should pay close attention to what is at stake in these debates for two reasons.

First, without an awareness of these important arguments, health promotion is liable to accept a default utilitarianism that it inherits from public health, which in turn the latter shares with orthodox economics. It is argued here that this unacknowledged utilitarianism is in direct contradiction to two profound moral intuitions that form the core of health promotion values: that it is wrong to increase overall health at the expense of the least well off; and, that human dignity and personal autonomy are overriding values.

Second, important developments in these debates are directly relevant to concerns with acting on the social determinants of health. Recent arguments have been refined concerning why and how we should address the issue of health inequity, both within and between societies (Nussbaum 2000; Anand et al. 2004). As health promoters, charged with the responsibility to advocate, enable, and mediate for equity in health, we should be armed with the very best arguments supporting our position.

We argue here that the one of the most promising theoretical developments in political philosophy that have implications for health promotion are in the evolving 'capabilities' approach to social justice and equity (Nussbaum and Sen 1993; Nussbaum 2000, 2006). This approach most nearly matches the health promotion approach to health as a 'resource for everyday living'; according to this doctrine, the 'social bases for health' would count as a primary good, or capability that should, by right, be provided to all citizens (and by extension, all human beings) at a minimum standard (Nussbaum 2006). It cannot be assumed that arguments for equity in health are unassailable and intuitively obvious for two reasons. First, without some substance behind what is meant by equity and what kinds of resources are to be distributed equitably, the demand for equity in health can be dismissed as either empty or naïvely utopian. Second, differing conceptions of what is just will lead to different outcomes in terms of actions to promote health. For example, unless we are very clear, 'equal opportunity' can be understood in an absolute minimalist sense and can

allow powerful institutions to continue to support vast inequities in resources for everyday living. As we will see next, how we conceive of social justice has a profound impact on the types of actions we can imagine as solutions to the gross inequities in health we find across the world. Specifically, it has become apparent that the goals of health promotion are intimately related to the goals of a socially just global development agenda. Next we focus on an important element that tends to be lost in much of the discourse on health equity: what are the political, economic, and social mechanisms by which the social determinants of health are reproduced as unequal resources for health? This is followed by considering one emerging global initiative aimed at policy solutions in this area: the *Health in All Policies* agenda.

The political economy of health promotion

Much good work has been done on the 'political economy of health' (the analysis of how different politico-economic social structures affect health outcomes), yet an enormous amount is still required (Navarro and Shi 2001; Navarro 2002; Langille 2003; Raphael 2003). Furthermore, there have been some excellent analyses of the 'political economy of healthcare' (the analysis of the effect of different political and economic arrangements on the quality and differential access to health services). In this subsection, we outline a different question: what is it about our contemporary political and economic structures that vitiates against the implementation of health promotion strategies and actions as they are conceived in the *Ottawa Charter*?

To begin to answer this question, we need to consider the three fundamental dimensions to health promotion: empowering communities and individuals, building health public policy, and creating supportive environments.

As has already been argued, empowerment is a key dimension of health promotion. By its very nature, empowerment aims to rebalance existing power arrangements by enabling those currently without power to gain access to the resources necessary to live fulfilled and happy lives. In order to do this with any success, health promoters must do two things: they must have a clear-headed view of existing power structures and relations; and,

they must recognize, as professionals, how they themselves fit into those power relations and how they help, often unconsciously, to reproduce them. One way of seriously addressing this issue is to pay more attention to the concepts of class and status mentioned on the section on social justice and health promotion.

Health promotion, to be successful, must rely on concerted action by governments around the world, both within their own territories and in cooperation to address needs that require global action, such as on climate change. However, while these wishes are often articulated (WHO 2005), seldom are we offered an analysis of the structure and dynamics of the contemporary state system in a global context. A more reflective perspective is important here, as theorists of the state argue that certain issues and certain groups, using specific strategies, are more or less likely to be successful changing the nature of hegemonic projects and reversing the direction of state policies (Jessop 2002). Health promotion must become more strategic in how it operates vis-à-vis the state; it must recognize in an explicit way the limitations and opportunities available and integrate theoretical perspectives and practical actions in regard to one of its key areas: building health public policy.

Finally, creating a supportive environment is even more wrapped up in the dynamics of global capitalism than all the other areas combined. The fundamental prerequisites for health as outlined in the *Ottawa Charter* are: peace, shelter, education, food, income, a stable ecosystem, sustainable resources, social justice, and equity. These are the elements that, when in adequate supply, make up many of the properties of a supportive environment for health. Yet, each of these elements is in large part determined by the particular structure and dynamics of our global socioeconomic system.

It is necessary to develop an awareness of the fundamental political and economic drivers behind the dynamics of contemporary societies in both the developed and developing worlds. If equity in health relies on the fundamental fairness of social, political, and economic institutions, then ignoring these basic realities is no longer an option for a serious approach to health promotion. These important insights should no longer be gained in an ad hoc way but should be seen as part of what should constitute core knowledge for competent health promoters.

Social theory and health promotion

Recently, particularly following the report of the WHO Commission on Social Determinants of Health (2008), there has been a raft of activity by both academia and governments aimed at developing knowledge about the social determinants of health. There has also been a marked (and needed) shift of emphasis to what types of *action* societies need to take to address the consequences of the iniquitous distribution of resources that support health. Less prominent are attempts to confront theoretically the main causal mechanisms that produce and reproduce the social inequalities that lead to health inequities. While there have been recent attempts by some health promotion researchers to re-engage with fundamental debates in social theory (McQueen et al. 2007), for the most part the public health community as a whole, and health promotion by proxy, is still mainly dominated by the narrowly focused methodological lens of epidemiological science although there is increasing use of theory in health promotion as can we seen in recent editions of Health Promotion International (see for example Harris et al. 2017). This reluctance to confront the ghost of social theory's past is most apparent in the way that the categories of *class* and *status* are dealt with in health promotion and public health discourse. Where they are mentioned at all, they tend to be conceptualized as epidemiological variables that measure an individual or group's socioeconomic attributes or properties (e.g. income, education level, wealth, job status), not as social processes that reproduce structural disadvantages for most and accumulate power and privilege for a few (see Scambler (2012) for an excellent review of this problem). The key elements of social theory that should be at the forefront of debate in health promotion circles are considered more extensively in two recent sources (McQueen et al. 2007; Carroll 2012).

The *Health in All Policies* movement

Launched as Finland's main theme of its 2006 European Presidency, *Health in All Policies* (HiAP) is a strategy to help link health and other policy sectors in an overarching intersectoral approach to improving health and well-being and reducing health inequity (Leppo et al. 2013). It is aimed at reinvigorating the original emphasis in the *Ottawa Charter* on healthy public policy, and to follow up on the work done at the WHO conference in

Adelaide, Australia in 1988 (WHO 1988). Since that meeting, there has been much reflection on both the details of how to implement healthy public policy and the particular challenges posed by integrated, coherent, intersectoral action for health. More recently, the WHO reconvened in Adelaide to produce an international statement on HiAP called the 'Adelaide statement on health in all policies: moving toward a shared governance for health and well-being' (WHO 2010). It is no coincidence that it lays heavy emphasis on 'governance'. This was followed by two important publications in 2012, also focusing on governance (Kickbusch and Gleicher 2012; McQueen et al. 2012). The central challenge of implementing HiAP has been the difficulty of managing the governance and accountability structures necessary to sustain both vertical (levels of government, non-governmental, and private sector) and horizontal (cross-ministry, interdepartmental) intersectoral collaboration. We are only now beginning in public health and health promotion circles to appreciate the different type of knowledge base required to assess the effectiveness of policy implementation. We are still largely stuck in an outmoded attempt to squeeze what are really matters for political science, economic sociology, and the sociology of the state into the narrow methodological confines of standard epidemiological research designs. Increasingly (Lawless et al. 2012; Carroll et al. 2013; Baum et al. 2017) HiAP Research applies political theory case studies of HiAP in order to better understand the key mechanisms underlying success and failure. Clearly, HiAP is required to move forward on the intersectoral action agenda outlined in the *Ottawa Charter*; however, improved knowledge on implementation and sustainability are required the future.

Complexity, context, and causality in health promotion research

Over the past 15 years, a series of publications have charted the specific methodological challenges for evaluating the effectiveness of health promotion interventions (IUHPE 1999; Rootman et al. 2001; Zaza et al. 2005; McQueen and Jones 2007). There has been some scepticism about applying the methodological protocols of evidence-based medicine (EBM) and RCTs as the gold standard because of the problems of complexity and context (McQueen 2007). One emerging alternative has been to use

different methods for synthesizing evidence, such as the realist (Pawson et al. 2005) or meta-narrative (Greenhalgh et al. 2005) review approaches. A more detailed treatment of these latter approaches is beyond the scope of this chapter; however, one further potential for advancing beyond traditional EBM-type methods, is in the use of so-called 'systems thinking' or 'complexity science' to understand the rich complexity and contextual subtlety of the settings within which health promotion interventions take place, and of the interventions themselves.

Key to understanding the critique of EBM and some of the proposed alternative strategies is the different nature of how causality is conceptualized. The realist alternative has a direct, philosophical critique of the underlying empiricist-positivism of EBM's approach to causality (Bhaskar 2008) and has been applied to health promotion Baum et al. 2019). Conversely, systems thinking and complexity science approaches are more concerned with EBM's inability to take account of the interactive, emergent, and non-linear dynamics of causation that are crucial to understanding health promotion interventions as complex adaptive systems that intervene in the context of settings that are themselves complex adaptive systems.

Some health promotion researchers have started to take seriously the potential for a complexity or systems approach to health promotion interventions (Rickles et al. 2007; Shiell et al. 2008; Hawe et al. 2009; Trickett et al. 2011). These emergent attempts to apply complexity science have yet to show fruit (though systems thinking in public health has a longer pedigree), yet they hold much potential to transcend the current impasses in health promotion effectiveness research.

Conclusion

We have attempted to offer the reader a chance to reflect on a set of core conceptual issues that underlie the health promotion problematic. The five key messages we want to impart about health promotion are listed here:

Health promotion is a complex, often ambiguous concept and set of practices. Health promotion finds its core values and principles in the Ottawa Charter which bears careful examination to comprehend the essence of health promotion.

Health promotion has an intimate connection to health education, with many of its most important and prolific thinkers having a health education background. The revolution in health education practice is directly connected to the birth of health promotion but beyond this, health promotion has its roots in the deep history of public health and has been invigorated by contemporary social movements.

Health promotion is fundamentally about ethics, values, and social justice. Only secondarily is it about technical strategies for behaviour change. The foundational principles of health promotion are *equity*, *participation*, and *empowerment*.

Health promotion is a professionally dominated movement. This requires health promotion professionals to be critical and reflexive in their practice; they must acknowledge power imbalances that favour professional dominance and work to restore power to individuals and communities.

Health promotion must take its duty to enable people to control the determinants of their health seriously. To do this it must engage more directly with contemporary arguments in political philosophy and it must be aware of the dynamics of the global political economy and its effect on the potential for health promotion.

Some of these issues are well known, such as the problem of professional dominance; while others, such as the political economy of health promotion, or the engagement with political philosophy, are not addressed or require much deeper reflection.

We have argued that, at its heart, health promotion is about a radical shift in values for public health. It is not that public health was never concerned with equality or alleviating the misery of the poor; arguably, the so-called ‘golden age’ was driven by exactly these moral questions. However, these intuitive commitments were not sufficiently followed through when it came to not just *what* outcomes to change but *how* to change them. Too much of public health, for too long, was driven by a benevolent paternalism that, particularly when it came to dealing with chronic diseases and with vulnerable populations, ended up being counterproductive. Indeed, not only

was this paternalism ineffective in many areas, it was unethical. It assumed the authority of experts and professionals, not only to determine technical solutions, but to determine needs. If we are to take the concepts of equity and empowerment seriously, they have profound implications for how we do public health interventions. We have learned that by addressing needs without first establishing a participatory framework that enables individuals and communities to determine those needs for themselves, we fatally undermine one of the most crucial capacities for health: human dignity and self-respect. This is particularly so in communities that have suffered historical social injustices. As Richard Sennett says, people subjected to this disempowering process, experience 'that peculiar lack of respect which consists of not being seen, not being accounted as full human beings' (Sennett 2003, pp. 12–13).

We hope to have demonstrated that there are many barriers to realizing this change in power relations; yet, there are also very important opportunities, such as with the Millennium Development Goals and the Commission on the Social Determinants of Health, where there is an increasing clamour for action to redress health inequities through empowering processes. It is notable that even the World Bank, often the subject of brutal criticism for exacerbating inequalities (Stiglitz 2003), has made significant moves towards recognizing the importance of reducing inequity in human development and has integrated an empowerment approach (World Bank 2006). It remains to be seen whether these gains can be translated into major policy changes and effective implementation; nevertheless, it is at this level where health promoters and all public health practitioners and researchers must have a strong advocacy position.

We hope that it is apparent that, in our interpretation, health promotion is much more than a set of technical public health interventions aimed at revamping traditional health education for the twenty-first century. We cannot let go of the core competencies built up by health education and other contributing fields, but we cannot be limited in our vision either. Health promotion has to face up to the fact that, while it may only be a junior partner in the global struggle to develop a more just and equitable world, when it comes to a key human capability and resource, *health*, it must take a lead role in making the argument for equity, develop and present the evidence for what *action* is necessary to achieve equity in health, and finally, to hold the powerful accountable where they fail to live

up to the demands of justice for health. Embedded in health promotion is an imperative to act ethically and justly. In this case, unlike most, there is no choice.

ACKNOWLEDGEMENTS

Text extracts from Hartrick, G. et al. Family nursing assessment: meeting the challenge of health promotion, *Journal of Advanced Nursing*, Volume 20, Issue 1, pp. 85–91, Copyright © 1994, reproduced with permission from John Wiley & Sons, Inc.

Text extracts from World Health Organization, *Health Promotion Glossary* (WHO/HPR/HEP/98.1), WHO, Geneva, Switzerland, Copyright © 1998, reproduced with permission from the World Health Organization, http://www.who.int/healthpromotion/about/HPR%20Glossary%201998.pdf?ua=1.

REFERENCES

Anand, S., Peter, F., and Sen, A. (eds.) (2004). *Public Health, Ethics, and Equity*. Oxford: Oxford University Press.

Aveneri, S. and de-Shalit, A. (eds.) (1992). *Communitarianism and Individualism*. Oxford: Oxford University Press.

Baum, F. (2016). *The New Public Health*. Melbourne: Oxford University Press

Baum, F. (2007). Cracking the nut of health equity: top down and bottom up pressure for action on the social determinants of health. *Promotion & Education*, 14, 90–5.

Baum, F., Delany, T., MacDougall, C., Lawless, A., van Eyk, H., and Williams, C. (2017). Ideas, actors and institutions: lessons from South Australian health in all policies on what encourages other sectors' involvement. *BMC Public Health*, 17(811), 1–16.

Baum, F., et al. (2019). To what extent can the activities of the South Australian Health in All Policies initiative be linked to population health outcomes using a program theory-based evaluation? *BMC Public Health*, 19(88), 1–16.

Baum, F., Sanders, D., and Narayam, R. (2020). The global people's health movement. What is the people's health movement? *Saude Debate*, 44(1), 11–23.

Bhaskar, R. (2008). *A Realist Theory of Science*. New York: Routledge.

Carroll, S. (2012). Social theory and health promotion. In: Rootman, I., Dupéré, S., Pederson, A., and O'Neill, M. (eds.) *Health Promotion in Canada* (3rd ed.), pp. 33–52. Toronto: Canadian Scholar's Press.

Carroll, S., Hills, M., Miller, G., Geneau, R., Mitic, W., and Foster, L. (2013). *Intersectoral Action for Health: The Case of BC.* Presentation at the 21st IUHPE World Conference on Health Promotion, Pattaya, Thailand, 28 August.

Commission on the Social Determinants of Health (2008). *Closing the Gap in a Generation: Health Equity Through Action on the Social Determinants of Health*. Geneva: WHO. [Online] Available at: http://www.who.int/social_determinants/final_report/en

Craig, G. (2005). *Community Capacity-Building: Definitions, Scope, Measurements and Critiques*. OECD Paper, Prague, Czech Republic, 8 December 2005.

Dupéré, S., et al. (2007). Conclusion: the rhizome and the tree. In: O'Neill, M., Pederson, A., Dupéré, S., and Rootman, I. (eds.) *Health Promotion in Canada*, pp. 371–88. Toronto: Canadian Scholar's Press.

Evans, R., Barer, M., and Marmor, T. (eds.) (1994). *Why are Some People Healthy and Others not?* New York: Aldine de Gruyter.

Evans, R. and Stoddart, G. (1990). Producing health, consuming health care. *Social Science & Medicine*, 31, 1347–63.

Fraser, D. (1973). *The Evolution of the British Welfare State*. London: Macmillan Press Ltd.

Freire, P. (1972). *Pedagogy of the Oppressed*. Harmondsworth: Penguin.

Godin, G. (2007). Has the individual vanished from Canadian health promotion? In: O'Neill, M., Pederson, A., Dupéré, S., and Rootman, I. (eds.) *Health Promotion in Canada*, pp. 367–70. Toronto: Canadian Scholar's Press.

Green, L. and Kreuter, M. (2005). *Health Promotion Planning: An Educational and Ecological Approach*. Toronto: McGraw-Hill Higher Education.

Green, L. and Raeburn, J. (1988). Health promotion. What is it? What will it become? *Health Promotion International*, 3, 151–9.

Greenhalgh, T., Robert, G., Macfarlane, F., Bate, P., Kyriakidou, O., and Peacock, P. (2005). Storylines of research in diffusion of innovation: a meta-narrative approach to systematic review. *Social Science & Medicine*, 61, 417–30.

Hamilton, N. and Bhatti, T. (1996). *Population Health Promotion: An Integrated Model of Population Health and Health Promotion.* Ottawa: Health Promotion Development Division, Public Health Agency of Canada. [Online] Available at: https://www.canada.ca/en/public-health/services/health-promotion/population-health/population-health-promotion-integrated-model-population-health-health-promotion.html

Harris, P., et al. (2017). Creating 'healthy built environment' legislation in Australia; a policy analysis. *Health Promotion International*, 33(6), 1090–100.

Hartrick, G., Lindsey, A., and Hills, M. (1994). Family nursing assessment: meeting the challenge of health promotion. *Journal of Advanced Nursing*, 20, 85–91.

Hawe, P., Shiell, A., and Riley, T. (2009). Theorising interventions as events in systems. *American Journal of Community Psychology*, 43(3–4), 267–76.

Held, D., et al. (1999). *Global Transformations*. Cambridge: Polity Press.

Hills, M. (1998). Student experiences of nursing health promotion practice in hospital settings. *Nursing Inquiry*, 5, 164–73.

Hills, M. and Carroll, S. (2009). Health promotion, health education, and the public's health. In: Detels, R., McEwen, J., Beaglehole, R., and Tanaka, H. (eds.) *The Oxford Textbook of Public Health* (4th ed.). Oxford: Oxford University Press, pp. 752–66.

IUHPE (1999). *The Evidence of Health Promotion Effectiveness: Shaping Public Health in a New Europe*. Brussels: European Commission.

IUHPE (2019). *WAIORA: Promoting Planetary Health and Sustainable Development for All*. [Online] Available at: http://www.iuhpe2019.com/PicsHotel/iuhpe/Brochure/Rotorua%20Statement%20For%20Closing%20Plenary%20Revised%20(1).pdf

Jessop, R. (2002). *The Future of the Capitalist State*. Cambridge: Polity Press.

Kemm, J. (2006). Health impact assessment and Health in All Policies. In: Ståhl, T., Wismar, M., Ollila, E., Lahtinen, E., and Leppo, K. (eds.) *Health in all Policies: Prospects and Potentials*.

Helsinki: Ministry of Social Affairs and Health, pp. 189–208.
Kickbusch, I. (1986). Health promotion: a global perspective. *Canadian Journal of Public Health*, 77, 321–6.
Kickbusch, I. (1989). Back to the future: moving public health into the 90's. In: *Proceedings from National Symposium on Health Promotion*, pp. 1–5. Victoria: BC Ministry of Health.
Kickbusch, I. (1994). Introduction: tell me a story. In: Pederson, A., O'Neill, M., and Rootman, I. (eds.) *Health Promotion in Canada: Provincial, National & International Perspectives*, pp. 8–17. Toronto: W.B. Saunders Canada.
Kickbusch, I. (2007). Health promotion: not a tree but a rhizome. In: O'Neill, M., Pederson, A., Dupéré, S., and Rootman, I. (eds.) *Health Promotion in Canada*, pp. 363–6. Toronto: Canadian Scholar's Press.
Kickbusch, I. and Gleicher, D. (2012). *Governance for Health in the 21st Century*. Copenhagen: WHO.
Kok, G., Van Den Borne, B., and Dolan Mullen, P. (1997). Effectiveness of health education and health promotion: meta-analyses of effect studies and determinants of effectiveness. *Patient Education and Counseling*, 30, 19–27.
Kymlicka, W. (1990). *Contemporary Political Philosophy: An Introduction*. Oxford: Oxford University Press.
Labonté, R. (1993). Community development and partnerships. *Canadian Journal of Public Health*, 84, 237–40.
Labonté, R. (1994). Death of program, birth of metaphor: the development of health promotion in Canada. In: Pederson, A., O'Neill, M., and Rootman, I. (eds.) *Health Promotion in Canada: Provincial, National & International Perspectives*, pp. 72–90. Toronto: W.B. Saunders Canada.
Labonté, R. (1997). The population health/health promotion debate in Canada: the politics of explanation, economics and action. *Critical Public Health*, 7(1 and 2), 7–27.
Labonté, R., Mohindra, K., and Schrecker, T. (2011). The growing impact of globalization for health and public health practice. *Annual Review of Public Health*, 32, 263–83.
Lalonde, M. (1974). *A New Perspective on the Health of Canadians*. Ottawa: Health and Welfare Canada.
Langille, D. (2003). The political determinants of health. In: Raphael, D. (ed.) *Social Determinants of Health: Canadian Perspectives*, pp. 283–96. Toronto: Canadian Scholar's Press.
Lawless, A., Williams, C., Hurley, C., Wildgoose, D., Sawford, A., and Kickbusch, I. (2012). Health in all policies: evaluating the South Australian approach to intersectoral action for health. *Canadian Journal of Public Health*, 103(Suppl.1), S15–19.
Leppo, K., et al., (eds.) (2013). *Health in All Policies: Seizing Opportunities, Implementing Policies*. Finland: Helsinki, Ministry of Social Affairs and Health.
MacDonald, J. (1992). *Primary Health Care: Medicine in its Place*. London: Earthscan Publications Ltd.
McKeown, T. (1976). *The Role of Medicine: Dream, Mirage or Nemesis*. London: Nuffield Provincial Hospitals Trust.
McQueen, D. (2001). Strengthening the evidence base for health promotion. *Health Promotion International*, 16, 261–8.
McQueen, D. (2007). Evidence and theory. In: McQueen, D. and Jones, C. (eds.) (2007). *Global Perspectives on Health Promotion Effectiveness*. New York: Springer Publications, pp. 281–303.
McQueen, D. and Jones, C. (eds.) (2007). *Global Perspectives on Health Promotion Effectiveness*. New York: Springer Publications.
McQueen, D., et al. (2007). *Health and Modernity: The Role of Theory in Health Promotion*. New York: Springer Publications.

McQueen, D., Wismar, M., Lin, V., Jones, C., and Davies, M. (2012). *Intersectoral Governance for Health in All Policies: Structures, Actions and Experiences*. Copenhagen: WHO, European Observatory on Health Systems and Policies.

Minkler, M. (1990). Improving health through community organization. In: Glanz, K., Lewis, F., and Rimer, B. (eds.) *Health Behavior and Health Education*. San Francisco, CA: Jossey-Bass, pp. 257–85.

Navarro, V. (ed.) (2002). *The Political Economy of Social Inequalities: Consequences for Health and Quality of Life*. Amityville, NY: Baywood Press.

Navarro, V. and Shi, L. (2001). The political context of social inequalities and health. *Social Science & Medicine*, 52, 481–91.

Nussbaum, M. (2000). *Women and Human Development*. Cambridge: Cambridge University Press.

Nussbaum, M. (2006). *Frontiers of Justice: Disability, Nationality, Species Membership*. Cambridge, MA: Harvard University Press.

Nussbaum, M. and Sen, A. (eds.) (1993). *The Quality of Life*. Oxford: Clarendon Press.

Pawson, R., Greenhalgh, T., Harvey, G., and Walshe, K. (2005). Realist review—a new method of systematic review for complex policy interventions. *Journal of Health Services and Policy Research*, 10(Suppl 1), 21–34.

Poland, B., et al. (1998). Wealth, equity, and health care: a critique of a population health perspective on the determinants of health. *Social Science & Medicine*, 46, 785–98.

Povall, S.L., et al. (2014). Health equity impact assessment. *Health Promotion International*, 29(4), 621–33.

Raeburn, J. and Rootman, I. (1998). *People-Centred Health Promotion*. Chichester: John Wiley.

Raeburn, J. and Rootman, I. (2007). A new appraisal of the concept of health. In: O'Neill, M., Pederson, A., Dupéré, S., and Rootman, I. (eds.) *Health Promotion in Canada*. Toronto: Canadian Scholar's Press, pp. 19–32.

Raphael, D. (ed.) (2003). *Social Determinants of Health: Canadian Perspectives*. Toronto: Canadian Scholar's Press.

Raphael, D. and Bryant, T. (2002). Putting the population into population health. *Canadian Journal of Public Health*, 91, 9–12.

Rawls, J. (1971). *A Theory of Justice*. Cambridge, MA: Harvard University Press.

Rickles, D., Hawe, P., and Shiell, A. (2007). A simple guide to chaos and complexity. *Journal of Epidemiology and Community Health*, 61(11), 933–7.

Rootman, I., Goodstadt, M., Hyndman, B., McQueen, D., Potvin, L., and Springett, J. (2001). *Evaluation in Health Promotion: Principles and Perspectives*. Copenhagen: WHO.

Rootman, I. and Raeburn, J. (1994). The concept of health. In: Pederson, A., O'Neill, M., and Rootman, I. (eds.) *Health Promotion in Canada: Provincial, National & International Perspectives*. Toronto: W.B. Saunders Canada, pp. 72–90.

Scambler, G. (2012). Review article: health inequalities. *Sociology of Health & Illness*, 34(1), 130–46.

Sennett, R. (2003). *Respect: The Formation of Character in an Age of Inequality*. London: Penguin.

Shiell, A., Hawe, P., and Gold, L. (2008). Complex interventions or complex systems? Implications for health economic evaluation. *British Medical Journal*, 336(7656), 1281–3.

Ståhl, T., Wismar, M., and Ollila, E. (eds.) (2006). *Health in All Policies: Prospects and Potentials*. Helsinki: Ministry of Social Affairs and Health.

Stiglitz, J. (2003). *Globalization and its Discontents*. New York: W.W. Norton & Company.

Terris, M. (1992). Concepts of health promotion: dualities in public health theory. *Journal of Public Health Policy*, 13, 267–76.

Tones, K. (1993). Changing theory and practice: trends in methods, strategies and settings in health education. *Health Education*, 52, 125–39.
Tones, K. (2002). Health promotion, health education, and the public health. In: Detels, R., McEwen, J., Beaglehole, R., and Tanaka, H.(eds.) *The Oxford Textbook of Public Health* (4th ed.). Oxford: Oxford University Press, pp. 865–76.
Tones, K. and Green, L. (2004). *Health Promotion: Planning and Strategies*. London: Sage.
Tones, K. and Tilford, S. (1994). *Health Promotion: Effectiveness, Efficiency and Equity*. London: Chapman and Hall.
Trickett, E., et al. (2011). Advancing the science of community level interventions. *American Journal of Public Health*, 101(8), 1411–19.
Turner, B. (ed.) (2000). *The Blackwell Companion to Social Theory*. Malden, MA: Blackwell Publishers.
Wallerstein, N. and Bernstein, E. (1988). Empowerment education: Freire's ideas adapted to health education. *Health Education Quarterly*, 15, 379–84.
Wallerstein, N. and Hammes, M. (1991). Problem posing: a teaching strategy for improving the decision-making process. *Health Education*, 22, 250–3.
White, K.L. (1991). *Healing the Schism: Epidemiology, Medicine, and the Public's Health*. New York: Springer-Verlag.
Wise, M. and Nutbeam, D. (2007). Enabling health systems transformation: what progress has been made in re-orienting health services? *Promotion & Education*, Suppl. 2, 23–27.
World Bank (2006). *World Bank Development Report 2006: Equity and Development*. Washington, DC: World Bank.
World Health Organization (WHO) (1946). *Constitution*. Geneva: WHO.
World Health Organization (WHO) (1978). *Declaration of Alma-Ata.* International Conference on Primary Health Care, Alma-Ata, USSR, 6–12 September 1978. Geneva: WHO.
World Health Organization (WHO) (1984). *Health Promotion: A Discussion Document on the Concept and Principles*. Copenhagen: WHO, European Regional Office for Europe.
World Health Organization (WHO) (1986). *Ottawa Charter for Health Promotion*. Ottawa: WHO.
World Health Organization (WHO) (1988). *The Adelaide Recommendations on Healthy Public Policy*. Geneva: WHO.
World Health Organization (WHO) (1998). *Health Promotion Glossary*. Geneva: WHO.
World Health Organization (WHO) (2005). *The Bangkok Charter for Health Promotion in a Globalized World*. Geneva: WHO.
World Health Organization (WHO) (2010). *The Adelaide Statement on Health in All Policies*, Geneva: WHO.
World Health Organization (WHO) (2018). *The Astana Declaration on Primary Health Care*. Geneva: WHO
Zaza, S., Briss, P., and Harris, K. (2005). *Guide to Community Preventive Services: What Works to Promote Health?* New York: Oxford University Press.

6.5

Development and evaluation of complex multicomponent interventions in public health

Rona Campbell and Chris Bonell

Introduction

There have been notable advances in developing the evidence base in public health to support decision-making in policy and practice (Academy of Medical Sciences 2017; United Kingdom Clinical Research Collaboration 2018) including successful funding initiatives in the United Kingdom specifically designed to enhance the quality and quantity of public health research and to support research capacity development (United Kingdom Clinical Research Collaboration 2018; National Institute for Health Research School for Public Health Research n.d.; White 2018). Nevertheless, at a time when health systems worldwide are struggling to respond to the burden of disease and demand for healthcare, the need for innovative, cost-effective public health interventions remains (Corbett et al. 2018). Several challenges to developing and evaluating such interventions are apparent.

Firstly, policy interventions which have the greatest capacity to improve population health (for example, increasing the price of unhealthy food or beverages to tackle obesity), are more difficult to formulate and implement than interventions merely promoting individual behaviour change because of the political commitment required (Swinburn et al. 2011). Furthermore, policy interventions are often introduced with little thought given to evaluation, missing opportunities for evaluators to work with stakeholders to identify what questions an evaluation could usefully address, collection of baseline data, identification of suitable control conditions, and the possibility of randomizing to intervention and control conditions.

A second challenge is the need for evaluations of public health interventions to attend to the contexts in which interventions are

implemented and received (Craig et al. 2018) and how this might affect their intended and unintended impacts, the latter including effects on the broader functioning of the social 'systems' operating within different contexts (Rutter et al. 2017), which we shall consider in more detail to follow. A third challenge is the need for the development and evaluation of interventions to involve collaboration between researchers, the public and public health policymakers, and practitioners, to maximize the relevance, acceptability, and rapid scale-ability of interventions. In this chapter, we consider how these challenges can be addressed as we describe methods for developing and evaluating complex public health interventions. We begin by examining what complexity means in the context of public health.

Definition and scope of the concept of a 'complex intervention'

A simple intervention is generally characterized as involving a s single component with effects unlikely to vary by context (Petticrew 2011). By contrast, complex interventions involve several synergistic interacting components whose effects are irreducible to a single 'active ingredient'(Medical Research Council 2000). More recent literature suggests that complexity can arise from the number and difficulty of behaviours required by those delivering or receiving the intervention, the number of groups or organizational levels targeted by the intervention, the number and variability of outcomes, and the degree of flexibility or tailoring of the intervention permitted(Craig et al. 2008; Skivington et al. 2021). Accordingly, all public health interventions must be regarded as complex.

Complex interventions or complex settings, contexts, and systems

In contrast to the aforementioned literature, it has also been recently suggested that complexity, rather than describing the characteristics of a particular intervention, actually describes how interventions interact with local contexts or 'systems' (e.g. family, school, community, country) (Hawe et al. 2009; Skivington et al. 2021).

Interventions with several interrelated parts, addressing multiple levels or requiring multiple new behaviours, it has been suggested, may be more

helpfully thought of as complicated rather than complex. Complexity, so the argument runs, is a property of the system that an intervention is trying to change, not an inherent feature of the intervention itself (Shiell et al. 2008). Complex systems have properties emerging from the mutual interactions of multiple ‘subunits (that are often composites themselves) whose repeated interactions result in rich, collective behaviour that feeds back into the behaviour of the individual parts’(Rickles et al. 2007, p. 933). Interventions, whether simple or complicated, can be thought of as interruptions to such complex systems, with the implication that their local impact will be context-dependent and affected by recipient or provider factors (Lewin et al. 2017). For example, Project Northland in the United States used schools as a setting for a health promotion intervention to discourage alcohol use by young people. It involved making changes to elements of the school curriculum, changing the school environment, and engagement with parents and the wider community to restrict both access to and demand for alcohol (Perry et al. 2002). The intervention is clearly a complicated one, but its impacts within schools are also complex. Schools were not simply the setting for delivery but were each complex systems, involving a whole series of different types of subunits (e.g. students, teaching staff, classes, buildings), which the intervention aimed to ‘interrupt’ in multiple ways in order to reduce alcohol use. The schools were also part of a wider community, another complex system, on which the intervention was also intended to have an impact.

When developing a public health intervention, it can be instructive to understand something of the nature of complex systems, and to think of an intervention as something that is intended to disrupt the system. Doing so can help ensure that the likely mechanisms for achieving change, the potential consequences of disrupting the system (both intentional and unintentional), and the interaction between the intervention and the context in which it is being implemented (Moore et al. 2015) are all carefully considered from the outset.

This increasing interest in systems science and how it might inform the development and evaluation of public health interventions results has several important implications. Firstly, those developing public health interventions should focus not merely on engendering individual behaviour change but also on transforming entire social systems (for example, urban planning, housing, and welfare policies) so that these promote rather than damage

health (Rutter et al. 2017). Secondly, a focus on one or two primary health outcomes might be broadened so that evaluations explore broader impacts on system functioning and how this affects those involved in these systems. For example, evaluations of school-based health curricula might continue to examine direct effects on students' health knowledge and behaviours but also explore how delivering the curricula affects staff workload and mental wellbeing, or student engagement with education. Thirdly, we might revise how we think of interventions; rather than discrete entities with clear starting points and modes of delivery which can 'manualized', public health interventions are often routine, constantly adjusted, and intermingled, so disentangling the effect of any part is difficult. Fourthly, while it remains useful to estimate intervention fidelity and effects within rigorously conducted studies, it is important to appreciate that the same fidelity and effects might not manifest when the intervention is delivered across other systems that vary across time and space. The overall implication of this is that evaluations will be most useful when they enable development and refinement of theories about how intervention can interrupt systems in ways that promote health' (Rutter et al. 2017) and how this will vary across contexts' (Egan et al. 2019, p. 14)

But such theories are always likely to be imperfect because of the properties of systems. Firstly, systems are adaptive and will likely adjust in response to the implementation or even suggestion of a new intervention. For example, when the UK government announced the introduction of a tax on sugar-sweetened beverages, some producers began to reformulate the sugar content of their drinks before the legislation came into effect. Secondly, as already mentioned, systems will have emergent properties arising from dynamic interactions between their components. For example, herd immunity, where a critical mass of individuals is immune to an infection such as through vaccination, is a property of a population system arising from infection-host interactions which confers benefits on individuals who might themselves lack individual immunity. Thirdly, systems often include feedback loops, where one change enhances, balances, or diminishes another. For example, the introduction of compulsory charges on plastic bags in shops in the United Kingdom to reduce plastic pollution has led directly to a huge decline in their use and shifted cultural norms towards reuse, perhaps contributing towards the popularity of other forms of reuse and recycling. Fourthly, systems are characterized by non-linear dose-responses. For

example, in an attempt to induce system level change in the secondary prevention of key non-communicable diseases in the United Kingdom, since 2009 there has been very substantial investment in the NHS Health Check programme. All those aged 40–74 years are invited to attend their general practice for checks relating to stroke, kidney disease, heart disease, type 2 diabetes, and dementia, and for measures to reduce risk of adverse outcomes to be taken. However, despite the substantial financial input (or dose) the response has been disappointing with a recent evaluation concluding that 'program performance was substantially below national and international targets' (Chang et al. 2016).

Standardized or flexible interventions?

When the UK MRC first published a framework for the development and evaluation of complex interventions (Medical Research Council 2000) it adapted the phases undertaken in drug development and testing and proposed that 'once the RCT has begun, the intervention must not evolve, as the RCT results will be unusable if later participants experience a different intervention than earlier ones. Fidelity to treatment protocols or quality standards may be essential to effectiveness' (Medical Research Council 2000, p. 11). This emphasis on testing a standardized intervention, implemented with a high degree of fidelity, caused some concern because, it was thought, public health interventions might be maximally effective if they actively encouraged appropriate tailoring to local circumstances. Such a view led some to conclude that randomized trials were therefore not a suitable method for evaluating public health and health promotion interventions (McNulty et al. 2013). As Hawe and colleagues (Hawe et al. 2004) pointed out, however, this concern was based on a misconception of what is standardized. She argued that what should be standardized is the function of the intervention and the steps in the process by which change is to be achieved, not that all the components of an intervention have to be delivered in exactly the same way to each person, or in each setting. Hawe and colleagues did not reject the randomized trial as a method of evaluation but argued that 'instead of mimicking trial phases which assume that the 'best' or the 'ideal' comes from the laboratory and gets progressively compromised in real world applications, community trial design would start by trying to understand communities themselves as complex systems and

how the health problem or phenomenon of interest is recurrently produced by that system' (Hawe et al. 2004, p.1562). A pragmatic solution to the problem of artificial standardization, however, is to simply embrace the idea that the interventions we develop and evaluate have adaptation and evolution built into them. This matter is discussed in further detail in the later subsection on fidelity.

Simple or complex research questions and questions from different research perspectives

Evaluations can address simple or complex questions (Petticrew 2011). For example, a local authority commissioner might wish to evaluate an obesity prevention programme comprising multiple components: health education for pregnant women to encourage breastfeeding and healthy weaning; provision of healthy food and opportunities for physical activity among children under the care of nursery schools, child minders, and day care; classroom-based education on nutrition and physical activity in primary schools; and local spectator sports clubs encouraging fathers to be physically active and eat healthily with their children. A commissioner may simply wish to know whether or not the overall programme has been effective in reducing rates of obesity. Or she may wish to know: for which subpopulations of children and in what settings it was effective or ineffective; whether there were important secondary outcomes such as reductions in adult obesity; whether certain components interacted to achieve particularly strong effects; or whether there were broader benefits or harms, such as a reduction in the number of fast-food outlets in the locality due to a drop in demand.

The research question might also be informed by the research perspective(s) of those doing the evaluation. The latest draft MRC guidance on the development and evaluation of complex interventions suggests four different perspectives on a continuum from efficacy to systems considerations each asking different but not mutually exclusive questions: '*Efficacy perspective*—To what extent does the intervention produce the intended outcome(s) in experimental or ideal settings? *Effectiveness perspective*—To what extent does the intervention produce the intended outcome(s) in real-world settings? *Theory-based Perspective*—What works

in which circumstances, and how? and *Systems perspective*—How do the system and intervention adapt to one another?' (Skivington et al. 2021)

Intervening at different levels and within systems to improve the health of populations

The importance of multilevel and multifaceted interventions was set out in both the 1986 *Ottawa Charter for Health Promotion* (WHO 1986) and the later *Bangkok Charter* (Anonymous 2006) so that interventions should aim both to address individual, 'downstream' determinants of health such as knowledge and attitudes, as well as community and societal 'upstream' determinants, such as exposure to health benefiting and health harming goods and services, and social inequalities in wealth and power. When designing complex public health interventions, the socioecological approach to public health described by McLeroy and colleagues (Kenneth et al. 1988) with its origins in the work of Bronfenbrenner (Bronfenbrenner 1974) on ecological systems theory, can provide a useful framework for thinking about the different levels of intervention. This socioecological framework was designed to focus attention on social and environmental factors as targets for health promotion. It identifies five complementary levels of change, targets, and approaches which are set out in Table 6.5.1.

Historical evidence suggests that the greatest improvements in public health have been achieved through non-medical interventions, such as improved sanitation, welfare benefits, and improved workplace safety (McKeown 1976). More recent evidence also indicates that multifaceted and multilevel interventions are more effective than those operating at a single level (Greaves et al. 2011), and that environmental level interventions are more likely to be cost-effective than individually targeted clinical or non-clinical prevention interventions (Chokshi and Farley 2012). Reviews of the use of socioecological models in health promotion since the late 1980s have, however, suggested that published evaluations still tend to focus on one or two levels of intervention (Richard et al. 2011; Golden and Earp 2012), although there has been a notable increase, in interventions with multiple and more 'upstream' targets, such as the community or political environment (Richard et al. 2011; United Kingdom Clinical Research Collaboration 2018).

Table 6.5.1 Socioecological framework

Level of change	Target	Approach
Intrapersonal	Individuals	Changing individuals by changing their knowledge, attitudes, values, motives, skills, and behaviour
Interpersonal	Small groups such as the family, work colleagues, friendship groups	Achieving change through formal and informal social networks and peer groups by changing group norms or the group culture
Institutional	Organizational settings such as schools, colleges, and workplaces	Achieving change via organizational policies and rules, and by creating healthier organizational cultures and environments
Communities as mediating structures, relationships among organizations, and as geographical entities with political power	Neighbourhoods, community organizations	Achieving change through collective action by working with community organizations that are sources of social capital, coordination and coalition building between organizations, and increasing the representation of disadvantaged population groups on community decision-making bodies
Public policy	Whole populations at local, regional, national, and international level	Achieving structural change through the development and implementation of health promoting local-, regional-, country-, and international-level policy, and legislation

Adapted with permission from McLeroy, K.R. et al., An Ecological Perspective on Health Promotion Programs, *Health Education Quarterly*, Volume 15, Issue, pp. 351–77, Copyright © 1988 by Society for Public Health Education, DOI: 10.1177/109019818801500401.

(United Nations General Assembly, 2015).

Health inequalities

Public health has the dual goals of improving population health and reducing health inequalities between different populations. The United Nations' *2030 Agenda for Sustainable Development* has health and wellbeing as one of 17

goals the achievement of which is integrally linked to ending poverty and hunger, planetary health, maintaining peace, ensuring prosperity, and achieving dignity and equality for all (United Nations General Assembly 2015). One approach to tackling health inequality is to target interventions at disadvantaged groups at high risk of morbidity or mortality. This can be problematic, both because it can sometimes be stigmatizing and even harmful, but also because even if effective in reducing these groups' risk it may be insufficient in bring about substantial population-level reductions in risk because, as Geoffrey Rose pointed out (Rose 1981), where disease risks follow a bell-shaped curve (as is the case with almost all non-communicable disease), the majority of cases of a disease will arise from those at low or medium risk (because there are simply more of them). However, universalist approaches which do not address the 'upstream' determinants of disease and only address individual behaviour may have the potential for widening inequalities because of the 'inverse prevention law' (Gordon et al. 1999) whereby socioeconomically advantaged groups often benefit disproportionately from universal interventions. In particular, interventions requiring action from recipients, such as interventions to reduce obesity requiring participation in weight-loss schemes or interpretation of nutrition labelling, are likely to benefit those most able to take such action, therefore more likely to widen inequalities. Interventions less demanding of individual actions, such as governmental measures to reduce the sugar and fat content of food or increase the affordability of healthy food (Adams et al. 2016; Rutter et al. 2017) are more promising means of benefiting disadvantaged populations. Progressive universalism has been advocated, for example, by the *Marmot Review*, as a way of avoiding these difficulties by ensuring that action on health inequalities both addresses 'the causes of the causes' but also by ensuring that the scale and intensity of more downstream interventions are proportionate to the level of disadvantage (Marmot 2010).

These matters must be examined empirically within evaluations. Evaluations of the free-swimming initiative for children in England and of the introduction of universal access to free leisure facilities in the north-west of England are good examples of evaluations examining whether the inverse prevention law was operating. In both cases findings showed that physical activity could be increased and inequalities reduced. (Audrey et al. 2012; Higgerson et al. 2018).

When developing public health interventions, it is therefore important to consider their potential for narrowing or widening health inequalities, and to ensure that evaluations assess whether factors such as socioeconomic status, gender, or ethnicity moderate (statistically interact with) intervention effects. Both the evaluations of physical activity interventions described earlier used large administrative datasets which provided the statistical power to for such moderator analyses. It is important to pursue such analyses, even in the absence of sufficient statistical power, because the findings can contribute to the accumulation of evidence in systematic reviews of the likely impact of particular interventions on inequalities (Tinner et al. 2018).

The PROGRESS acronym (Evans and Brown 2003) indicates disadvantage may be related to place of residence, race/ethnicity, occupation, gender, religion, education, socioeconomic status, and social capital and is a helpful prompt to thinking about which factors are relevant to the population for which the intervention is being designed. The Health Inequalities Assessment Toolkit (NIHR CLAHRC North West Coast 2017) is another useful resource for this purpose.

Stages in the development of complex interventions

Several guides relevant to the process of intervention development have been published recently. These include a taxonomy of approaches to developing health improvement interventions (O'Cathain et al. 2019), guides to the process itself (French et al. 2012; Wight et al. 2016; Hawkins et al. 2017) guidance on considering context in relation to population health interventions (Craig et al. 2018) and a framework for reporting the details of complex interventions in a standardized way for replication, transfer and evidence synthesis purposes (Hoffmann et al. 2014). The taxonomy of approaches to intervention development is based on a systematic review and synthesis of relevant literature from which the authors identify and describe eight categories of approach to intervention development and 18 different actions which can be taken within each approach (O'Cathain et al. 2019). The strengths and weaknesses of each approach are identified. This compendium is very detailed and covers the development of clinical and patient-based interventions as well as those with a population health focus. Next we outline the key stages in the process of complex public health intervention development.

Review of existing quantitative and qualitative evidence

A first practical step in developing a complex intervention, whether designing something new or modifying an existing approach, is to review the available evidence. If there are already relevant systematic reviews, it may only be necessary to read these or update them. But if not, a new review may be needed (see Chapter 5.13 for further information about systematic reviews). While systematic reviews of quantitative evidence of intervention effects will summarize what is known about the effectiveness of a particular intervention, it is also useful to review relevant qualitative or quantitative research on implementation, which may provide important knowledge about the contexts in which the intervention has previously been used and how context might influence the feasibility, acceptability and therefore potential impact of the intervention (Craig et al. 2018).

Public involvement from the outset

Including potential beneficiaries in intervention development is vital in ensuring that the intervention is relevant and acceptable. In the health and social care sectors, pre-existing groups representing particular patient and client groups can be engaged as partners in research, though the extent to which they are representative of all potential beneficiaries is often uncertain. In public health, identifying representative groups, particularly for universalist interventions, may be even more challenging. There is, however, guidance available on how to approach this and examples of good practice. INVOLVE is part of the National Institute for Health Research in the United Kingdom and its role is to encourage, support and set standards for public involvement in health research (INVOLVE 2017). Its website provides ample guidance for researchers (http://www.invo.org.uk) and the public. Involvement in public health research can and should, where relevant, include children and young people (Bird et al. 2013) and there are specialist groups set up who can provide such input. The ALPHA group (Advice Leading to Public Health Advancement) is one such example. This specially trained group of young people meets monthly to provide input into research undertaken by the DECIPHer public health research centre and other research organizations (DECIPHer, n.d.).

Involving policy and practice partners from the outset

Consulting or collaborating with those who determine policy about or actually deliver similar interventions at this early stage will also provide useful insights. Involving such stakeholders from the outset is also more likely to result in its successful translation from research into routine practice if evaluation of the intervention suggests it is effective and cost-effective. System mapping is increasingly being used in stakeholder consultations to help identify what factors are thought to contribute to the public health problem to be tackled, to consider how and where to intervene, and what resources might be leveraged and deployed to address the problem. Such system maps can also be used to guide evaluations as in the example to follow (Fig. 6.5.1), which was used to inform the design of an evaluation of the health impacts of the UK Treasury Soft Drinks Industry levy (SDIL) (White 2017).

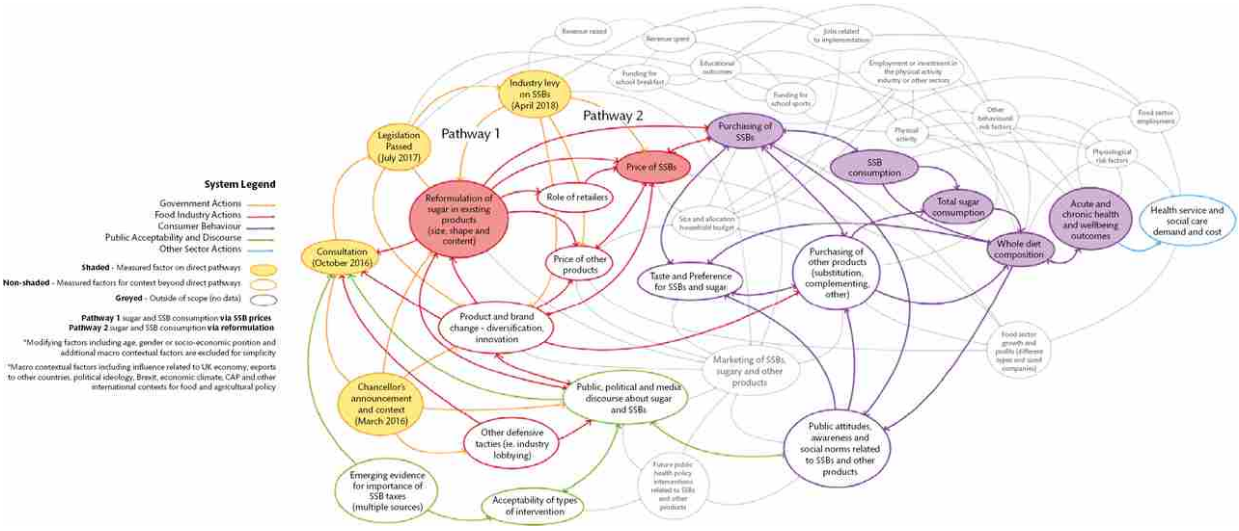


Fig. 6.5.1 System map for the evaluation of the health impacts of the UK Treasury Soft Drinks Industry Levy.

Importance of theory

Pawson and Tilley (1997) suggest all interventions are 'theories incarnate' because they assume that a certain action (or intervention) will trigger mechanisms which, in combination with the conditions that prevail at the time (context), generate particular impacts (or outcomes). However, such assumptions may remain implicit and the opportunity for evaluation to test them and thus contribute to the refinement of middle-range theory may be lost (Moore and Evans 2017). Guidance on developing complex interventions advocates using explicit theories of change to inform their design (Campbell et al. 2007; Craig et al. 2008; Glanz and Bishop 2010; Skivington et al. 2021). While interventions formulated without recourse to theory can and do work, there is evidence that interventions are more likely to be effective when informed by theory (Noar et al. 2008; Glanz and Bishop 2010). The ASSIST peer-led smoking prevention intervention is an example of a successful theory-based smoking prevention intervention (Campbell et al. 2008). Students aged 12–13, identified by classmates at secondary school as being influential, are trained to have conversations with their peers about the benefits of not smoking (Audrey et al. 2004). The intervention was informed by diffusion of innovations theory (Rogers 1995) with the peer supporter's role being to diffuse new attitudes and new norms of behaviour relating to smoking throughout the year group. A cluster RCT conducted in south Wales and south-west England showed a 22% overall reduction in the odds of being a smoker in intervention compared with control schools. More detailed follow-on research on the peer supporters found that those who are asked to work informally, rather than under the supervision of teaching staff, engaged more effectively with the task they have been asked to undertake and more effectively diffused the health-promotions messages, in this case smoking prevention (Audrey et al. 2008).

The ASSIST intervention drew on a theory from the social sciences. Taking theories 'off the shelf' to inform intervention development is a common but potentially problematic approach because of a tendency for researchers to choose popular psychological or behavioural theories which may not be altogether pertinent to the mechanisms through which the problem is actually sustained (Moore and Evans 2017). For example, the Gatehouse whole-school intervention aimed to promote mental and physical health among adolescents by rendering school environments more inclusive and engaging but was informed by Bowlby's attachment theory which focuses on dynamics between mothers and infants (Patton et al. 2003). This

tendency is confirmed by reviews of which theories have been utilized most frequently in public health interventions. These show that the field is dominated by psychological theories (Shahidullah 1995; Glanz and Bishop 2010; Davis et al. 2014). However, as we have already indicated, effective public health intervention is likely to require action at more than just the level of the individual, needing to be informed by a clear understanding of the workings of the social system that sustains the public health problem to be tackled. Public health intervention research may therefore need to draw on theories from a variety of disciplines. There are useful resources available which provide compendia of theories but the range of theories included in each reflects the disciplines involved or the contexts examined (Glanz and Rimmer 1997; Conner and Sparks 2005; Nutbeam et al. 2010; Michie et al. 2014). It also means that stakeholders (including the public) must be included in a process of co-creating interventions to ensure that their experiential knowledge of the problem and its context can be factored into the theory of change (Moore and Evans 2017).

Development of intervention logic models

Logic models are diagrammatic representations of the intervention theory of change and of the logical relationships, for example, between an intervention's inputs, activities, outputs, and outcomes, although there is scope for flexibility regarding the precise categories used (Cooksy et al. 2001; W.K. Kellogg Foundation 2004). Intervention theory and logic models are useful tools in clarifying how an intervention is to be delivered and to what ends, and thereby provide a focus for evaluating process (and outcomes). Figure 6.5.2 provides an example of a recent logic model. It is for the Learning Together (INCLUSIVE) intervention which was designed to reduce bullying and aggression in secondary schools (Bonell et al. 2018a). The intervention's theory of change is informed by Markham and Aveyard's theory of human functioning and school organization (Markham and Aveyard 2003). Its tenets inform a diagrammatic logic model consisting of inputs, processes, institutional changes, intermediate impacts, and health outcomes. Inputs include training staff in 'restorative practice' (RP) as an approach to preventing and managing student aggression and misbehaviour, as well as an external facilitator to convene meetings of students and staff to review data on student needs in order to plan and deliver school-specific

actions. These inputs are intended to enable processes, such as improved personal, social, and health education (PSHE) and the use of RP in schools, as well as the review and revision of school policies and rules by staff and students working collaboratively. In turn these processes are intended to modify various 'boundaries' within the school, such as the quality of staff–student relationships, and how schooling is framed, for example, to render decision-making more student-centred. These processes are ultimately theorized as transforming the ethos of the school so that more students use their sense of agency to commit to the school's instructional order (teaching) and the regulatory order (discipline and school sense of community) and are thus less likely to decide to commit to antischool peer groups and behaviours such as bullying and substance use which can function symbolically as alternative, antischool markers of identity and status.

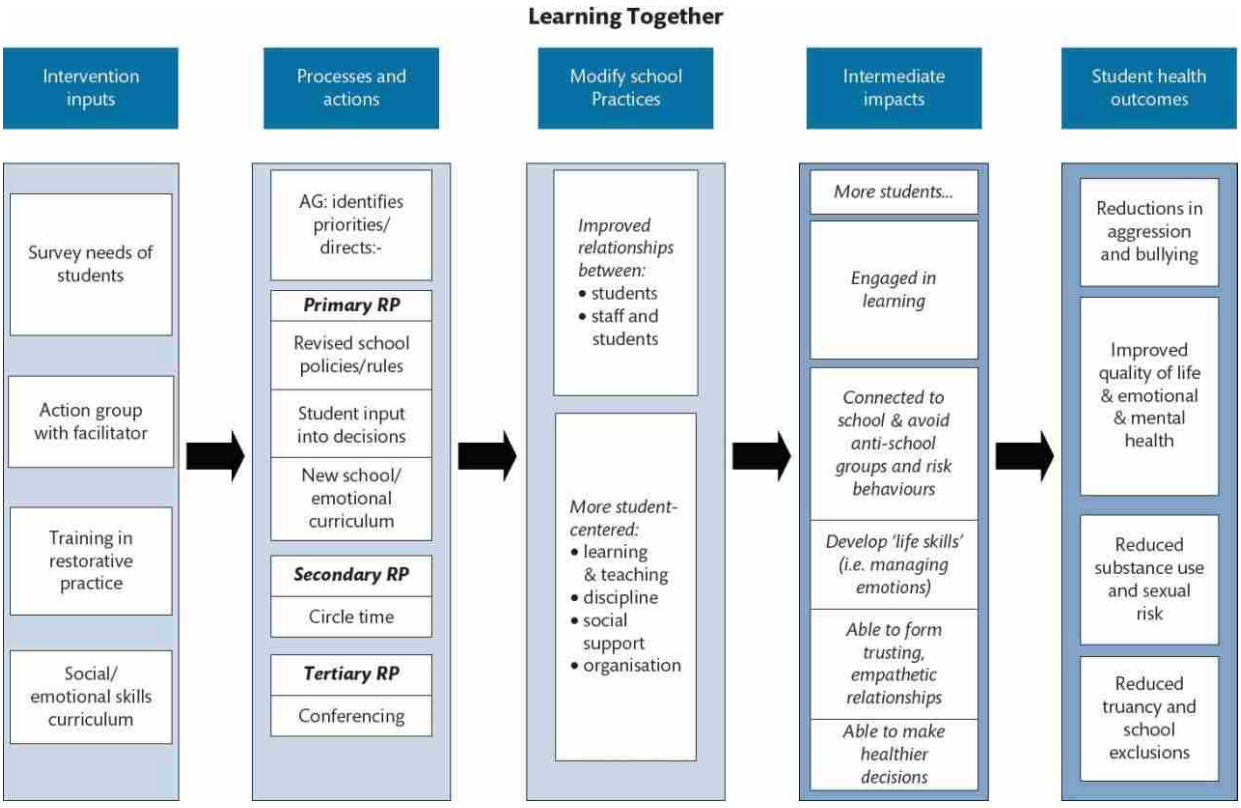


Fig. 6.5.2 Logic model for the Learning Together (INCLUSIVE) intervention.

The intervention has been tested in a cluster RCT in 40 secondary schools in England and was shown to have a modest, but in public health terms an

important, impact on bullying but no effect on aggression (Bonell et al. 2018a).

While useful, logic models have limitations. They describe linear chains of causation, offering no indications of how intervention implementation or the mechanisms by which outcomes are generated might interact with and therefore vary by context. In the case of INCLUSIVE, the linear logic model was augmented by additional context-mechanism-outcome (CMO) hypotheses suggesting how contextual factors and intervention mechanisms might interact to generate different outcomes in different settings. A further problem with logic models is that they can sometimes lead to the variables used to measure phenomena being confused with the phenomena themselves, thus reducing social, psychological, or biological mechanisms to chains of variables with little discussion of what processes connect these. So, while logic models provide a useful summary and guide to empirical assessment, they should build on a more nuanced textual description of the theory of change (Skivington et al. 2021). In the case of INCLUSIVE, for example, the theory of change invoked concepts such as agency to make decisions as well as the symbolic meaning of risk behaviours such as bullying, to explain intervention mechanisms (Bonell et al. 2018b).

Public health interventions, complex or otherwise, can also have unintended consequences and the potential to harm (Lorenc and Oliver 2014). Theorizing an intervention's potential iatrogenic mechanisms and harmful effects and developing so called 'dark logic' models to represent these ensures that they are considered in any evaluation with a view to understanding how they might be mitigated or avoided in future (Bonell et al. 2015b).

Evaluation of complex interventions

Piloting and feasibility studies

Before embarking on an RCT (or some other evaluation of the effects of a public health intervention), it is wise to undertake a pilot or feasibility study. Guidance on how to conduct these has been lacking but the main report from the GUEST study (Guidance for F**e**asibility Studies of Complex Public Health Interventions) will be published shortly (Moore et al. 2018). The terms pilot and feasibility study have tended to be used interchangeably

(Arain et al. 2010) but Eldridge and colleagues (Eldridge et al. 2016) suggest that a pilot study is just one form of feasibility study. A pilot study is in effect a miniature version of the main randomized trial (or whatever design is anticipated to examine effectiveness) and is designed to test that both the intervention and the evaluation can be implemented as intended. If the pilot is successful and no changes are made, then data from it can be incorporated into the main study (this is termed an internal pilot as opposed to an external one where data cannot be incorporated into the main study because the intervention or the evaluation methods are refined between pilot and main study). In contrast, feasibility studies are intended to examine specific, discrete areas of uncertainty to decide whether an effectiveness study is feasible or worthwhile, such as the willingness of participants to be randomized, response rates to questionnaires collecting outcome data, or the standard deviation of the primary outcome measure required for the sample size calculation. Feasibility studies also provide the opportunity to test out and optimize the intervention. Such studies need not pilot every aspect of a main study, including random allocation, and will use whatever methods are required to address the identified areas of uncertainty.

Setting predetermined criteria for progression from a feasibility study or a pilot trial to the main study of effectiveness helps bring objectivity to whether or not to proceed and may be required by research funders. Criteria may inform binary stop/go decisions or more nuanced red/green/amber with amber denoting progression contingent on some refinements. Next in Table 6.5.2 are the criteria set for progression to a full-scale RCT to test the cost-effectiveness of a secondary-school-based intervention to improve teacher well-being. Part of the intervention involved training a cadre of teachers in intervention schools in mental health first aid so that they could provide peer support to colleagues. Not all the criteria were met but the full trial was funded because the study authors successfully argued that what they had learnt in the feasibility study would enable them to improve the recruitment of schools and peer supporters, and increase questionnaire response rates from teachers.

Table 6.5.2 Criteria for progression from a pilot to a full RCT in the WISE (Wellbeing in Secondary Schools) Study and the outcome.

Criterion	Outcome
Recruit and maintain peer supporter teams assuming ratio of 1 peer supporter: 15 staff members	Mean ratio of 1 peer supporter: 17 staff members
50% schools willing to be recruited and randomized	Wrote to 33 eligible schools: 9 requested further information, 7 met with the research team, 6 took part in the feasibility study
Response rate of at least 75% in staff and student follow up questionnaires	Response rate at follow up 52.7% all staff (66.2% teachers) and 85.7% students
Average of at least one intervention a month delivered by each peer supporter	Mean of 3 people per supporter per 2 weeks BUT some supporters as many as 12 per 2 weeks, while others provided no support

(Kidger J. 2019, personal communication).

Process evaluation methods

Process evaluation is as important as outcome evaluation. Process evaluation can provide valuable insight not only within feasibility and pilot studies, but also within full RCTs as well as in studies examining implementation of scaled-up interventions after effectiveness trials have been completed. Process evaluations can examine how interventions are planned, delivered, and received (Bonell et al. 2006; Oakley et al. 2006).

There is no single, agreed definition of process evaluation. The UK MRC guidance on the subject states that a process evaluation 'is a study which aims to understand the functioning of an intervention, by examining implementation, mechanisms of impact, and contextual factors. Process evaluation is complementary to, but not a substitute for, high quality outcomes evaluation' (Moore et al. 2015). Other priority areas have been identified. One of the most widely cited frameworks is that of Steckler and Linnan (2002) which identified priority areas for process evaluation, including fidelity (the extent to which the intervention is delivered as intended), reach (whether an intervention is received by all those it targeted), dose delivered (the amount or number of units of intervention offered to

participants, dose received (the extent of participants' active engagement in the scheme), and context (local factors which impact implementation). Thus, process evaluation can explore whether those whom an intervention targets actively engage with it, an important consideration for social interventions. The following subsections consider some of these priority areas in more depth.

Intervention delivery and fidelity

The importance of process evaluation was increasingly recognized during the 1980s, informed by discussion of 'type 3 error', defined as concluding that an intervention is ineffective because of a failure of its theory of change when lack of effect reflects its not being implemented as intended (Basch et al. 1985). Initially, the main aim of process evaluations was therefore to assess intervention fidelity to determine whether outcomes evaluation could represent valid assessment of an intervention's underlying theory. Fidelity is defined variously. Moncher and Prinz (1991) define fidelity as confirmation that manipulation of the independent variable occurred as planned. Within the context of an RCT or other comparative evaluation, this requires assessment of the integrity of both delivery within the intervention arm and its differentiation from what occurred in the control arm. Later definitions of fidelity, for example, by Lichstein and colleagues (Lichstein et al. 1994) saw this as also encompassing receipt (participant uptake of the intervention) and enactment (the extent to which participants apply the lessons of interventions into practice in their lives). But such definitions begin to blur the useful distinction between fidelity and factors such as dose, acceptability, and even outcomes, so most researchers continue to define fidelity more narrowly in terms of delivery.

Assessments of fidelity generally involve quantitative measurement, for example, using observations, video-recording, or provider self-assessments to audit delivery against what was intended. However, fidelity should not be viewed only in terms of ticking boxes on simple checklists. Steckler and Linnan (2002) argue that when evaluating fidelity, it is also important to assess whether the 'spirit' of an intervention is delivered in practice, and not just its mechanics. For example, when evaluating motivational interviewing, which is counselling aimed at enabling clients to recognize their own motivations for behaviour and to modify behaviour where this is not in line

with their own goals, it is important to evaluate fidelity not merely as a cookbook list of techniques, but also whether the underlying empathic, patient-centred spirit of the approach is achieved (Miller and Rollnick 2009).

Another important consideration when evaluating fidelity is the degree to which interventions should be standardized. Advocates of strict fidelity (Mihalic 2004) argue that this is essential if effective interventions are to be replicated, especially when an intervention's 'active ingredient' may not be known (or is thought to arise from synergies between all components). They present evidence which suggests that high fidelity is associated with greater impact for some interventions (Mihalic 2004). In contrast, advocates of local adaptation (Dane and Schneider 1998) argue that interventions need to be tailored to local circumstances. Durlak and DuPre (2008) propose a compromise whereby an intervention's 'core components' should be delivered in standard form but less central intervention components or features can be modified to fit local needs. They present research suggesting that a balance between fidelity and adaptation is likely to be most effective, with the precise balance dependent on the specific intervention. As noted earlier in this chapter, Hawe and colleagues go further to argue that with complex interventions characterized by interactions between multiple components and between intervention and context, too much attention has been paid to standardizing the *form* of intervention components (what precise actions are delivered) and more attention should be given to standardizing the *function* of intervention components (what the aim of each component is and how it contributes to the overall aim of the intervention) (Hawe et al. 2004). Consider the case of the INCLUSIVE intervention described earlier. Fidelity of form might be measured by determining whether the training in RP was delivered as planned and whether the facilitator convened a sufficient number of staff–student meetings. In contrast, fidelity of function might be measured by determining whether teachers were equipped to deliver RP and whether staff and students felt involved in the process of reviewing school policies.

Intervention reach, dose, and acceptability

Steckler and Linnan conceive of evaluating intervention reach and dose, and participants' responses to an intervention largely in quantitative terms (Steckler and Linnan 2002). Reach and dose are commonly examined

quantitatively using methods such as questionnaire surveys exploring participants' exposure to and satisfaction with an intervention. However, receipt can also be seen in qualitative terms as exploring participants' reports of an intervention in their own terms. Qualitative research can be useful in examining how participants perceive an intervention in unexpected ways which may not be fully captured by researcher-developed quantitative constructs. Qualitative research can also explore how providers or participants enact 'agency' (willed action) in engaging with the intervention rather than merely receiving it passively (Oakley et al. 2006). For example, process evaluation of a youth work intervention to reduce teenage pregnancy in England reported how some youth workers and teachers responded to onerous recruitment and retention targets by transforming the intervention from one delivered out of school hours to one delivered as a partial alternative to normal school time, which may have explained its ineffectiveness (Wiggins et al. 2006).

Context

Context is a critical aspect of process evaluation. Although there is no consistent definition (Craig et al. 2018) context can be taken to mean anything that was in place before the intervention itself was implemented. This could include factors such as the needs of participants, the infrastructure within which interventions will be delivered, the skills and attitudes of providers, and the attitudes and cultural norms of potential participants. In Steckler and Linnan's (2002) framework, context is largely considered in terms of factors which may help or hinder implementation. For example, an intervention may be delivered poorly in some areas, but well in others, because of better provider capacity or more receptive community norms in some areas. Others consider context also in terms of how it influences outcomes. For example, 'realist' evaluators place great emphasis on whether intervention mechanisms generate beneficial outcomes in some contexts more than others. Context can be measured quantitatively in order to inform 'moderator' analyses within or between trials, though the latter is rarer (Bonell et al. 2012). Qualitative research allows for a different understanding of the importance of context, for example, examining how intervention providers or recipients describe the interaction between their

context and their own agency in explaining their actions (Oakley et al. 2006).

Intervention mechanism of action

Assessing an intervention's mechanisms of effects involves assessing the validity of the theory of change in explaining how outcomes are generated. Such analysis can explain why an intervention is found to be effective or ineffective within an outcome evaluation. This might be critically important in refining an intervention found to be ineffective or in understanding the potential generalizability of interventions found to be effective. Quantitative data can be used to undertake mediator analyses to assess whether intervention outputs or proximal outcomes appear to fully or partially account for intervention effects on distal health outcomes (Gardner et al. 2006). Qualitative data can be also used to examine such pathways in terms of how they are perceived and described by those involved, and this is particularly useful when the pathways in question have not been comprehensively examined using quantitative data, as well as when pathways are too complex (e.g. using multiple steps or feedback loops) to be assessed adequately using quantitative analyses. This was, for example, the case with an assessment of possible causal pathways underlying an intervention to modify school environments in order to reduce student substance use, which appeared to involve feedback loops whereby student participation within certain intervention activities increased student commitment to the intervention which in turn increased student participation within other components (Bonell et al. 2010).

However, such analyses can be challenging. First, as suggested earlier, quantitative analyses require evaluators to have correctly anticipated what data are needed to examine causal pathways and to have collected these. For example, the evaluation of the youth work intervention aiming to prevent teenage pregnancy reported a higher rate of pregnancies among those receiving the intervention. It was hypothesized that this apparent adverse outcome resulted from the intervention bringing together young people engaged in risk behaviours so that pro-risk social norms were amplified. However, data on these were not collected so that this hypothesis could not be tested (Wiggins et al. 2008). A second challenge involves using qualitative alongside quantitative data to understand causal pathways. If

qualitative data are analysed in order to explain quantitative findings, this may bias interpretation (Wight and Obasi 2002; Campbell et al. 2015). Furthermore, quantitative and qualitative methods originate from different research traditions. Qualitative research focuses on participants' accounts, drawing out generalizations about meanings and agency using largely inductive methods. They offer rich insights into how actors understand social phenomena rather than trying to test hypotheses about causal associations. As Munro and Bloor (2010) put it: 'process evaluations do indeed enrich our understanding of the social processes involved in the delivery and receipt of complex interventions', they provide 'interpretations to which a degree of indeterminacy is always attached that cannot be expressed in probabilistic terms. The continuing tension between positivist and interpretive paradigms within that research design pose problems of understanding and reporting. Not so much "paradigm peace" as "paradigm truce" '.

While we advocate process evaluation as an important evaluative tool it should not be seen as a panacea for answering all the questions an outcome evaluation cannot, nor is its use unproblematic. The range of possible domains it can focus on (delivery, fidelity, reach, dose, acceptability, context, sustainability) is wide, and resources may not be available to collect the data required to examine them all. Some may be a greater priority than others depending on the nature of the intervention and its stage of development. For example, the acceptability of an intervention is something that should predominantly be explored in a pilot or feasibility study rather than leaving it till the definitive trial, where understanding mechanism of effects may be a more pressing matter for a process evaluation to examine.

Implementation issues

Sustainability

The implementability and sustainability of new public health interventions at a scale capable of population impact is an important consideration. For example, it may be that a group of health promotion specialists, substance misuse experts and youth workers have been specially assembled to deliver a new intervention in secondary schools to reduce substance and alcohol misuse. This intervention is shown in a cluster RCT to be effective. The

issue then becomes how can that delivery model be replicated and sustained outside of the trial so that effectiveness is maintained. The 'RE-AIM' framework (reach, effectiveness, adoption, implementation, maintenance) is useful in conceptualizing this (Glasgow et al. 1999). Its originators argue that those commissioning public health services need to know not only that an intervention is effective but that it is capable of being adopted by provider agencies and implemented by them as planned, and will reach sufficient people to achieve a population health impact that can be maintained over time (Glasgow et al. 2001).

From research into practice

The impact of research on policy and practice is increasingly recognized to occur through a gradual and indirect process of enlightenment rather than via direct effects of single studies or even reviews on single decisions (Weiss 1986; Graham et al. 2006). However, while empirically accurate, this is normatively disappointing for those advocating a more direct role for evidence in informing policy (Davies et al. 2000). Now there is recognition that to be influential, research needs to engage reciprocally with policy and practice. Earlier in this chapter we highlighted the need to consult key policy and practice stakeholders and the public early in the development and evaluation of complex public health interventions. This approach is now being extended by some into a process whereby evidence is co-produced through a knowledge exchange partnership between researchers, policymakers, and practitioners using transdisciplinary action research (Daker-White et al. 2014). This method involves academics from a variety of relevant disciplines working with policymakers, practitioners, and the public to identify what research is required, develop research questions, conduct evaluations, synthesize evidence, and implement and maintain evidence-based public health practice. The advantage of this approach is that if evaluations indicate that changes to practice are producing the desired outcomes then the capacity-building necessary to sustain that whole system change is built into the approach. There are a number of examples of promising interventions (de Silva-Sanigorski et al. 2010; Chittleborough et al. 2013), including the INCLUSIVE intervention in England (Bonell et al. 2015a; Bonell et al. 2018a) and the Gatehouse project in Australia, where this approach has been taken (Bond et al. 2001, 2004).

Discussion and future direction

Public health interventions can comprise policy-level initiatives implemented at a national or international level, for example the World Health Organization's Healthy Cities programme (World Health Organization, n.d.), through to brief individual behaviour change interventions to encourage reduced alcohol consumption delivered in primary care populations (Kaner et al. 2007). Interventions may be programmatic in nature with many different components such as Project Northland in secondary schools to reduce alcohol consumption, or it may be simpler, as in the example of human papilloma virus vaccination which involves three injections over a 12-month period to reduce substantially the chances of a woman having cervical cancer. Given this huge variability there is no 'one-size-fits-all' approach to their development and evaluation. What we have identified in this chapter are the key issues to be considered and the range of methods available to tackle them.

We conceive of the development and evaluation of complex public health interventions as distinct elements but ideally part of an overall symbiotic and cyclical process rather than sequential phases. Knowledge exchange partnerships between researchers and public health practitioners and policymakers offer a way of achieving this ideal, but, as is sometimes the case with the interventions themselves, the evidence suggests that the effectiveness of knowledge translation varies according to context (LaRocca et al. 2012).

REFERENCES

Academy of Medical Sciences (2017). The Future of Public Health Research: Summary report of a workshop held 20–21 July 2017. Academy of Medical Sciences.

Adams, J., Mytton, O., White, M., and Monsivais, P. (2016). Why are some population interventions for diet and obesity more equitable and effective than others? The role of individual agency. *PLoS Medicine*, 13, e1001990.

Anonymous (2006). The Bangkok charter for health promotion in a globalized world. *Health Promotion International*, 21(Suppl 1), 10–14.

Arain, M., Campbell, M.J., Cooper, C.L., and Lancaster, G.A. (2010). What is a pilot or feasibility study? A review of current practice and editorial policy. *BMC Medical Research Methodology*, 10, 67.

Audrey, S., Cordall, K., Moore, L., Cohen, D., and Campbell, R. (2004). The development and implementation of a peer-led intervention to prevent smoking among secondary school students using their established social networks. *Health Education Journal*, 63, 266–84.

Audrey, S., Holliday, J., and Campbell, R. (2008). Commitment and compatibility: teachers' perspectives on the implementation of an effective school-based, peer-led smoking intervention. *Health Education Journal*, 67, 74–90.

Audrey, S., Wheeler, B.W., Mills, J., and Ben-Shlomo, Y. (2012). Health promotion and the social gradient: the free swimming initiative for children and young people in Bristol. *Public Health*, 126, 976–81.

Basch, C.E., Sliepcevich, E.M., Gold, R.S., Duncan, D.F., and Kolbe, L.J. (1985). Avoiding type III errors in health education program evaluations: a case study. *Health Education & Behavior*, 12, 315–31.

Bird, D., Culley, L., and Lakhanpaul, M. (2013). Why collaborate with children in health research: an analysis of the risks and benefits of collaboration with children. *Archives of Disease in Childhood—Education and Practice*, 98, 42–8.

Bond, L., Glover, S., Godfrey, C., Butler, H., and Patton, G.C. (2001). Building capacity for system-level change in schools: lessons from the Gatehouse Project. *Health Education & Behavior*, 28, 368–83.

Bond, L., Patton, G., Glover, S., Carlin, J.B., Butler, H., Thomas, L., and Bowes, G. (2004). The Gatehouse Project: can a multilevel school intervention affect emotional wellbeing and health risk behaviours? *Journal of Epidemiology & Community Health*, 58, 997–1003.

Bonell, C., et al. (2018a). Effects of the Learning Together intervention on bullying and aggression in English secondary schools (INCLUSIVE): a cluster randomised controlled trial. *The Lancet*, 392, 2452–64.

Bonell, C., et al. (2015a). Initiating change locally in bullying and aggression through the school environment (INCLUSIVE): a pilot randomised controlled trial. *Health Technology Assessment*, 19, 1–109, vii–viii.

Bonell, C., Fletcher, A., Morton, M., Lorenc, T., and Moore, L. (2012). Realist randomised controlled trials: a new approach to evaluating complex public health interventions. *Social Science & Medicine*, 75, 2299–306.

Bonell, C., Jamal, F., Melendez-Torres, G.J., and Cummins, S. (2015b). 'Dark logic': theorising the harmful consequences of public health interventions. *Journal of Epidemiology and Community Health*, 69, 95.

Bonell, C., Moore, G., Warren, E., and Moore, L. (2018b). Are randomised controlled trials positivist? Reviewing the social science and philosophy literature to assess positivist tendencies of trials of social interventions in public health and health services. *Trials*, 19, 238.

Bonell, C., Oakley, A., Hargreaves, J., Strange, V., and Rees, R. (2006). Assessment of generalisability in trials of health interventions: suggested framework and systematic review. *British Medical Journal*, 333, 346–9.

Bonell, C.P., et al. (2010). Pilot multimethod trial of a school-ethos intervention to reduce substance use: building hypotheses about upstream pathways to prevention. *Journal of Adolescent Health*, 47, 555–63.

Bronfenbrenner, U. (1974). Developmental research, public policy, and the ecology of childhood. *Child Development*, 45, 1–5.

Campbell, N.C., et al. (2007). Designing and evaluating complex interventions to improve health care. *British Medical Journal*, 334, 455–9.

Campbell, R., et al. (2015). Intervention fidelity in a school-based diet and physical activity intervention in the UK: Active for Life year 5. *International Journal of Behavioral Nutrition and Physical Activity*, 12, 141.

Campbell, R., et al. (2008). An informal school-based peer-led intervention for smoking prevention in adolescence (ASSIST): a cluster randomised trial. *Lancet*, 371, 1595–602.

Chang, K.C.-M., et al. (2016). Impact of the National Health Service Health Check on cardiovascular disease risk: a difference-in-differences matching analysis. *The Canadian Medical Association Journal*, 188, E228–38.

Chittleborough, C.R., Nicholson, A.L., Young, E., Bell, S., and Campbell, R. (2013). Implementation of an educational intervention to improve hand washing in primary schools: process evaluation within a randomised controlled trial. *BMC Public Health*, 13, 757.

Chokshi, D.A. and Farley, T.A. (2012). The cost-effectiveness of environmental approaches to disease prevention. *New England Journal of Medicine*, 367, 295–97.

Conner, M., and Sparks, P. (2005). The Theory of Planned Behaviour and Health Behaviour. In: Conner, M. and Norman, P. (eds.) *Predicting Health Behaviour: Research and Practice with Social Cognition Models*. Buckingham: Open University Press.

Cooksy, L.J., Gill, P., and Kelly, P.A. (2001). The program logic model as an integrative framework for a multimethod evaluation. *Evaluation and Program Planning*, 24, 119–28.

Corbett, J., D'Angelo, C., Gangitano, L., and Freeman, J. (2018). Future of health: findings from a survey of stakeholders on the future of health and healthcare in England. *Rand Health Quarterly*, 7, 1–1.

Craig, P., Dieppe, P., Macintyre, S., Michie, S., Nazareth, I., and Petticrew, M. (2008). Developing and evaluating complex interventions: the new Medical Research Council guidance. *British Medical Journal*, 337, a1655.

Craig, P., et al. (2018). Taking account of context in population health intervention research: guidance for producers, users and funders of research. Canadian Institute of Health Research and the UK National Institute for Health Research. [Online]. Available at: https://www.repository.cam.ac.uk/bitstream/handle/1810/278765/876%20-%20FullReport-CIHR-NIHR-01.pdf?sequence=1&isAllowed=y

Daker-White, G., Donovan, J., and Campbell, R. (2014). Redefined by illness: meta-ethnography of qualitative studies on the experience of rheumatoid arthritis. *Disability and Rehabilitation*, 36(13), 1061–71.

Dane, A.V. and Schneider, B.H. (1998). Program integrity in primary and early secondary prevention: are implementation effects out of control? *Clinical Psychology Review*, 18, 23–45.

Davies, H.T.O., Nutley, S.M., and Smith, P.C. (2000). Learning from the past, prospects for the future. In: Davies, H.T.O., et al. (eds.) *What Works? Evidence-based Policy and Practice in Public Services*. Bristol: Policy Press, pp. 351–62.

Davis, R., Campbell, R., Hildon, Z., Hobbs, L., and Michie, S. (2014). Theories of behaviour and behaviour change across the social and behavioural sciences: a scoping review. *Health Psychology Review*, 8, 323–44.

De Silva-Sanigorski, A., et al. (2010). Evaluation of the childhood obesity prevention program Kids—'Go for your life'. *BMC Public Health*, 10, 288.

Decipher (n.d.). *ALPHA—Advice Leading to Public Health Advancement*. [Online] Available at: http://decipher.uk.net/en/content/cms/about-decipher/public-involvement/alpha%20researchers/

Durlak, J.A. and Dupre, E.P. (2008). Implementation matters: a review of research on the influence of implementation on program outcomes and the factors affecting implementation. *American Journal of Community Psychology*, 41, 327–50.

Egan, M., et al. (2019). Guidance on Systems Approaches to Local Public Health Evaluation. *Part 2: What to Consider when Planning a Systems Evaluation*. London: National Institute for Health Research School for Public Health Research.

Eldridge, S.M., et al. (2016). Defining feasibility and pilot studies in preparation for randomised controlled trials: development of a conceptual framework. *PLoS One*, 11, e0150205.

Evans, T. and Brown, H. (2003). Road traffic crashes: operationalizing equity in the context of health sector reform. *Injury Control and Safety Promotion*, 10, 11–12.

French, S.D., et al. (2012). Developing theory-informed behaviour change interventions to implement evidence into practice: a systematic approach using the Theoretical Domains Framework. *Implementation Science*, 7, 38.

Gardner, F., Burton, J., and Klimes, I. (2006). Randomised controlled trial of a parenting intervention in the voluntary sector for reducing child conduct problems: outcomes and

mechanisms of change. *Journal of Child Psychology and Psychiatry*, 47, 1123–32.
Glanz, K. and Bishop, D.B. (2010). The role of behavioral science theory in development and implementation of public health interventions. *Annual Review of Public Health*, 31, 399–418.
Glanz, K. and Rimmer, B. (1997). *Theory at a Glance: A Guide for Health Promotion Practice*. Bethesda, MD, National Cancer Institute.
Glasgow, R.E., Mckay, H.G., Piette, J.D., and Reynolds, K.D. (2001). The RE-AIM framework for evaluating interventions: what can it tell us about approaches to chronic illness management? *Patient Education and Counselling*, 44, 119–27.
Glasgow, R.E., Vogt, T.M., and Boles, S.M. (1999). Evaluating the public health impact of health promotion interventions: the RE-AIM framework. *American Journal of Public Health*, 89, 1322–27.
Golden, S.D. and Earp, J.A.L. (2012). Social ecological approaches to individuals and their contexts: twenty years of health education & behavior health promotion interventions. *Health Education & Behavior*, 39, 364–72.
Gordon, D., Shaw, M., Dorling, D., Davey-Smith, G., and Townsend, P. (1999). *Inequalities in Health: The Evidence Presented to the Independent Inquiry into Inequalities in Health, chaired by Sir Donald Acheson*. Bristol: Policy Press.
Graham, I.D., et al. (2006). Lost in knowledge translation: time for a map? *Journal of Continuing Education in the Health Professions*, 26, 13–24.
Greaves, C., et al. (2011). Systematic review of reviews of intervention components associated with increased effectiveness in dietary and physical activity interventions. *BMC Public Health*, 11, 119.
Hawe, P., Shiell, A., and Riley, T. (2004). Complex interventions: how 'out of control' can a randomised controlled trial be? *British Medical Journal*, 328, 1561–3.
Hawe, P., Shiell, A., and Riley, T. (2009). Theorising interventions as events in systems. *American Journal of Community Psychology*, 43, 267–76.
Hawkins, J., et al. (2017). Development of a framework for the co-production and prototyping of public health interventions. *BMC Public Health*, 17, 689.
Higgerson, J., Halliday, E., Ortiz-Nunez, A., Brown, R., and Barr, B. (2018). Impact of free access to leisure facilities and community outreach on inequalities in physical activity: a quasi-experimental study. *Journal of Epidemiology and Community Health*, 72, 252–8.
Hoffmann, T.C., et al. (2014). Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. *British Medical Journal*, 348, g1687.
Involve (NIHR) (2017). *National Standards for Public Involvement*. [Online] Available at: https://www.invo.org.uk/posttypepublication/national-standards-for-public-involvement/
Kaner, E.F., et al. (2007). Effectiveness of brief alcohol interventions in primary care populations. *Cochrane Database of Systematic Reviews*, (2), CD004148.
Kenneth, R.M., Bibeau, D., Steckler, A., and Glanz, K. (1988). An ecological perspective on health promotion programs. *Health Education Quarterly*, 15, 351–77.
Larocca, R., Yost, J., Dobbins, M., Ciliska, D., and Butt, M. (2012). The effectiveness of knowledge translation strategies used in public health: a systematic review. *BMC Public Health*, 12, 751.
Lewin, S., et al. (2017). Assessing the complexity of interventions within systematic reviews: development, content and use of a new tool (iCAT_SR). *BMC Medical Research Methodology*, 17, 76.
Lichstein, K.L., Riedel, B.W., and Grieve, R. (1994). Fair tests of clinical trials: a treatment implementation model. *Advances in Behaviour Research and Therapy*, 16, 1–29.
Lorenc, T. and Oliver, K. (2014). Adverse effects of public health interventions: a conceptual framework. *Journal of Epidemiology and Community Health*, 68, 288–90.
Markham, W.A. and Aveyard, P. (2003). A new theory of health promoting schools based on human functioning, school organisation and pedagogic practice. *Social Science and Medicine*,

56, 1209–20.
Marmot, M. (2010). *Fair Society, Healthy Lives: The Marmot Review: Strategic Review of Health Inequalities in England post-2010*. London. [Online] Available at: http://www.instituteofhealthequity.org/resources-reports/fair-society-healthy-lives-the-marmot-review
McKeown, T. (1976). *The Role of Medicine: Dream, Mirage, or Nemesis?* London: Nuffield Provincial Hospitals Trust.
McNulty, C.A.M., et al. (2013). Increasing chlamydia screening tests in general practice: a modified Zelen prospective cluster randomised controlled trial evaluating a complex intervention based on the theory of planned behaviour. *Sexually Transmitted Infections*, *90*, 188–94.
Medical Research Council (2000). *A Framework for Development and Evaluation of RCTs for Complex Interventions to Improve Health*. London: Medical Research Council.
Michie, S., West, R., Campbell, R., Brown, J., and Gainforth, H. (2014). *ABC of Behaviour Change Theories*. London: Silverback Publishing.
Mihalic, S. (2004). The importance of implementation fidelity. *Emotional & Behavioral Disorders in Youth*, 4, 83–105.
Miller, W.R. and Rollnick, S. (2009). Ten things that motivational interviewing is not. *Behavioural and Cognitive Psychotherapy*, 37, 129–40.
Moncher, F. and Prinz, R. (1991). Treatment fidelity in outcome studies. *Clinical Psychology Review*, 11, 247–66.
Moore, G.F., et al. (2015). Process evaluation of complex interventions: Medical Research Council guidance. *British Medical Journal*, 350, h1258.
Moore, G.F. and Evans, R.E. (2017). What theory, for whom and in which context? Reflections on the application of theory in the development and evaluation of complex population health interventions. *SSM—Population Health*, 3, 132–5.
Moore, L., et al. (2018). *Extending the MRC Complex Guidance on Developing and Evaluating Complex Health Interventions. Extending the MRC Complex Guidance on Developing and Evaluating Complex Health Interventions*. London: Medical Research Council.
Munro, A. and Bloor, M. (2010). Process evaluation: the new miracle ingredient in public health research? *Qualitative Research*, 10, 699–713.
National Institute for Health Research School for Public Health Research (n.d.). *Research Themes*. [Online] Available at: http://sphr.nihr.ac.uk/research
NIHR CLAHRC North West Coast (2017). *Health Inequalities Assessment Toolkit.* [Online] Available at: http://www.hiat.org.uk/introduction.html
Noar, S.M., Chabot, M., and Zimmerman, R.S. (2008). Applying health behavior theory to multiple behavior change: considerations and approaches. *Preventive Medicine*, 46, 275–80.
Nutbeam, D., Harris, E., and Wise, W. (2010). *Theory in a Nutshell: A Practical Guide to Health Promotion Theories*. Sydney: McGraw-Hill.
O'Cathain, A., et al. (2019). Taxonomy of approaches to developing interventions to improve health: a systematic methods overview. *Pilot and Feasibility Studies*, 5, 41.
Oakley, A., Strange, V., Bonell, C., Allen, E., and Stephenson, J. (2006). Process evaluation in randomised controlled trials of complex interventions. *British Medical Journal*, 332, 413–16.
Patton, G., Bond, L., Butler, H., and Glover, S. (2003). Changing schools, changing health? design and implementation of the Gatehouse Project. *Journal of Adolescent Health*, 33, 231–9.
Pawson, R. and Tilley, N. (1997). *Realistic Evaluation*, London: Sage Publications.
Perry, C.L., et al. (2002). Project Northland: long-term outcomes of community action to reduce adolescent alcohol use. *Health Education Research*, 17, 117–32.
Petticrew, M. (2011). When are complex interventions 'complex'? When are simple interventions 'simple'? *The European Journal of Public Health*, 21, 397–8.

Richard, L., Gauvin, L., and Raine, K. (2011). Ecological models revisited: their uses and evolution in health promotion over two decades. *Annual Review of Public Health*, 32, 307–26.
Rickles, D., Hawe, P., and Shiell, A. (2007). A simple guide to chaos and complexity. *Journal of Epidemiology and Community Health*, 61, 933–7.
Rogers, E.M. (1995). *Diffusion of Innovations*. New York: Free Press.
Rose, G. (1981). Strategy of prevention: lessons from cardiovascular disease. *British Medical Journal*, 282, 1847–51.
Rutter, H., et al. (2017). The need for a complex systems model of evidence for public health. *The Lancet*, 390, 2602–4.
Shahidullah, M. (1995). The sisterhood method of estimating maternal mortality: the Matlab experience. *Studies in Family Planning*, 26, 101–6.
Shiell, A., Hawe, P., and Gold, L. (2008). Complex interventions or complex systems? Implications for health economic evaluation. *British Medical Journal*, 336, 1281–3.
Skivington K, Matthews L, Simpson SA, et al. A new framework for developing and evaluating complex interventions: update of Medical Research Council guidance. BMJ 2021;374:n2061. doi:10.1136/bmj.n2061
Steckler, A. and Linnan, L. (2002). *Process Evaluation for Public Health Interventions and Research*. San Francisco, CA: Jossey-Bass.
Swinburn, B.A., et al. (2011). The global obesity pandemic: shaped by global drivers and local environments. *The Lancet*, 378, 804–14.
Tinner, L., et al. (2018). Examining subgroup effects by socioeconomic status of public health interventions targeting multiple risk behaviour in adolescence. *BMC Public Health*, 18, 1180.
United Kingdom Clinical Research Collaboration (2018). *UKCRC Public Health Research Centres of Excellence Final Report 2018*. [Online] Available at: http://ukcrc-org.stackstaging.com/wp-content/uploads/2019/06/UKCRC-report-with-infographic-2.pdf
United Nations General Assembly (2015). *Transforming Our World: The 2030 Agenda for Sustainable Development*. [Online] Available at: https://sustainabledevelopment.un.org/post2015/transformingourworld
W.K. Kellogg Foundation (2004). *Logic Model Development Guide*. Battle Creek, MI: W.K Kellogg Foundation.
Weiss, C.H. (1986). The circuitry of enlightenment: diffusion of social science research to policymakers. *Knowledge: Creation, Diffusion, Utilization*, 8, 274–81.
White, M. (2017). Evaluation of the health impacts of the UK Treasury Soft Drinks Industry Levy (SDIL). National Institute for Health Research Public Health Research Programme. [Online] Available at: https://europepmc.org/grantfinder/grantdetails?query=pi%3A%22White%2BM%22%2Bgid%3A%22PHR%2F16%2F130%2F01%22%2Bga%3A%22National%20Institute%20for%20Health%20Research%20%28Department%20of%20Health%29%22
White, M. (2018). Celebrating 10 years of funding some of the best public health research in the world. [Online] Available at: https://www.nihr.ac.uk/blog/celebrating-10-years-of-funding-some-of-the-best-public-health-research-in-the-world/11025
Wiggins, M., et al. (2008). *Young People's Development Programme Evaluation: Final Report*. London: Social Science Research Unit, Institute of Education, University of London.
Wiggins, M., Austerberry, H., Burchett, H., Strange, V., and Bonell, C. (2006). *Evaluation of the Young People's Development Programme Second Interim Report*. London: Social Institute of Education.
Wight, D. and Obasi, A. (2002). Unpacking the black box: the importance of process data to explain outcomes. In: Stephenson, J., Imrie, J., and Bonell, C. (eds.) *Effective Sexual Health Interventions: Issues in Experimental Evaluation*. Oxford: Oxford University Press.
Wight, D., Wimbush, E., Jepson, R., and Doi, L. (2016). Six steps in quality intervention development (6SQuID). *Journal of Epidemiology and Community Health*, 70, 520.

World Health Organization (WHO) (1986). Ottawa Charter for Health Promotion. First International Conference on Health Promotion Ottawa, 21 November 1986.
World Health Organization (WHO) (n.d.). *Types of Healthy Settings: Healthy Cities*. [Online] Available at: http://www.who.int/healthy_settings/types/cities/en/index.html

6.6

Economic appraisal in public healthcare: assessing efficiency and equity

David Parkin, Stephen Morris, and Nancy Devlin

What is economic appraisal?

Planning healthcare services to meet the health needs of a population must always recognize that availability of resources with which to meet them is a constraint on the achievement of its goals. There are many choices that must be made reflecting that constraint. For example, which public health and healthcare services should be provided, how they will be provided, and to whom they will be provided. Economic appraisal (EA) is a systematic way of analysing the consequences of such choices. It is one element of health technology assessment (HTA), the uses of which include among other things healthcare pricing and reimbursement decisions, planning the location of healthcare facilities, and setting of priorities for confronting health problems. EA is not the only element of HTA, but it is an important one.

Similarly, the discipline of economics is not the only contributor to EA. It also requires other public health relevant disciplines such as epidemiology, statistics, and operational research. Economics contributes the rationale for EA, an overall framework for analysis, and some measurement principles and methods.

The term ‘economic appraisal’ describes a set of techniques that weigh up the costs of an action, such as providing a particular type of healthcare to a patient or demographic group, against the benefits that it provides. The term ‘economic evaluation’ is also used. The distinction between these terms is that appraisal is undertaken before the action is taken, to help in deciding whether or not and how the action is to be done, and evaluation is undertaken after the action, to judge its effects. This distinction is often forgotten, and these terms are used interchangeably in much of health

economics. The expression ‘evaluation’ is now most commonly used in the health economics literature, but appraisal is used here because that more accurately describes the main purpose for which economic analysis in healthcare is applied in practice. The usual label given to this activity in economics is cost-benefit analysis (CBA). However, that term is often reserved in healthcare EA for a particular technique, which is explained later in this chapter.

The principles of EA can be applied to all kinds of health interventions, from individual patients’ treatments to population-based programmes, all of which may be of interest to public health practitioners. This chapter therefore does not restrict itself to activities conventionally labelled as public health, and also includes interventions that are not usually thought of as healthcare, such as health promotion. It highlights the special issues that EA faces in evaluating public health activities and finishes with a specific examination of those issues.

The economics of economic appraisal

EA should not be confused with *financial* appraisal. A financial appraisal is concerned with the monetary implications of a proposed action, such as investment in a healthcare facility or the provision to a population of one or more interventions. It examines cash expenditures and available funding to ensure that the actions taken are viable and provide a reasonable return. These cash flows are measured using accounting principles. EA shares some of the methods and principles of financial appraisal, but its aim is a more general assessment of costs and benefits and has measurement principles derived from economics concepts. The most important of these are *opportunity cost*, *social cost*, *marginal cost*, *efficiency*, and *equity*.

In practice, these theoretical principles are hard and occasionally impossible to adhere to, but they provide a guide as to the best sources of cost and benefit data and to the reliability of the numbers used.

Opportunity cost

As with any goods and services that are created within an economy, production of healthcare requires resources such as personnel (often referred to as labour), equipment and buildings (often referred to as capital), land,

and raw materials. There are finite amounts of such resources available to be converted into goods and services at any particular time, so it is not possible to produce everything that everyone in a population might want. Resources are therefore scarce relative to the demands for them. In the economy as a whole, there are not enough scarce resources to meet all of the wants that people have, so we have to choose which wants are met and which are not met. Similarly, in the healthcare system there are not enough healthcare resources to meet all of the health needs that people have, so we have to choose which needs are met and which are not met. Of course, it might be argued that if we cannot meet all health needs, resources should be diverted from other parts of the economy, and it is legitimate for public health practitioners to advocate that. But it would not be possible to do that indefinitely, and the consequences might be a reduction in education, welfare, and other goods and services in the public and private sectors of the economy that are not only desirable in themselves, but also contribute to maintaining and improving the health of the population.

Economics claims that scarcity and the resulting necessity to choose are ubiquitous and unavoidable. This leads directly to an important and fundamental economics concept, *opportunity cost*. If scarce resources are used to produce a good or service, those resources cannot be used to produce other goods or services. There is a lost opportunity to enjoy the benefits that those other goods or services would have provided. Because there are many possible uses for resources, the opportunity cost of using resources in a particular way is defined as the benefits that would have resulted from their *best* alternative use.

When economists refer to costs, they usually mean opportunity costs. This is quite different to the more familiar concept of financial costs, the costs of goods and services and of scarce resources in terms of money. Very often, financial costs are used to measure opportunity costs, but this is not always the case. Opportunity costs and financial costs are different ways of thinking about costs, rather than separate elements of overall costs; it would make no sense to calculate them separately and add them together, for example.

Measuring opportunity costs is often difficult in practice. It is preserved in good EA practice by the separation of costing services into measuring the quantity of resources required to provide that service and calculating the value of those resources, by, for example, examining the prices of each type of resources. This is also good practice for ensuring that research can be

generalized to other settings where quantities and prices may differ. For example, different countries may have very different ways of providing services, with different combinations of different kinds of personnel, facilities, and locations, as well as differing relative costs of those resources.

Social costs and benefits

A financial appraisal of a proposed action takes into account only costs that are incurred by the person or organization taking the action and benefits to the person or those who are the organizations' clients. These are referred to as private costs and benefits. However, in some kinds of EA it is in principle essential to take account of all costs and benefits to society, whatever they are, and whoever incurs them. In particular, it is essential to take account of *social costs and benefits* in a CBA that examines the impact of the action on the economy as a whole. For example, the total costs of a medical intervention would include not only costs incurred within the health sector, but also those incurred in subsequent personal care and support, whether by organizations or private carers. However, it is arguable that such wider costs should not be included for some of the EA methods discussed next, for example, in a cost-effectiveness analysis (CEA), in which the aim is to maximize healthcare outcomes within a limited budget.

A special example of the difficulties involved in this is accounting for *transfer payments*, examples of which are social security payments and other state benefits paid out of taxation. Except for the costs involved in administering them, these are not costs to society as a whole. Rather, they are transfers between different members of society, since the recipients derive benefits from them that match the costs to taxpayers. However, from the state's point of view such expenditure may be regarded as a use of tax revenues which has an opportunity cost in terms of other government expenditure.

Again, the principle of including all social costs and benefits is almost impossible to follow perfectly because of information limitations. However, it is preserved in good practice in health EA by identifying costs and benefits according to who bears them, in three categories: the health service, patients and their families, and the rest of society. It is also embodied in the recommendation that any analysis should make clear the perspective from

which the analysis is conducted, for example, society, the government, or the health sector.

Walker et al. (2019) have argued that adoption of an abstract 'social perspective' is not helpful in undertaking EA in practice and have developed a framework by which these issues can be considered in a systematic way. This is further discussed next in the context of EA of Public Health interventions, where it is of special importance.

Marginal costs and benefits

The total cost of any activity will depend on how much of it is carried out. For example, the total cost of a vaccination scheme will depend on the number of people to be vaccinated. If the total cost of the activity and the number of units of activity are known, it is possible to calculate an average cost per unit of activity, which is what accounting data usually provide. However, in some circumstances average costs may be a misleading estimate of the cost of a unit of activity, and what are required are *marginal costs*.

Marginal cost is the change in cost that is caused by the smallest possible change in another variable to which it is related. The marginal cost of a good or service is usually more specifically defined as the extra cost incurred in producing one more unit of it. Note that marginal does not mean small or unimportant; instead it means at the margins of an existing condition, for example, the cost that will be incurred by changing the allocation of resources slightly. That cost could actually be large, even though the change in the amount of the good or service is small. As an extreme example, suppose that the service is a screening programme and the screening centre processing the test material has reached full capacity. Screening an extra person could only be carried out if a new screening centre was built, so its marginal cost would be very high. By contrast, the marginal cost of the last test performed within the existing capacity may have been quite small, simply the cost of performing the test. This demonstrates another important point. Marginal cost may vary considerably with respect to the same size of change in the other variable, depending on the level of that other variable, in this case the number of people screened.

An example of the importance of looking at marginal costs is the impact of schemes designed to lower hospital inpatient surgical costs by reducing length of stay through earlier discharge. Hospitals may have information on

the average cost of an inpatient stay, which can be used to calculate an average cost per day. However, the costs of an inpatient day are not constant. In particular, they may be much smaller than the average towards the end of the stay, because the average includes a share of the costs of treatment and perhaps of high dependency care that took place towards the beginning of an inpatient stay. So, reducing the number of low dependency days at the end of the stay will save far fewer resources than would be expected by examining the average. Marginal costs calculated with respect to an increase or decrease in the number of days would give a correct estimate of the likely savings.

The same principles apply to benefits. Marginal benefits are the extra benefits obtained by consuming one more unit of a good or service. An example of the importance of looking at marginal rather than average benefits is a screening programme that can be carried out with different numbers of sequential tests. The more tests, the more cases are detected. A programme that used two sequential tests might yield 11 cases per 10,000 people tested, which could look reasonable. But suppose that a programme that used only one test would discover ten cases. The marginal benefit of including the second test is only one case per 10,000, which does not look so good.

To give a real example, Torgerson and Spencer (1996) illustrated the importance of marginal analysis by analysing data on the cost-effectiveness of biochemical screening for Down's syndrome as a replacement for screening by maternal age. An estimate had been published that introducing biochemical screening in a population of 20,000 pregnant women would cost £413,501 and might prevent 11 affected births, giving a cost-effectiveness ratio of £37,591 per affected foetus identified. Torgerson and Spencer pointed out that for the same number of women, existing antenatal screening procedures based on maternal age and additional tests already cost £79,500 with four births potentially prevented. So, the true cost-effectiveness of biochemical screening was (£413,500—£79,000)/(11–4) = £334,500/7 = £47,786, a much higher cost per affected foetus identified than for the estimate based on average cost.

Efficiency and equity

Economic analysis is usually employed to judge the way in which scarce resources are employed according to two main criteria: efficiency and equity. These two concepts have technical definitions, which will be described in the following subsections, but in very broad terms efficiency refers to obtaining the greatest output for a given set of resources and equity refers to a fair distribution of that output among the population. The term 'output' is very general and in our context we usually mean the additional amount of healthcare or health produced.

Efficiency

It is necessary to give a warning about definitions of efficiency. Economists, as specialists in the analysis of efficiency, largely agree about what it means and how different types of efficiency may be defined, but the labels they give to those types vary. The same concept may be given to different names and the same name given to different concepts. Here we use the labels and definitions given in Morris et al. (2012).

Knapp (1984) gave a very broad definition of efficiency as 'the allocation of scarce resources that maximizes the achievement of aims'. This is a useful start, because it suggests the benign nature of the desire to achieve efficiency. If we have scarce resources and competing uses for them, our aim will be to obtain the best set of uses, with 'best' defined in whatever way we want. If we decide that the aim of the health system is to improve the health of the population and we have a fixed healthcare budget, we will obtain the biggest health gain if the allocation of scarce healthcare resources is efficient.

For practical purposes, we require more precise and technical definitions of efficiency. However, it is useful first to understand a slightly abstract idea, called *Pareto efficiency* which is also, though not consistently, called *allocative efficiency*. This arises from an attempt to create a widely acceptable criterion for judging different allocations of resources to different ends. The suggestion is that we would be able to say that one allocation is better than another if at least one person is better off under the first allocation and no one is worse off. This is called the *Pareto criterion*. Consider a change from one allocation of resources to another, for example making changes to the kind of care that is made available in the healthcare system. If some people benefit and no one is made worse off, this is described as a *Pareto improvement*. If it is not possible to make any such

changes, then we have achieved a *Pareto optimum*. Another way to view the Pareto optimum is that it is a position where it is not possible to make anyone better off without making someone else worse off.

If we aim to make people as well off as possible and are not concerned about whether some people are better off than others, then a Pareto optimum is efficient. We cannot further improve the achievement of our aim because even if we can make someone, or even many people, better off, we do not know if this is outweighed by those who are made worse off. This is not a unique position; there are many allocations of resources that would be Pareto optimal, some of which would imply great inequalities between different people. If our aims also took account of the undesirability of inequalities, then we might not view all Pareto optimums as efficient. Nevertheless, ideas of efficiency in economics derive from this concept of Pareto efficiency. Here, we define three types: *technical*, *economic*, and *social*.

Technical efficiency is a concept that is used in considering how health and healthcare, or 'outputs', are produced from scarce resources, or 'inputs'. There is technically efficient production if the most output possible is being produced from a given amount of inputs or the fewest possible inputs are used to produce a given amount of output. For example, the number of people that can be vaccinated against influenza in a general practice clinic depends on the number of nursing staff that are available and other inputs. If the most that can be provided by three practice nurses is 100 vaccinations each day, then it is technically inefficient if 99 vaccinations are provided by three nurses or 100 vaccinations by more than three nurses.

Another way of viewing technical efficiency, which is consistent with Pareto efficiency, is that the practice cannot vaccinate more people without employing at least one more nurse. More generally, production is technically efficient for a given set of inputs if it is only possible to produce more by employing more of at least one input.

In practice, EA does not directly address technical efficiency, though as will be explained it is an essential underlying assumption in most cases.

Economic efficiency is given different labels, including cost-effectiveness, though that term should be used carefully, as will be explained next. Technical efficiency is defined with respect to the physical numbers of inputs, but economic efficiency is interested in the cost of those inputs. Economic efficiency is achieved if we are producing most output for a given cost or producing a set amount of output at the lowest possible cost. Using

the earlier example, suppose that people could be vaccinated by different grades of nurses, but at least one senior nurse must be part of the team. It might be equally technically efficient for 100 people to be vaccinated each day by employing one senior and two junior nurses, one junior and two senior nurses, or three senior nurses. But senior nurses are more expensive to employ than junior nurses, so it will be economically efficient to use the first of these staff mixes. This demonstrates that although achieving technical efficiency is necessary to achieve economic efficiency, not all technically efficient ways of producing are economically efficient.

Another way of viewing economic efficiency, which is consistent with Pareto efficiency, is that given the costs of employing nursing staff, the practice cannot undertake more treatments without them costing more to provide. More generally, production is economically efficient for a given set of input prices if it is only possible to produce more by incurring greater costs.

In practice, economic efficiency is the dominant type of efficiency analysed in EA. However, there are different ways in which this is done, as will be explained next.

Social efficiency is a much broader concept than technical and economic efficiency. If there is economic efficiency in production in every part of the economy, there is allocative efficiency in production for the economy as a whole. There is an equivalent concept relating to the consumption of goods and services called allocative efficiency in consumption where, given prices of goods and services, consumers maximize the benefit that they derive from those goods and services. If both of these are achieved, then there is allocative efficiency in the economy as a whole, which is also known as social efficiency. This is the same as the Pareto efficiency described earlier.

Obviously, social efficiency according to this strict definition is not a concept likely to be of practical use in health economics. However, it is important because it underlies CBA, whose aim in theory is to move the economy closer to social efficiency.

Equity

Equity is an important criterion for allocation of resources, and it is particularly important in health systems because it is observable that people attach more importance to fairness in health and healthcare than to the distribution of many other goods and services. Almost every healthcare

system in the world has equity as an important policy objective. However, economic analysis of equity is less clear than that of efficiency, except for how equity might be measured, and there is less agreement among economists about it.

Equity means fairness, and in the context of healthcare this means fairness in the distribution of health and healthcare between people and in the burden of financing healthcare. It is not a synonym for equality, which means an equal distribution. It may not always be fair to be equal, for example, it might be thought to be unfair if both healthy and sick people are given the same amount of healthcare. But equality is often used to define equity, as, for example, in the principles of horizontal equity—equal treatment of those who are equal—and vertical equity—unequal treatment of those who are not equal. An example often used in healthcare is the proposal that there should be equal access to healthcare for people with the same level of healthcare needs.

That is not the only possible definition of healthcare equity, and more generally there is no single definition of what equity is. Moreover, ideas about equity are not mainly founded on economics principles, rather on philosophy, ethics, law, and other forms of enquiry. An example is the equity principle of procedural justice, which in this context might mean that the process used to decide on the allocation of healthcare resources between people should be fair. This is an important principle in deciding how HTA is applied, but it is not especially based in economics. Culyer and Bombard (2012) have suggested that one way to incorporate the plurality of approaches to equity in HTA is to use a ‘checklist’ to ensure that decision-makers take all aspects into account.

Cookson et al. (2017) have developed a framework for examining both equity and efficiency, in terms of cost-effectiveness, in a systematic and consistent way. It uses two analytical concepts, *health equity trade-offs* and *net equity impacts*. It uses these to describe two approaches to conducting cost-effectiveness analyses that incorporate equity concerns, *equity impact analysis*, and *equity trade-off analysis*, which are described next.

Health equity trade-offs result from the possibility that efficiency and equity may conflict under some circumstances. For example, a screening programme may be cost-effective in terms of the cost of detecting treatable cases, leading to an overall increase in health for the population as a whole, but if take-up is lower among groups who have worse overall health, this

will lead to a widening of health inequalities. There are therefore four possibilities, as in Table 6.6.1:

Table 6.6.1 Health equity trade-offs

		Net health impact	
Positive	**Negative**		
Health equity impact	Positive	1 Cost-effective Improves equity	4 Cost-ineffective Improves equity
	Negative	2 Cost-effective Harms equity	3 Cost-ineffective Harms equity

The suggestion is that healthcare that falls into box 1 should be pursued and those that fall into box 3 should be avoided. There is however a trade-off for healthcare in boxes 2 and 4 that should be explicitly considered by decision-makers.

Net equity impacts are based on the fact that the resources used in healthcare have opportunity costs. A healthcare intervention will generate a distribution of benefits between different groups and individuals within the population. However, alternative uses of the resources used by the intervention will also have benefits that are distributed within the population. Both of these must be considered in giving a full account of the impact of the intervention on equity.

It is necessary to consider what it is that we wish to be equitable about. Three areas in which equity may be considered are the finance of healthcare, the distribution of healthcare, and the distribution of health.

Analysis of equity in the finance of healthcare mainly concentrates on vertical equity, with a particular emphasis on whether or not healthcare is financed according to people's ability to pay. Horizontal equity in financing considers whether or not people, who have the same income, and therefore the same ability to pay for healthcare, make the same payments.

Horizontal equity in the distribution of healthcare mainly examines whether or not people with the same need for healthcare make the same use

of healthcare services. Vertical equity in the distribution of healthcare is usually interpreted to mean whether individuals with different levels of ill-health have different levels of use that are appropriate to that difference.

Equity in the distribution of health is almost always expressed in terms of inequalities in health. Health inequalities, particularly those that demonstrate that health levels vary systematically and inversely with socioeconomic status, are always of some importance in health policy debates and a major concern of some governments, depending on their political preferences. It may be argued that this is in fact the only real equity concern, since a concern for equity in healthcare derives solely from a concern about the distribution of health.

Formulating an economic appraisal

One way to look at how an appraisal is carried out is to consider it as a process with stages, as follows:

Defining the problem to be analysed. The problem is always the starting point of an appraisal. This may seem obvious, but in fact it is often solutions that are, wrongly, regarded as the starting point. From an appraisal point of view, the problem is not whether a fluoridation programme is cost-effective, but what the most cost-effective way is to deal with the problem of dental caries.

Specifying the alternatives. These are different ways of dealing with the problem defined in the first stage. It is important to select the most relevant alternative whatever it is, for example, promotion of tooth-friendly nutrition among children, or doing nothing.

Assessing costs and benefits which has three elements:

- Enumeration of costs and benefits, which means the drawing up of a descriptive list of the costs and benefits that are to be included in the appraisal.
- Measurement, which means obtaining data to describe the levels of costs and benefits for the different alternatives.
- Valuation, which is required to convert the data into values. For example, data on resource use should be converted into costs by applying to those data the value of the resources.

Calculating results means combining the data on costs and benefits into the results that will be presented. The exact nature of this will depend on the type of appraisal that is carried out and how uncertainty over the results is dealt with.

Deciding. The endpoint is when a decision is taken on the basis of the appraisal. This will not be taken by an appraiser, but by someone who has the responsibility to decide. However, appraisers often do recommend a decision which is based on their own appraisal or imply such a recommendation. This is legitimate if the relevant decision rules are known. For example, there might be an accepted decision rule that states that if the benefits of a health intervention exceed its cost then it should be provided to the population. A finding, for a particular intervention, that its benefits did exceed its costs may then lead to a recommendation that it should be provided.

One important factor in this is that an appraisal should be formulated and carried out with full awareness of which viewpoint is being taken. This arises because different actors in the health system—patients, health

professionals, and government, for example—may have different interests and concerns, so that what is efficient from one point of view may not be from another. This is important because it determines which type of appraisal is to be used, what constitutes a benefit and a cost, and how these are to be valued. Most appraisals will take the viewpoint of the health service, since that embodies most of those who have the legitimacy to decide which healthcare is to be provided. However, they may also take the viewpoint of society as a whole, since that defines the recipients of healthcare as well as the ultimate funders of it. Whatever viewpoint is adopted, researchers should be explicit about which it is. In EAs of public health interventions, a broader perspective is frequently taken, to account for costs and benefits that occur outside the health system.

Types of economic appraisal

Although there are several different types of EA, we will concentrate here on those most likely to be encountered in healthcare and other public health activities. Unfortunately, there are different ways in which these types can be classified. This leads to occasional disputes about how results of a particular type of appraisal should be interpreted or used. One way is to classify them according to what kind of efficiency they analyse. A second way is to classify them according to which costs and benefits are measured and how this is done. A third way is to classify them according to the type of decision that they apply to. Although we will discuss all three types here, the most widespread and popular classification in health economics is given by Drummond et al. (2005), which uses both the measurement and decision-type principles.

In what follows, we refer to what is being appraised as alternatives, reflecting the use of EA in healthcare in decision-making about alternative ways of using healthcare resources. This includes many kinds of decisions, including different ways of delivering healthcare, different types of healthcare, and different treatment options.

All of the types of EA discussed here are concerned only with the appraisal of efficiency. Variants of these that also assess equity are discussed in a separate section next.

Cost-benefit analysis

In economics more generally, CBA starts with an inventory of all of the costs and benefits of each of the alternatives, whatever they are, and whoever incurs them. This can be regarded as a balance sheet in which costs and benefits are weighed up against each other.

However, in healthcare appraisals CBA is usually defined, following Drummond et al. (2005), as a technique in which all costs and benefits are measured in terms of money. The rationale for this is that it is only possible to weigh up all of the costs and benefits if they are measured in the same unit. Although in principle any common unit could be used, in practice money is the obvious and natural choice, as it is the measure of value most used in modern economies. The result of such a CBA would be to establish which of the alternatives has the greatest net benefit, the difference between benefits and costs, which could of course be negative.

The main decision rule for CBA is that an activity should be undertaken if the sum of the benefits is greater than the sum of the costs or, identically, if the net benefit is positive. If only one activity with a positive net benefit can be undertaken (because, for example, there are limited funds), then the rule is to choose the activity with the highest net benefit.

CBA is based on economics theory, because if all of the costs and benefits are measured in the correct way—for example, all costs measure their true opportunity costs—and an alternative has a net benefit, it will lead to a Pareto improvement. The way that this has been interpreted by Drummond et al. (2005) is that CBA is therefore appropriate for answering questions about whether or not a healthcare programme should be implemented, or a treatment should be used, rather than which of a number of alternative programmes or interventions is the most efficient.

Because CBA requires benefits to be measured in financial or monetary terms and to cover a very wide range of costs and benefits, it is rarely used. An example in public health is Wang et al. (2005), who estimated a limited range of costs and benefits of using bike and pedestrian trails to reduce healthcare costs arising from inactivity. They calculated that the annual cost per person of using the trails was US$209.28 and the resulting reduction in medical costs was $564.41. This gives a net benefit of $355.13 per person.

Investment appraisal

Investment appraisal in the context of healthcare can be explained using the example of a specific technique called by the United Kingdom's finance ministry (the Treasury) option appraisal. This is described in the guidance that it gives to public bodies about how they should appraise and evaluate projects that are to be paid for from public funds; the distinction between appraisal and evaluation is carefully kept to in this case. It is a process in which different options for meeting an objective, defined by the aim of meeting some public need, are generated; CBA is applied to these options; and the best solution for meeting the aims is chosen on the basis of the results. For example, an appraisal of a project for setting up a new national programme for measuring obesity among schoolchildren would not start as a comparison of the new programme with the existing circumstances. It would start with the need to understand the extent of obesity among schoolchildren, generate different options for dealing with that need, which might include a new measurement programme, examine the costs and benefits of each alternative, and from this derive the economic case for the new programme, if it was found to be the best alternative. The economic case for the project would simply be part of a more general project appraisal including affordability and achievability and various types of impact assessment including on health, the environment, and health and safety.

Cost-consequences analysis

Cost-consequences analysis (CCA) is a form of CBA which does not try to put all of the costs and benefits into the same units. In particular, it accepts that there are different types of benefits that cannot be measured in the same units. This distinguishes it from CEA, which is discussed in the following subsection. The assumption is that in making decisions based on a CCA, different decision-makers will place their own weights on the different benefits and on costs, implicitly if not explicitly. CCA is of particular interest in public health because it is argued that it is more suitable for population health issues involving a wide range of agencies and of differing and incommensurable outcome indictors. For example, the National Institute for Health and Care Excellence (NICE) in England permits CCA to be used for public health interventions, unlike other healthcare. CCA is often referred to as a disaggregated approach, because the benefits and costs are

not combined into a single indicator such as net benefit or a cost-effectiveness ratio, which are defined next.

Cost-effectiveness analysis

Cost-effectiveness derives from the analysis of economic efficiency, where one alternative is preferred to another if it provides greater benefit at the same or lower cost, or lower cost for the same or greater benefit. This definition leaves open the question of which of two alternatives is more efficient if one provides greater benefit than the other, but at a lower cost. However, under certain conditions, for example, that the alternatives can be reduced or increased in size to produce any level of total cost or benefit, such comparisons can be made. A cost-effectiveness ratio (CER), defined as costs divided by benefits, can be calculated in order to do this. The CER most often used in health economics is called an incremental CER, or ICER, where the costs and benefits of each alternative are calculated compared with their next best alternative, rather than with a common alternative.

The measurement principle for CEA is that costs are measured in terms of money, but that benefits are measured in units other than money. However, unlike CCA, all of the benefits are measured in the same units, usually because only one type of benefit is considered. The obvious consequence of this is that costs and benefits are not weighed against each other. The ICER simply shows the cost of obtaining a unit of benefit. Whether that cost is worth incurring is a different question. So, unlike CBA, which tells us whether or not health programmes or treatment are an efficient use of resources, CEA tells us which of the possible ways of providing those is the most efficient.

Drummond et al. (2005) restrict the use of the term CEA in two ways. First, they regard CEA as being applicable only when both the costs and benefits differ between alternatives. The case where benefits are the same but costs differ is called cost-minimization analysis (CMA). In practice, this term is rarely used to describe analyses that have been carried out, although such analyses may well be quite widespread. Secondly, they restrict the measurement of benefits to be in what they call ‘natural’ units, such as numbers of cases detected, changes in clinical measures like blood pressure, or changes in undesirable biological markers. This contrasts with the measurement of benefits in terms of changes in health-related quality of life,

which is often given the label cost-utility analysis (CUA), discussed in the following subsection.

The first decision rule for this kind of CEA is that we should reject any alternatives that are dominated by another alternative or combination of alternatives. This is economic efficiency in the strict sense described earlier, where the dominated alternative has a greater cost with no greater benefits or lower benefits with the same or greater costs. The choice between non-dominated alternatives is more complex. Where only one alternative can be chosen, that with the lowest ICER should be chosen, but only if it is below a ceiling ratio, which is a level of the ICER which any alternative must meet if it is to be regarded as cost-effective. Where combinations of more than one alternative can be used, it is in principle necessary to calculate the ICERs for every possible combination to decide which is most efficient and if any meets the ceiling ratio requirement.

These issues are often illustrated using a cost-effectiveness plane diagram (Black 1990) as in Figure 6.6.1. ICERs are presented graphically as a combination of the costs and the effects of a health intervention, described in Figure 6.6.1 as a treatment, compared to some alternative. Costs are conventionally placed on the north-south axis and effects on the east-west axis. In both cases, these effects can be negative, zero, or positive.

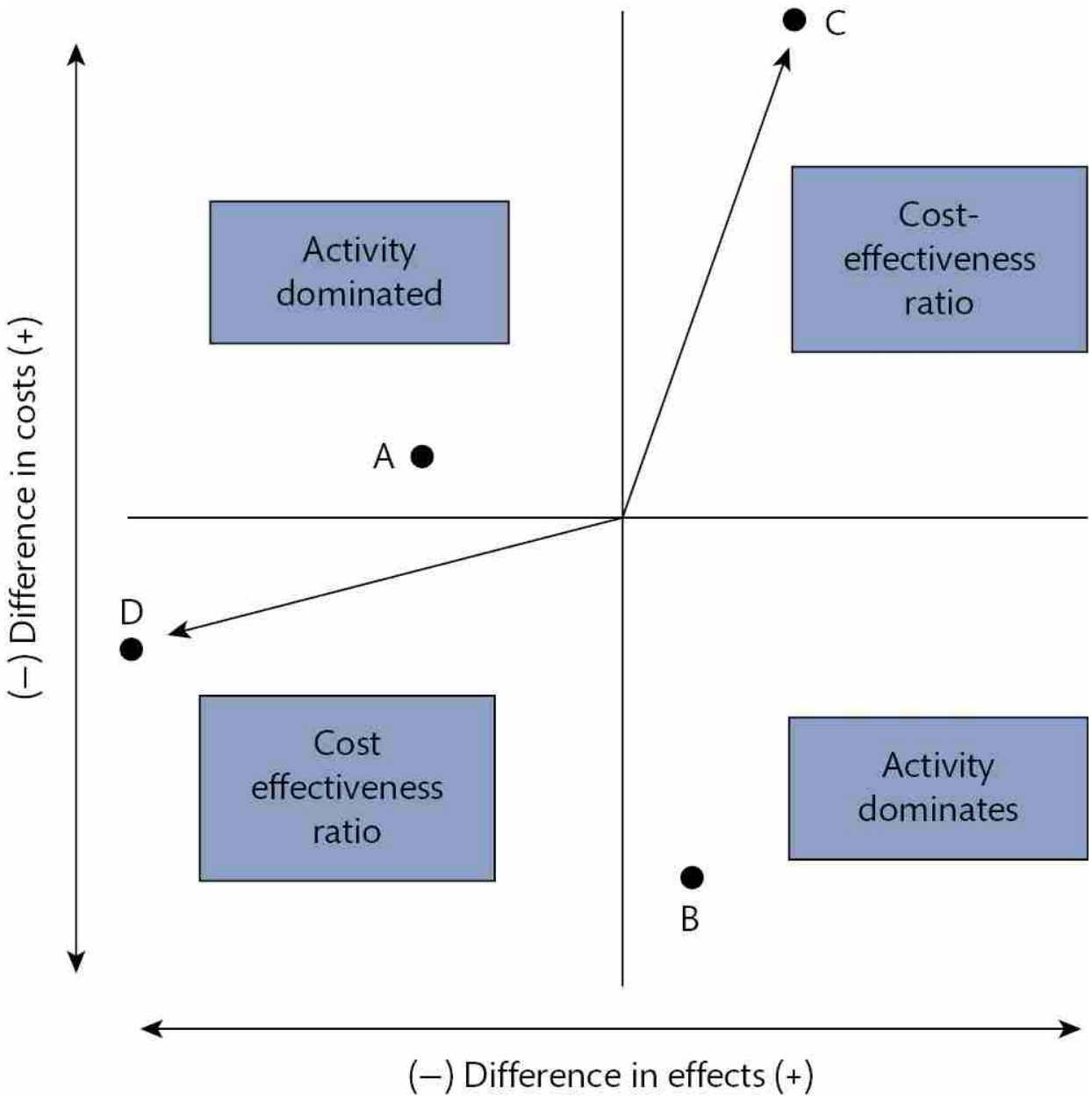


Fig. 6.6.1 The cost-effectiveness plane. An intervention can be placed anywhere on this diagram according to its incremental costs and benefits. If it lies in the northwest quadrant, such as point A, the costs of the intervention are higher than the alternative, and its benefits are lower. It is therefore unambiguously worse and is described as being dominated by the alternative. Similarly, in the southeast quadrant, at a point such as B, costs are lower, and benefits are higher, so the treatment dominates its alternative. In the northeast quadrant, at a point such as C, higher benefits are gained at a net cost over the alternative. So, we can calculate an ICER, the cost per unit of effect gained, measured as the slope of the line from the origin to the point. In the southwest quadrant, at a point such as D, lower costs are possible, but at the expense of lower benefits. Again, we can calculate an ICER, although this now refers to a cost-saving per unit of effect lost, which is again measured as the slope of the line from the origin to the point.

The ceiling ratio can also be demonstrated using a cost-effectiveness plane diagram, where it is often referred to as demonstrating *cost-effectiveness acceptability*, as in Figure 6.6.2.

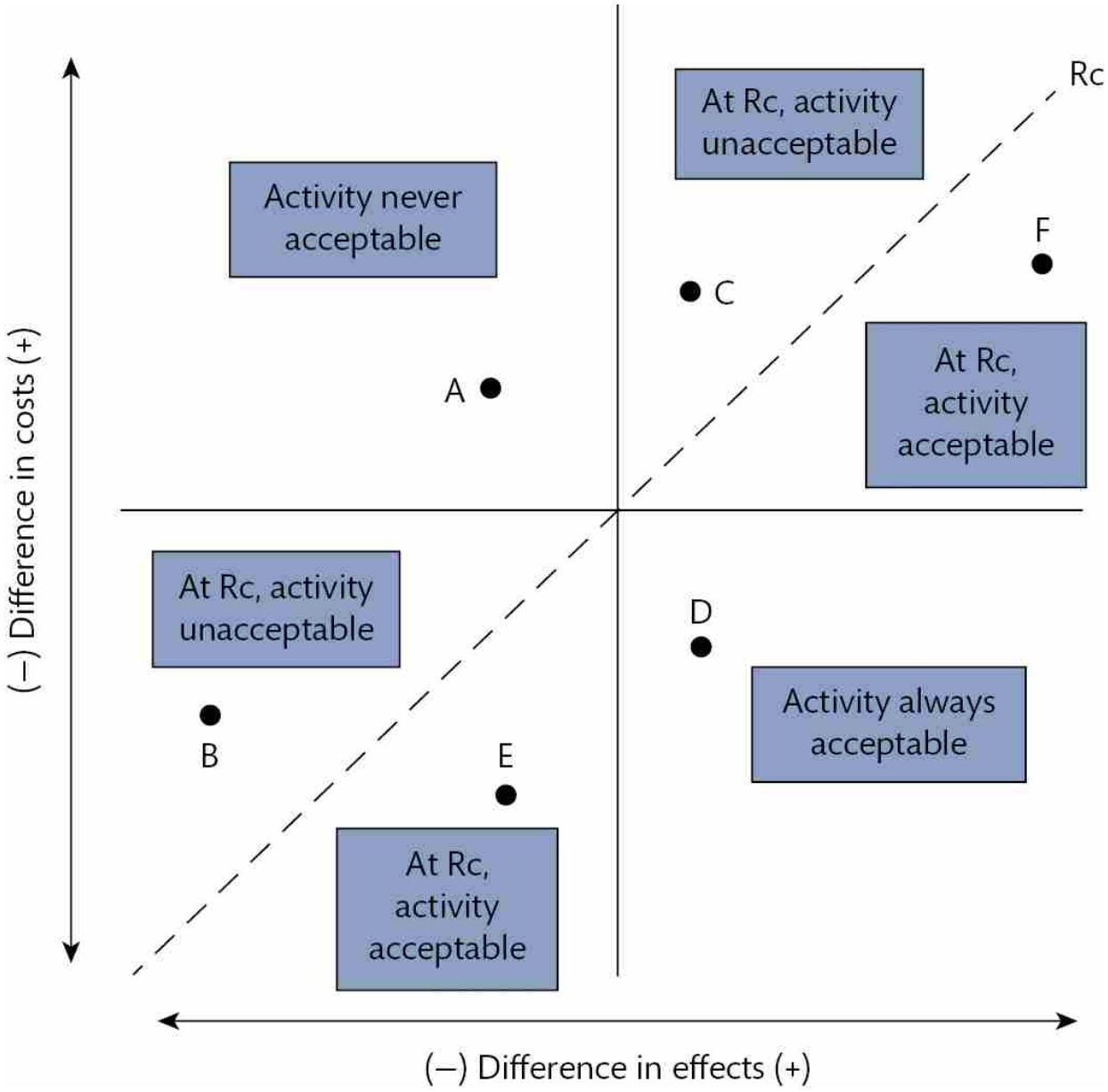


Fig. 6.6.2 Cost-effectiveness acceptability. In this diagram, the dotted diagonal line marked Rc represents the ceiling ratio. If an intervention lies above the line, it will not be acceptable on cost-effectiveness grounds. This is either because it is dominated by the alternative, whatever the ceiling ratio is, as in point A, or its ICER does not satisfy the actual ceiling ratio, as in points B and C. It should be noted that in the south-west quadrant, this means that the ICER is below the ceiling ratio, as in point B, while in the north-east quadrant it is above, as in point C. Below the line it will be acceptable. This is either because it dominates the alternative, as in point D, or its ICER satisfies the ceiling ratio, as in points E and F. In this case, the ICER is above the ceiling ratio in the south-west quadrant, as in point E, or below it in the north-east quadrant, as in point F.

Cost-effectiveness acceptability is also important because it is one way in which uncertainty in EA is dealt with, which is explained later in this

chapter.

Cost-utility analysis

CUA is a term often used in the United Kingdom and elsewhere, though rarely in the United States, to refer to a special form of CEA in which health benefits are measured in terms of quality-adjusted life years (QALYs). QALYs are a composite measure of gains in life expectancy and health-related quality of life, described in more detail next, along with disability-adjusted life years (DALYs). The distinctive outcome of a CUA is the calculation of an ICER in terms of the extra cost per QALY gained (CQG), which may also be thought of as the price of obtaining one extra QALY using a particular healthcare intervention or programme. An example is described in Box 6.6.1.

The rationale for CUA is more complex than that of CEA more generally. QALYs in themselves are regarded by many as a better outcome indicator than the 'natural' units of CEA when appraising therapeutic alternatives for the same condition. This is because they deal with efficiency in the production of health itself rather than simply of healthcare. But CUA also offers something that CEA more generally cannot, which is the possibility of comparing across treatments for different conditions. In principle, it is possible to compare interventions for, say, eating more healthily to reduce weight with, say, reducing alcohol consumption, to determine which is the most efficient at producing health gain in the form of QALYs. Of course, this is controversial. For a CEA restricted to different treatments for a particular condition, it can be assumed that the patients receiving the treatments are the same people, but this will not be true of comparisons across conditions.

Box 6.6.1 A cost-utility analysis of vaccinating schoolgirls against HPV infection

Anonychuk et al. (2009) investigated the cost-effectiveness of vaccinating 12-year-old girls against human papillomavirus (HPV) types 16 and 18 to prevent cervical cancer. The vaccine also cross-protects against other HPV types, and herd immunity effects, the consequent reduction in the risk of infection to non-vaccinated people, determine its overall effectiveness. For each 100,000 girls vaccinated, between 390 and 633 cases of cervical cancer and 168 and 175 deaths would be

prevented, depending on whether or not herd immunity and cross-protection effects are included. This would generate QALY gains of 0.0083 to 0.0095 per person over the group's lifetime. There would be an increase in cost of Can$256 to Can$263 per person. For the base case of no herd immunity or cross-protection effects, Table 6.6.2 shows the costs and effects with and without vaccination.

This gives an extra cost per QALY gained (CQG) of Can$263/0.0083 = Can$31,687. With cross-protection, the CQG was Can$26,947; with herd immunity Can$27,849; and with both Can$18,672. The authors concluded that the vaccine reduced cancer cases and mortality and was cost-effective, though that conclusion does depend on an assumption about the size of the ceiling ratio.

Source: data from Anonychuk et al., A cost-utility analysis of cervical cancer vaccination in preadolescent Canadian females, *BMC Public Health*, Volume 9, Issue 1, pp. 401, Copyright © 2009 Anonychuk et al. licensee BioMed Central Ltd. Reproduced under the Creative Commons Attribution License 2.0.

CUA has, in principle, the ability to compare across all health interventions. A contentious implication of this is that it can be used to allocate healthcare resources between different health programmes, and therefore be used to determine healthcare priorities. That would overcome the principal problem that CEA has compared with CBA, because it would indicate whether or not the healthcare benefits from a particular type of healthcare should be realized, rather than simply which is the best way of achieving those benefits. Some people therefore regard CUA as a limited form of CBA rather than a special form of CEA. Viewed in this way, CUA is better than CBA, because it avoids having to measure benefits in terms of money.

To explain this further, recall that a decision rule for CEA is that the ICER of the best alternative is compared to a ceiling ratio. In the case of CUA the ICER is the CQG, so the ceiling ratio is the amount that we think it is reasonable to pay to gain a QALY. This is sometimes referred to as a CQG threshold, because it is the dividing line between healthcare that is regarded as cost-effective and that which is not. NICE has a threshold, which it regards as deriving from the limited budget that the National Health Service has, which determines what can be afforded. This is known as the shadow price of a QALY. Alternatively, it could be set according to what the population is willing to pay for health gain. This is known as the social value of a QALY. But either way, if this is known, it can in principle be used as a means of converting costs to QALYs or vice versa. This puts costs and benefits into the same units, enabling the calculation of net benefits in the same way as for CBA.

Net benefits can be calculated in terms of either money or QALYs. Suppose that an intervention has incremental costs of £30,000 and incremental benefits of 2 QALYs. The ceiling ratio can be regarded as the price that it is acceptable to pay for a QALY, and therefore represents the money value of a QALY. Suppose that the ceiling ratio is £20,000 per QALY gained. If we decide to analyse cost-effectiveness in monetary terms, we first multiply the QALY gain by the ceiling ratio to give the money value of the QALY gain, in this case 2 × £20,000 = £40,000. The net cost is then subtracted from this to give a net monetary benefit of £40,000—£30,000 = £10,000. Alternatively, we could convert the costs to their QALY gain equivalent by dividing the ceiling ratio by them. This would give a net cost equivalent of £30,000/£20,000 = 1.5 QALYs, and an overall net health benefit of 2—1.5 = 0.5 QALYs. Because these are positive—and they must always have the same sign—the intervention is deemed cost-effective.

Of course, this is exactly the same result as if we had calculated the ICER to be £30,000/2 = £15,000 and observed that it is below the ceiling ratio of £20,000. However, this net benefit approach does have an advantage if we do not know the ceiling ratio, since we can calculate net benefits for any value of the ceiling ratio.

Measuring costs

As suggested earlier in the chapter (see ‘Formulating an economic appraisal’), the process of costing involves three steps: identifying and describing resource use changes; quantifying them in physical units; and valuing them.

In general, there are two types of costing: macro- or top-down costing and micro- or bottom-up costing. These are distinguished by the level of disaggregation at which individual resources are measured and valued. In its pure form, macro-costing starts with the total costs incurred by, for example, a hospital, and calculates the costs of, for example a specialty, by allocating a proportion of the total costs to it using some indicator of the costliness of that specialty compared with others. For example, if a specialty had 20% of the hospital’s cases, 20% of the costs would be allocated to it. At the extreme, micro-costing would examine every item of service and every consumable used by every patient. Each item of service would be costed by examining all of the time spent on it by every member of staff as well as the

time spent using every piece of equipment. In practice, these extremes are not found. Moreover, it is common for mixed methods to be used for different elements of cost.

An important consideration is when resources are used, as not all costs are incurred at the same time. This has two important elements. First, if costs occur in different years, then they should be adjusted for any inflation or deflation that has occurred. This is to make the figures comparable over different years and to ensure that it is the value of resources that is considered, which may be unaffected by the rate of inflation. The best practice is to select one year's unit costs and apply them to different years' resource use. But that is rarely possible for many items of cost, so the procedure is to deflate yearly costs by a cost index. This also means that if there are any cost projections, these should not take account of projected inflation.

Secondly, although adjusting for inflation means that each year's costs are measured in the same units of value for each year, this does not mean that every year's costs are of the same value when viewed from a particular year. In particular, people in general prefer to postpone costs if they can, and therefore a cost now has a higher weight than exactly the same money cost incurred in the future. Looked at another way, costs in the future have a lower value now. This is dealt with by discounting, which is a time weight applied to costs, which diminishes the further in the future the cost is. In practice, what is done is to apply a discount rate, which can be viewed as the inverse of an interest rate. Applying an interest rate to an investment means that the value grows over time even if it is fixed. Applying a discount rate means that the value falls over time. In an EA, if costs are summed over time, then discounting all future costs back to the present gives a present value.

The issue of timing may be very important in healthcare, because of the long-lasting effects of many interventions. (Benefits in the future may also have a lower value now, requiring the same approach; this is discussed later in the chapter, see 'Measuring benefit in terms of health gain'). Of particular public health interest is the problem that prevention incurs costs now, and even if it reduces costs in the future by a greater amount, there may not be a net cost-saving if the future costs are discounted. Even if there is a net cost-saving, it will have a lower value than if discounting was not used. In the extreme, measures to deal with maternal nutrition to reduce illness in the

later adult life of their babies, as suggested by the Barker hypothesis (Barker 1992), might see cost reductions so many years in the future that discounting would reduce them to a negligible value.

There is some dispute about exactly what the discount rate to be used should be, and how it should be obtained. There are two competing theories to guide this. One is the social opportunity cost approach, which assumes that public and private investments compete for resources, so the public sector should use market rates of borrowing. The other is the social rate of time preference, which measures what people are willing to receive in compensation for delaying consumption from one year to the next.

Measuring benefits

Measurement of benefits in terms of 'natural' units in a CEA does not require any special economic analysis, and the measurement of diverse benefits within a CCA is too diverse to be dealt with here. This section therefore concentrates on the measurement of health gain for use in a CUA, and the measurement of monetary gain for use in a CBA.

However, a third technique, which does not fit neatly into these two categories, is becoming increasingly important, the Discrete Choice Experiment (DCE). In this, respondents are offered a choice between different interventions, which are described according to attributes that may include health outcomes and price. The choices made reveal the relative values of different attributes. This might be used to ascribe a value to different interventions, or to calculate values in terms of money or to obtain values for different health states. The last of these is described next.

Measuring benefit in terms of health gain

There are many different ways in which health and improvements in health can be measured. Different concepts and measurement techniques are appropriate for different uses. The approaches that are adopted for use in EA are largely determined by the requirements that their use for appraisal purposes dictates.

First, appraisal is attempting to obtain an unambiguous measure of benefit. This implies that the measure of health should be a single number representing all relevant aspects. Secondly, because appraisal looks at the

use of scarce resources that could be devoted to all sorts of healthcare, the measure of health should be capable of comparing those different uses of resources. This implies that a generic measure of health should be used. Thirdly, appraisal compares costs, which represent the value of resources used with benefits, which implies that the measure of health should also be capable of being interpreted as a value.

One measure that meets these requirements is the QALY. This combines length of life and quality of life into a single indicator. Depending on the type of health gain, QALYs can be thought of as a quality adjustment to years of life gained or the length of time for which quality of life is improved or both. It is important to remember that the benefit of a health intervention is the gain in QALYs that it produces. A similar measure is the DALY. This chapter does not describe DALYs fully; such details can be found in Chapter 5.17, which compares the two measures. Essentially, however, the DALY is a population-based measure that combines life expectancy in a population with the prevalence of disability to generate an estimate of a 'health gap'.

Figure 6.6.3 illustrates how in principle QALYs are calculated, for a highly stylized description of prognoses for a disease with and without treatment. For example, suppose that with a particular disease, the prognosis is that a person will live for a further 10 years. During the first two of those, their health-related quality of life (HRQOL) will deteriorate at a constant rate from 80% of full health to 60% and during the last eight will deteriorate more slowly to 30% by the time of their death. This is represented in the diagram by the solid line, with HRQOL values of 0.8 at time t, 0.6 at (t + 2) and 0.3 at (t + 10). With treatment, life expectancy will rise to 12 years, in each of which the patient will enjoy full health. This is shown by the dashed line with HRQOL = 1.0 from time t to (t + 12).

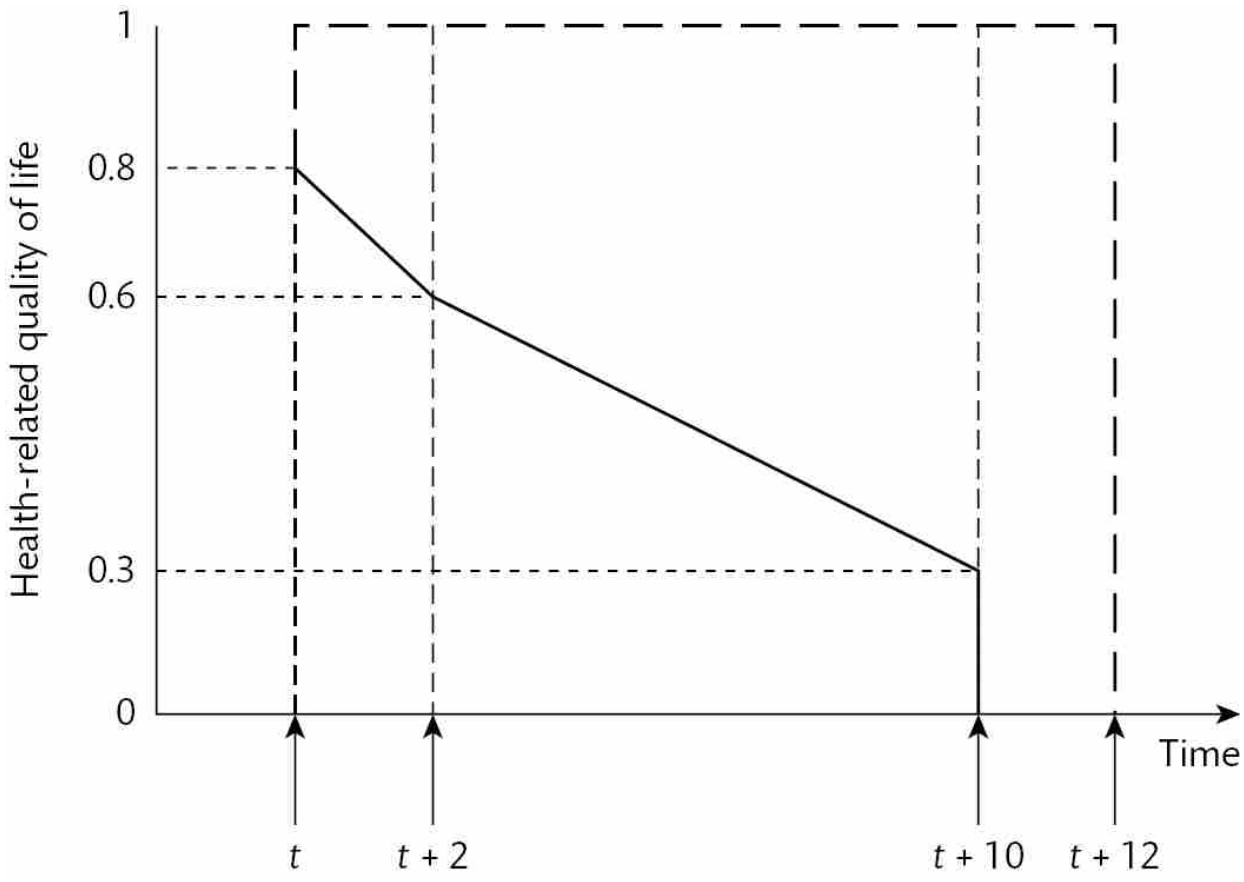


Fig. 6.6.3 Calculating gains in quality-adjusted life years.

The QALY gain can be calculated by subtracting the total QALYs for an untreated patient from the total that they would have if treated. Without the treatment, the patient has 2 years at an average HRQOL of 0.6 = 1.4 QALYs, plus 8 years at an average of 0.45 = 3.6 QALYs, making a total of 5 QALYs. With the treatment, the patient has 12 years at full health, which is 12 QALYs. The gain is therefore 7 QALYs. Equivalently, we can calculate the total gain by summing the gains within each time period. The first 2 years have a HRQOL improvement from an average of 0.7 to 1 = 0.3, giving a QALY gain of 0.6. The next 8 years have an improvement from an average of 0.45 to 1 = 0.55, giving a QALY gain of 4.4. The final 2 years are additional years at full health, which is a gain of 2 QALYs. The total gain is as before 7 QALYs.

One of the big advantages of the QALY is that it can, in principle, be applied to any kind of intervention, whether it raises life expectancy without improving quality of life, improves quality of life without affecting longevity, improves both or improves one at the expense of lowering the

other. However, to some people this is a weakness of QALYs, as they argue that length and quality of life cannot or should not be compared in the same metric.

Before looking at how QALYs are calculated in practice, it should be noted that if QALYs are gained, or indeed lost, at different times, then in principle they should be discounted if they are to be added together, in exactly the same way as costs, if they are to be used in an economic evaluation. There has been a considerable debate about this and there are arguments in favour of discounting both costs and QALYs at the same rate, discounting costs but not QALYs, and discounting them both, but at different rates.

In what follows, the concern is only with the quality of life element of the QALY, since the measurement of life expectancy has no special economic aspects attached to it except for discounting.

In carrying out an appraisal, QALYs can be calculated in a number of different ways. One is to measure the value that patients attach to their quality of life directly. Another is to apply a conversion factor to other indicators of health that will estimate from the indicator a quality of life value. However, an approach which has become most popular, and indeed is recommended by NICE, is to obtain from patients a measure of their health state using a generic health status measure, such as the EQ-5D (Brooks 1996; Herdman et al. 2011), and to apply to the resulting health profile data a value for each state that has been measured for a standard population. The intention is that those values should represent the values of society as a whole.

These sets of values are derived from population surveys. There are different techniques that are used to obtain values, of which the main four are rating or visual analogue scales (VAS), time trade-off (TTO), standard gamble (SG), and the DCE. A VAS usually consists of a single line drawn with verbal and numerical descriptors at each end, describing the meaning of the two ends, such as ‘Best possible health’ and ‘Worst possible health’. Scale markers are often added to the line to denote distance along the line, and these are sometimes also numbered. Respondents are presented with a set of health states and are asked to rate the desirability of each by placing it at some point on the line on or between the two end points.

In the TTO method, respondents are offered a choice between a number of years in full health and a number of years in a particular health state and

asked to choose between them. Slightly different versions of this are used depending on whether the states are chronic or acute state and whether the respondent believes them to be better or worse than being dead. For a chronic health state preferred to death, different numbers of years in full health are offered for a fixed number of years in a health state, until the respondent cannot choose between the alternatives. The value for the health state is then determined by the ratio of the number of years in full health to the number of years in the health state. This calculation is based on the assumed equivalence of the number of QALYs generated by the two alternatives.

In the SG method, respondents are asked to choose between two options. One is that they will be in a specified health state and remain in that state for a given amount of time. The other is that there is a chance that they will either be able to enjoy full health during that time or die before it. Again, slightly different versions of this are used depending on whether the specified health states are chronic or acute and whether the respondent believes them to be better or worse than being dead. For a chronic health state preferred to death, different probabilities of being in full health are offered, until the respondent cannot choose between the alternatives. The value for the health state is equal to the probability at that point of indifference between alternatives. This calculation is based on the expected values of the two alternatives, which is discussed later.

DCEs applied to health state valuation are undertaken by offering respondents a choice between two, or sometimes more, health states and recording their choice. This technique does not measure any individual respondent’s valuations of health states but infers from the responses of a group their collective valuation.

Generic health state measures are usually accompanied by sets of values for each health state that they define. Because they have so many health states, the values are usually generated by a model, which is sometimes called a *multiattribute utility model*, though that label is not explicitly used in some cases and its use in some cases may be disputed. Essentially, health states are decomposed into attributes. Generic health measures usually describe health states using different dimensions, in which people can be at different levels, which form the attributes.

The attributes are combined using a mathematical function to give a single number representing the value for a health state. For example, the widely

used EQ-5D-3L has attached to it a number of value sets that have been calculated and published (Szende et al. 2007). The mathematical form usually used is additive, which means that the attributes are weighted and added together.

Measuring benefit in monetary terms

Monetary measures of benefits can be based on revealed preference or stated preference. Revealed preference techniques use observations from people's behaviour to infer monetary values. As an example, it is possible to examine the value that people place on the avoidance of death or injury by looking at the wages that are paid to people to undertake jobs that differ in the risk of these occurring but are otherwise identical. Stated preference techniques rely on surveys and experiments in which people are essentially asked what monetary value they place on healthcare, or health states, or on other aspects of health benefits. They are not usually asked that question directly, but are given choices between different alternatives, from which money values are inferred.

Such stated preference studies are sometimes called willingness-to-pay studies, although that description is not entirely accurate for all of them. The economic theory underlying these kinds of studies is the measurement of what are called compensating variation and equivalent variation. In the context of measuring the money value of health benefits, these mean monetary payments made by or to a person to compensate for a change in their health, either an improvement or a deterioration, and not having a change in their health, respectively. Willingness to pay refers to compensation paid by the person, either for obtaining an improvement in their health or not having their health deteriorate. Willingness to accept refers to compensation paid to the person, either for having their health deteriorate or not obtaining an improvement in their health.

Although stated preference methods used vary, best practice in economics more generally is to use the contingent valuation method. Contingent valuation requires the survey or experiment to describe a plausible market in which the choices made by respondents about goods reveal the values that are required. The values are said to be contingent on that description. One way to do this is to use a DCE in which price is one of the attributes that the goods have.

Economic appraisal models and uncertainty

In theory, it is possible that an EA could be carried out using one source of data, for example, a clinical trial in which resource use, costs, and quality of life are all measured for individual patients. This would have all of the advantages of a clinical trial in terms of the reliability of results. In practice, EAs are usually based on different sources of data, which have to be linked together. There are many calculations behind the estimates of costs and benefits, which are embodied in what is called an EA model, and in practice all healthcare EAs involve some modelling. In fact, economic modelling is regarded by many as superior to a trial-based EA as it may overcome the problems that trials have in terms of generalisability and speed of producing usable results. Of course, models may use data from trials, but only as one input.

A simple model might be regarded simply as a balance sheet in which the data used to calculate costs and benefits are combined to give figures for total costs and total benefits. However, many EAs cannot use such simple models, and these are no longer regarded as acceptable in the context of technology appraisals examined by bodies such as NICE. The reasons are, first, that treatment options are not always straightforward and clinical decision-making may involve strategies for management, rather than a simple choice between using and not using a single therapy; and secondly, that there is usually some uncertainty about both the model and the data.

As a result, modelling usually uses decision analysis (see, for example, Briggs et al. 2006). This deals both with the complexity issue, by imposing a structure on the implied clinical decision-making, and some aspects of uncertainty. However, there is one aspect of uncertainty that is not dealt with directly by decision analysis, which is uncertainty about the values that data take. A number of techniques have been developed in recent years to deal with this, which are briefly described.

The problem that is dealt with can be encapsulated by the observation that summaries of effectiveness data, such as changes in blood pressure resulting from some intervention, are usually presented as both a point estimate and an interval estimate, or confidence interval, reflecting variability in the data; we should have a similar summary for cost-effectiveness data. When this is done as part of a trial-based analysis it is called stochastic CEA. Unfortunately, there are two problems with this. First, many of the data used

in an EA model are not based on sampling, and therefore do not have a variance that can be used to create an interval estimate. Secondly, the summary index of cost-effectiveness is the ICER, which as a ratio has special problems, such that in general it is not possible to estimate a confidence interval even if all of the data that are used to calculate it do have variability. This second problem applies equally to trial-based and model-based EAs.

Sensitivity analysis

The usual way in which uncertainty is dealt with when using a model is sensitivity analysis. This is a set of techniques that analyse how sensitive the results are to changes in the model, for example, in the data that are contained within it or the way that the data are combined. Some sensitivity analysis techniques are described next. In these, the model that has the best assumed structure and data is called the base case, and the assumptions and data that we wish to test are referred to as the parameters of the model.

One-way sensitivity analysis means looking at how sensitive results are to changes in one parameter. If, for example, a new screening programme is appraised using a model, one of the parameters of the model may be the unit cost of the test used. The base case will use the best estimate that the analyst has of the unit cost as an input to the calculated ICER. One-way sensitivity analysis will calculate the ICER for different values of the unit costs, showing whether the cost makes a lot of or a little difference to the result.

Of course, what is meant by a lot or a little should really be defined. One way to do this is to use the ceiling ratio. If at every possible value of the parameter the ICER remains above or alternatively remains below the ceiling ratio, then the result remains the same, and the decision is not sensitive to the value of the uncertain parameter. In other cases, it may be that at some levels of the parameter for example when the unit cost is low—the ICER will be below the ceiling ratio, but at others—for example, when the unit cost is high the ICER will be above. At some point the ICER will be equal to the ceiling ratio, and at that point called a threshold—a change in the preferred option occurs. This is known as threshold analysis. If the base case and the threshold are very different, then the results are not sensitive to the values of the parameter.

But again, what does very different mean? One way to deal with this is to define upper and lower plausible values for the parameter; for example, with a base case cost of £100 per unit, it may be thought unlikely that the cost could be as high as £200 or as low as £10. This will give a plausible range to the calculated ICER, and if the threshold is within that range, then the results are sensitive to the assumption made.

This can be extended to vary more than one parameter at the same time and to observe the combined impact on the ICER. The ceiling ratio can again be used to define a threshold, and plausible ranges for the ICER can be calculated from the plausible values of the variables, taken in combination. Such multiway analysis is essential, because it is possible that separate one-way analyses of the parameters might suggest that the results are not sensitive to their value, but in combination they might be.

Statistical sensitivity analysis

Plausible range methods are not very satisfactory because their basis is often unclear and not testable. Using statistical methods is better because, as suggested, the base case cost-effectiveness results can be seen as a point estimate and sensitivity analysis as an interval estimate. When this is done as part of a decision analysis model, it is called probabilistic sensitivity analysis.

In this, Monte Carlo simulations are used to generate an ICER distribution. A distribution for each of the model's parameters is assumed, if it is not known. Samples are taken from each distribution, and the ICER is calculated from those simulated data using the model. This is repeated many times to generate an ICER distribution, from which, in principle, a variance might be calculated, and therefore a confidence interval.

Unfortunately, in practice that might still not be possible, because of another problem with ICERs. It is quite possible that the distribution of the ICERs generated by simulation might include both positive and negative values. These mean such different things—a negative value means dominance rather than a true ICER—that a confidence interval calculated from them is meaningless.

There are two possible solutions to this. One is essentially to get rid of the ICER by using the net benefit approach described earlier. Because net benefit is a single number, not a ratio, there is no problem with using

simulation to generate a distribution of it and to calculate a variance and therefore an interval estimate.

Another solution is to generate a cost-effectiveness acceptability curve (CEAC), which retains the CER, but uses a method of describing uncertainty which is different to the confidence interval (van Hout et al. 1994). The simulated ICER values are each compared with a ceiling ratio, and a count is obtained of the proportion of simulated values that are acceptable at that ratio. This is repeated for each possible value of the ceiling ratio, and the proportion that is acceptable will be different for each of those values. A CEAC plots these together, as in Figure 6.6.4. In this illustrative example, the proportion 25% of the simulations of cost and outcomes for a particular therapy are acceptable if the ceiling ratio is £4,500 per QALY gained; 50% if £11,500 and 75% if £20,500. At the ceiling ratio £30,000, 89% are acceptable. Under certain conditions, which are beyond the scope of this chapter to describe, the proportions may be interpreted using a Bayesian approach as probabilities, such that at a ceiling ratio of £30,000, there is a 0.89 probability that the therapy is cost-effective.

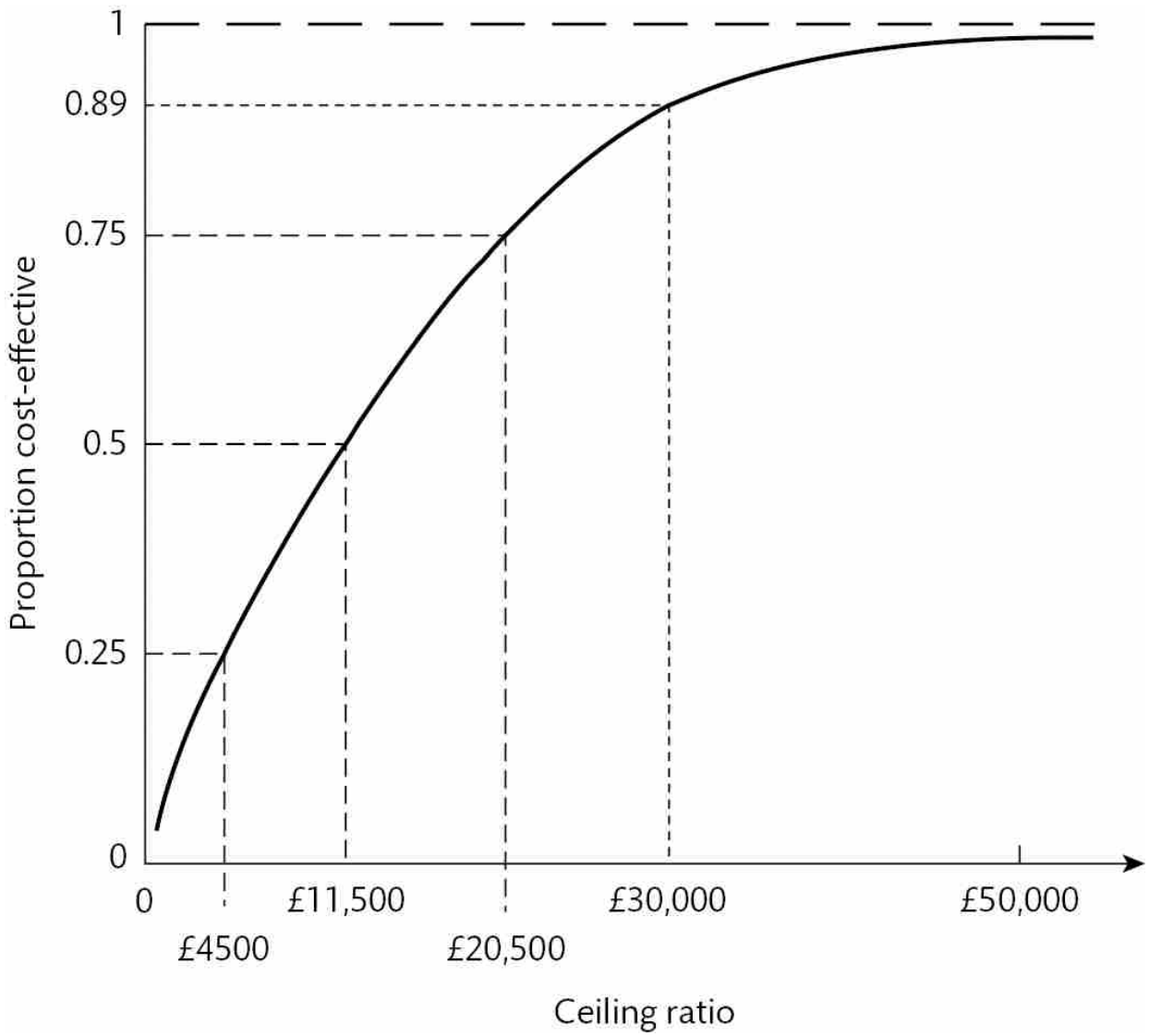


Fig. 6.6.4 Cost-effectiveness acceptability curve.

Decision analysis

Decision analysis is used when the outcomes of decisions are not certain, but the probabilities of different outcomes are known. For example, the treatment of cases detected by a screening programme may not be effective in every case but is known to work in 80% of cases. It is of particular value when decisions involve strategies with different actions. For example, if some members of the population do not respond to a whole population screening initiative, then they may be targeted to increase uptake. Decision

analysis is a way of structuring the different options in order to calculate which of the strategies is best.

Decision analysis relies on the concept of expected values. We can think of the expected value of something as its average value when it is repeated a large number of times. For example, if a treatment produces a gain of 2 QALYs per person in 80% of cases and a gain of 1 QALY per person in the remaining 20%, then out of 100 patients the total gain will be (2 x 80) + (1 x 20) = 180, giving an average of 1.8 QALYs gained per patient. This is the expected value. In reality the nature of uncertainty is such that it is unlikely that the probabilities of 0.8 and 0.2 would produce exactly 80 and 20 in a population of 100, which is why an assumption of large numbers is made. More generally, the expected value is calculated as the sum over all possible outcomes of the probability of the outcome multiplied by the value of the outcome.

Decision analysis calculates the expected values of both costs and benefits and also, though this is a more complicated issue, ICERs. It usually takes the following form, which may be compared with the more general formulation of an EA given earlier:

Structure the problem by constructing a mathematical model of decision-making, usually a decision tree. Identify the decision alternatives; list the possible outcomes of each alternative; and specify the sequence of events.
Assign probabilities to chance events.
Assign values to all possible outcomes of chance events.
Calculate the expected value of each strategy.
Perform sensitivity analyses by systematically changing the assumptions in (1)—(3) to see the impact on the result of (4).

An extension of decision analysis, which is however not yet well established in practical decision-making, is value of information analysis. Decision analysis is based on choosing the alternative that has the highest expected value given current information. The value of information approach analyses the consequences of making the wrong choice, with the aim of assessing how much it is worth spending to improve information so that we do not make the wrong choice. If an incorrect choice is made, the loss incurred is the outcome from the option chosen minus the expected value of the best alternative option. Multiplying this loss by the probability that the incorrect option is chosen gives its expected value. This gives the value of perfect

information; it is worth paying up to this value to avoid making the incorrect choice and therefore investing in research into probabilities of events, quality of life, and costs. Using statistical information on levels of significance and power, it is also possible to estimate the value of sample information, which gives the amount that it is worth paying for a sample of given size to improve information.

Economic appraisal and equity

In discussing equity, we noted that Culyer and Bombard (2012) have suggested using a checklist to ensure that decision-makers take all aspects of equity into account. Here we describe some equity analyses related to EAs and concerning the distribution of healthcare resources and outcomes between people.

The impact of healthcare interventions is ultimately on individual people, and it is always relevant to examine equity at that level. However, in practice inequities are mostly observed between population groups defined by particular characteristics, for example, people who have serious ill-health or are poor, aged, or children. This is particularly important for public health interventions, which are usually conducted at the population level.

Similarly, although it is always relevant to look at the distributional consequences of decisions about every kind of healthcare, it may be of particular importance for public health interventions, since they may be designed specifically to reduce inequalities in population health. It may be that some interventions that target individuals or groups who are disadvantaged would not be regarded as cost-effective relative to many other interventions. It may also be that some public health interventions, such as those designed to enhance health-promoting lifestyles, may be less effective for those who are disadvantaged, thereby increasing health inequalities. In each case an analysis of the implied trade-off between efficiency and equity is essential.

Despite the near-universal acceptance that assessment of equity is as important as efficiency in EA, this has not been matched by practice. In a systematic review, Sassi et al. (2001) were unable to find any satisfactory equity analyses as part of EAs of healthcare in general; Weatherly et al. (2009) found the same for public health interventions.

Cookson et al. (2009) proposed four possible approaches to analysing equity:

Reviewing existing inequalities related to the topic being appraised. This is not in itself evaluative, but it provides background information and therefore some context to decision-making that is based on an efficiency-oriented EA.
Measuring the redistributive impact of a decision about healthcare, that is a quantitative description of who gains and who loses, and the size of gains and losses. Baeten et al. (2010) proposed that for interventions targeted on a particular group, appraisal might assess outcomes in absolute terms for the targeted group but also relative to groups who are not targeted. Sassi et al. (2001) made similar suggestions for assessing the distributional effects for subgroups: the effect of competing interventions; the effect of selectivity; and the effect of prioritizing over other interventions.
If an intervention generates greater equity, but at the expense of a larger health gain overall, measuring its opportunity cost. This cost is not measured in terms of money, but of health, for example, the number of QALYs that would be lost for the population overall by targeting a particular group.
Cost-effectiveness with equity weights applied to the outcomes for different groups. An example which is in practice applied is the 'end-of-life premium' used by NICE in England, which weighs more highly QALYs generated by interventions where survival without the intervention is short, even if the length of the gain from the intervention is itself short. Baeten et al. (2010) also proposed a variant of this in the form of multicriteria decision analysis (MCDA). In MCDA, decision-makers consider many relevant characteristics of an intervention and explicitly incorporate trade-offs within and between them in the form of relative weights. The proposal is that both efficiency and equity consequences are included in characteristics of intervention, thereby taking both into account. Unlike the other three proposed approaches, explicit weighting requires an extra source of data to generate the weights, rather than using data available from an efficiency-based appraisal.

Baeten et al. (2010) illustrated their proposal with an assessment of equity for breast cancer screening. They applied the redistributive impact assessment and both variants of the equity-weighting approach; these gave very similar conclusions in favour of selectivity on both efficiency and equity grounds.

Cookson et al. (2017) distinguish two types of economic analyses that that address equity-related consequences of health and other policy interventions, *equity impact analysis*, and *equity trade-off analysis*. Equity impact analysis quantifies how the costs and effects of alternative policy options are distributed within and between different subgroups of interest to policymakers. Two examples of this are described next, *Extended cost-effectiveness analysis* (ECEA) and *distributional cost-effectiveness analysis* (DCEA). There are other methods, such as benefit-incidence analysis, that do not have a focus on economic appraisal.

Equity trade-off analysis quantifies trade-offs between the two objectives of improving total health and promoting equity. It is used where the outcomes are either a cost-effective intervention that reduces equity or an equity-improving intervention that is not cost-effective, as in boxes 2 and 3 of the health equity trade-off table described earlier. Two examples of this are *Equity Constraint Analysis* and *Equity-Weighting Analysis*. Equity Constraint Analysis assesses how much health is lost by pursuing a fairer option instead of a more cost-effective option. Such an approach treats equity as a constraint on the aim of maximizing overall health. Equity-Weighting Analysis seeks to quantify the trade-off and enable a decision to be taken using a single criterion. It would assess the relative weights that decision-makers put on the criteria of efficiency and equity, quantifying the strength of concern for equity that would lead them to choose whether or not to implement interventions for which a trade-off exists.

Extended cost-effectiveness analysis

ECEA was originally developed by Verguet, Laxminarayan, and Jamison (2015) for low- and middle-income countries as part of the University of Washington's Disease Control Priorities project. It is similar to a CCA, but is specifically aimed at health policy interventions, and includes effects not just on health but also on other societal concerns, in particular the financial consequences of ill-health, referred to as financial risk protection (FPR). The emphasis is on the distribution of these effects over subgroups of a country's population. ECEA is more of a framework for analysis than a set of techniques and its exact content depends on the nature of the policy and the region to which it is applied (Verguet and Jamison 2016).

Health benefits are defined very widely and could include any measures of mortality and morbidity, including life years, QALYs, DALYs, and cases of individual diseases. FPR mainly covers private expenditures and costs by individual members of the population, including out of pocket payments for medical care, other non-medical out of pocket costs such as transport costs to access care, and time and productivity losses, measured by income foregone. Both efficiency and equity are covered, efficiency by considering the net cost of the policy overall and equity by analysing the impact on the distribution of health and financial benefits on subgroups, in particular lower income groups for FPR.

Distributional cost-effectiveness analysis

DCEA was developed by Asaria et al. (2014) as a framework for incorporating equity concerns into CEA. Its aim is to appraise any type of healthcare intervention that has an explicit health inequality reduction objective. As with ordinary CEA, it focuses on health benefits and healthcare costs, rather than including non-health benefits and costs to other sectors of the economy. It therefore appraises effects on the distribution of health and does not consider other dimensions of equity.

Health benefits are disaggregated according to whichever social groupings are thought to be relevant to decision-makers, such as age, sex, social deprivation, and ethnicity. Costs are converted to a health benefit equivalent using a value reflecting the opportunity cost of the resources used in terms of the health benefits that could have been obtained from their use elsewhere in the healthcare system. These are similarly assigned to the different social groups, resulting in the calculation of a net benefit or loss to each group. By comparing the distribution of health before and after the intervention, an index can be calculated of the change in inequality. An equity trade-off analysis can then be undertaken by comparing the resulting equity change index with the total net health benefits.

Asaria et al. (2015) conducted a DCEA on the UK Bowel Cancer Screening programme, comparing different options, of which there were two that dominated in cost-effectiveness terms, screening plus a reminder to all non-respondents and screening plus a reminder targeted on high-risk groups. The results showed that the targeted reminder resulted in a smaller improvement in health but a larger reduction in health inequalities. The

overall judgement would depend on the degree of aversion to inequalities that decision-makers hold, with greater aversion favouring the targeted reminder option. This would in effect mean that the DCEA included an Equity-Weighting Analysis.

Economic appraisal of public health interventions

This chapter has highlighted some issues that arise in applying EA to specifically public health interventions. The state of the art in addressing these is currently less advanced than for other healthcare programmes. But there are many examples that demonstrate both the value of EA in public health and the value of public health interventions themselves.

In reviewing the state of the art in EA of public health interventions, Weatherly et al. (2009) outlined four key methodological challenges, one of which, the specific concern that public health has for equity, we have already discussed. Of the others, one is the kind of evidence that public health interventions are based on. Guidelines for HTA of interventions with individual patients emphasize that the best source of scientific data is the randomized controlled trial (RCT). However, not only are there fewer RCTs for public health interventions, they are inherently more difficult to conduct. We discussed one of the reasons for this in relation to discounting (see ‘Measuring costs’); the impact of public health interventions may be on a longer time scale than an RCT could plausibly encompass. Other problems include that it may be infeasible to establish controls for population-based interventions such as screening and that pathways of causation between intervention and outcome are often less direct than individual treatments, for example, health education programmes. It is therefore necessary to use other ways of establishing evidence, and to rely on extensive use of modelling.

A second concerns the way in which outcomes are specified and measured. Cost-effectiveness analyses use QALYs or other health measures as their outcome measure and it is arguable that is the correct measure where health gain to individual patients is the main aim. However, in public health interventions not only is health gain difficult to measure, because of the difficulties of establishing evidence that we have discussed, but also because the intervention may have additional goals, or important side effects. These include improving public knowledge and awareness, and the benefit of

reassurance, as well as possible unintended adverse consequences such as false reassurance and raised anxiety over possible health problems.

The third concerns the boundaries given to the costs and benefits that are included in an EA. Previously we stated that a social perspective would include all of these (see ‘Social costs and benefits’), whatever they are, and whoever bears them, but it may be essential to restrict their scope depending on the perspective that is taken. For example, as our discussion of the NICE threshold implies (see ‘Cost-utility analysis’), an HTA for interventions that are paid for by a health agency that has a fixed budget is justified in restricting the scope of costs to those incurred by that agency. However, public health interventions typically have much wider effects on the public sector than the health sector, and indeed on the economy as a whole. In turn, those could impact on health; Weatherly et al. (2009) suggest as an example that ‘improvements in housing could reduce illness and injuries, with consequent reductions in healthcare utilization’.

Walker et al. (2019) have outlined a structured way of considering these issues using an ‘impact inventory’. In this the multiple social concerns of interest to decision-makers are specified as different dimensions that might be affected by a policy decision. The inventory contains the impacts of a policy within the different dimensions on different individuals or social groups, both outcomes and costs. This enables aggregate effects to be summed either within dimensions or within individuals or groups. Weighting of dimensions and of individuals or groups would enable a judgement to be made about the policy’s overall net benefit or cost, incorporating any equity concerns.

Despite these problems, there are reasons to be optimistic both about the possibilities of undertaking EA in public health and also about the results from such analyses. Owen et al. (2012) examined the economic analyses that formed the evidence base for public health guidance by NICE issued from 2006 to 2010, an example of which is described in Box 6.6.2; this was extended to the period 2011 to 2016 by Owen et al. (2018). Of 200 cost-effectiveness analyses identified from 2006 to 2010, 30 of the interventions both improved outcomes and lowered costs, in other words dominated the alternative and were in the south-east quadrant of the cost-effectiveness plane. One hundred and forty-one had CERs below NICE’s official stated ‘lower threshold’ of £20,000 per QALY gained, with a median of £1,020; a further seven were below the ‘upper threshold’ of £30,000, with a median of

£25,150; and 11 were above that, with a median of £90,786. Eleven were dominated, that is in the north-west quadrant where costs are higher and benefits lower. For 2011 to 2016, 138 CERs were estimated, with a substantially higher median ICER than for 2006 to 2010 (£7,843 compared with £1,053). There was also greater variability, with both a higher proportion that were cost-saving, but also a higher proportion that were over the £20,000 per QALY threshold. Overall, 63% of public health interventions assessed from 2006 to 2016 were judged to be cost-effective.

Box 6.6.2 Economic analysis of interventions to prevent smoking among schoolchildren

An example is the evidence review underlying NICE guidance on public health interventions to prevent the uptake of smoking among schoolchildren (Jit et al. 2009). A literature review was undertaken and no published economic evaluations of school-based smoking prevention programmes in the United Kingdom were identified. It was argued that studies based in other countries were of limited relevance to the United Kingdom because of differences in the populations being studied and the methodological framework used. In addition, most studies assumed that school-based smoking prevention programmes achieved a lasting reduction in smoking prevalence beyond school age, and this finding was not supported by evidence used to produce the guidance. As a consequence, the authors undertook new research to evaluate whether or not the implementation of a school-based smoking prevention initiative would be cost-effective.

Survey data were used to quantify the association between the age at which smoking was initiated and the probability of quitting smoking quit in later life (older age at initiation was significantly associated with an increased likelihood of quitting in later life). A systematic review was undertaken to quantify the impact of smoking prevention programmes on decreasing smoking uptake in the long term and delaying the onset of smoking, and whether these effects varied by the age of the participants.

These data were combined with cost and other data in an economic model and used to calculate the cost-effectiveness of school-based smoking prevention programmes. Results were presented for a generic programme delivered to 11-year-old students. Assuming an intervention cost of £38.50 per participant the CQG was estimated to be £2,030–12,700, depending on assumptions made about the persistence of effect of the programme over time. The results were fairly robust to changes in key model parameters concerning the association between age of smoking initiation and probability of smoking in later life, mortality in smokers and non-smokers, health-related quality of life in smokers, lifetime medical costs of smokers, and the quality of the studies used to inform the models. In few cases was the extra CQG greater than the ceiling ratio, which was assumed to be £20,000 per QALY gained.

Of course, these samples are only from England, and the coverage may not be representative of the broad range of public health interventions; 127 of the original 200 were related to smoking and other interventions may not

have been selected for evaluation if they were considered unlikely to be cost-effective. Nevertheless, they did cover very different types of interventions to very different sections of the population. An important finding was that interventions aimed at the population as a whole produced the lowest costs per QALY gained.

Masters et al. (2017) conducted a systematic review of published economic appraisals of public health interventions, restricted to high-income countries with universal healthcare, including the Western Europe, the United States, Canada, Japan, Australia, and New Zealand. Their focus was on papers that enabled the calculation of a *benefit-cost ratio*, called by the authors a cost-benefit ratio (CBR), or a *return on investment (ROI)*, which are alternatives to net benefit as a way of summarizing a CBA. The ROI is a concept derived from financial appraisal, measuring the gains derived from an investment as a ratio of gains to cost. In this context it is the ratio of net benefits to cost, equal to the benefit-cost ratio minus 1.

They identified 52 studies from which a ROI could be calculated. For public health interventions as a whole, the median ROI was 14.3, and median CBR was 8.3. As with the NICE study, interventions at the national level produced a greater return (ROI 27.2, CBR 17.5) than those at the local level (ROI 4.1, CBR 10.3). One of the issues with financing public health interventions is that there is often a long time from incurring costs to realizing benefits, in many cases 10–20 years. However, the returns identified were in some cases realized within a short period, for example falls prevention interventions had a ROI of 20.6, much of which was achieved within 6–12 months.

Efficiency and equity in the economics sense are not the only considerations for decisions in public health or healthcare in general, and EA is only one of the types of evidence that are relevant to them. However, EA is a powerful way of assessing evidence about important consequences of such decisions, and in particular helps in considering how to manage scarce healthcare resources. In advocating that public health should receive a fair share of those resources, it is essential to base arguments that are as strong as those for other types of healthcare and to use equally powerful techniques for analysing them.

Table 6.6.2 Incremental costs and QALY gains from vaccination

	Non-vaccinated	Vaccinated	Difference
Total cost (Can$)	476	739	263
Total QALYs	30.5015	30.5098	0.0083

Source: data from Anonychuk et al., A cost-utility analysis of cervical cancer vaccination in preadolescent Canadian females, *BMC Public Health*, Volume 9, Issue 1, pp. 401, Copyright © 2009 Anonychuk et al. licensee BioMed Central Ltd. Reproduced under the Creative Commons Attribution License 2.0.

REFERENCES

Anonychuk, A.M., Bauch, C.T., Merid, M.F., Van Kriekinge, G., and Demarteau, N. (2009). A cost-utility analysis of cervical cancer vaccination in preadolescent Canadian females. *BMC Public Health*, 9(1), 401.

Asaria, M., Cookson, R., and Griffin, S. (2014). Incorporating health inequality impacts into cost effectiveness analysis. In: Culyer, A. (ed.) *Encyclopedia of Health Economics*. San Diego, CA: Elsevier, pp. 22–6.

Asaria, M., Griffin, S., Cookson, R., Whyte, S., and Tappenden, P. (2015). Distributional cost-effectiveness analysis of health care programmes—a methodological case study of the UK Bowel Cancer Screening Programme. *Health Economics*, 24, 742–54.

Baeten, S., Baltussen, R., Uyl-de Groot, C., Bridges, J., and Niessen, L. (2010). Incorporating equity-efficiency interactions in cost-effectiveness analysis—three approaches applied to breast cancer control. *Value in Health*, 13(5), 573–9.

Barker, D.J.P. (ed.) (1992). *Fetal and Infant Origins of Adult Disease*. London: BMJ Books.

Black, W.C. (1990). The cost-effectiveness plane: a graphic representation of cost-effectiveness. *Medical Decision-Making*, 10, 212–15.

Briggs, A., Sculpher, M., and Claxton, K. (2006). *Decision Modelling for Health Economic Evaluation*. Oxford: Oxford University Press.

Brooks, R. (1996). EuroQoL: the current state of play. *Health Policy*, 37, 53–72.

Cookson, R., Drummond, M., and Weatherly, H. (2009). Explicit incorporation of equity considerations into economic evaluation of public health interventions. *Health Economics, Policy and Law*, 4(2), 231–45.

Cookson, R., et al. (2017). Using cost-effectiveness analysis to address health equity concerns. *Value in Health*, 20, 206–12.

Culyer, A. and Bombard, Y. (2012). An equity framework for health technology assessments. *Medical Decision-Making*, 32(3), 428–41.

Driessen, J., Olson, Z.D., Jamison, D.T., and Verguet, S. (2015). Comparing the health and social protection effects of measles vaccination strategies in Ethiopia: an extended cost-effectiveness analysis. *Social Science & Medicine*, 139, 115–22.

Drummond, M.F., Sculpher, M.J., Torrance, G.W., O'Brien, B., and Stoddart, G.L. (2005). *Methods for the Economic Evaluation of Health Care Programmes* (3rd ed.). Oxford: Oxford University Press.

Herdman, M., et al. (2011). Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). *Quality of Life Research*, 20, 1727–36.

Jit, M., Barton, P., Chen, Y.-F., Olalekan, U., Aveyard, P., and Meads, C. (2009). *School-Based Interventions to Prevent the Uptake of Smoking Among Children and Young People: Cost-Effectiveness Model*. Birmingham: West Midlands Health Technology Assessment Collaboration.

Knapp, M. (1984). *The Economics of Social Care*. London: Macmillan.

Masters, R., Anwar, E., Collins, B., Cookson, R., and Capewell, S. (2017). Return on investment of public health interventions: a systematic review. *Journal of Epidemiology and Community Health*, 71, 827–34.

Morris, S., Devlin, N., Parkin, D., and Spencer, A. (2012). *Economic Analysis in Health Care* (2nd ed.). London: Wiley & Sons.

Owen, L., Morgan, A., Fischer, A., Ellis, S., Hoy, A., and Kelly, M.P. (2012). The cost-effectiveness of public health interventions. *Journal of Public Health*, 34(1), 37–45.

Owen, L., Pennington, B., Fischer, A., and Jeong, K. (2018). The cost-effectiveness of public health interventions examined by NICE from 2011 to 2016. *Journal of Public Health*, 40(3), 557–66.

Sassi, F., Archard, L., and Le Grand, J. (2001). Equity and the economic evaluation of health care. *Health Technology Assessment*, 5(3), 1–138.

Szende, A., Oppe, M., and Devlin, N. (2007). *EQ-5D Value Sets: Inventory, Comparative Review and User Guide*. Dordrecht: Springer.

Torgerson, D. and Spencer, A. (1996). Marginal costs and benefits. *British Medical Journal*, 312, 35–6.

van Hout, B.A., Al, M.J., Gordon, G.S., and Rutten, F.F.H. (1994). Costs, effects and C/E ratios alongside a clinical trial. *Health Economics*, 3, 309–19.

Verguet, S., Laxminarayan, R., and Jamison, D.T. (2015). Universal public finance of tuberculosis treatment in india: an extended cost-effectiveness analysis. *Health Economics*, 24(3), 318–32.

Verguet, S. and Jamison, D. T. (2016). Health policy assessment: applications of extended cost-effectiveness analysis methodology in disease control priorities. In: Jamison, D.T., et al. (eds.) *Disease Control Priorities* (3rd ed.) (*Volume 9*). Washington, DC: World Bank.

Walker, S., Griffin, S., Asaria, M., Tsuchiya, A., and Sculpher, M. (2019). Striving for a societal perspective: a framework for economic evaluations when costs and effects fall on multiple sectors and decision makers. *Applied Health Economics and Health Policy*, 17(5), 577–90.

Wang, G., Macera, C., Scudder-Soucie, B., Schmid, T., Pratt, M., and Buchner, D. (2005). A cost-benefit analysis of physical activity using bike/pedestrian trails. *Health Promotion Practice*, 6, 174–9.

Weatherly, H., et al. (2009). Methods for assessing the cost-effectiveness of public health interventions: key challenges and recommendations. *Health Policy*, 93(2–3), 85–92.

6.7

Behavioural economics and health

Alison Buttenheim and Harsha Thirumurthy

Introduction

Among the many determinants of health outcomes around the world, human behaviour plays a central role. While resources allocated to the health sector and access to health products or services are certainly important, it is often the case that individual decision-making influences the burden of disease and the effectiveness of health services available in any given country. Examples of why human behaviour is relevant are numerous in both low- and high-income settings. Behaviours such as smoking and exercise have an impact on the burden of non-communicable diseases (NCDs) globally. Utilization of vaccination services and infant feeding behaviours have an impact on child health outcomes. Understanding human behaviour—specifically how people make health-related decisions—is therefore essential for explaining health outcomes around the world and also for developing solutions to leading challenges in global health. This understanding can be advanced and enriched using behavioural economics.

Behavioural economics blends theories from the disciplines of economics and psychology to uncover key insights about human decision-making. In recent years, a rapidly expanding literature has applied behavioural economics specifically in the health domain. Behavioural economics insights have been useful for understanding why proven health interventions are not taken up at the ‘last mile’, or why it is difficult to sustain healthy behaviours like smoking cessation or medication adherence. Relevant for health policy, the field of behavioural economics has also brought to light a number of interventions that can contribute meaningfully to efforts to improve population health.

This chapter introduces several prominent theories from behavioural economics and reviews examples of how these theories can be useful for studying global health and developing policies or programmes to improve health. We begin by reviewing the 'standard' theory or model of decision-making that features prominently in economics. This model posits a form of rational decision-making and offers valuable insights about what drives individuals' decisions in various domains of life, from consumption to health. We discuss several policy implications that follow from this theory and make a case for why these are important for global health policy. We then turn to theories and principles from behavioural economics and adjacent fields that emerged from a recognition that the standard economic model of decision-making does not adequately explain many aspects of human behaviour. We discuss these anomalies—departures from rational decision-making models—and summarize behavioural economics theories that better explain human behaviour. In parallel, we draw upon empirical evidence from around the world and highlight actionable behaviour change interventions that can be useful for students of global health and practitioners alike.

Given the breadth of behavioural economics theories and empirical evidence, we are necessarily selective in our review of this discipline and its relevance for global health. We focus on a few prominent theories of decision-making that are key to understanding and modifying human behaviour. We introduce the concept of present bias, which is particularly relevant for understanding health behaviours in which the consequences of one's actions are not realized immediately. Present bias in individuals' decision-making offers a compelling rationale for offering incentives—a widely-studied topic for which we summarize key studies and lessons. We also review the concepts of loss aversion, status quo bias, limited attention, and social influence, all of which have direct relevance for health behaviour and health policy. Our chapter concludes by reviewing more recent theories of the link between poverty and decision-making that offer new ways of understanding health behaviour and new opportunities for interventions that improve health.

Empirical examples in this chapter come from diverse settings, including both low- and middle-income countries (LMICs) and high-income countries, and from diverse areas of health, from infectious diseases such as

HIV and malaria for which the burden is greatest in low-income regions of the world to NCDs that are common globally.

Standard economic theories of decision-making

To explain human behaviour and develop behaviour change interventions, the field of economics has commonly relied on a model grounded in the assumption that individuals are rational, self-interested decision makers who seek to maximize their own well-being. This workhorse model in economics offers a number of relevant insights for global health. The model posits three factors that drive choices: preferences that characterize tastes and values, the relative prices faced for the various choices available, and the constraints on available resources such as time or money. Preferences describe individuals' likes and dislikes as well as the extent to which they value the present relative to the future and the degree of risk or uncertainty they are willing to tolerate—all of which can influence one's propensity to make specific choices. Prices reflect the costs of purchasing products and services. And finally, resource constraints are typically denominated in income or wealth, but can also appear as time constraints. We discuss each of these in more detail next.

Preferences are an important determinant of human behaviour that vary from person to person. For a number of health behaviours, it is easy to see why preferences might be relevant. Some individuals may place a greater value than others do on future costs or benefits of actions they take today, and as a result may be less likely to smoke cigarettes or more likely to adhere to medications. These temporal preferences have been shown to be associated with both health behaviours and outcomes in several studies (Thirumurthy et al. 2015; Bradford et al. 2017). Preferences towards risk are another relevant decision-making factor. Heterogeneity in risk preferences and the association between risk preferences and risky health behaviours have been documented in many studies (Anderson and Mellor 2008). Behavioural economics research discussed next raises doubts about the within-person stability of time and risk preferences, but standard economic theory treats risk and time preferences as not easily malleable.

Beyond time and risk preferences, individuals' propensity to use various health services or to engage in healthy behaviours can also be influenced by

the information they possess about associated benefits (or harms). This has motivated a large literature testing whether information provision to individuals can influence decision-making, and a common approach to public health campaigns of raising awareness about the benefits of healthy behaviours or the harms of unhealthy behaviours. Empirical evidence on the effects of information provision is mixed. Some studies have documented large impacts of one-time information provision. In Kenya, informing teenage girls about the risk of unsafe sex with older men who have a higher likelihood of being HIV-positive led to sexual behaviour change and reduced pregnancy risk for the girls (Dupas 2011). Larger-scale testing of this approach in other countries has shown more muted effects, however. In the United States, researchers and policymakers have sought to influence food choices by providing caloric and other nutrition information on food items. Again, there is some evidence that this type of information provision can alter people's preferences and ultimately, their behaviour (Wisdom et al. 2010). These examples underscore the importance of paying attention to individuals' preferences when designing public health interventions.

The second determinant of health behaviour—prices—is highly relevant for understanding health behaviour and continues to shape health policy in important ways. A large literature has shown that prices of health products and services influence demand. Price experiments in high- and low-income countries have shown that demand for healthcare—in much the same way as demand for consumer products—is highly sensitive to prices. The RAND health insurance experiment showed long ago that healthcare utilization is driven by the prices people face (Manning et al. 1987). In low-income countries, a series of randomized trials has demonstrated that people's demand for life-saving health products and services is highly sensitive to prices even when prices are extremely low. One study from Zambia showed that demand for water treatment products declined substantially when going from free distribution to prices just above zero (Dupas 2014). Similarly, a recent study in Zimbabwe showed there was a 25% decline is demand for HIV self-tests among people offered self-tests for $0.50 instead of being offered them for free (Chang et al. 2019). A key lesson that emerges from these studies in low-income countries is that there is a compelling rationale for subsidizing healthcare products and services instead of charging user fees for them. This is especially the case for health products and services

associated with infectious diseases, where usage by one individual can influence health outcomes of other individuals.

In addition to public subsidies, two other policy measures also follow directly from the recognition that prices influence people's decisions. One is the use of taxes to raise the prices faced by individuals and thereby curb certain unhealthy behaviours. Another is the provision of financial incentives to increase benefits associated with a certain health behaviour and thereby promote health behaviours. Taxes are a widely-used policy measure not only for raising government revenue but also for changing human behaviour. Cigarette taxes are perhaps the most prominent example of taxes as a lever for behaviour change in global health. Multiple studies have shown that taxes are an effective way to reduce smoking rates. More recently, in response to growing concern about rising obesity, some governments have begun to impose taxes on sugar-sweetened beverages. Studies from Mexico as well as individual cities in the United States have suggested that this is one way to reduce sugar-sweetened beverage consumption (Colchero et al. 2016; Silver et al. 2017; Roberto et al. 2019). Financial incentives (which we discuss in the next section in the context of present bias) have a clear and compelling rationale even in the standard model of decision-making. Just as subsidies lower the prices faced by consumers and thereby boost demand, incentives can offset various opportunity costs associated with seeking health services and bring down the full price faced by consumers.

Resources available to individuals are the third and final determinant of health behaviour in standard models of decision-making. A core prediction of economic theory is that all else equal, individuals facing financial or time constraints will be less able to access healthcare than those without such constraints. This provides further motivation for interventions such as price subsidies or health insurance; in recent years there have been a number of efforts in LMICs to expand access to health insurance. Another policy response to curb the effects of poverty on health has been the provision of unconditional cash transfers. Many countries in Africa and Asia have expanded unconditional cash transfer programmes with the goal of both alleviating poverty and improving health outcomes. Later in this chapter we discuss additional ways in which poverty can influence health outcomes by summarizing recent research on the psychological consequences of poverty.

Behavioural economics insights and related interventions

Behavioural economics theories have gained prominence in the study of health behaviour because the standard economic models of decision-making outlined earlier, while useful for explaining many facets of decision-making, fail to explain some commonly-observed patterns of behaviour. For example, even when life-saving health technologies and services are available for free or at very low prices, individuals do not utilize them at the high rates that would be predicted by standard economic theory and keenly desired by health officials. Similarly, despite medications for chronic diseases having large benefits that people recognize as important and valuable, medication adherence remains far lower than desired. These and many other behavioural 'anomalies' have underscored the limitations of standard models of decision-making and given rise to the behavioural economics theories that we discuss next.

Present bias and the rationale for incentives

Perhaps the most prominent and widely discussed decision-making context is situations that involve trade-offs between present and future costs and benefits. Seeking preventive services or exercising regularly, for example, requires incurring immediate economic costs, time investment, or other disutility in order to realize health benefits in the future. For such decisions, pervasive in the health domain, individuals do not discount future costs and benefits according to a consistent discount rate, but rather do so in ways that are *present-biased*. Specifically, the immediacy of costs and benefits drives behaviour to a much greater extent that would be expected from the rational tendency to discount the future. Thus, even with large benefits to adopting certain health behaviours such as quitting smoking, exercising, taking medications, or seeking preventive services, individuals may fail to adopt if the benefits are delayed and the costs (in the form of economic costs, inconvenience, or psychological burdens) are immediate. Present-biased preferences have been documented in lab and field settings and go a long way towards explaining why behaviour change can be difficult when there are intertemporal trade-offs.

An important implication of present-biased decision-making is that people can be guided towards engaging in more forward-looking, healthy

behaviours if the immediate benefits of engaging in the behaviour are enhanced or the immediate costs are lowered. This insight motivates the widespread use of financial and non-financial incentives for healthy behaviours in many parts of the world. Since the future benefits of engaging in a healthy behaviour are often insufficient for promoting the behaviour, the provision of small, immediate rewards can accelerate benefits into the present and make the behaviour more likely. They can also offset the disproportionate role of immediate costs or hassle factors. A vast empirical literature has examined the effects of incentive-based interventions on behaviour change around the world. The provision of incentives by employers and governments is also widespread. A few examples illustrate the ways in which incentives and highlight the strengths and limitations of such interventions to promote better health. An early and prominent example of deploying incentives to motivate health behaviours in LMICs is the *Progresa* programme implemented in Mexico beginning in the 1990s and later renamed the *Oportunidades* programme. In this programme, low-income households could receive cash transfers from the government contingent on children in those households attending school regularly and obtaining preventive healthcare. While present-biased decision-making was not the sole rationale for the conditional cash transfers—the programme crucially also sought to alleviate poverty—the interventions succeeded in part by offering more immediate rewards to households for engaging in behaviours that would generate long-term benefits. Various studies have evaluated the Progresa programme and demonstrated that it was largely successful in promoting behaviour change (Paul Schultz 2004; Fernald et al. 2009).

A growing number of randomized trials have also examined the effects of very small incentives in promoting individual health behaviours such as HIV testing, medical male circumcision for HIV prevention, and vaccination uptake. In many of these cases, even though the health services were provided at no cost to individuals, utilization was well below desired levels and incentives succeeded in increasing utilization rates. One study conducted in rural India examined the effects of bringing reliable, monthly immunization services closer to communities and providing caregivers non-financial incentives worth approximately $1 per immunization administered (Banerjee et al. 2010). Rates of full immunization were significantly higher in the villages randomized to receive reliable services and incentives (39%)

as compared to villages that only received reliable services (18%) and to control villages (6%). Another randomized trial conducted in rural Malawi showed that small financial incentives ranging from $0.20 to $3 led to significant increase in HIV testing uptake among adults (Thornton 2008). Individuals who received some incentive were more than twice as likely to obtain their HIV test results as those who did not. Finally, several studies have also tested whether the provision of relatively small non-financial incentives increases uptake of medical male circumcision in settings with a high prevalence of HIV. Even though medical male circumcision reduces men's risk of HIV acquisition by 60%, uptake has been low in settings where free circumcision services have been made available. Present-biased decision-making is one of several reasons why uptake may be low, as several costs of getting circumcised are relatively immediate (taking time off from work, fear of pain, and sexual abstinence for up to 6 weeks) whereas the benefits are distributed over a future time horizon. Studies in Kenya and Zambia have shown that small incentives worth as little as $9 can significantly increase uptake among adult men. One randomized trial found that uptake in a 2-month period was below 2% among men offered no incentives but considerably higher among those offered incentives of $9 and $15 (6.6% and 9.0%) (Thirumurthy et al. 2014). The use of incentives to promote specific health behaviours is by no means widespread in LMICs, but there are high-profile examples of such incentive programmes being implemented at scale. In India, for example, the government has introduced several schemes that provide pregnant women cash transfers contingent on their use of prenatal health services and contingent on facility-based deliveries.

Incentives to promote health behaviours have not always succeeded, however, and there are a number of lessons to be learned from these examples as well. Incentives have generally been less effective in promoting sustained behaviour change. For example, medication adherence, which often requires frequent, nearly daily pill-taking is more difficult to promote with incentives that less-frequent, one-time health behaviours. A recent study in Uganda, for example, found that financial incentives did not increase viral suppression rates among HIV-positive adults taking antiretroviral therapy (ART) (Thirumurthy et al. 2019) while some other studies have shown that incentives do have promise as an adherence intervention (Linnemayr et al. 2017; McCoy et al. 2017). In practice, for

incentive interventions to succeed it is important that incentive payments be delivered as close in time to the behaviours that is of interest. For example, an incentive to promote daily pill-taking is less likely to be effective if payments are not made on a nearly daily basis and instead made at the end of the month. With rising ownership of mobile phones and new technologies that make it possible to monitor behaviours like medication adherence or daily exercise (such as physical activity monitors), the potential for devising more effective incentive-based interventions is likely to grow. Costs and feasibility of implementing these interventions at scale in low-income settings remain a limitation, however. Further below, we draw attention to interventions that rely on social incentives to promote behaviour change rather than financial incentives.

Loss aversion and the endowment effect

Another finding in the behavioural economics literature that goes against traditional economic theory is the endowment effect, which predicts that we value something we own (or have been 'endowed' with) more than we value the same item if we are asked to pay for it. The general principle underlying the endowment effect is loss aversion, or our tendency to weight losses more heavily than a comparable gain (Kahneman et al. 1991). We are made more unhappy by the loss of $20 than we are made happy by the gain of $20. This asymmetry also stands in stark contrast to traditional economic theory, which predicts that we should value that $20 equivalently whether it's headed into our wallet or out of it.

Loss aversion can be leveraged in the design of financial incentives to promote behaviour change. Instead of earning incentives prospectively (for example, receiving a weekly payment for a negative cotinine test as evidence of smoking cessation, or for meeting physical activity targets), participants can receive incentives up front (an endowment of sorts) but lose some of the incentive if they fail to meet targets. To leverage loss aversion even further, deposit contract or commitment contract schemes ask participants to put up their own funds, which they risk losing if behaviour goals are not met. In high-income settings, these types of incentives that leverage loss aversion have been shown repeatedly to be more effective than standard, gain-framed (prospectively-earned) incentives (Halpern et al. 2015). In the Philippines, the CARES programme offered smokers the

opportunity to deposit funds into a savings account. If a urine test at six months showed abstinence from smoking, the savings were returned; if not, the accumulated funds were donated to charity. Among smokers who took the offer and were followed up at 6 months, 35% were biochemically confirmed to have quit smoking. While only 11% of those offered the programme took it up, quit rates among those offered the programme were 3 percentage points higher than those not offered the programme (Giné et al. 2010). The selection into such commitment contracts is of course an important aspect of these programmes, and may reveal higher levels of readiness to quit, self-awareness, or optimism about having a successful quit attempt.

Loss aversion can also inform the design of public health information campaigns and other information-provision interventions. People respond differently to the same information when framed as a loss compared to a gain frame. In general, people tend to exhibit more risk aversion in response to gain frames and are more risk-seeking or more tolerant of risk in response to loss frames (Kahneman and Tversky 1979; Rothman et al. 1999). In a multi-country test of household use of water treatment technologies, messages that included both loss and gain messages were more motivating than gain-framed only messages (Luoto et al. 2014). Loss aversion may also shape how consumers respond to prices. For example, discounts on health-promoting items may feel like a gain and may therefore be less persuasive than taxes on unhealthy items, which may feel like a loss (Cawley 2011). Similarly, choices about purchasing or joining health insurance schemes may be shaped by the relative cost burden of premiums (a price, which may not invoke loss aversion) vs. deductibles, which must be paid following care and may feel like a loss (Rice 2013).

Just as loss aversion can be leveraged to strengthen the effectiveness of financial incentives, other insights from behavioural economics are also relevant when designing incentive interventions. The tendency for individuals to overestimate the likelihood of low-probability events (known as base rate neglect) has been used to motivate lottery-based rewards for engaging in healthy behaviours. Several studies have tested whether lottery-based incentives are effective. One randomized trial in Lesotho found that providing lottery-based rewards to individuals if they did not obtain sexually transmitted infections (STIs) led to a large and significant reduction in HIV incidence (Björkman Nyqvist et al. 2018). Another study

conducted in rural Uganda found that lottery-based incentives worth about $1 per person were more effective in increasing HIV testing uptake than standard, gain-framed incentives worth the same amount (Chamie et al. 2018). The implication therefore is that incentives may be more effective if they are offered in the form of lotteries—with some individuals receiving a large reward and others receiving a small reward or no reward. That said, the context in which incentives are provided and the specific behaviours being targeted are also relevant, as several studies have shown that lottery-based rewards are not more effective than no incentives or standard incentives (Thirumurthy et al. 2016; Buttenheim et al. 2018).

Status quo bias and defaults

Humans have a strong tendency to stick with the status quo. We prefer the option placed in front of us—the default—over having to invest cognitive and decision-making effort to choose a different option. This so-called status quo bias explains many behaviours, from continuing magazine subscriptions that automatically renew each year to accepting the side dishes that automatically come with a fast-food meal. Status quo bias is another adaptive heuristic that can lead to suboptimal health behaviours and decisions. Status quo bias is adaptive in that it conserves limited attention span and bandwidth: if we had to thoroughly evaluate all available options for every choice we are presented with in the course of decision-rich day, we would quickly fatigue. Status quo bias can be maladaptive, however, when it causes us to remain with a preselected option purely because it is the default.

The pervasiveness of status quo bias makes the design of default options very important in health interventions of all kinds. Of all the popular nudge approaches, defaults have consistently been shown to be the most effective and powerful. While status quo bias is likely the most powerful engine of defaults (it is simply easier to stay with the default option, and inertia will likely keep us from investing time and effort to switch), defaults are powerful nudges for other reasons as well: default options may also carry some implicit level of endorsement from the choice architect—if this is default option, someone must have thought it was a good idea. Two other behavioural economic concepts may also contribute to the power of defaults: because switching to another option from the one you already have

may be perceived as a potential original option, loss aversion may come into play. Additionally, omission bias predicts that we prefer doing nothing to doing something, particularly when there may be negative consequences to our (in)action. Omission bias will make the default option seem even more attractive compared to taking action to switch to a different choice (Johnson and Goldstein 2003).

There are many compelling empirical examples of successful defaults to change health behaviour in both high-income and LMIC settings. All opt-out screening or testing programmes employ defaults, as the patient will be automatically screened or tested unless they proactively decline. Opt-out testing for HIV has been widely scaledup across several countries in sub-Saharan Africa in diverse care settings and for different populations. In general, opt-out testing is highly effective in boosting testing rates (Steen et al. 2007) and is also feasible and acceptable (April et al. 2009). Some studies suggest that opt-out HIV testing can reduce HIV-related stigma or may be most effective in increasing testing among those with higher perceived stigma (Kaufman et al. 2017). Test-and-treat policies, in which those diagnosed with HIV immediately begin treatment with ART regardless of viral load, CD4 count, or length of time living with infection, also harness the power of defaults. When immediate initiation of treatment is framed as the default, newly diagnosed individuals do not have to make complex decisions about returning for follow-up care, repeated viral load testing, or the optimal time to begin treatment (Linnemayr and Stecher 2015).

Limited attention: availability heuristic and salience bias

Traditional economic theories of decision-making described earlier assume that individuals have complete information about choices and can attend to and process information without constraint. In reality we live in a very information-rich environment, with many demands on our attentional and cognitive resources. To deal with this information overload, we have adaptive strategies to filter out much of the information directed at us and selectively pay attention to signals. These strategies produce predictable biases and heuristic thinking that may lead to less healthy decisions, but that can also be leveraged in health-promoting interventions.

The *availability heuristic* prompts us to estimate the likelihood of an event based on how easily we can think of a recent example of that event (Tversky and Kahneman 1973). Immediately following an infectious disease outbreak, for example, people may overestimate the likelihood of future outbreak or of their own susceptibility to the disease. On the flip side, a community may underestimate the prevalence of a common but 'silent' or stigmatized condition such as mental illness or HIV if individuals cannot easily recall a specific instance of the illness in their network. These biased estimates of risk can then affect decisions about prevention behaviours, including screening or vaccination. A related phenomenon is salience bias, which predicts that we will focus our attention on information that is most unusual, striking, or vivid—that is, more salient. Information (visual or otherwise) that stands in contrast to the surrounding environment is particularly salient—think, for example, of a single red word on a page or a cell phone ringing in a quiet library.

Public health interventions that rely on the provision of information to change behaviour will be most effective if they take the availability heuristic and salience bias into account. This has been shown in particularly compelling way in multiple water, sanitation, and hygiene (WASH) interventions (Cole 2008; Kremer et al. 2009; Luoto et al. 2014). A study in rural India made the safety of household drinking water more salient for households in rural India by conducting a visible chemical test of water contamination in the household's water storage containers and showing households the results (Hamoudi et al. 2012). Dangerous levels of bacterial contamination cannot necessarily be detected by sight or smell, but a chemical test strip showing presence of faecal bacteria makes this information highly salient. Households receiving this intervention were more likely to spend additional financial resources on clean water than controls, although, importantly, they did not invest more time in home-based water handling or treatment measures.

Salience bias can also inform the design and delivery of interventions that use SMS reminders to promote attendance at appointments or adherence to medications. Messages that grab attention with relevant, personalized information that are delivered at the optimal time to promote action have been shown to be highly effective. One study conducted in Kenya, for example, examined the effects of daily and weekly text messages sent to HIV-positive adults taking ART (Pop-Eleches et al. 2011).

Since adherence is important for realizing the health benefits of ART and achieving viral suppression, messages in the study were intended to remind patients to take their medications and also increase their motivation to stay adherent. Over a 12-month period, the study found that patients who received weekly text message reminders were significantly more likely to have medication adherence exceeding 90%. Another study, also conducted in Kenya, found that two-way text messages sent to ART patients on a weekly basis led to significantly higher self-reported adherence as well as higher viral suppression rates (Lester et al. 2010). These promising examples of mHealth interventions have given rise to many other interventions. While such interventions are not uniformly successful, they represent a low-cost, easily testable way to promote behaviour change.

Social norms and social influence

We are very social creatures who look to the actions of others to guide our own decisions and behaviours. This is another adaptive heuristic mode of thinking that kicks in particularly when information is limited, choices are many, and cognitive bandwidth and attentional resources are constrained. This tendency means that social norms—standards for behaviour shared within a social group—can strongly influence behaviour, far beyond what standard economic theories of decision-making might predict. Social norms theory helpfully distinguishes between descriptive norms (what we see other people do), and injunctive norms (what others believe we ought to do) (Reid et al. 2010). For example, if you decide to take your child to be vaccinated because you observe everyone else in your community doing the same, you are acting on a descriptive norm. If you are told by others in your network that everybody ought to vaccinate their children, and you fear there will be social consequences if you don’t, you are acting on injunctive norms.

The potential for social norms to influence behaviour is evident in an extensive literature describing correlations between descriptive norms and behaviour. An observational study of household participation in a Chagas disease vector control campaign in Peru found that participation decisions were associated with whether neighbours also participated, controlling for other predictors of participation that might also be correlated with neighbour participation (e.g. socioeconomic status, infestation status)

(Buttenheim et al. 2014). Importantly, perceived norms (even if incorrect) can also predict behaviour. Men in a South African township overestimated the prevalence of risky sexual behaviour and underestimated prevalence of condom use among their peers. These behavioural norms correlated with the respondents' own risky behaviours (Carey et al. 2011). Similarly, a study in Uganda found that village residents vastly underestimated the prevalence of HIV testing among other village members (HIV testing was actually a normative behaviour); underestimation was also associated with never having tested (Perkins et al. 2018). Multiple studies have also found a link between social norms (both descriptive and injunctive) and child marriage (Cislaghi et al. 2019).

Of course, these relationships may reflect mere correlation or even reverse causation. Experimental trials of interventions that leverage social norms to change behaviour have had mixed results in LMICs. Two interventions informed by the descriptive work on 'contagious' participation in the vector control campaign in Peru that leveraged social norms in multiple ways were not able to boost household participation beyond the existing campaign (Buttenheim et al. 2018). In a randomized trial of promotion of water treatment products in Bangladesh, willingness to pay for products was not increased by observing neighbour purchases of similar products (Luoto et al. 2012). It should be noted that many of the same successful interventions mentioned earlier to improve hygiene and sanitation behaviours and reduce child marriage have also incorporated social norms as part of complex, multicomponent interventions, making the direct of effect of changing perceived norms or making norms more salient difficult to disentangle.

Generally, there are three opportunities to deploy social norms in an intervention design to effect behaviour change. First, the intervention can make others' behaviour more visible or salient, hopefully correcting misperceptions that shape individuals' evaluations of the risks and benefits of the behaviour. Second, an intervention can make one's own behaviour more observable by others, such that individuals are more likely to comply with descriptive and injunctive norms for fear of sanction. Third, respected peers or opinion leaders can promote, endorse, or request the behaviour, providing a strong signal that the behaviour is normative and will garner social approval. Recognizing the substantial variation in the effects of social norms-based interventions, Theory of Normative Spectrum posits that

norms vary in their influence on behaviour along a spectrum, with the strongest norms making a behaviour obligatory and the weakest norms making a behaviour merely possible (Cislaghi and Heise 2018).

Scarcity theory and the psychology of decision-making

This chapter has presented two overarching perspectives on human behaviour that are both useful in the real world: the rational decision-making perspective and the behavioural economics perspective. Several studies conducted in recent years provide a more nuanced view on decision-making and provide additional insight on the two perspectives we have summarized so far. By studying how being in poverty affects individuals' decision-making, this work represents a bridge between the perspective that poverty affects behaviour because of pure constraints (not having sufficient income to access healthcare, for example) and the perspective that heuristics and biases in decision-making explain behaviour (people are present-biased in their decision-making, for example). Lab and field experiments have shown that in fact, the very condition of being poor–facing scarcity–gives rise to various features of decision-making central to behavioural economic theory such as present bias and limited attention.

One study has shown that poverty leads to increased temporal discounting—the degree to which future costs and benefits are devalued—and also to greater present-biased decision-making (Haushofer and Fehr 2014). Using data from a randomized trial in Kenya in which some households received large cash transfers, this study was able to examine differences in time discounting and risk-taking behaviour among individuals who did and did not face extreme poverty. This study and others have also shown that poverty increases stress (Haushofer and Fehr 2014) and impedes cognitive function (Mani et al. 2013). By reducing mental bandwidth (Mani et al. 2013), poverty makes it more difficult for individuals to undertake actions that are beneficial in the long-term but lack clear short-term benefit. These and other studies revealing the psychological consequences of poverty have important implications for many domains of health in LMICs and for antipoverty programmes more generally. If poverty contributes to limited attention, cognitive capacity, and worse decision-making regarding health and other future-oriented

behaviours, then programmes and policies should at the very least reduce the cognitive burdens associated with processing health information or consider using interventions such as incentives that address present-biased decision-making. Interventions such as reminders to counter the effects of limited attention also become more necessary for some behaviours such as medication adherence.

An even stronger implication of research on the psychology of poverty is that interventions such as unconditional cash transfers may have a broad array of favourable consequences for decision-making and for health outcomes. Rather than merely bringing households out of poverty and eliminating resource constraints that limit consumption expenditures, cash transfer programme can also promote forward-looking decision-making. Recent years have seen growing consideration as well as implementation of unconditional cash transfer interventions in LMICs, sometimes under the rubric of universal basic income policies. Several evaluations of unconditional cash transfer programmes have shown largely positive effects on a broad array of health and economic outcomes. One randomized trial examined the effects of unconditional cash transfers of $400-$1,500 provided by a non-governmental organization, GiveDirectly, to rural Kenyan households and found large increases in consumption, psychological well-being, and asset holdings (Haushofer and Shapiro 2016). Several other studies have examined the effects of unconditional cash transfers provided by governments to low-income households. These studies have documented improvements in sexual and reproductive health outcomes in addition to changes in other forms of decision-making.

Moving the field forward

Applying behavioural economics insights to global health programmes offers great potential to address the ‘last mile’ behavioural challenges to achieving better health outcomes. Researchers and practitioners from several disciplines have generated promising results in diverse settings through focused laboratory or basic science studies and through field experiments. Two important questions have been answered in the strong affirmative: first, can behavioural economics and behavioural science more broadly move the needle on behaviour change that is relevant for global

health? Yes—we hope this chapter is persuasive in identifying successful interventions that leverage behavioural science to boost or supercharge existing health interventions. Second, is there continuing demand among health ministries, funders, and bilateral and multilateral development agencies for bringing a behavioural economics lens to intervention and programme development? We say ‘yes’ to this as well: a quick search of ‘behavioural insights’ and ‘health behaviour change’ will yield a rich array of initiatives in many countries, at all levels and at diverse scales, that are focused on tackling persistent global health challenges with the behavioural economics toolkit.

That said, practitioners and researchers alike have many outstanding questions about how to incorporate behavioural economics into their global health work. Similarly, funders and government units must set strategic priorities and allocate budgets. We argue that the field is poised at a unique moment to take stock of progress and chart a path for the coming decade that will enhance and amplify this work further. As a starting point, we propose this focused list of recommendations for opportunities to move this exciting work forward:

Continue to support and conduct rigorous field trials of promising interventions. As we write this chapter, a team of economists has just won the 2019 Nobel Prize in Economics for their efforts to use randomized controlled trials to test policies and interventions to alleviate global poverty and improve well-being. We will always need efficacy and effectiveness trials to answer the crucial 'Did it work?' question for interventions informed by behavioural insights. While health ministries and communities are understandably skeptical and sometimes reluctant to randomize interventions to participants and may not view evaluation investments as their highest priority, we need to ensure that we scale the interventions that work and abandon or revise those that don't. While it's important of course to learn about average treatment effects, heterogeneous treatment effects are also vital to answer crucial secondary questions: 'for which subgroups, settings, and behaviours did this intervention work best?' It is harder to power, fund, and carry out studies that pre-specify subgroup analyses, but we will learn faster how to target and tailor our behaviourally informed interventions if we do.

Build capacity in behavioural science across sectors and disciplines. In our experience, frontline health programme managers and government ministry staff are eager to learn and apply behavioural science. Few will receive formal training in the theory or practice of behavioural economics, but focused capacity-building and 'on the job' training can infuse a behavioural science perspective vertically and horizontally across work units responsible for health programmes and outcomes. This capacity-building can be layered on to existing topical conferences, funder convenings, and other high-profile efforts for maximum leverage. While degree programmes in applied behavioural science are proliferating globally, there is also enormous scope for certifications, short courses, institutes, and initiatives that allow frontline staff and managers to get up to speed rapidly.

Invite implementation science perspectives to the table. There is an extensive literature documenting the behavioural science evidence from focused efficacy and effectiveness trials. Less studied (and less resourced) are efforts to systematically study the implementation and scaling of the evidence-based interventions that emerge from those trials. Implementation science encourages us to investigate the strategies that enable behavioural economics interventions to successfully and sustainably scale. More work is needed on how to bake this perspective into existing approaches and

initiatives. For example, a key strategy to implement a behaviourally informed intervention may be to build stakeholder demand and increase stakeholders' positive appraisal of behavioural economic or nudge approaches. Empirical work that formulates and tests implementation strategies of this type is sorely needed (Trujillo et al. 2014).

Consider a Nudge Unit or Behavioural Insights Team approach. The formation of units embedded within government agencies or consulted by government agencies and other health sector practitioners can foster real-world application of behavioural economics. These units, sometimes referred to as Behavioural Insights Teams or Nudge Units, typically comprise behavioural science experts who are tasked with actively engaging practitioners and jointly developing and testing low-cost interventions to improve outcomes. Governments in the United Kingdom and United States have formed behavioural insights teams to improve the performance of government agencies (Halpern and Sanders 2016). Several health systems have also established such teams to improve healthcare delivery (Patel et al. 2018). These units work best when members of the health system—from leaders to frontline healthcare workers—have opportunities to obtain basic training in behavioural insights and intervention design and can propose and prototype nudges and other interventions. Members of the nudge units themselves can further guide intervention development.

Bring behavioural insights to additional health domains in LMICs. Many of the most pressing challenges in global health remain understudied from a behavioural economics perspective. A glaring case of this is the growing burden of NCDs is in LMICs. Relatively few studies or programmes have applied behavioural insights to promote the behaviour change that is needed to curb the rise of NCDs. In contrast, behavioural economics applications to health in the United States and other high-income countries have long focused on behaviours such as smoking cessation and physical activity. Just as behavioural economics research and practice has furthered our understanding of how to control HIV/AIDS and other infectious diseases in low-income settings, we see considerable potential for similar approaches in the NCD space. Another important contributor to global health outcomes that has been relatively neglected by behavioural economics researchers is the supply side of healthcare. The vast majority of nudge studies conducted and nudge interventions that have been scaled-up have addressed the

demand side of health, which is to say they have focused on modifying health behaviours of patients or individuals in the general population. Provider behaviour can and should equally be a target of interventions that use behavioural economics insights.

Applying behavioural economics in isolation will not magically solve the persistent global health challenges faced by LMICs. Rather than a panacea, it is instead a promising approach that warrants further consideration and testing. No simple nudge will remove structural barriers to health behaviour change, such as limited access to healthcare or lack of affordability. While many successful behavioural interventions have been targeted at health providers, behavioural insights are not an obvious or easy path to fixing supply-side issues as well. In its most successful application, behavioural economics offers straightforward ways to boost or supercharge existing systems or interventions to promote better take up, improve engagement, reduce procrastination, or eliminate hassle factors. In the compelling example given of offering lentils and a metal plate set to parents who vaccinated their children in India, note that the other key intervention in that study was guaranteeing that vaccination camps would be functional and reliable. It is unlikely (though not tested in that study) that the incentives alone in the absence of guaranteed vaccination camps could have boosted coverage rates as much as they did.

While this same study had a substantial effect on vaccination, many behavioural interventions will move the needle by a fairly small amount. This increases the urgency to design low-cost, light-touch interventions, and to deploy them in settings where a small effect size is still clinically or programmatically relevant if sustained. The ‘behavioural revolution’ asks us to bring a simple but powerful perspective to existing health behaviour change programmes and initiatives. If we can pay careful attention to how people actually think and behave (vs. how we hope they might optimally behave), and recognize the constraints on cognitive resources, attention span, self-control, and future orientation that poverty imposes, we can design, test, and scale interventions that will help individuals meet their health goals, and generate sustained improvements in population health and lifespan.

REFERENCES

Anderson, L.R., and Mellor, J.M. (2008). Predicting health behaviors with an experimental measure of risk preference. *Journal of Health Economics*, 27, 1260–74.

April, M.D., et al. (2009). HIV testing rates and outcomes in a South African community, 2001–2006: implications for expanded screening policies. *Journal of Acquired Immune Deficiency Syndromes (1999)*, 51, 310.

Banerjee, A.V., Duflo, E., Glennerster, R., and Kothari, D. (2010). Improving immunisation coverage in rural India: clustered randomised controlled evaluation of immunisation campaigns with and without incentives. *British Medical Journal*, 340, c2220.

Björkman Nyqvist, M., Corno, L., De Walque, D., and Svensson, J. (2018). Incentivizing safer sexual behavior: evidence from a lottery experiment on HIV prevention. *American Economic Journal: Applied Economics*, 10, 287–314.

Bradford, D., Courtemanche, C., Heutel, G., Mcalvanah, P., and Ruhm, C. (2017). Time preferences and consumer behavior. *Journal of Risk and Uncertainty*, 55, 119–45.

Buttenheim, A.M., et al. (2014). Is participation contagious? Evidence from a household vector control campaign in urban Peru. *Journal of Epidemiology and Community Health*, 68, 103–9.

Buttenheim, A.M., et al. (2018). Increasing participation in a vector control campaign: a cluster randomised controlled evaluation of behavioural economic interventions in Peru. *BMJ Global Health*, 3, e000757.

Carey, K. B., et al. (2011). Community norms for HIV risk behaviors among men in a South African township. *Journal of Behavioral Medicine*, 34, 32–40.

Cawley, J. (2011). *The Oxford Handbook of the Social Science of Obesity*. Oxford: Oxford University Press.

Chamie, G., et al. (2018). Comparative effectiveness of novel non-monetary incentives to promote HIV testing: a randomized trial. *AIDS*, 32, 1443–51.

Chang, W., et al. (2019). Effect of prices, distribution strategies, and marketing on demand for HIV self-testing in Zimbabwe: a randomized clinical trial. *JAMA Network Open*, 2, e199818.

Cislaghi, B., and Heise, L. (2018). Four avenues of normative influence: a research agenda for health promotion in low and mid-income countries. *Health Psychology*, 37, 562.

Cislaghi, B., Mackie, G., Nkwi, P., and Shakya, H. (2019). Social norms and child marriage in Cameroon: an application of the theory of normative spectrum. *Global Public Health*, 14, 1479–94.

Colchero, M.A., Popkin, B.M., Rivera, J.A., and Ng, S.W. (2016). Beverage purchases from stores in Mexico under the excise tax on sugar sweetened beverages: observational study. *British Medical Journal*, 352, h6704.

Cole, M. (2008). Compliance and infection control guidelines: a complex phenomenon. *British Journal of Nursing*, 17, 700–4.

Dupas, P. (2011). Do Teenagers Respond to HIV Risk Information? Evidence from a Field Experiment in Kenya. *American Economic Journal: Applied Economics*, 3, 1–34.

Dupas, P. (2014). Getting essential health products to their end users: subsidize, but how much? *Science*, 345, 1279–81.

Fernald, L.C., Gertler, P.J., and Neufeld, L.M. (2009). 10-year effect of Oportunidades, Mexico's conditional cash transfer programme, on child growth, cognition, language, and behaviour: a longitudinal follow-up study. *Lancet*, 374, 1997–2005.

Giné, X., Karlan, D., and Zinman, J. (2010). Put your money where your butt is: a commitment contract for smoking cessation. *American Economic Journal: Applied Economics*, 2(4), 213–35.

Halpern, D., and Sanders, M. (2016). Nudging by government: progress, impact, & lessons learned. *Behavioral Science & Policy*, 2, 52–65.
Halpern, S.D., et al. (2015). Randomized trial of four financial-incentive programs for smoking cessation. *The New England Journal of Medicine*, 372, 2108–117.
Hamoudi, A., Jeuland, M., Lombardo, S., Patil, S., Pattanayak, S.K., and Rai, S. (2012). The effect of water quality testing on household behavior: evidence from an experiment in rural India. *The American Journal of Tropical Medicine and Hygiene*, 87, 18–22.
Haushofer, J., and Fehr, E. (2014). On the psychology of poverty. *Science*, 344, 862–7.
Haushofer, J., and Shapiro, J. (2016). The short-term impact of unconditional cash transfers to the poor: experimental evidence from Kenya. *The Quarterly Journal of Economics*, 131, 1973–2042.
Johnson, E.J., and Goldstein, D.G. (2003). Do defaults save lives? *Science*, 302, 1338–9.
Kahneman, D., Knetsch, J.L., and Thaler, R.H. (1991). Anomalies: the endowment effect, loss aversion, and status quo bias. *Journal of Economic Perspectives*, 5, 193–206.
Kahneman, D., and Tversky, A. (1979). Prospect theory: an analysis of decision under risk. *Econometrica*, 47, 263–91.
Kaufman, M.R., et al. (2017). The differential effects of an opt-out HIV testing policy for pregnant women in Ethiopia when accounting for stigma: secondary analysis of DHS data. *Prevention Science*, 18, 245–52.
Kremer, M., Miguel, E., Mullainathan, S., Null, C., and Zwane, A.P. (2009). Making water safe: price, persuasion, peers, promoters, or product design. *Unpublished manuscript*.
Lester, R.T., et al. (2010). Effects of a mobile phone short message service on antiretroviral treatment adherence in Kenya (WelTel Kenya1): a randomised trial. *Lancet*, 376, 1838–45.
Linnemayr, S., and Stecher, C. (2015). Behavioral economics matters for HIV research: the impact of behavioral biases on adherence to antiretrovirals (ARVS). *AIDS and Behavior*, 19, 2069–75.
Linnemayr, S., Stecher, C., and Mukasa, B. (2017). Behavioral economic incentives to improve adherence to antiretroviral medication. *AIDS*, 31, 719–26.
Luoto, J., Levine, D., Albert, J., and Luby, S. (2014). Nudging to use: achieving safe water behaviors in Kenya and Bangladesh. *Journal of Development Economics*, 110, 13–21.
Luoto, J., et al. (2012). Learning to dislike safe water products: results from a randomized controlled trial of the effects of direct and peer experience on willingness to pay. *Environmental Science & Technology*, 46, 6244–51.
Mani, A., Mullainathan, S., Shafir, E., and Zhao, J. (2013). Poverty impedes cognitive function. *Science*, 341, 976–80.
Manning, W.G., Newhouse, J.P., Duan, N., Keeler, E.B., and Leibowitz, A. (1987). Health insurance and the demand for medical care: evidence from a randomized experiment. *The American Economic Review*, 77, 251–77.
McCoy, S.I., et al. (2017). Cash vs. food assistance to improve adherence to antiretroviral therapy among HIV-infected adults in Tanzania. *AIDS*, 31, 815–25.
Patel, M.S., Volpp, K.G., and Asch, D.A. (2018). Nudge units to improve the delivery of health care. *The New England Journal of Medicine*, 378, 214.
Paul Schultz, T. (2004). School subsidies for the poor: evaluating the Mexican Progresa poverty program. *Journal of Development Economics*, 74, 199–250.
Perkins, J.M., et al. (2018). Actual versus perceived HIV testing norms, and personal HIV testing uptake: a cross-sectional, population-based study in rural Uganda. *AIDS and Behavior*, 22, 616–28.
Pop-Eleches, C., Thirumurthy, H., Habyarimana, J. P., and Al., E. (2011). Mobile phone technologies improve adherence to antiretroviral treatment in a resource-limited setting: a randomized controlled trial of text message reminders. *AIDS*, 25, 825–34.

Reid, A.E., Cialdini, R.B., and Aiken, L.S. (2010). Social norms and health behavior. *Handbook of Behavioral Medicine: Methods and Applications*, 3, 263.

Rice, T. (2013). The behavioral economics of health and health care. *Annual Review of Public Health*, 34, 431–47.

Roberto, C.A., et al. (2019). Association of a beverage tax on sugar-sweetened and artificially sweetened beverages with changes in beverage prices and sales at chain retailers in a large urban setting. *The Journal of the American Medical Association*, 321, 1799–810.

Rothman, A.J., Martino, S.C., Bedell, B.T., Detweiler, J.B., and Salovey, P. (1999). The systematic influence of gain-and loss-framed messages on interest in and use of different types of health behavior. *Personality and Social Psychology Bulletin*, 25, 1355–69.

Silver, L.D., et al. (2017). Changes in prices, sales, consumer spending, and beverage consumption one year after a tax on sugar-sweetened beverages in Berkeley, California, US: a before-and-after study. *PLoS Medicine*, 14, e1002283.

Steen, T.W., et al. (2007). Two and a half years of routine HIV testing in Botswana. *Journal of Acquired Immune Deficiency Syndromes*, 44, 484–8.

Thirumurthy, H., et al. (2015a). Time preferences predict mortality among HIV-infected adults receiving antiretroviral therapy in Kenya. *PLoS One*, 10, e0145245.

Thirumurthy, H., et al. (2014). Effect of providing conditional economic compensation on uptake of voluntary medical male circumcision in Kenya: a randomized clinical trial. *The Journal of the American Medical Association*, 312, 703–11.

Thirumurthy, H., et al. (2016). The effects of providing fixed compensation and lottery-based rewards on uptake of medical male circumcision in Kenya: a randomized trial. *Journal of Acquired Immune Deficiency Syndromes*, *72*(Suppl 4), S309–15.

Thirumurthy, H., et al. (2019). Financial incentives for achieving and maintaining viral suppression among HIV-positive adults in Uganda: a randomised controlled trial. *Lancet HIV*, 6, e155–63.

Thornton, R. (2008). The demand for, and impact of, learning HIV status. *American Economic Review*, 98, 1829–63.

Trujillo, A.J., Glassman, A., Fleisher, L.K., Nair, D., and Duran, D. (2014). Applying behavioural economics to health systems of low-and middle-income countries: what are policymakers' and practitioners' views? *Health Policy and Planning*, 30, 747–58.

Tversky, A. and Kahneman, D. (1973). Availability: a heuristic for judging frequency and probability. *Cognitive Psychology*, 5, 207–32.

Wisdom, J., Downs, J.S., and Loewenstein, G. (2010). Promoting healthy choices: information versus convenience. *American Economic Journal: Applied Economics*, 2, 164–78.

6.8

Governance and management of public health programmes

Zhiyuan Hou and Na He

Introduction

Public health programmes are diverse in scale, coverage, and complexity. They may be organized by local organizations, non-government organizations (NGOs), regional or national governments, and even involve globally (Yeoh 2015). Due to their diversity and complexity, governance and management of public health programmes become very important, and are essential to the health system performance and health outcomes (Lauer 2001; WHO 2002a, 2007; Mills et al. 2006; Institute on Governance 2011). The literature has increasingly showed the strong association between governance and health system performance (Olafsdottir 2012), and effective governance can achieve the goals of public health organizations or programmes (Institute on Governance 2011). The cross-country studies by both the World Bank and World Health Organization (WHO) found strong causal relationships between better governance and health outcomes (Kaufmann et al. 1999; Lauer 2001; Mills et al. 2006). In contrast, Gupta et al. (2000) found that corruption indicators were negatively associated with child and infant mortality, immunization coverage, and low-birthweight. Furthermore, Wagstaff and Claeson (2004) showed that health spending reduced child mortalities only in countries with high governance measures, and more spending in countries with medium and low governance measures would not reduce child mortality. The association between governance and health can operate at multiple domains, including the broader governance environment, health and general public policies, and the effectiveness of organizations within the health system (Savedoff 2011).

Worldwide, health systems work well in some countries but not well in others. To answer the question for health system performance, the WHO

builds a framework of strengthening health systems to improve health outcomes, which includes six function components: governance and leadership, healthcare financing, health workforce, medicines and technologies, information and research, and service delivery (WHO 2007). In the WHO health systems framework, governance and leadership is the central component to make the other function components effectively work together. Governance in the health system refers to 'a wide range of steering and rule-making related functions carried out by governments/decisions makers as they seek to achieve national health policy objectives that are conducive to Universal Health Coverage (UHC)' (WHO 2007). Good governance can enable the effective use of healthcare finances, health workforce, medicines and technologies, information, and research to deliver better health services and outcomes (WHO 2016). Overall, effective governance is central to improve the health systems performance and achieve UHC (WHO 2016).

As the world become increasingly globalized, public health programmes are sometimes cross-border, and their governance also need to be considered at the global level. Development assistance for health has globally increased from US$7.2 billion in 1990 to US$36.4 billion in 2015 (Dieleman et al. 2016). Funds for these development assistance programmes are increasingly distributed through NGOs and bilateral assistance to governments instead of the traditional United Nations (UN) agencies. Although development assistance for health increased quickly, their effects depend on the quality of local governance and policies, especially in the resource constrained settings (Fryatt et al. 2017). On one side, many international donors question whether these development assistance programmes can reap benefits of improved health services and outcomes. On the other side, these global programmes may attract health workforce to leave local health systems, conflict with the goals of local health systems, and distort the local health priorities (Yeoh 2015). These may lead to negative effects on local health systems. Therefore, global health governance is necessary to address these problems at the global level. Coronavirus Disease 2019 (COVID-19) spread across the world in 2020, which emphasis on the emergence of strengthening health governance in the global context (Phelan et al. 2020).

In this chapter, we will discuss the following sections. First, we introduce the concepts of governance, health governance, and global health governance. Second, we introduce governance theories and a framework of

health governance system, covering the actors, tools, and levels of health governance. Third, we describe in detail three governing actors—government, civil society, and private sector—and their roles in health governance. Fourth, tools of health governance are introduced to carry out the health governance functions and enable relationships among actors. Fifth, we discuss four levels of health governance: supranational, national, organization, and programme. Sixth, we summarize the assessment frameworks of health governance. Finally, we take cases for better understanding health governance.

Concepts of governance, health governance, and global health governance

The concept of governance has evolved in the changing context of government and society. Originally, governance is described as 'the act or manner of governing, of extending control or authority over the actions of subjects; a system of regulations' in the *Oxford English Dictionary.*

In the domain of health, WHO first defined governance as 'stewardship' which refers to 'the careful and responsible management of the well-being of the population' (WHO 2000). This definition clarifies the importance of health governance and its role in health system and population health (Mikkelsen-Lopez et al. 2011). However, because of incoherent nomenclature and the changing context, no conceptual agreement on health governance has been reached (Kickbusch and Gleicher 2012).

The United Nations Development Programme (UNDP), defines governance as 'the exercise of political, economic administrative authority of a country's affairs at all levels' (UNDP 1997). This definition reveals that governance consists of the complex mechanisms, processes, and institutions through which citizens and organizations articulate their interests and rights. And it identifies the political and economic nature of social life, and the roles and interdependencies of the society, such as the government, private sector, and civil society. Dodgson and his colleagues consider health governance as ''the actions and means adopted by a society to organize itself in the promotion and protection of the health of its population' (Dodgson et al. 2002). WHO refines health governance as 'involves ensuring that strategic policy frameworks exist and are combined with effective oversight,

coalition building, regulation, attention to system-design and accountability' (WHO 2007). It characterizes the health governance as a set of processes formally or informally applied to distribute accountability among actors of a health system (Brinkerhoff and Bossert 2008). These definitions adapt a system perspective to understand the mechanisms in government, the private sector, and civil society, their complex nature and relationships, and how they influence health outcomes (Yeoh 2015).

More recently, the necessity of strengthening health governance at the global level has been increasingly discussed due to a series of globalizing challenges. The response to global health challenges has become fragmented and ineffective without global leadership. Global health governance is then defined as 'the action and means adopted by a society to promote collective action and deliver collective solutions in the pursuit of common goals' (Dodgson et al. 2002). This definition admits the whole range of cooperation in global society, from informal arrangements and agreements by communities to the formalized rules and procedures institutionalized in governments. This definition also pays attention to the equal roles of the governments, private sectors, and civil society in global health governance, because health governance has become beyond the scopes of governments (Yeoh 2015). A single emphasis on the government role in health governance, neglecting the important roles of civil society and the private sector in health system, will undermine the governance capacity (Ollila 2005; Renukumar n.d.). Contributions from all governing actors for the collective actions in the pursuit of common goals are critically necessary in health governance.

Theory and framework of health governance system

Theory of governance

There are two main theories to approach systems-level governance: principal-agent theory from the economics discipline, and theory of network governance from the public administration discipline (Greer et al. 2016). Principal-agent theory focuses on market competition, contracting, outsourcing, and incentives; and the theory of network governance emphasizes the power and effectiveness of networks, cooperation, and partnership. Both theories aim to address the interactions between

organizations in governance system, including the shifting boundaries, unclear authority, and complexity of coordination, and have been widely used in health system governance.

Principal-agent theory. Principal-agent theory illustrates the contract relationship between principals and agents. The principle hires the agent to undertake an action and pays in some certain forms such as incentives. Agents usually have more relevant information on the service than principals, thus enables them with an advantage of 'information asymmetry' to favour their own interests at the cost of the principal. Another feature of the principal-agent theory is that the principal does not have full control over the agent and possess only partial information about the service behaviour of the agent (Stoker 1998). Because of the distinctive features, moral hazard, and adverse selection problems may appear. The agent tends to take its own preferred actions and hide some information the principal needs to make decisions for the agent's interest (Gailmard 2012). Taking account of the uncertainty and complexity of behavioural outcomes of the agent, principal-agent theory is used as an approach to analyse different actors in governance systems.

Health systems are featured as purchaser–provider splits, more use of markets and contracts to encourage private sector participation, greater management autonomy for providers, emphasis on procedure control and regulation, and outsourcing ancillary functions. In health system, there are various actors—policymakers, healthcare providers and citizens, and these actors and relationships constitute the health governance system. Principal-agent theory can be used to understand these actors and relationships, and various governance frameworks for health systems are intensively developed based on principal-agent theory (Brinkerhoff and Bossert 2014). However, the relationships are mostly more about fiduciary than agency, especially for government and policymakers, and government accountability has long been questioned. Therefore, the approach to analyse the relationships among organizations should be to in the round- in their interconnections, dependencies, and accountabilities. Despite widely used in health governance, the problem with principal-agent theory is that the complex web of relationships is oversimplified to principal-agent relationship.

Theory of network governance. Theory of network governance aims to address the complexity that principal-agent theory neglects and provide a solution to coordination dilemmas among multisector actors, in which

clients, providers, and producers are linked together as co-producers. Network governance focuses on interdependence among organizations, interactions among network members, and management joint responsibilities (Rhodes 1996). Fixed organizational boundaries and roles are replaced by a new rationality of the shared aims and culture (Considine and Lewis 2003). Sørensen and Torfing (2007) defined 'governance network' as a number of interdependent and autonomous actors who interact through negotiations under a relatively self-regulating institutionalized framework and contribute to the production of public purpose.

Network governance is promoted for the potential contribution to efficient governance because of its following distinctive features. First, network governance comes with a potential for identifying early stage problems and making proactive adjustment. Besides, aggregation of information, knowledge, or opinions help qualify decisions or consensus building. Finally, networks of organizations greatly reduce the risk of implementation resistance (Sørensen and Torfing 2007). While network governance is advocated to promote cooperation among multi-actor systems, there are two empirical problems with regarding governance solely as partnerships and networks. On one hand, networks and partnerships can be everywhere and nowhere, and so finding networks that actually govern and align organizations and forces can be hard. On the other hand, network governance does not simply displace hierarchy and law. In some situations, it seems that networks function best under the shadow of hierarchy to oblige them with legal force. Cross-sectional legislation could offer a general guidance for network governance in activation. However, legislation and hierarchy pose a risk to the autonomy and flexibility of the network. Still, the questions regarding decision-making and accountability can hardly be addressed by legislation (Saikku and Karjalainen 2012). The merits and problems of network governance mean that they are unlikely to be the sole basis for health system governance.

Framework of a health governance system

A population's health is determined together by individual factors, social and community networks, and social, economic, cultural, and environmental conditions (Commission on Social Determinants of Health 2008). The most important determinants of health are from other sectors, rather than the

health sector itself. Most health challenges originate beyond healthcare, but they also depend on the global processes that drive disease outbreaks, food insecurity, climate change, urbanization, and economic inequity. This calls for a whole-of-society governance approach for health. The governments, the private sectors, communities, and citizens are required to increasingly engage in governance for health. Thus, health governance involves a wide range of actors at the level of the governments, civil society, and private sectors.

In health governance system, the government, private sector, and civil society are the governing triangle (actors), and they are seeking public health collectively (Abimbola 2020). Each governing actor has its structures, roles in governance, governing tools, and mechanisms for interactions within and between the actors. They play health governance functions at four levels to pursue health goals: the supranational or national or organizational or programme levels. Overall, the three governing actors (triangle), their governance functions and mechanisms, and governance relationship, form a health governance system and a whole-of-society governance approach for health. A framework for health governance system can help to understand the whole-of-society governance approach for health, covering the governing actors, governance mechanisms and tools, their governance relationship, and governance at different levels (Yeoh 2015). Figure 6.8.1 summarized this framework for health governance system.

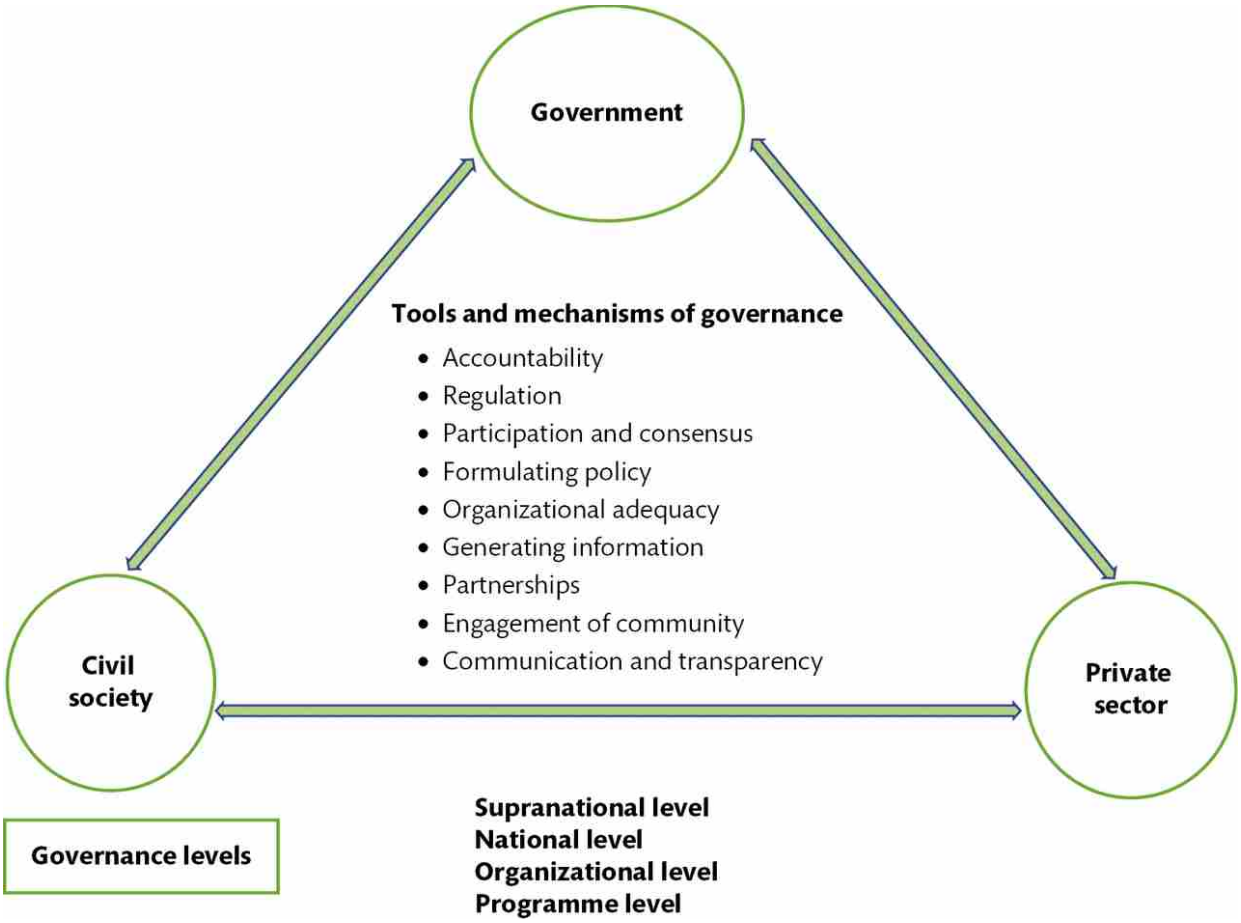


Fig. 6.8.1 Framework for a health governance system.

WHO proposed effective health system governance to achieve UHC, and the governing triangle can be the government, health service providers, and the citizen or civil society. The governing triangle is linked through the governance relationships, mainly principal-agent relationships (Brinkerhoff and Bossert 2014). They make or change governance rules, and these governance rules in turn influence each other's actions and relations (Abimbola 2020). In the relationship between civil society and the government, civil society acts as the principal to express health needs and preferences to the government, and the government works as the agent to respond to these needs and preferences. In the governance relationships between the government and providers, health policymakers are principals setting up objectives and standards, providing resources, and conducting overseeing providers as their agents. In exchange for the resources, providers take health service delivery activities to meet the requirement of policymakers. In the governance relationships between the civil society and

providers, civil society expresses health services needs to providers, who provide health services in turn. But this relationship is challenged by information asymmetries, capacity gaps, and moral hazard between the civil society and providers. Theories of principal-agent and network governance can help to address these challenges and complexity of networks, and better understand health governance at system level (Greer et al. 2016).
According to the health governance system framework, the following sections will introduce each governing actor and its role in governance, the governance tools available, and the levels of governance.

Actors and roles in health governance

Health governance system is the triangle of government, private sector, and civil society. This section will focus on each of governing triangle and its roles in health governance (Table 6.8.1).

Table 6.8.1 Actors and roles in health governance

Actors	Roles and functions in health governance
Government	Regulation Policy development Resource stewardship Coordination and partnership
Private sector	Health services provision Financing Health workforce training Regulation
Civil society	Shaping policies and standards Health services provision Strengthening the health system Community mobilization

Government and its roles in health governance

The government is the leading actor of national health governance, and the key actor of other levels of health governance. There are two forms of

government: federal and unitary, which determine the form of health governance. The federal government owns all the basic powers, but powers are further distributed to local governments of states and authorities, which is decentralized in nature (Elazar 1997). The federal government is not essentially more powerful than those of its states, and usually, unilaterally assigning itself new powers cannot success without the process of constitutional amendment (Fisch 2018). In contrast, the key principle in unitary government is the hierarchical and efficient control and organization of power (Elazar 1997). The federal and unitary forms of governments have their own strengths and weaknesses. Federal government allows the promotion of diversity, and unitary government creates a sense of belongings and uniformity. In addition, decentralization of power accommodates variations suitable for regional problems, while unitary governments make laws only by a single government, and enforce throughout the whole nation without territorial distinction. Federalism increases political participation for different regions and allows states to decide policies reflecting their own needs, while for unitary government, there is less regional level political participation. Besides, federalism may encounter obstructs when acting on national issues due to various interests of states, whereas centralized unitary government determines all regional, national, and international affairs without any disruption (Perera 2020).

In the health field, the federal government may have the financing responsibility while delivering and regulating health services may be the provincial or local responsibility. However, various responsibilities may not be clearly distributed across different levels of government. This ambiguity creates challenges for the coordination of public health responses among different levels of government, and gaps and duplications in carrying out public health responsibilities become potential problems (Office of the Auditor General of Canada 1999). In unitary government, coordination of public health responses and conflicts across different levels of government does not have the challenges of federal forms, while there may be the organizational and managerial problems (Yeoh 2015).

The government plays the dominant role in health governance system no matter of federal or unitary government. The traditional roles of government include regulation, policy development, and resource stewardship (Carlson et al. 2015). Since more actors increasingly engage in health governance, the role of government is expanding and it should also play a role of fostering

coherence, collaboration, and partnership, with civil society and the private sector (St-Pierre 2008).

Regulation. The regulation, as all mechanisms of intentional and unintentional control, sets up the roles, responsibilities, obligations, and functions of the governing body, health officer, and agency staff by the laws and rules (Saltman and Russe 2002). Powers of regulations accrue to the three branches of government, the executive, legislative, and judicial, and can be in forms of administrative decree, legislation, and judicial orders. Many important public health policies are enacted in laws. Regulations on drugs and biological products seek to protect the public, ensuring the safety and quality of these products. The regulatory function may be administered by the government directly or a public agency or the professionals.

Policy development. The development of policies aims to protect, promote, and improve public health. The government should develop the evidence-based policies that support public health agency goals; develop and regularly updating vision, mission, goals, measurable outcomes, and values statements; set short- and long-term priorities and strategic plans; regularly evaluate public health policies to ensure that they are based on the best available evidence for public health practice (Carlson et al. 2015).

Resource stewardship. Resource stewardship aims to assure the availability of adequate resources (legal, financial, human, technological and material) to perform public health services, including resource collection and allocation (Carlson et al. 2015). First, taxation is one way of resource collection, and providing tax incentives or disincentives would influence targeted behaviours (Deber et al. 2004). Provision of tax exemptions for NGOs and charitable foundations has encouraged their growths in civil society, many of whom are engaged in health services. Tax disincentives, such as tobacco tax, have effectively discouraged the smoking behaviour. Second, resource allocation can be cash or in-kind support, and can address the inequity across regions. For examples, contracts and direct payments to NGOs to operate public health programmes, or providing subsidies for vaccines to encourage vaccination uptake.

Coordination and partnership. In modern society, empowered citizens demand more information and influence, and a broad consensus of stakeholders should be achieved (Jakubowski and Saltman 2013). There are strong needs for improved central coordination across stakeholders and decentralizing information and decision-making to citizens. The government

should build and strengthen community partnerships through education and engagement to ensure the collaboration of all relevant stakeholders in promoting and protecting the community's health. A flexible partnership mechanism can be networks of government agencies with the private sector and civil society, for example public–private partnership.

The private sector and its roles in health governance

The private sector discussed in this section refers to all the privately-owned groups and organizations participating in health governance: profit and not-for-profit, formal and informal, domestic and international (Clarke et al. 2019). In most countries, the private sector is generally large and heterogeneous, ranging from medicine stores and individual clinical practitioners to corporate hospital chains and international private insurers (Mackintosh et al. 2016). They play a very important role in health governance.

Almost all countries have mixed health systems, with health products and services provided by both the public and private sectors. However, the private sector may live in the different positions across countries (Mackintosh et al. 2016). The private sector can be dominant in some countries' health systems, such as India and Nigeria, or be complementary in countries like Sri Lanka and Thailand. The private sector may be at the top of a stratified health system in countries like Argentina and South Africa, or the stratified private sectors are shaped by low incomes and public sector characteristics in some African countries. There is also the highly commercialized public sector in some countries.

A large proportion of health expenditure goes to the private sector (Forsberg and Montagu 2014), which reaches nearly 50% in low- and middle-income countries (WHO 2011). The private sector has become an essential actor with abundant resources and powerful influences to deal with health issues. The private sector not only provides lots of jobs but also influences public policy in governance system. They are usually diverse in organization forms (Stallworthy et al. 2014), which allow them rapid response to the diversified demand from people and changing environments. Although the private sector has these advantages, challenges surrounding conflicts of interests are ubiquitous, and there is also a lack of conceptual clarity about the role of the private sector in health systems. The other governing actors engaging the

private sector in healthcare may have different understandings of key terms and concepts of the private sector (WHO 2018).

These resource and characteristics enable the private sector to play multiple roles in the health governance system. These governing roles mainly include services provision, financing, training, and regulations. The provision of health services and products from the private sector occupies an important position in the health system. A review of data from six countries displays a significant trend of greater use of private sector to deliver care in Asia over the last decade (Forsberg and Montagu 2014). With the rapid development of the financing role, the private sector has become a necessary condition for participation in the social insurance programme in some countries such as America. Private corporations and foundations also run many public health programmes around the world, which train health workforce, create job opportunities, and bring the latest technology to local health systems. The private sector has been also increasingly cooperating and authorized by the government for regulations. The Open Working Group Proposal for Sustainable Development emphasizes that the implementation of Sustainable Development Goals depends on global partnerships with the active engagement of governments, civil society, and the private sector (Kunkle et al. 2016). In countries receiving large development assistance from international private organizations, private donors have a strong influence on policymaking and priorities setting of local health systems. Moreover, the private sector and the public sector are influenced by each other (Mackintosh et al. 2016). Public–private partnership therefore creates an integral part of government policies which enables mobilizing and ensuring efficient use of resources, promoting improved delivery systems, and developing localized models (Stallworthy et al. 2014).

Civil society and its roles in health governance

In the absence of agreed definition, civil society is generally understood as the social arena existing between the state and the individual or household to empower social individuals and groups to be involved in the decision-making process and to build platforms for civil actions, which theoretically comprises a diversity of spaces, actors, and institutional forms varying in their degree of formality, autonomy and power (WHO 2001). Civil society creates the space for collective actions and practices based on common

purposes, interests, and values, which is generally believed to be non-state, not-for-profit, and voluntary (The Commission on Social Determinants of Health 2007). Civil society organizations (CSOs) usually include philanthropic foundations, labour unions, non-governmental organizations, business associations, trade unions, professional associations, academic institutions, think tanks, religious groups and organizations, cooperatives and community groups, social movements, lobby groups, women's networks, youth associations, relief organizations, coalitions and advocacy groups (WHO 2002b). Thus, there exists numerous differentiations among civil society actors that primarily determine the heterogeneous and complex nature of it, as well as its various ways of operation (Commission on Social Determinants of Health 2007). In some circumstances, the boundaries between civil society and government or commercial actors can become blurred due to the state or market links to some CSOs (Commission on Social Determinants of Health 2007).

Despite the complexity and heterogeneity of civil society, it plays a prominent role in health governance, which includes four aspects: shaping policies and standards; health services provision; strengthening the health system; community mobilization.

Shaping policies and standards. Civil society has engaged in developing health policies and standards through advocating the stakeholders and the public, creating dialogues with policymakers, or promote the legalization of health policies (WHO 2001). Civil society has mastered and applied the skills of communicating with governments and stakeholders, such as using the clear and strong languages the policymakers preferred, which remarkably promotes their engagements in health policymaking. The policies and standards in which the civil society has been involved cover various health issues such as prices and access to drugs, breast milk substitutes, tobacco control, environment, and policies or standards in debates (WHO 2001). Besides, civil society also provides new health evidences to policymakers to support or oppose some health policies and standards through conducting or supporting a great number of researches (Lencucha et al. 2011). Civil society enables the participation of communities in health policymaking, and also represents the voices and perspectives of some excluded or marginalized social groups.

Health services provision. Civil society has had a long history of providing health services and has become more involved in recent years. Civil society

provides services to targeted population in response to their health needs and adapted to local conditions (WHO 2002b). Civil society focuses on the vulnerable or disadvantaged groups poorly served by government health facilities and provides more cost-effective health services to them (WHO 2001). Civil society also assists governments in major diseases control programmes, treatment campaigns, and health resources distribution, and is playing a more effective role in facilitating the interaction among governments, communities, and clients. Besides, civil society supplies a range of technical supports for health services provision. It is worth noting that civil society carries competitive advantages in health services provision when they own the access to health resources that is not available to government.

Strengthening the health system. There is an increasing number of CSOs contributing to support governments to strengthen local health systems. In many developing countries, CSOs have provided assistances and shared expertise to support Ministries of Health (MOH) to strengthen health systems at all levels (Barbazza and Tello 2014). To build strong partnerships with local MOHs, many CSOs have tailored their actions to the goals and strategic plans of local authorities. Meanwhile, CSOs also focus on the communication, transparency, and flexibility when administrating and implementing joint programmes with local authorities for health systems strengthening. Health workforce and infrastructure are the two key areas that civil society have engaged in for strengthening health system.

Community mobilization. CSOs are effective channels of communications between individuals and business and the state, and promote the participation of community on health campaigns and actions (Van der Rijt and Pang Pangestu 2015). Civil society raises the community preferences in public resources allocation and management that better reflect community needs.

Tools of health governance

Health governance must be equipped with applicable tools to enable the governance function and to steer the health system toward defined goals. Tools to govern are central to carrying out the health governance function and enabling relationships among actors. Tools of governance can be classified into nine governance dimensions: (1) Control mechanisms and

accountability; (2) Regulation; (3) Participation of stakeholder and consensus in decision-making; (4) Formulating policy and strategic direction; (5) Organizational adequacy or system design (capacity building); (6) Generating information and intelligence; (7) Partnerships, coordination, and collaboration; (8) Engagement of community; (9) Communication and transparency (Barbazza and Tello 2014). The first six dimensions mainly involve to play the subfunction of health governance, and the last three dimensions involve to enable relationships or networks among actors. Table 6.8.2 described the dimensions and tools of health governance.

Table 6.8.2 Dimensions and tools of health governance

Dimensions	Tools of health governance
Control mechanisms and accountability	• Governance structure • Fair competitive elections, systems of judicial redress, procedures to combat corruption • Regular internal/external audits • Performance-based contracts
Regulation	• Laws and declarations • Rules, procedures, decrees • Codes of conduct • Performance standards, accreditation, licensing/certification • Statutory bodies (e.g. national regulatory agency)
Participation of stakeholder and consensus in decision-making	• Sector-wide approaches • Consensus conferences • Forma consultations in drafting legislation • Stakeholder identification and voice • User-friendly complaint system • National health forums • Open meetings, public workshops, national forums, citizen advisory committees, citizen juries • Satisfaction surveys, feedback, responsiveness
Formulating policy and strategic direction	• National health plan • National targets, goals, and performance measures • Multiyear strategic plan • Policies, operational guidelines, training manuals, protocols • Medium-term expenditure frameworks and national health accounts • Targets-and priority-setting techniques

Dimensions	Tools of health governance
Organizational adequacy or system design (capacity building)	• Implementation or annual operational plan • Monitoring and evaluation plan • Training/personnel manuals • Mission/vision statement, organizational chart • On-the-spot training measures • Peer-to-peer learning programmes
Generating information and intelligence	• Project evaluation • Conducting research • Commissioned reports by independent researchers • Interministerial fact-finding commissions or taskforce • Parliamentary investigations • Periodic audits, public expenditure, and performance reviews • Internationally recognized tools for conducting a situation analysis • Geographic information systems, or health needs assessment
Partnerships, coordination, and collaboration	• Interministerial and interdepartmental committees • Professional advisory committees • Public–private task force • Partnerships with civil society or agencies • Integrated budgets and accounting, co-funding arrangements, or earmarked grants • Cross-cutting information education systems • Common workforce training, qualifications, support, and advice • Ad-hoc, initiative-specific committees
Community engagement	• Outreach, consult, involve, collaborate, share leadership
Communication and transparency	• Watch dog committees • Releasing financial or performance information to providers and the public • Inspectorates, fact-finding commissions • Opportunities for public consultation or feedback • Mass media • Conflict-of-interest boards

Reproduced from Barbazza E, Tello JE. (2014). A review of health governance: definitions, dimensions and tools to govern. *Health Policy*, 116(1), 1–11 with permission from Elsevier.

Accountability means that decision-makers in government, the private sector, and CSOs involved in health are accountable to the public, as well as to institutional stakeholders (Siddiqi et al. 2009). Accountability tools include governance structure, grievance redress mechanisms, regular internal/external audits, and performance-based contracts. Regulation means that legal frameworks pertaining to health should be fair and enforced impartially to support strategies, goals, and priorities. Regulation tools involve legislative process; interpretation of legislation to regulation and policy; enforcement of laws, and regulations (Siddiqi et al. 2009).

Participation and consensus involve generating dialogue and receiving feedback among various actors (Barbazza and Tello 2014). The participation and consensus tools include sector-wide approaches, national health forums, open meetings, citizen juries, and consensus conferences, among health governance actors.

Every organizations should formulate policy and strategic direction, and ensure organization adequacy or the appropriateness of the system design, through capacity building and various training programmes (Barbazza and Tello 2014). Intelligence and information are essential to provide evidence for informed decisions that influences the behaviour of different interest groups. Tools to generate intelligence and information include project evaluation, conducting research, public expenditure, and performance reviews.

Partnerships, coordination, and collaboration should be established across government, civil society, and the private sector, to facilitate action which improves health and to generate support for health policies (Barbazza and Tello 2014). Coordination can enable interactions that are supportive of the joint management of activities through formal processes, whereas collaboration is established on shared values and goals but with more informal interaction.

Community engagement is the process of working collaboratively with and through groups of people affiliated by geographic proximity, special interest, or similar situations to address issues affecting public health (CDC 1997). It can achieve social mobilization to recruit, mobilize, and motivate individuals and groups to actively participate. Communication with the public is intended to simply allow dialogue and improve the exchange of information,

and the free flow of information can improve transparency. And tools in the communication and transparency dimension include the releasing financial or performance information to providers and the public, mass media, and opportunities for public consultation or feedback (Demand for Good Governance Team 2010).

Levels of health governance

Health governance occurs at four levels: (1) supranational or global level; (2) national or local levels; (3) organization level; (4) programme level. The governing actors can play health governance functions at each of the levels to collectively pursue health goals, and the mechanisms and governance at each level need to be examined. This section will introduce the different levels of health governance: supranational health governance, national health governance, governance of public health organizations, and management of public health programmes.

Supranational health governance

Health governance at supranational level is distinct from governance at other levels: there is no overarching political authority at the global level (Frenk and Moon 2013). The traditional tools for health governance at national level, such like democratic decision-making procedure or regular enforcement of laws and regulations, are absent at the global level (Lee and Kamradt-Scott 2014). The global nature of health determinants restricts the capacity of national governments to address them on their own, and therefore the international and global strategies are required. The global trends, such like changing population demographics, financial uncertainty, deepening social inequity, outbreak of new infectious diseases, further call for the effectively collective action, agreement and enforcement on rules, and achievement of policy coherence to address global health affairs (McInnes et al. 2012). Therefore, more actors gradually join in the actions of global health governance to improve health, along with the rules and norms governing their interactions in the new reality of pluralism (Szlezák et al. 2010).

The primary actors involving in the supranational health governance include: United Nations agencies; national governments; multilateral development

banks; global health initiatives; philanthropic organizations; and CSOs (Frenk and Moon 2013; Ng and Ruger 2014). Table 6.8.3 describes the primary actors in the health governance at supranational level, and gives some examples.

Table 6.8.3 Primary actors in health governance at supranational level

Type of actors	Examples
United Nations system	World Health Organization
	United Nations Drug Control Programme
	United Nations Population Fund
National governments	Ministries of health
Multilateral development banks	World Bank
Global health initiatives	GAVI Alliance
Philanthropic organizations	Bill and Melinda Gates Foundation
Civil society organizations	International Diabetes Federation

WHO is established by UN as 'the directing and coordinating authority on international health work' (WHO 2013a). WHO is developed based on the universal membership of all recognized nation states, who play the central role in global health system to produce profound influences on a range of difference actors to enable them to contribute more effectively to the well-beings of all people (WHO 2013a). WHO operates as a scientifically grounded, technically focused institution guided by a humanitarian ethic that viewed health as central rather than basing on treaty-based value (WHO 2013a; Ng and Ruger 2014). WHO has gain great achievements during the last decades. However, there exists consensus that WHO, as the sole authority on global health, is unable to provide sufficient leadership for global health and effectively respond to vital global health threats. It is imperative for WHO to reform its architecture and framework to regain the capacity and authority on global health. Also important to the supranational health governance system are other UN multilateral agencies. The World

Bank has superior resources in health development, financing, and policy advising of the main global multilateral agenda (Ruger 2005).

The national governments are still considered as the core player to take ultimate responsibility in global health governance affairs, even in low- and middle-income countries, since they decide what is implemented domestically and internationally and provide support to organizations like WHO (Kohlmorgen 2005; Fidler 2010). Moreover, supranational health governance significantly depends on the capacity of nation states (Gostin and Mok 2009). National governments coordinate their responses to global health challenges through various mechanisms (Frenk and Moon 2013). Although rich and powerful states could set priorities in global policies and initiatives (Ng and Ruger 2014), emerging countries such as China and Brazil, increasingly engaged in global health governance through financial and technical assistances (Ruger and Ng 2010).

In addition to the well-established and long-standing actors in supranational health governance, new players present increasing engagements in supranational programmes and initiatives, producing both competition and cooperation among old actors (Kohlmorgen 2005; Fidler 2010). The Global Alliance for Vaccines and Immunization (GAVI) play a vital part in functioning global markets of health commodities for target diseases control (McColl 2008). Philanthropic organizations and CSOs potentially become the essential partners on global health activities due to their organizational flexibility, cost-effectiveness, and access to communities, especially in remote and difficult areas (Ollila 2005; Doyle and Patel 2008; Kruk 2012). The proportion of development assistance for health channelled via UN agencies and development banks has decreased, while global health initiatives, NGOs and the private sector have accounted for an increasing share of development assistance (Ravishankar et al. 2009). The significance of the changes demonstrates the fact that these UN agencies have to compete with NGOs for a share of the voluntary contributions.

The health governance at supranational level carry out three main functions to achieve health goals. The first function is to provide overall strategic direction to the global health system so that all other functions can be implemented sufficiently (Frenk and Moon 2013). The second function is to mobilize global solidarity including financial development, technical cooperation, and humanitarian assistance for achieving global agenda (Frenk and Moon 2013). The last function is to generate global public goods in a

variety of forms including research and development, standards, and guidelines, and comparative evidences and analysis (Stuckler and Mckee 2008; Frenk and Moon 2013). Good global governance for health is considered to demonstrate the following characteristics: ‘effectiveness, equity, and efficiency in achieving outcomes, as well as credibility and legitimacy in decision-making procedures’ (Koplan et al. 2009; Frenk and Moon 2013).

However, there exists three challenges that impede the global governance for health. First of all, in the absence of a clear leadership of global health, it turns to be very difficult for individual actors to achieve the sustainable agreement on their respective contributions in response to the common health threats (Koplan et al. 2009; Kickbusch 2016). Given the diversity of the challenges in global health and the growing number of actors, the regime of supranational health governance is complex and even competing when engage multiple players addressing different problems through diverse principles and processes (Stuckler and Mckee 2008; WHO 2013b). As a result, the collective actions and works at global level are hampered, which become more straitened on a circumstance of uneven distribution of health resources and the conflict interests of various actors. Second, health is increasingly influenced by the governance in other global arenas, such as trade, migration, environment, education. One major consequence of these growing recognition of the connection between health and many other areas of social and economic policy is a growing demand for intergovernmental, rather than purely technical processes, in order to reach durable and inclusive agreements (Fidler 2010). However, these policymaking arenas across multiple sectors have not sufficiently taken health concerns into account (Frenk and Moon 2013). Last, there is in the absent of definite mechanism for the accountability of non-state actors for the global impacts of their actions (Gostin and Mok 2009). The growing competition between NGOs and local service providers for funding and human resources could make the host country more difficult to provide sustainable services (Weishaar et al. 2012). It is feared that the encroachment of non-state actors will hamper the ownership and control of local actors on health governance.

As discussed, the global governance challenges will continue to weaken the capacities and efforts to deal with the complicated and critical global health issues. Therefore, it becomes urgent to take more endeavours to address the structural governance challenges. Also innovative attempts for developing

more inclusive governance for health at global level are required. The involved global actors should make new arrangements on governance to enable it more effective, accountable, and efficient. Meanwhile, more assessments on the past experiences of governance reform are needed through rigorous and systematic analysis and researches.

National health governance

The social and environmental determinants of health are mostly beyond the health sector and health policies, and thus the effect on health should be considered across sectors. Up to date, the national health governance has experienced the evolution of three waves according to the expansion of health policy—from intersectoral action, to healthy public policy, and to the Health in All Policies (Kickbusch and Gleicher 2012). The WHO defines Health in All Policies as 'an approach to public policies across sectors that systematically takes into account the health implications of decisions, seeks synergies, and avoids harmful health impacts in order to improve population health and health equity' (WHO 2014). It improves accountability of policymakers for health impacts at all levels of policymaking, and needs joint action of health and non-health sectors, of public and private actors and of citizens for public health.

WHO points out that the public health agenda has not been given the priority in some countries, and this has been attributed to insufficient political commitment, marginalization of health department in government, and the limited capacity of public health to influence policymakers (WHO 2002a). National health governance needs strong government commitment to public health, and all administrative departments of government work together to improve public health. WHO has characterized the distinguishing features of national health governance as follows: 'is multisectoral and professional; combines knowledge and action; has epidemiology at its core; is influential across all health determinants; involves public health reporting, leading to health strategy development; communicates with politicians, professionals, and the public; is influential organizationally and financially; lies at the heart of the civic society' (WHO 2002a).

The Ministry of Health assumes responsibility for public health, through the following channels: formulating health policies; developing strategies; assessing health needs and determining priorities; defining regulatory

frameworks; commissioning, monitoring, and evaluating public health programmes; and collecting and seeking health data, information, and intelligence. But many social determinants of health are not in the scope of the health ministry. All related ministries, such as ministries of finance and social security, need to collaborate to improve public health, and the cross-departmental coordination should be in place to achieve the health in all policy.

To achieve the Health in All Policies, National government should conduct the development of Health in All Policies; the ministerial leadership role in influencing other relevant government ministries; integrated policymaking mechanisms; and horizontal integration at local, provincial, national, and supranational levels. Local government should implement a national policy framework, organizational changes for coordination of providers and networks, and funding measures to enable integrated provision of public health.

In addition, horizontal mechanisms are necessary to coordinate functions and relationships among government agencies, the private sector, and civil society (Yeoh 2015). Because public engagement in decision-making is useful to design and deliver the localized health services, community governance has become an important mechanism (Pillora and McKinlay 2011). Community governance requires public engagement in consultation and decision-making processes, and strong civic leadership to mobilize communities and related stakeholders. Community governance and public engagement are particularly applicable for disadvantaged communities and to implement public health programmes.

Governance of public health organizations

The governance of organizations can be defined as ‘a set of processes and tools related to decision-making in steering the totality of institutional activity, influencing most major aspects of organizational behaviour and recognizing the complex relationships between multiple stakeholders’ (Saltman and Duran 2015). As an essential part of organization leadership, governance pertains to decision-making in an organization—what positions have authority or control over specific decisions. Governance refers to the management and controls within organizations to ensure the protection of stakeholder’s interests. Effective governance in public health organizations

is vitally important for programmes implementation to improve the human well-being.

In practice, public health organizations are diverse in the size, organizational formality, and legal incorporation status. Different public health organizations have their own governance systems and operation patterns. The legally established organizations have usually formal board structures, policies, and procedures, whereas the majority of community groups and networks have no formal governance structure and they act collectively to achieve the common objective.

Some theories are useful to guide the governance of public health organizations no matter their organizational formality. The theories are the agency theory, stewardship theory, resource dependency theory, and stakeholder theory (Chambers et al. 2012). Agency theory (Keasey et al. 1997) assumes that the owners and managers of an organization would have different interests. According to this theory, the main function of the board is to control managers through the monitoring and compliance. The boards can reduce the problems of managers pursuing their own interests or performing poorly through monitoring processes such as external audit and reporting requirements. In public health organizations, the public objectives are at risk of managers pursuing their own interests, and the board plays a key role to monitor managers and ensure their compliance in the organization objectives. Stewardship theory (Muth and Donaldson 1998) has opposite assumptions to agency theory, which assumes that managers and owners share a common agenda and work ‘side-by-side’. Hence, the main role of the board is not monitoring managers to ensure compliance or conformance, but to develop strategy to improve organizational performance. For public health organizations, the theory holds some intuitive appeal, with trustees (the public) and managers both sublimating their personal interests in pursuit of an organization’s philanthropic goal.

In addition, resource dependency theory (Pfeffer and Salancik 1978) minimizes the uncertainty caused by external environmental factors, and it considers organizations as interdependent with their environment. The main function of the board is to maintain good relations with key external stakeholders to ensure the flow of resources into and from the organization, and to help the organization respond well to external change. For public health organizations, relationships with funders, media, and various local representatives, play an important part in securing long-term funding and

other forms of support. Stakeholder theory (Hung 1998) assumes that organizations should be responsible to a range of groups in society other than just the owners of organizations. It is expected that organizations are more likely to respond to the broader social interests than the narrow interests of only one group. Therefore, different stakeholders are incorporated on boards, and boards play a political role to negotiate and resolve the potentially conflicting interests of different stakeholder groups in order to determine the objectives of the organization. In public health organizations, the principle of stakeholder involvement is less controversial, and the theory legitimates the boards because stakeholders represent society and users the organizations are serving.

In the governance of public health organizations, boards play a crucial role to ensure public accountability and good performance (Cornforth 2003). In general, there are two main functions of the boards: conformance and performance. Conformance refers to external accountability, which contains compliance with legal and regulatory requirements, accountability to shareholders or other stakeholders, supervision of management through oversight, monitoring performance, and ensuring adequate internal controls. In contrast, performance is about driving the organization forward to better achieve its mission and goals, which consists of policy formulation and strategic thinking (Garratt 1996).

Management of public health programmes

According to the WHO, public health programmes are managerial units with objectives, directors, managers, lines of command, budgets, and action plans, which are an integral part of health systems (Blas and Kurup 2010). Programmes targeting any of the determinants of health can be defined as health programmes. Health programmes not only focus on some aspect of the physical environments in which people live and work, but also address issues about people's behaviours, their biology, the social factors that affect them, or the health services they receive (Longest 2014). Public health programmes can range in scale, scope, and complexity, from local programmes in communities to global programmes, and the effective public health programmes need to be adapted to the local and broader contexts.

There are three different types of work in health programmes: direct work, support work, and management work. Management work involves

establishing the mission and objectives of a health programme and creating conditions for direct work and support work to accomplish that mission and fulfil the objectives. Programme management can be defined as 'the activities through which the mission and objectives of a programme are established and pursued by means of various processes using human and other resources' (Longest 2014). Programme management comprises planning, implementation, and evaluation, and specially, it contains two sets of activities: one is core activities, including developing/strategizing, designing, and leading; the other is facilitative activities, including decision-making, communicating, managing quality, marketing, and evaluating. Developing/strategizing means developing a new programme or improving services or products of ongoing programmes, which can result in programme adjustment decision and help the programmes adapt to continuously changing external environments (Ginter et al. 2013). Designing is to establish and change the intentional patterns of relationships among human and other resources within a programme, or the relationship of the programme to its external environment. Leading is the work of influencing other participants to contribute to the programme goals. Decision-making is simply making a choice between two or more alternatives, and continual decision-making is necessary due to the dynamic context. Communicating means that senders conveying ideas, intentions, and information to receivers, and the senders and receivers can be individuals, groups, or organizations. Managing quality is important to people who work in programmes or receive services, and there are six aims for quality improvement according to Institute of Medicine: safe, effective, patient centred, timely, efficient, and equitable (Institute of Medicine 2001). Marketing is a means bringing voluntary exchanges with people outside a programme to cross the boundary between a programme and its external environment. Evaluating is an essential programme activity to provide evidence of improving programmes (Longest 2014).

The environments of implementing public health programmes may change continuously, and implementation strategies need to be modified accordingly. Implementation science has been developed to ensure the implementation and external validity of programmes. In public health field, a widely used framework of implementation science is the Reach, Effectiveness, Adoption, Implementation, and Maintenance (RE-AIM) framework (Gaglio et al. 2013). Reach is the proportion of individuals who

are willing to participate in public health programme. Effectiveness is the impact of public health programme on outcomes, including potential negative effects, quality of life, and economic outcomes. Adoption is the proportion of settings and intervention agents who are willing to initiate a programme. Implementation refers to the intervention agents' fidelity to the various elements of an intervention's protocol, which includes consistency of delivery as intended and the time and cost of the intervention. Maintenance is the extent to which a programme becomes institutionalized or part of the routine organizational practices. The implementation evaluation would help to improve the practice of public health programmes and promote them in a wider context.

Programme evaluation is an essential complement of public health programme management, and it gathers necessary information to improve and account for programme effectiveness. Evaluation needs to be integrated in routine programme practice, and involves all programme staff and stakeholders, not only evaluation experts. Center for Disease Control and Prevention of the US developed a Framework for Programme Evaluation in Public Health, to guide effective programme evaluation (Center for Disease Control and Prevention 1999). The framework is composed of six evaluation steps, and a set of 30 standards to assess the quality of evaluation activities. The six evaluation steps are to engage stakeholders, describe the programme, focus evaluation design, gather credible evidence, justify conclusions, and ensure use and share lessons. These steps are often completed in a cycle, and earlier steps provide the basis for subsequent progress. The first step is engaging stakeholders involved in or influenced by the programme. Stakeholders should be engaged, and their perspectives or concern are essential for programme implementation and outcomes. The second step is sufficiently describing the programme, which covers need, expected effects, activities, resources, stage, context, and logic model about the programme. The third step is focusing the evaluation design, and the evaluation must focus on the issues of greatest concern to stakeholders. The fourth step is gathering credible evidence, and indicators, sources, quality, quantity, and logistics are main aspects of evidence that can affect perceived credibility. The fifth step is justifying conclusions on the basis of evidence, which includes standards, analysis and synthesis, interpretation, judgement, and recommendations. The last step is ensuring use and sharing lessons learned, and translate lessons into informed decisions and appropriate actions. Five

elements are critical for ensuring use of an evaluation, which consist of design, preparation, feedback, follow-up, and dissemination. In addition, the second part of the framework is 30 standards to assess evaluation quality, and the 30 standards can be organized into four groups: utility, feasibility, propriety, and accuracy. Utility standards aim to satisfy evaluation users' needs, feasibility standards ensure the viability and practicability of the evaluation, propriety represents the ethics of the evaluation, and accuracy ensures the correctness of the evaluation findings (Lavinghouze and Snyder 2013).

Assessment framework of health governance

Major actors of health governance have increasingly paid attention to the assessment of health governance to ensure effective governance for better health outcomes (Barbazza and Tello 2014). Several institutions have developed principles of effective governance, and this section focuses on the five widely used frameworks of assessing health governance. They are the WHO's domains of stewardship (WHO 2000), the World Bank's six fundamental aspects of governance (Kaufmann et al. 1999), the UNDP's five governance principles (UNDP 1997), the WHO Eastern Mediterranean Regional Office's ten governance principles (Siddiqi et al. 2009), and the WHO Regional Office for Europe's TAPIC framework (Greer et al. 2016).

The WHO's domains of stewardship. The World Health Report 2000 identified four key functions which determine the performance and outcomes of health system—financing, resource generation, service delivery, and stewardship (WHO 2000). Stewardship focuses on the governments' role to take responsibility for guiding health system and improving population health. The objects of stewardship involve actors in provision and financing; the performance of stewards themselves; users and consumers, and non-health systems that affect health. WHO proposed six domains of stewardship: 'generation of intelligence, formulating strategic policy direction, ensuring tools for implementation, building partnership, ensuring a fit between policy objectives and organization structure and culture, and ensuring accountability' (WHO 2000). As a concretely applicable function for health system strengthening, stewardship, or responsibility of health sectors has been specifically allocated by legislation and law. However, there remain some gaps which call for ongoing collaboration among relevant

sectors, such as political and power dynamics, effective incentives for stewardship, good governance agenda and reform, as well as domestic resource mobilization (Brinkerhoff et al. 2019).

The World Bank's six fundamental aspects of governance. The World Bank measured six fundamental aspects of governance by subjective perceptions (Kaufmann et al. 1999). It defined governance broadly as 'the traditions and institutions by which authority in a country is exercised'. This includes '(1) the process by which governments are selected, monitored and replaced; (2) the capacity of the government to effectively formulate and implement sound policies; and (3) the respect of citizens and the state for the institutions that govern economic and social interactions among them' (Kaufmann et al. 1999). They are assessed from six basic governance aspects: voice and accountability, political instability and violence, government effectiveness, regulatory burden, rule of law, and graft. 'Voice and accountability' measures political process, civil participation, and political rights, and 'political instability and violence' refers to the possibility that the current government could be overthrown by unconstitutional and/or violent movements. 'Government effectiveness' focuses on the prerequisites for producing and implementing good policies, and 'regulatory burden' considers more about the policies themselves. 'Rule of law' refers to the rules of society being trusted and abided by agents, and 'graft' measures a straightforward focus of corruption. The cross-county comparison study provided a strong causal relationship from better governance to better development outcomes.

The UNDP's five governance principles. The UNDP defined governance as the exercise of political, economic, and administrative authority in the management of a country's affairs at all levels (UNDP 1997). The UNDP identified five principles for effective governance. The first principle is 'legitimacy and voice' through participation and consensus orientation to emphasize the voice of all individuals in decision-making and agreement based on the best interest. The second principle is 'direction' through strategic vision of leaders and the public, which is grounded on long-term perspective and understanding the complex historical and cultural context. The principle 'performance' refers to responsiveness along with effectiveness and efficiency, and 'accountability' incorporates accountability of policymakers, private sectors, and civil organizations with transparency built on free and direct information access. The last principle 'fairness'

means equity and inclusiveness together with rule of law to be fair and enforced equally.

The WHO Eastern Mediterranean Regional Office's ten governance principles. The WHO Eastern Mediterranean Regional Office developed an assessment framework as a gateway for effective governance of health systems by incorporating the three aforementioned frameworks (Siddiqi et al. 2009). They adapted the UNDP governance principles, and incorporated health-specific aspects of governance. Finally, they proposed ten governance principles for assessing health governance: strategic vision, participation and consensus orientation, rule of law, transparency, responsiveness, equity and inclusiveness, effectiveness and efficiency, accountability, intelligence and information, and ethics. Specifically, they summarized intelligence as an independent principle because of the unique nature of information asymmetry in health, which requires more intelligence to assess its creation, flow, and use. The ten principles have been disaggregated into 63 broad questions which relate to context, description, process, and outcomes of governance. The assessment framework has been applied in many low- and middle-income countries.

The WHO Regional Office for Europe's TAPIC framework. The WHO Regional Office for Europe proposed a robust framework with five key aspects of governance named TAPIC in 2016 (Greer et al. 2016). The TAPIC framework includes the five mutually exclusive elements of governance—transparency, accountability, participation, integrity, and capacity—which are inducted from a variety of literature. None of these elements are essentially a good thing, for they can be excessive or wrongly implemented. For each element, it can relate to a series of different policy and administrative tools, and be properties of both systems and individual organizations. The five elements are also mutually reinforcing on a broader scale. Improvement of one attribute could link to the development of other aspects as a result. This framework highlights the effectiveness of networks, the complexity of accountability, and the difficulty of distinguishing among different types of governance. But even if policies made in the name of transparency, accountability or participation does not guarantee the effectiveness or productivity, and in some cases can conflict with other attributes such as flexibility or capacity.

The five governance elements of TAPIC framework are introduced as follows. First, transparency means that institutions inform the public and

other actors of both upcoming decisions and decisions that have been made, and of the process by and the rationale on which decisions are being made. Transparency mechanisms include ‘watch dog committees, inspectorates, regular reporting, performance assessment and reporting, Freedom of Information legislation, and clear and useful public information’ (Greer et al. 2016). Second, accountability means that one actor must inform and explain the others of decisions, and can be mandated or sanctioned. The mechanisms of accountability include ‘contracts, other financial mechanisms such as pay for performance, laws that specify objectives, reporting and mechanisms, organizational separation, competitive bidding, conflict of interest policies, standards, codes of conduct, regulation, delegated regulation, standards, codes of conduct, and choice mechanisms that create horizontal accountability by letting marginal users vote with their fees’. Third, participation means that relevant parties acquire a meaningful stake and power in the process of decision-making in the institution. Participation mechanisms include ‘stakeholder forums, elections, consultations, appointed representatives, choice mechanisms, legal remedies, advisory committees, surveys, partnerships, joint budgets, joint workforce, etc., and more radically democratic innovations, such as participatory budgeting’. Fourth, integrity means that procedure of representation, decision-making and implement should be clearly specified, and roles and responsibilities should be clearly allocated. Integrity mechanisms include ‘solid and well-rewarded internal career trajectories, internal audit and censor, personnel policies, legislative mandate, budget, audit, procedures (e.g. document management, board behaviour, minuting meetings), and organizational roles and purposes’. Fifth, capacity refers to the ability to propose the policy which is fitting to resources in pursuit of goals. Mechanisms to improve policy capacity include ‘intelligence on performance, intelligence on process, research/analysis capacity, staff training, to improve their technical policy capacity, hiring procedures, to improve the quality of the policy bureaucracy, procedures to incorporate specialist advice into policy formulation and recommendations, and good purchase/make decisions’ (Greer et al. 2016).

Overall, effective health governance is a hybrid of principles and a dynamic process beyond specific policies, sector, and actors. The application of assessment frameworks of health governance requires the comprehensive understanding of the governance concept and the changing context around it.

Cases for health governance

Case for supranational health governance

Ebola epidemic in West Africa epitomes global health governance. Due to high population mobility across borders and international travel, this epidemic smouldered from sporadic rural areas to densely populated urban areas, from Guinea into Liberia, Siera Leone, and Nigeria. Ebola epidemic varied from previous local outbreaks in Africa, and led to bilateral or multilateral cooperation, instruction, coordination, and regulation (WHO 2015). Many types of actors responded to this epidemic together, including UN systems (WHO, the UN, and UNICEF), international CSOs, foreign governments, local governments, and communities, etc. (Table 6.8.4).

Table 6.8.4 Actors and their functions of governance in the Ebola epidemic of West Africa

Actors		Functions of governance
UN systems	WHO	Leadership and coordination, technique/laboratory/expert support, 'Roadmap' construction, case tracing, epidemic report, surveillance system building etc.
	UN	Health mission construction, priority airline support
	UNICEF	Early isolation and safe burial promotion, primary health construction
	US, UK, etc.	Surveillance system construction, treatment centre construction, worldwide epidemic prevention
International civil society organizations	Médecins Sans Frontières	Clinical support, psychological support, case tracing
	World Food Programme	Food, nutrition, and logistics support
	InterAction	Overall complement functions: health centres construction, clinical/psychological support, safety body disposal, etc.
Foreign governments Local governments	Ministry of Health, etc.	Case tracing, clinical services, epidemic investigation, and response
Community	Religious leaders, etc.	Health promotion

UN systems

As the most significant coordinator and supervisor, WHO with other partners engaged in case detection and management, and contact tracing from the very beginning. WHO prepared for case diagnosis by mobilizing its collaboration laboratories. Besides experimental support, WHO sent several multidisciplinary teams, which are specialised in epidemiology, laboratory services, risk communication, and social mobilization, to work with local staffs and community leaders. Meanwhile, WHO publicized the number of mortality and morbidity, and conveyed related concerns and advocations at

diverse meetings and assemblies, keeping the rest of the world aware of the latest epidemic situation. Furthermore, with the WHO leadership and more extricate and integrated partnership (local governments, partners, Ebola survivors, representative of airline and mining companies, and financial donors etc.), key priorities were identified to ensure the primary objectives and imperative moves for diverse partnerships. In addition to technical guidance and material support, WHO also urged governments to take actions on national laws and policy to control the outbreaks. An emergency committee was also organized to assess the Ebola situation under the provisions in the International Health Regulations with a declaration to further arouse international alert to the potential international spread. By launching 'Roadmap' for response to the epidemic, many actors (including US/UK government, UN Security council and NGOs) were coordinated and focused on the most urgent needs. WHO supported the three most affected countries to rebuild their disease surveillance and health service systems after 2014.

In 2014, UN declared its first-ever emergency health mission called UN Mission for Ebola Emergency Response (UNMEER) (UN 2015) as a temporary measure to meet immediate needs of the fight against Ebola. As a supplement of WHO, the UNMEER managers remain in the countries under the oversight of WHO and with the support of UN Resident Coordinators and UN country teams, quickly founding an air bridge to facilitate the flow of staff and materials and devising response plans based on 'Roadmap'. UNICEF devised Humanitarian Action for Children (HAC) framework to reduce transmission of Ebola and strengthen primary healthcare system (UNICEF 2016).

International civil society organizations

Under the WHO's leadership, international CSOs play various and harmonious roles in fighting against Ebola epidemic. Ahead of WHO, Médecins Sans Frontières (MSF) was one of the first international organizations involved in the Ebola outbreaks. Cooperating with WHO, MSF mainly provided services such as psychological support, health promotion, surveillance, and contact tracing. And MSF admitted over 10,000 patients to its Ebola management centres (MSF 2017). IFRC Red Cross teams worked with community elders and religious leaders to publicize fundamental Ebola knowledge by a range of communication channels, which

are useful for curbing unhygienic conventions and achieving early treatment. The World Food Programme (WFP) also provided food, nutrition, transport, and logistics support to people harshly attacked by the outbreak (WFP 2017). Nearly all of the organizations released the efficacy assessment of their responses during Ebola epidemic to ensure transparency and future improvement.

More than 60 NGOs engaged in the responses to the West Africa Ebola outbreak (UN 2014). Aligning with UN systems, NGOs made up for the shortage of health services handled by local governments. Meanwhile, NGOs complimented the role of the UN and other actors to evolve itself into a global coordination structure. In addition to running treatment units and tracing cases, NGOs were also involved in proper burial procedures and outreach programmes. Conducting more than 400 projects in the affected countries, NGOs served as a crucial echo chamber with clear objectives based on joint goals.

Foreign governments

The US and UK governments spared efforts to control Ebola epidemic. In Liberia, US Centre for Disease Control and Prevention, WHO and other partners assisted the MOH at locals to construct a national incident management system to improve coordination of response efforts (Marston et al. 2017). In Sierra Leone, civilian and military personnel and resources from UK bolstered the coordination of national and local responses to Ebola outbreak. These actions also strengthened these countries' future surveillance abilities, and built treatment centres in Liberia and Sierra Leone. Other countries, like China, addressed public health emergencies of international concerns through multilateral channels to keep epidemic from spreading into other parts of the world, and expanded their engagement to cooperation with international actors around the world (Huang 2017).

Local governments and community engagement

West African countries without previous experience of Ebola outbreak, were poorly prepared for this unexpected disease. However, their relatively rapid actions, timely reports, and cooperation with international aid teams prohibited the epidemic to some extent. The head of the Meliandou health post first informed district health officials when five fatal cases were

happening. With the partnership of experts from MSF and WHO, the pathogen has been finally narrowed down to Zaire species, the most lethal virus in the Ebola family within the first 6 months. As the main force of aid teams, local health officials accepted trainings offered by international organizations and NGOs, and assisted international experts to trace cases, treat patients, and disease surveillance, etc. Additionally, local community and religious leaders worked with experts to stress the hazard of high-risk cultural practices and traditional beliefs to promote scientific preventive measures in the general population.

Case for national health governance

Country/region-level health governance is also one significant dimension of governance. When it comes to public health within one country, the country's national strength usually dominates its roles and interactions among various actors. Here, two countries, the US and Ukraine, are taken as cases to illustrate the potential interactions and tools used to conduct national health governance.

Faced with diverse health threats, the US has already constructed ample cooperation and collaboration, involving governments at all levels, NGOs, communities, and even individual residents. Dominated by Health Departments, the public health system, community partners, and workforce all devote material and intelligence to improve national operational capacity, impacting every community programme and public health activity and further leading to better health outcomes, disparity reduction and better preparedness (Center for Disease Control and Prevention 2013). The federal government plays a crucial role in providing leadership (technique supports and formulation of public health goals, etc.), contributing operational and financial resources, financing research and higher education. State Governments and Local Health Departments further specify their duties as laboratory services, disease surveillance, community health assessment and assurance. Meanwhile, other organizations and institutes also serve as indispensable coordinating and complementary roles regarding public health, which includes NGOs, community coalitions, churches, schools, media, community services, business, and so on. Under such circumstance, the government is dominantly in charge of its national health affairs with the assistance of other parts of the society.

However, for countries without strong economic support, they are still able to achieve effective public health governance with systematic governance tools. The Ukrainian Swiss Perinatal Health Programme exemplifies how to balance interests of multisector stakeholders to promote sustainable development of a public health programme (Bodiang 2005).

During the Soviet era, disease prevention in Ukraine followed a centrally designed top-down strategy, which means that most preventive duties were delegated to residents, weighing much more on service's quantity rather than quality. After the Soviet Union collapsed, Ukraine went through a comprehensive reform transition, which also presented in the healthcare system. Meanwhile, Ukraine faced a negative population growth due to relatively high infant mortality rates, reduced life expectancy, health-related behaviour, and HIV epidemic. Stressed by its internal economic depressions and harsh health situation, the Ukrainian Swiss Perinatal Health Programme under the Ukrainian Ministry of Health (MOH), funded by the Swiss Agency for Development and Cooperation (SDC) and with technical assistance from the Swiss Centre for International Health (SCIH) of the Swiss Tropical Institute introduced a special pillar of health promotion in the realm of perinatal health of four of oblasts (Donetsk, Volyn, Ivano Frankyvsk, and Rivne).

At the very beginning, there was neither public fund nor civil engagement in decision-making in Ukraine, due to its previous absolute centrality governance. According to principles addressed in the Ottawa Charter of Health Promotion, approaches should be based on partnership creation and multiple communication strategies. However, due to disagreements between the two main parties (Ukrainian and Swiss), lengthy negotiations were taken. Finally on the basis of long-lasting trust, the MOH agreed to enlarge the range of health promotion actors and to promote a transparent tendering process based on its leadership and supervision position. With bilateral compromises, multisectoral collaborations between governmental and non-governmental actors at decentralized levels were constructed and programmes were handed over to local qualified NGOs, strengthening the local ownership for the modern approach to health promotion. With the foundation of both organization and techniques, local oblast government secured a funding commitment for future sustainable programmes without participants from Swiss.

To ensure the transparency and validity of these programmes, several workshops were launched to define local leading NGOs, topics of each oblast and the procedure of selection criteria. Training workshops were also held and contracts were signed to clarify accountabilities. Evaluations were processed through baseline analysis in order for further adjustments and improvements. With final evaluation workshops, further independent health promotion programmes under the charge of Ukrainian organization were guaranteed. Gains from these programmes laid foundations for future public health programmes.

Case for public–private partnership in public health

Quoting Reich as saying that ‘A public–private partnership (PPP) includes three points. First, the partnership involves at least one private for-profit organisation with at least one not-for-profit organisation. Second, the core partners provide a joint sharing of efforts and of benefits. Finally, partnerships in public health are committed to the creation of social value (improved health), especially for disadvantaged population’ (Reich 2000). PPPs have sprung up based on the global trend of health issues, especially those disturbing low- and middle-income countries. Both non-profit and for-profit organizations figured out their limitations that both of them were incapable of tackling with global health problems on their own.

Addressing the demand of products for combating disease associated with poverty, product development (PD) PPPs were flourished (Widdus and White 2004). The first PD PPP is International AIDS Vaccine Initiative (IAVI) for vaccine development, and Rockefeller Foundation sponsored the first several meetings, which laid the foundation of IAVI. The first drug development PD PPP is Medicines for Malaria Venture (MMV). With the first two organizations founded, Rockefeller Foundation and Bill & Melinda Gates Foundation have played as the major investors in promoting a number of product development ventures.

For PD PPPs, it is pivotal to manage effectively with other companies and institutions. Not like traditional pharmaceutical tycoons, which primarily maintain internal connections in product development, PD PPPs are usually small and implement their work though multilevel cooperation, thus management of interfaces with other actors are greatly significant to PD PPPs’ strategy.

Disease-endemic countries could help build ownership of potential products and also are served as integral components of the products development process, providing the end-user perspective, clinical trial capacity, and assessment. Effective clinical trials require physical sites, ethical review capacity, and the appropriate regulatory bodies to support trials and approve the products for use. Therefore, PD PPPs need to work closely with developing countries without losing their own focus and take measures to raise attention from international organizations like WHO to work together with local governments.

PD PPPs also need to work with other public sectors to ensure the availability of financing, public policy, and infrastructure once the product has been successfully invented. For example, if there are no existing regulatory bodies with sufficient capacity, it will definitely hinder the product's entry into usage. In addition, PD PPPs could also promote product-directed researches in academic institutions by investing related studies, for basic research community is the source of novel drugs, which is critical in product pipeline.

Additionally, private sector, both a source of candidate products and access to technical expertise, is increasingly important actors interacted with PD PPPs. Innovative licensing agreements with private sector is seen as one promising area of opportunity. This method is suitable for situation that the products of the company have both commercial and public health market. In this way, PD PPPs benefit from licensing high-potential products with technical support, while reserving the commercial gaining for the private company. One agreement between International Partnership for Microbicides and Tibotec (affiliated with the Johnson & Johnson companies) is just an example.

Conclusions

Governance is central to improve the health systems performance and achieve UHC. Good governance can enable the effective use of health finances, workforce, medicines, and information to deliver better health services and outcomes. Theories of principal-agent and network governance are introduced to better understand health governance at system level. In health governance system, the government, private sector, and civil society are the governing triangle (actors) to collectively pursue health goals, which

forms a whole-of-society governance approach for health. Each governing actor has its structures, roles in governance, governing tools, and mechanisms for interactions within and between the actors. They play health governance functions at four levels: supranational health governance, national health governance, governance of public health organizations, and management of public health programmes. At each level, tools to govern are central to carry out the health governance function and enable relationships among actors. Tools of governance can be classified into nine governance dimensions: accountability, regulation, participation, and consensus in decision-making, formulating policy/strategic direction, organizational adequacy/system design, generating information/intelligence, partnerships for coordination and collaboration, engagement of community, communication, and transparency. Five assessment frameworks for health governance are further introduced to help achieving effective governance. And several cases of health governance are introduced for better understanding.

ACKNOWLEDGEMENTS

Text extracts from World Health Organization, *Everybody's Business: Strengthening Health Systems to Improve Health Outcomes: WHO's Framework for Action*, World Health Organization, Geneva, Switzerland, Copyright © 2007, reproduced from the World Health Organization, available from http://www.who.int/healthsystems/strategy/everybodys_business.pdf.

Text extracts from World Health Organization, *Public Health Management: World Health Organization—University of Durham meeting report*, World Health Organization, Geneva, Switzerland, Copyright © 2002, reproduced from the World Health Organization, available from http://www.who.int/chp/knowledge/publications/PH_management7.pdf.

Text extracts from Kaufmann, D. et al., *Governance Matters*, Policy Research Working Paper 2196, The World Bank Development Research Group, Macroeconomics and Growth and World Bank Institute Governance, Regulation and Finance, Copyright © 1999, reproduced from the World Bank, available from http://info.worldbank.org/governance/wgi/pdf/govmatters1.pdf.

Text extracts from Yeoh E.K. (2015). Governance and management of public health programmes. In: Detels, R., Gulliford, M.R., Karim, Q.A., and Tan, C.C. (eds.) *Oxford Textbook of Global Public Health* (6th ed.). Oxford: Oxford University Press.

REFERENCES

Abimbola, S. (2020). Health system governance: a triangle of rules. *British Medical Journal Global Health*, 5(8), e003598.

Barbazza, E. and Tello, J.E. (2014). A review of health governance: definitions, dimensions and tools to govern. *Health Policy*, 116(1), 1–11.

Blas, E.B.E. and Kurup, A.S. (2010). *Equity, Social Determinants and Public Health Programmes*. Geneva: World Health Organization.

Bodiang, C.K. (2005). *New Approaches to Health Promotion and Prevention-the Ukrainian Swiss Perinatal Health Programme*. Basel: Swiss Centre for International Health/Swiss Tropical Institute.

Brinkerhoff, D.W. and Bossert, T.J. (2008). *Health Governance: Concepts, Experiences, and Programming Options*. United States Agency for International Development.

Brinkerhoff, D.W. and Bossert, T.J. (2014). Health governance: principal-agent linkages and health system strengthening. *Health Policy and Planning*, 29, 685–93.

Brinkerhoff, D.W., Cross, H.E., Sharma, S., and Williamson, T. (2019). Stewardship and health systems strengthening: an overview. *Public Administration and Development*, 39(1), 4-–10.

Carlson, V., Chilton, M.J., Corso, L.C., and Beitsch, L.M. (2015). Defining the functions of public health governance. *American Journal of Public Health*, 105(Suppl 2), S159–66.

CDC/Agency for Toxic Substances and Disease Registry (ATSDR) Committee on Community Engagement (1997). *Principles of Community Engagement*. Atlanta. [Online] Available at: https://www.atsdr.cdc.gov/communityengagement/

Center for Disease Control and Prevention (1999). *Framework for Program Evaluation in Public Health*. *MMWR Morbidity & Mortality Weekly Report*.

Center for Disease Control and Prevention (2013). *United States Public Health 101*. *Center for Disease Control and Prevention*. [Online] Available at: https://www.cdc.gov/stltpublichealth/docs/usph101.pdf

Chambers, N., et al. (2012). Assessing governance theory and practice in health-care organizations: a survey of UK hospices. *Health Services Management Research*, 25(2), 87–96.

Clarke, D., et al. (2019). The private sector and universal health coverage. *Bulletin of the World Health Organization*, 97, 434–5.

Commission on Social Determinants of Health (2007). *The Civil Society Report*. Geneva: Commission on Social Determinants of Health. [Online] Available at: http://www.who.int/social_determinants/resources/cs_rep_2_7.pdf

Commission on Social Determinants of Health (2008). *Closing the Gap in a Generation: Health Equity Through Action on the Social Determinants of Health. Final report of the Commission on Social Determinants of Health*. Geneva: WHO.

Considine, M. and Lewis, J.M. (2003). Bureaucracy, network, or enterprise? Comparing models of governance in Australia, Britain, the Netherlands, and New Zealand. *Public Administration Review*, 63(2), 131–40.

Cornforth, C. (2003). *The Governance of Public and Non-Profit Organizations*. Routledge. [Online]. Available at: http://oro.open.ac.uk/15872/

Deber, R., Topp, A., and Zakus, D. (2004). *Private Delivery and Public Goals: Mechanisms for Ensuring that Hospitals Meet Public Objectives*. Toronto: University of Toronto.
Demand for Good Governance Team (2010). *Demand for Good Governance in the World Bank: Conceptual Evolution, Frameworks and Activities*. Washington, DC: World Bank Social Development Department.
Dieleman, J.L., et al. (2016). Development assistance for health: past trends, associations, and the future of international financial flows for health. *The Lancet*, 387, 2536–44.
Dodgson, R., Lee, K., and Drager, N. (2002). *Global Health Governance: A Conceptual Review*. Geneva: WHO and London School of Hygiene and Tropical Medicine.
Doyle, C. and Patel, P. (2008). Civil society organisations and global health initiatives: problems of legitimacy. *Social Science and Medicine*, 66(9), 1928–38.
Elazar, D.J. (1997). Contrasting unitary and federal systems. *International Political Science Review*, 18(3), 237–51.
Fidler, D.P. (2010). *The Challenges of Global Health Governance*. *Council on Foreign Relations Working Paper*. [Online] Available at: https://www.cfr.org/report/challenges-global-health-governance
Fisch, J. (2018). The case for effective environmental politics: federalist or unitary state? comparing the cases of Canada, the United States of America, and the people's republic of China. *University of Michigan Journal of Law Reform 777*. [Online] Available at: https://repository.law.umich.edu/mjlr/vol51/iss4/4
Forsberg, B.C. and Montagu, D. (2014). Further advances in knowledge on the role of the private sector in health systems. *Health Policy & Planning*, 29, i1–i3.
Frenk, J. and Moon, S. (2013). Governance challenges in global health. *The New England Journal of Medicine*, 368, 936–42.
Fryatt, R., Bennett, S., and Soucat, A. (2017). Health sector governance: should we be investing more? *British Medical Journal Global Health*, 2(2), e000343.
Gaglio, B., Shoup, J., and Glasgow, R. (2013). The RE-AIM framework: a systematic review of use over time. *American Journal of Public Health*, 103(6), e38–e46.
Gailmard, S. (2012). Accountability and principal-agent models. In: Bovens, M., Goodin, R.E., and Schillemans, T. (eds.) *The Oxford Handbook of Public Accountability*. Oxford: Oxford University Press.
Garratt, B. (1996). *The Fish Rots from the Head: The Crisis in our Boardrooms: Developing the Crucial Skills of the Competent Director*. London: HarperCollins.
Ginter, P.M., et al. (2013). *Strategic Management of Health Care Organizations* (7th ed.) San Francisco, CA: Jossey-Bass.
Gostin, L.O. and Mok, E.A. (2009). Grand challenges in global health governance. *British Medical Bulletin*, 90, 7–18.
Greer, S.L., Wismar, M., and Figueras, J. (2016). *Strengthening Health System Governance: Better Policies, Stronger Performance. European Observatory on Health Systems and Policies Series*. [Online] Available at: https://www.euro.who.int/__data/assets/pdf_file/0004/307939/Strengthening-health-system-governance-better-policies-stronger-performance.pdf?ua=1
Gupta, S., Davoodi, H., and Tiongson, E. (2000). *Corruption and the Provision of Health Care and Education Services*. IMF Working Paper. Washington, DC: International Monetary Fund.
Huang, Y. (2017). China's response to the 2014 Ebola outbreak in West Africa. *Global Challenges*, 1, 1600001.
Hung, H. (1998). A typology of the theories of the roles of governing boards. *Corporate Governance: An International Review*, 6(2), 101–11.
Institute of Medicine (2001). *Crossing the Quality Chasm: A New Health System for the 21st Century*. Washington, DC: National Academies Press.

Institute on Governance (2011). *Governance*. [Online] Available at: http://iog.openconcept.ca/en/about-us/governance

Jakubowski, E. and Saltman, R. (2013). *The Changing National Role in Health System Governance: A Case-Based Study of 11 European Countries and Australia*. WHO Observatory Studies Series, No. 29. [Online] Available at: http://studiorum.org.mk/h2020mkd_upload/upload/The%20Changing%20National%20Role%20in%20Health%20System%20Governance.pdf

Kaufmann, D., Kraay, A., and Zoido-Lobaton, P. (1999). *Governance Matters*. Policy research working paper. Washington, DC: The World Bank.

Keasey, K., Thompson, S., and Wright, M. (1997). The corporate governance problem –competing diagnoses and solutions. In: Keasey, K., Thompson, S., and Wright, M. (eds.) *Corporate Governance: Economic and Financial Issues*. Oxford: Oxford University Press, pp. 1–17.

Kickbusch, I. (2016). Global health governance challenges 2016—are we ready?. *International Journal of Health Policy Management*, 5(6), 349–53.

Kickbusch, I. and Gleicher, D. (2012). *Governance for Health in the 21st Century*. Copenhagen: WHO Regional Office for Europe.

Kohlmorgen, L. (2005). *International Organisations and Global Health Governance. The Role of the World Health Organization, World Bank and UNAIDS. The Salzburg Seminar on the Governance of Health*. Salzburg: German Overseas Institute.

Koplan, J.P., et al. (2009). Towards a common definition of global health. *The Lancet*, 373, 1993–5.

Kruk, M.E. (2012). Globalisation and global health governance: implications for public health. *Global Public Health*, 7(S1), S54–62.

Kunkle, S., Christie, G., Hajat, C., and Yach, D. (2016). The role of the private sector in tilting health systems toward chronic disease prevention. *Global Heart*, 11(4), 451–4.

Lauer, J.A. (2001). Determinants of performance. In: Murray, C. and Frenk, J. (eds.) *Health System Performance: Concepts, Measurement and Determinants*. Geneva: WHO, pp. 1–16.

Lavinghouze, S.R. and Snyder, K. (2013). Developing your evaluation plans: a critical component of public health program infrastructure. *American Journal of Health Education*, 44(4), 237–43.

Lee, K. and Kamradt-Scott, A. (2014). The multiple meanings of global health governance: a call for conceptual clarity. *Global Health*, 10, 28.

Lencucha, R., Kothari, A., and Labonté, R. (2011). The role of non-governmental organizations in global health diplomacy: negotiating the Framework Convention on Tobacco Control. *Journal of Health Policy Planning*, 26(5), 405–12.

Longest, B.B. (2014). *Health Program Management* (2th ed.). San Francisco, CA: Jossey-Bass.

Mackintosh, M., et al. (2016). What is the private sector? Understanding private provision in the health systems of low-income and middle-income countries. *The Lancet*, 388, 596–605.

Marston, B.J., et al. (2017). Ebola response impact on public health programs, West Africa, 2014–2017. *Emerging Infectious Diseases*, 23(13), S25–32.

McColl, K. (2008). Europe told to deliver more aid for health. *The Lancet*, 371, 2072–3.

McInnes, C., et al. (2012). Framing global health: the governance challenge. *Global Public Health*, 7, S83–S94.

Médecins Sans Frontières (MSF) (2017). *Ebola. Medecins Sans Frontieres*. [Online] Available at: http://www.msf.org/en/diseases/ebola

Mikkelsen-Lopez, I., Wyss, K., and de Savigny, D. (2011). An approach to addressing governance from a health system framework perspective. *BMC International Health and Human Rights*, 11, 13.

Mills, A., Rasheed, F., and Tollman, S. (2006). Strengthening health systems. In: Jamison, D.T., et al. (eds.) *Disease Control Priorities in Developing Countries*. Washington, DC: The World Bank, pp. 87–102.

Muth, M. and Donaldson, L. (1998). Stewardship theory and board structure: a contingency approach. *Corporate Governance: An International Review*, 6(1), 5–28.
Ng, N.Y. and Ruger, J.P. (2014). Global health governance at a crossroads. *Global Health Governance*, 3(2), 1–37.
Office of the Auditor General of Canada (1999). National Health Surveillance—Diseases and Injuries. *1999 September Report of the Auditor General of Canada.* Ottawa: Office of the Auditor General of Canada and the Commissioner of the Environment and Sustainable Development.
Olafsdottir, A.E. (2012). *Governance and Health Systems Performance: Exploring the Association and Pathways*. Brunel University Thesis (Dr.P.H.).
Ollila, E. (2005). Global health priorities—priorities of the wealthy? *Global Health*, (1), 6.
Perera, L. (2020). *Federal and Unitary forms of Government in a Plural Society* [Online] Available at: https://ssrn.com/abstract=3542312
Pfeffer, J. and Salancik, G.R. (1978). *The External Control of Organizations: A Resource Dependence Perspective*. New York: Harper & Row.
Phelan, A.L., Katz, R., and Gostin, L.O. (2020). The novel coronavirus originating in Wuhan, China: challenges for global health governance. *Journal of the American Medical Association*, doi: 10.1001/jama.2020.1097.
Pillora, S. and McKinlay, P. (2011). *Evolution in Community Governance: Building on What Works: Literature Review*. Sydney: Australian Centre of Excellence for Local Government, Sydney UoT.
Ravishankar, N., et al. (2009). Financing of global health: tracking development assistance for health from 1990 to 2007. *The Lancet*, 373(9681), 2113–24.
Reich, M.R. (2000). Public–private partnerships for public health. *Nature Medicine*, 6, 617–20.
Renukumar, N. (n.d.). *Good Governance: Concepts and Components. [Online] Centre for Good Governance.* [Online] Available at: http://www.scribd.com/doc/45724301/Good-Governance
Rhodes, R.A.W. (1996). The new governance: governing without government. *Political Studies*, 44(4), 652–67.
Ruger, J.P. (2005). The changing role of the World Bank in global health. *American Journal of Public Health*, 95(1), 60–70.
Ruger, J.P. and Ng, N.Y. (2010). Emerging and transitioning countries' role in global health. *Saint Louis University Journal of Health Law and Policy*, 3(2), 253–89.
Saikku, P. and Karjalainen, V. (2012). Network governance in activation policy—health care as an emergent partner. *International Journal of Sociology & Social Policy*, 32(5/6), 299–311.
Saltman, R.B. and Duran, A. (2015). Governance, government, and the search for new provider models. *International Journal of Health Policy & Management*, 1(1), 33–42.
Saltman, R. and Russe, R. (2002). Balancing regulation and entrepreneurialism in Europe's health sector theory and practice. In: Saltman, R., Russe, R., and Mossialos, E. (eds.) *Regulating Entrepreneurial Behaviour in European Health Care Systems*. Buckingham: Open University Press, pp. 1–52.
Savedoff, W.D. (2011). *Governance in the Health Sector: A Strategy for Measuring Determinants and Performance*. Social Insight. World Bank Policy Research Working Paper 5655.
Siddiqi, S., et al. (2009). Framework for assessing governance of the health system in developing countries: gateway to good governance. *Health Policy*, 90(1), 13-25.
Sørensen, E. and Torfing, J. (2007). *Theories of Democratic Network Governance*. Basingstoke: Palgrave Macmillan.
Stallworthy, G., Boahene, K., Ohiri, K., Pamba, A., and Knezovich, J. (2014). Roundtable discussion: what is the future role of the private sector in health? *Globalization and Health*, 10, 55.
Stoker, G. (1998). Governance as theory: five propositions. *International Social Science Journal*, 68(227–8), 15–24.

St-Pierre, L. (2008). *Governance Tools and Framework for Health in All Policies*. Montréal: National Collaborating Centre for Healthy Public Policy.
Stuckler, D. and Mckee, M. (2008). Five metaphors about global-health policy. *The Lancet*, 372, 95–7.
Szlezák, N.A., et al. (2010). The global health system: actors, norms, and expectations in transition. *Plos Medicine*, 7(1), e1000183.
United Nations Development Programme (UNDP) (1997). *Governance for Sustainable Human Development: A UNDP Policy Document*. [Online] United Nations Development Programme. Available at: http://mirror.undp.org/magnet/policy/
United Nations (UN) (2014). *NGOs' Response to the West Africa Ebola Outbreak* (as of 30 Sep 2014). UN Office for the Coordination of Humanitarian Affairs. [Online] Available at: https://reliefweb.int/sites/reliefweb.int/files/resources/ebola_crisis_response_141001.pdf
United Nations (UN) (2015). *UN Mission for Ebola Emergency Response (UNMEER)*. United Nations. [Online] Available at: https://ebolaresponse.un.org/un-mission-ebola-emergency-response-unmeer
United Nations Children's Fund (UNCEF) (2016). *Evaluation of UNICEF's Response to the Ebola Outbreak in West Africa 2014–2015. United Nations Children's Fund*. [Online] Available at: https://www.unicef.org/evaldatabase/files/2232-UNICEF-Ebola_Eval_report_web.pdf
Van der Rijt, T. and Pang Pangestu, T. (2015) Governance within the World Health Assembly: a 13-year analysis of WHO Member States' contribution to global health governance. *Health Policy*, 119, 395–404.
Wagstaff, A. and Claeson, M. (2004). *Rising to the Challenge: The Millennium Development Goals for Health*. Washington, DC: World Bank.
Weishaar, H., Collin, J., Smith, K., Grüning, T., Mandal, S., and Gilmore, A. (2012). Global health governance and the commercial sector: a documentary analysis of tobacco company strategies to influence the WHO Framework Convention on Tobacco Control. *PLoS Medicine*, 9(6), e1001249.
Widdus, R. and White, K. (2004). *Combating Diseases Associated with Poverty-Financing Strategies for Product Development and the Potential Role of Public–Private Partnership. The Initiative on Public–Private Partnerships for Health.* [Online] Available at: http://www.who.int/intellectualproperty/topics/ppp/en/CombatingDiseases-Abridged.pdf?ua=1
World Food Programme (2017). *Ebola Crises. World Food Programme*. [Online] Available at: http://www.wfp.org/widgets-test/ebola-crisis.
World Health Organization (WHO) (2000). *World Health Report 2000. Health Systems: Improving Performance*. Geneva: WHO.
World Health Organization (WHO) (2001). *Strategic Alliances—The Role of Civil Society in Health*. Geneva: WHO.
World Health Organization (WHO) (2002a). *Public Health Management. University of Durham Meeting Report*. Geneva: WHO.
World Health Organization (WHO) (2002b). *Understanding Civil Society Issues for WHO*. Geneva: WHO.
World Health Organization (WHO) (2007). *Everybody's Business: Strengthening Health Systems to Improve Health Outcomes: WHO's Framework for Action*. Geneva: WHO.
World Health Organization (WHO) (2011). *National Health Accounts Database.* [Online] Available at: http://apps.who.int/nha/database/
World Health Organization (WHO) (2013a). *WHO's Role in Global Health Governance*. Geneva: WHO.
World Health Organization (WHO) (2013b). *Health Systems: Governance.* World Health Organization. [Online] Available at: http://www.who.int/healthsystems/topics/stewardship/en/
World Health Organization (WHO) (2015). *One Year into the Ebola Epidemic: A Deadly, Tenacious and Unforgiving Virus*. World Health Organization. [Online] Available at:

http://www.who.int/csr/disease/ebola/one-year-report/ebola-report-1-year.pdf?ua=1
World Health Organization, and Finland Ministry of Social Affairs and Health (2014). *Health in all Policies: Helsinki Statement. Framework for country action*. [Online] Available at: https://apps.who.int/iris/handle/10665/112636
World Health Organization (WHO) (2016). *Health Systems Strengthening for UHC: Building a Shared Vision. UHC 2030, Presentation*.
World Health Organization (WHO) (2018). *The Private Sector, Universal Health Coverage and Primary Health Care*. Geneva: World Health Organization Regional Office for Europe.
Yeoh, E.K. (2015). Governance and management of public health programmes. In: Detels, R., Gulliford, M., Karim, Q.A., and Tan, C.C. (eds.) *Oxford Textbook of Global Public Health* (6th ed.). Oxford: Oxford University Press.

6.9

Implementation science and translational public health

Wafaa M. El-Sadr, Judith Wasserheit, Bryan Wiener, Andrea Howard, Catherine Hankins, Patricia J. Culligan, and Katherine Harripersaud

Introduction

The pace and scope of innovation to improve health and the potential to increase health equity have never been greater. Yet translation, implementation, and scale-up of efficacious biomedical, behavioural, and socio-structural interventions into policies and their adoption into standards of care, or large-scale programmes has remained slow and uneven. In the United States, on average, it takes almost a generation—17 years—to transform less than 15% of original research into practices, programmes, or policies that improve patient care and population health (Balas and Boren 2000). In many low- and middle-income countries (LMICs), closing this 'know-do gap' or the 'knowledge to action gap' may take far longer, although in exceptional cases 'leapfrogging' has been noted with rapid adoption of innovation. Furthermore, interventions that prove efficacious in randomized trials may not be effective when introduced at large scale under 'real world' conditions without appropriate adaptation, or stakeholder and government buy-in, particularly in fragile, overburdened health systems or new social, political, or demographic contexts (Chen et al. 2004).

Implementation science health-related research is a rapidly emerging paradigm that addresses the urgent need to close the 'know-do gap' by providing new conceptual frameworks and tools to evaluate the factors that could influence the successful closure of this gap. Implementation Science uses rigorous methods from multiple disciplines and sectors to inform how best to implement and scale up evidence-based interventions, or intervention packages, in order to improve population health. Implementation science expands the opportunities and responsibilities of

health researchers beyond the laboratory, bench, clinic, and bedside, by asking them to also focus on how health system, health facility, community, national, and transnational contexts impact on closing this gap (Elzarrad et al. 2013). The primary focus of implementation science has evolved differently in the United States (US) in contrast to the global health context. With growing support from the US National Institutes of Health (NIH) during the past 15 years, implementation science in the health sciences blossomed with a central focus on the translation of biomedical advances into widespread clinical practice, often working with providers and patients at health facility level (Zerhouni 2003; Glasgow et al. 2012; Neta et al. 2015). Over the same period, the primary implementation science focus globally was on translating a broad array of biomedical, behavioural, information technology and policy interventions into large-scale programmes in resource-limited settings, frequently working with policymakers and programme managers at the national or transnational level. This approach to implementation science was shaped by support from multilateral and bilateral organizations such as the World Health Organization (WHO), the Joint United Nations Program on HIV/AIDS (UNAIDS), the US President's Emergency Plan for AIDS Relief (PEPFAR), the United States Agency for International Development (USAID), the Centers for Disease Control and Prevention (CDC), the World Bank, The Global Fund for AIDS, Tuberculosis and Malaria (GFATM), and the US NIH Fogarty International Center among other NIH institutes (Madon et al. 2007; The Global Fund to Fight AIDS Tuberculosis and Malaria USAID World Health Organization 2008; Padian et al. 2011).

At the same time, implementation science, as a science, continues to evolve, so much so that even its boundaries are continually being demarcated. The following sections describe current thinking around defining implementation science, its governing frameworks and research designs, and the critical role of stakeholder engagement in applying implementation science.

Definitions of implementation science

In new, rapidly evolving fields, consensus on their primary focus, core methods, boundaries, and terminology often emerge over time, as a critical mass of thought leaders are galvanized, engaged, and confronted with new

challenges. This is the case for implementation science (Odeny et al. 2015). This is manifested in the diversity of definitions for implementation science which reflects both the relative youth of this field and the different factors that influence implementation of this emerging field in high-income countries and in LMICs. One early and widely cited definition from researchers based in the United Kingdom and the US frames implementation science as 'the scientific study of methods to promote the systematic uptake of clinical research findings and other evidence-based practices into routine practice, and hence to improve the quality and effectiveness of health services' (Eccles 2006; Eccles et al. 2009). Others have offered broader definitions that go beyond a clinical, facility-based focus on quality improvement to call for 'applied research that aims to develop the critical evidence base that informs the effective, sustained and embedded adoption of interventions by health systems and communities' (Allotey et al. 2008) Some characterize implementation science even more broadly as the study of the determinants and processes that influence successful integration of evidence-based innovations in specific settings, and the results of implementation and scale-up, as well as related issues such as the degree of fidelity with which core components are translated and other components are adapted as these innovations are introduced and taken to large scale (Hirschhorn et al. 2007; Rabin et al. 2008; Peters et al. 2013; Peters and Adam 2013; National Implementation Research Network 2015). Common features of all of these definitions include the following: (1) applying rigorous, systematic research methods to guide the introduction, improvement or expanded use of interventions for which there is strong evidence of efficacy from earlier phase research; (2) a multidisciplinary or transdisciplinary approach; and 3) collaboration between researchers and relevant stakeholders, prominently including programme implementers and policymakers throughout the process (Sanders and Haines 2006; Bhattacharyya et al. 2009; Knapp and Anaya 2013; Bauer et al. 2015).

To accelerate improvements in global health and progress towards national or transnational goals such as the Sustainable Development Goals (SDGs), the authors of this chapter support the following broad definition, i.e. that *Implementation science is a systematic, scientific approach to ask and answer questions about how we get 'what works' to people who need it with greater speed, fidelity, efficiency, quality, sustainability and coverage.*

This definition promotes the application of at least 10 research approaches utilized by different disciplines as essential for understanding and improving the determinants, processes, and outcomes of implementation and, ultimately, scale-up and sustainability to achieve population-level health benefits (Kemp et al. 2018). These research approaches include impact evaluation, economic evaluation, qualitative health systems research, operations research, systems analysis and improvement, organizational assessment, stakeholder and policy analysis, dissemination research, social marketing, and surveillance and data systems. Although these 10 research methods are not exhaustive, together they provide a core set of tools to assess and iteratively improve implementation and scale-up of health interventions. These methods enable exploration of a synergistic continuum of three categories of outcomes for implementation science from most proximal such as acceptability, appropriateness, feasibility, cost, adoption, fidelity, penetration, scale-up, and sustainability; through intermediate *service delivery outcomes* such as efficiency, equity, timeliness, and patient-centredness; to distal *health outcomes*. While improvements in service delivery and health outcomes are the ultimate criteria for evaluating the effectiveness of implementation strategies, implementation outcomes may be early signals of successful implementation, and preconditions for achieving service delivery and health objectives.

The boundaries of the terminology of implementation science also remain to be defined, as are the boundaries of other closely related, emerging fields, such as improvement science, programme science, and delivery science, potentially creating confusion in terminology—with each aiming to improve health despite having slightly varying primary foci and approaches. For example, improvement science is primarily concerned with quality improvement, one of the 10 core implementation science methods (Foundation 2011). Programme science flips the standard paradigm of ‘getting research into practice’ and focuses on ‘getting research out of practice’ by working with programme staff to develop research questions and conduct research (Blanchard and Aral 2011). Delivery science prioritizes strengthening the capacity of health systems using intersectoral public policy and financing approaches to improve health and development outcomes (Kim 2018). Finally, the science of scale-up prioritizes policy and implementation research that focuses on the scaling of existing health tools

and delivery services (Kruk et al. 2016). As these fields mature, realignments or mergers may occur, leveraging the complementary strengths and learnings.

Implementation science frameworks

Implementation science offers several frameworks for guiding research and practice. In a 2012 review, Tabak et al identified 61 frameworks for supporting the implementation of evidence-based interventions (Tabak et al. 2012). Additional frameworks have been proposed since then. Selecting a framework can be challenging given the number of options available (Birken et al. 2017). Nilsen categorized implementation science frameworks into three types based on three common objectives for implementation science: (1) process models for describing or guiding the process of implementation; (2) determinants frameworks for identifying the factors that influence implementation; and (3) evaluation frameworks for assessing the degree of implementation success. Although specific frameworks can address more than one of these objectives, these general types of frameworks can help guide the selection of a framework for implementation research and practice (Nilsen 2015).

The first type of frameworks, process frameworks, specify the steps, stages, or phases of implementing evidence-based interventions. Process frameworks such as the Knowledge to Action (K2A) Framework (Wilson et al. 2011) and the Canadian Institutes of Health Research (CIHR) Model of Knowledge Translation (Canadian Institutes of Health Research) describe a sequence that begins with research on the effectiveness of interventions, then moves to the dissemination of those interventions to targeted audiences, and then moves to the implementation of those interventions in real-world settings. Other process frameworks, such as the Quality Implementation Framework (QIF) (Meyers et al. 2012), the Stetler Model (Stetler 2010), and the Ottawa Model (Logan 2010) offer practical advice for planning and carrying out an implementation effort. These frameworks are useful for guiding implementation because they not only highlight the key tasks that should be performed, but also prescribe the order in which they should be performed. The QIF, for example, delineates four phases of implementation: initial considerations regarding the host setting (Phase One), creating a structure for implementation (Phase Two), ongoing

structure once implementation begins (Phase Three), and improving future applications (Phase Four). The QIF further identifies 14 critical steps across these four phases that are believed to result in high-quality implementation. Although process models describe, or in some cases prescribe, a logical sequence of steps and depict an idealized view of the implementation process, actual implementation is often nonlinear and disordered.

Unlike process frameworks, which describe steps or stages of implementation, the second type of frameworks, determinants frameworks, identify factors that influence implementation, either positively in the case of facilitators or negatively in the case of barriers. Such frameworks are useful for assessing the context in which implementation occurs and, consequently, identifying strategies to overcome barriers or harness facilitators. They are also useful for explaining why implementation was successful or not in specific settings or why implementation success varied across multiple settings. The Consolidated Framework for Implementation Research (CFIR) is a commonly used determinants framework in implementation science, with growing use in global health (Damschroder et al. 2009). The CFIR synthesizes 19 implementation science frameworks by organizing 37 factors into five domains: intervention characteristics, outer setting, inner setting, individuals involved, and implementation processes. The attractiveness of the CFIR derives from its comprehensiveness, applicability, and flexibility; the CFIR's developers encourage users to 'select constructs from the CFIR that are most relevant for their particular study setting and use these constructs to guide diagnostic assessments of implementation context, evaluate implementation progress, and help explain findings in research studies or quality improvement initiatives' (Damschroder et al. 2009). The Theoretical Domains Framework (TDF) is another widely used determinants framework that, like the CFIR, synthesizes 33 psychological and behaviour change theories by organizing more than 100 constructs into 14 theoretical domains, such as knowledge, skills, intentions, goals, and social influences(Cane et al. 2012). While neither the CFIR nor the TDF hypothesizes causal relationships among determinants, both frameworks acknowledge that such relationships exist.

Thirdly, evaluation frameworks offer a conceptual structure and a set of metrics for gauging implementation success. The two best known evaluation frameworks in implementation science are the Implementation Outcomes Framework (IOF) (Proctor et al. 2011) and Reach, Effectiveness,

Adoption, Implementation, and Maintenance (RE-AIM) framework (Glasgow et al. 1999). The RE-AIM Framework is described here since the IOF was introduced earlier in this chapter. Developed for evaluating the public health impact of health promotion programmes, RE-AIM identifies five metrics that signal the applicability, scalability, potency, and sustainability of programmes or interventions in 'real world' settings (Glasgow et al. 1999). These metrics are as follows: the *reach* of the programme within the intended target population, the *effectiveness* of the programme, the *adoption* of the programme by target settings or staff, the *implementation* of the programme in terms of the consistency and quality of programme delivery, and the *maintenance* of programme delivery and programme effects. The RE-AIM Framework is particularly useful in global health because it balances classic concerns about programme effectiveness with considerations relevant to implementation and scale up, such as adoption among targeted staff, settings, or institutions and coverage (reach) within targeted populations.

Table 6.9.1 highlights the relative strengths and limitations of the three main types of implementation frameworks described by Nilsen. Two broad observations are noteworthy. First, the different types of frameworks have complementary strengths and limitations and, thus, work well together in implementation research or practice. For example, determinants frameworks such as the CFIR or the TDF can be employed in the pre-implementation stages of process models such as the QIF to assess needs and resources, fit of the evidence-based intervention and the setting, and capacity or readiness for implementation. Likewise, determinants frameworks and evaluation frameworks can supply the independent and dependent variables in research investigating the effects of context on implementation success or failure. Second, many, if not most, implementation science frameworks have been developed in high-income countries with Western cultures. Although frameworks like the CFIR have been used in implementation science in global health, adaptation is likely needed to improve their fit with and utility for studying implementation in LMICs and countries with non-Western cultures.

Table 6.9.1 Strengths and limitations of implementation science frameworks

Framework type	Strengths	Limitations
Process models (e.g. QIF, Ottawa model)	• Include actionable steps, useful outside research context • Prescribe a sequence for performing implementation tasks	• Describe a rational, ideal sequence rarely observed or practicable in implementation • Focus neither on determinants, nor outcomes
Determinants frameworks (e.g. CFIR, TDF)	• Include large number of potential determinants at multiple levels • Synthesize and harmonize prior frameworks, theories, and models • Include data collection and analysis tools • Useful for formative, concurrent, or summative evaluation	• Provide no guidance for implementation • Focus on determinants, not outcomes • Need adaptation for use in LMIC contexts and non-Western cultural contexts • Include fewer determinants at higher levels (e.g. community, policy) • Reliable, valid measures of many constructs lacking
Evaluation framework (e.g. IOF, RE-AIM)	• Include multiple outcomes for evaluating public health impact • Differentiate implementation outcomes from effectiveness • Specify how outcomes can be quantitatively assessed • Outcomes can be measured at multiple levels	• Provide no guidance for implementation • Focus on outcomes, not determinants • Reliable, valid measures for some outcomes lacking • Relations among some outcomes need clarification

Scale-up frameworks

Although implementation science frameworks could be used to design, guide, and evaluate scale-up efforts, Barker et al. identified 45 theories, frameworks, and models that have been developed specifically for scaling up evidence-based health interventions (Barker et al. 2016). Two well-known process frameworks for scale-up are the Management Systems International (MSI) Scaling Up Management Framework for Practitioners

and the ExpandNet/WHO Framework. The MSI Framework describes three main steps: developing a scale-up plan, establishing the preconditions for scaling up, and implementing the scale-up process. Within these steps, the framework offers guidance for performing 10 tasks: creating a vision, assessing scalability, filling information gaps, preparing a scaling-up plan, legitimizing change, building a constituency, realigning, and mobilizing resources, modifying and strengthening organizations, and coordinating action.

The ExpandNet/WHO Framework outlines a nine-step process for developing a scale-up strategy and offers practical advice for conducting each step. The ExpandNet/WHO Framework also could serve, as a determinants framework. It conceptualizes scale-up as a system of five inter-related elements: (1) the innovation that is being scaled up; (2) the resource team, or the individuals and institutions that seek to facilitate the innovation's wider use; (3) the user organization, meaning the institutions or organizations within which the innovations are expected to be widely adopted and implemented; (4) the scaling-up strategy, which designates the means by which the innovation is communicated, transferred, or otherwise promoted; and (5) the environment in which scaling up takes place. For each element, the Framework describes attributes of success, many of which correspond to the constructs found in the CFIR. For example, the attributes of success for the innovation include evidence strength and credibility, observability, relative advantage, compatibility, trialability, and complexity. The Framework's attributes of success can be viewed as determinants that hinder or facilitate scale-up. A distinctive feature of the Framework is its delineation of strategic choices regarding the type of scaling up used; the process of communicating the innovation and preparing users in its application; the organizational options selected in implementing the process; assessing costs and resource mobilization; and monitoring and researching the process and outcomes of scaling up. Like the ExpandNet/WHO Framework, Yamey's conceptual framework also identifies success factors (i.e. determinants) that fall into six categories representing different components of the scaling up process: attributes of the specific tool or service being scaled up, attributes of the implementers, the chosen delivery strategy, attributes of the 'adopting' community, the socio-political context, and the research context (Yamey 2011).

Although there are many approaches and methods for monitoring and evaluating scale-up, including impact evaluation, no evaluation framework has emerged in global health that offers a conceptual structure and general set of metrics for evaluating scale-up efforts. While the RE-AIM Framework was developed to evaluate the public health impact of health promotion programmes in the US context, it is suitable for use as an evaluation framework beyond the US because it balances traditional concerns about programme effectiveness with considerations relevant to scale up, such as programme coverage (reach) in a targeted population, programme spread (adoption) among settings and people expected to deliver the programme, consistency and quality of programme delivery (implementation), and programme integration (maintenance) into government policies and organizational practices.

Research designs

A wide variety of research designs have been employed in implementation science. In 2017, Brown et al. identified three broad categories of designs used in implementation science within the traditional translational research pipeline, including those that provide (1) within-site; (2) between-site; or (3) within- and between-site comparisons of implementation strategies, to determine how to make an effective clinical or preventive intervention work within community and/or service settings (Brown et al. 2017). As these designs focus on interventions that have already demonstrated effectiveness, the emphasis of such designs is on measuring (1) healthcare process, or activities undertaken by the health system, and (2) healthcare output, or observable measures of service delivery provided to the target population, as opposed to health outcomes or health-related behaviours of individuals who are the ultimate target of the intervention. Although many of these designs employ random assignment, and thus have high internal validity, some are observational or quasi-experimental, and are preferable in instances when randomization is not feasible or warranted, such as when policies are disseminated or implemented by law.

Within-site designs evaluate implementation successes or failures by studying changes that occur inside an organization, community, or system. The simplest design is a post design, which examines healthcare process and/or utilization after introduction of an implementation strategy aimed at

delivery of an evidence-based intervention in a novel health setting. On the other hand, in a pre-post design, the impact of a new implementation strategy for an existing intervention is inferred by comparing process and/or output measures prior to and following adoption of the strategy. A multiple-baseline times-series design is a variant of the pre-post design in which data are collected at multiple time points to examine whether the effects of the implementation strategy are sustained.

Between-site designs compare healthcare process and outputs among sites that have different exposures. In a new versus implementation-as-usual design, some sites are assigned to an innovative implementation strategy while others maintain the usual condition. Process and output measures for the two strategies are then compared over the same time period. When evaluating complex, multilevel implementation strategies, a cluster-randomized design is often employed, in which sites (health facilities, communities, or organizations) are randomly assigned to either the new implementation or usual-practice condition, and endpoints are assessed for all eligible individuals in both conditions. If the implementation strategy is designed to achieve impact at a smaller unit of the organization, then randomization can occur at the level of the ward, team, or provider within each site. Individual-level randomization can also be utilized, if the intervention is delivered at the patient or individual level.

A head-to-head randomized implementation trial design, also known as a comparative effectiveness implementation trial, tests which of two active implementation strategies is more successful in implementing an intervention, in terms of quality, quantity, or speed of intervention implementation. The same intervention is used for both study arms, and sites are randomly assigned to one of the implementation strategies. In a variant of this design, sites are randomized to one of two implementation strategies that focus on different patient populations of individuals or patients and target different outcomes. With this design, each site serves as both an active intervention and a control.

Factorial designs evaluate a combination of two or more implementation strategies simultaneously. Each experimental factor has two or more levels (e.g. presence or absence, or low, medium, or high intensity of implementation). With a 2 × 2 factorial design, units are randomly assigned to one of four conditions (i.e. strategy 1, strategy 2, strategy 1 and 2 combined, or neither strategy), and estimates of the effect of each factor by

itself and their interaction are derived. In an incomplete, or fractional, factorial design, one or more arms of a complete factorial are excluded. This design may be chosen if it is deemed unethical to withhold both implementation strategies from participating sites, or alternatively if it would be too difficult to implement both strategies at the same site.

A multiphase optimization strategy implementation trial (MOST) design enables consideration of a large number of components that comprise an implementation strategy, including those involving leadership, the clinic level, the provider level, and key processes such as planning, training, financing, restructuring, and quality management. Testing each combination of these components is often not feasible, the MOST design can be used to identify and test an optimal intervention. This design consists of three phases: preparation, optimization, and evaluation. In the preparation phase, components are selected and pilot tested using an explicit optimization criterion. In the optimization phase, a randomized trial is conducted to assess the effectiveness of these components, using a fractional factorial design. Based on the trial's results, the group of components that best meets the optimization criterion is identified. In the evaluation phase, a standard randomized implementation trial is conducted that compares the optimized implementation strategy with an appropriate comparison condition.

A sequential multiple assignment randomized implementation trial (SMART) design is an adaptive approach that involves multistage randomizations such that the site-level implementation process can be modified if unsuccessful. In the latter design, the adaptation probability within each site is constant. A variant of this design is the SMART design with adaptive randomization (SMART-AR), in which the randomization probability of each strategy is updated based on the outcomes in the other sites, so that the expected outcome is optimized.

A doubly randomized, two-level nested, or split-plot implementation design is used to assess whether experimental implementation factors directed towards two different hierarchical levels (e.g. provider and patient) have additive or synergistic effects.

Within- and between-site comparison designs are also known as rollout randomized implementation trial designs. In these designs, sites begin in one implementation condition and cross over to another, and the timing of the start of the new implementation strategy is randomly assigned. The simplest design is a stepped wedge, or dynamic wait-listed, design, in

which all sites start in the implementation-as-usual setting, and then cross over at randomly determined time intervals, until all are receiving the new implementation strategy. At each time interval, a between-site comparison of those sites assigned to the new implementation strategy and those remaining in the control condition is made. Additionally, a within-site comparison can be conducted when sites cross over from the usual practice setting to the new implementation strategy. In a head-to-head rollout trial, all sites start in the implementation-as-usual setting and then at a predetermined time, are randomly allocated to one of two new implementation strategies, which they continue until the end of the study.

Other research designs encompass elements beyond the traditional translational research pipeline, including hybrid designs, discrete choice experiments, and designs that involve simulation. Curran et al. described three effectiveness implementation hybrid design constructs, all of which focus a priori on assessing both implementation and clinical effectiveness (Curran et al. 2012). Type 1 hybrid designs primarily test the effects of a clinical intervention on health outcomes, while collecting information on its delivery and/or its potential for implementation in a real-world setting. The latter is usually achieved through a process evaluation that ascertains barriers and facilitators to delivering the intervention, as well as potential modifications to the intervention that could be made to maximize implementation. Type 2 hybrid designs simultaneously test a clinical intervention and an implementation strategy. Type 3 hybrid designs primarily test an implementation strategy, while observing or gathering information on the clinical intervention's impact on health outcomes.

Another research method is discrete choice experiment approach, a health economics tool that elicits preferences from decision makers over alternatives, with each alternative consisting of several attributes. It can be used pre-implementation to elicit information from a variety of stakeholder groups, including patients, caregivers, providers, and administrators, to assess the potential acceptability or appropriateness of health services, products, or implementation strategies, in order to improve implementation (Salloum et al. 2017).

Lastly, mixed methods designs that collect, analyse, and merge both quantitative and qualitative data are often utilized in implementation science. Typically, qualitative methods are used to identify strategies for facilitating implementation of an evidence-based intervention, or to obtain

an in-depth understanding of reasons for implementation success or failure, while quantitative measures are used to test a hypothesis and assess predictors of successful implementation (Palinkas et al. 2011).

In summary, choosing a research design for a specific context is a complex process that requires consideration of the research questions, existing knowledge base, stakeholder values, and available resources. Table 6.9.2 shows the relative strengths and limitations of various research designs.

Table 6.9.2 Strengths and limitations of research designs

	Strengths	Limitations
Within-site designs		
Post design of an implementation strategy to adopt an evidence-based intervention in a new setting	• Useful for assessing factors that predict programme adoption • Useful when new health guidelines or changes in policy occur	Lack of comparison group limits internal validity
Pre-post design of an implementation strategy of an intervention already in use	• Can be used to assess adoption of a guideline or policy by a healthcare system • Can examine whether an implementation strategy has sustained effects by collecting data at multiple time points	• Requires information about pre-implementation levels • Does not account for secular trends that may influence healthcare processes and outputs
Between-site designs		
New implementation strategy vs. usual practice implementation design	• Can determine whether new implementation strategy improves reach or utilization of a healthcare intervention, compared with standard of care • Use of a contemporaneous control condition and randomized assignment of the implementation conditions strengthens internal validity	• May be subject to assignment bias, if sites assigned to the new implementation strategy do not accept some of the strategy components • May be subject to contamination, particularly if randomization occurs at a lower level of the organization
Head-to-head randomized implementation trial design	• Can compare two implementation strategies • Use of contemporaneous comparison group and random assignment strengthens internal validity	Resource-intensive, as both implementation strategies must be executed with equivalent attention to fidelity

	Strengths	Limitations
Factorial designs for implementation	• Can assess whether there is an interaction between two implementation strategies • Use of contemporaneous comparison group and random assignment strengthens internal validity	Resource intensive Complex to design, implement, and evaluate
Multiphase optimization strategy implementation trial design (MOST)	• Enables consideration of a large number of implementation strategy components	Time and resource intensive Complex to design, implement, and evaluate
Sequential multiple assignment randomized implementation trial (SMART) implementation design	• Optimizes resource allocation, as additional strategies are only tested in sites that do not achieve desired outcomes	Time and resource intensive Complex to design, implement, and evaluate
Doubly randomized, two-level nested designs	• Can assess whether there is an interaction between two implementation factors directed towards two distinct hierarchical levels	Complex to design, implement, and evaluate
Within- and between-site comparison designs		
Stepped wedge design	• Can use when there are perceived advantages to a new implementation strategy, or it is unethical to withhold a new implementation strategy throughout the study	Not appropriate for all interventions Complex to design, implement, and evaluate

	Strengths	Limitations
Head-to-head rollout design	• Can compare two new implementation strategies against the standard of care	If no randomization, does not obviate risk of bias
Effectiveness implementation hybrid designs		
Hybrid type 1	• Implementation data can inform subsequent implementation research trials	Require additional research expertise and funding
Hybrid type 2	• Can facilitate more rapid translation of research findings into routine practice	Require additional research expertise and funding • Some design compromises will likely be necessary in one or both aims given funding limits
Hybrid type 3	• Useful if suspected that clinical intervention effects susceptible to change during implementation in new settings or under less controlled conditions	Primary data collection on patient outcomes may not be feasible with large multisite implementation study, thus may need to rely on subsample of patients, medical record review, and/or administrative data
Discrete choice experiment		
	• Provides stepwise method for selecting and tailoring an implementation strategy to a particular context • Guides stakeholders to consider multiple attributes of an implementation strategy, which enhances the precision with which strategies are tailored to a specific setting	Primary data may be necessary to establish attributes and attribute levels, particularly in low-income settings

Stakeholder and community engagement

The seeds for effective implementation science, whether it takes the form of a phase four trial, a population-level study, or other design, are sown during the conduct of the research that strengthened the scientific evidence propelling forward a line of investigation. In particular, whether effective stakeholder engagement was accomplished in the preceding trials or other research which could have a direct effect on the feasibility of conducting follow-on studies at the end of a clinical trial or other such studies.

Transparent, mutually beneficial, respectful, sustained relationships between research funders, sponsors, researchers, implementers, community members, and other stakeholders are critical to the success of any study. These relationships set the stage for effective collaboration and shared ownership of all aspects of the research process, starting at problem formulation and in the research design stage. In the context of implementation science research, they can directly influence the relevance of the research; the likelihood that interpretable, valid data and reliable results will be produced; and the eventual uptake of study outcomes.

The foundational principles underpinning effective stakeholder involvement, as now defined in Good Participatory Practice (GPP) guidance in several research fields, drew inspiration from the controversies and debates that arose in 2004 regarding pre-exposure prophylaxis (PrEP) trials for HIV prevention where national political leaders in Cambodia and Cameroon opposed these trials (Mills et al. 2005; Singh and Mills 2005; Global Campaign for Microbicides 2009a; Global Campaign for Microbicides 2009b). In response, the Joint United Nations Programme on HIV/AIDS (UNAIDS) convened the *Creating Effective Partnerships in Research* process in 2005 (UNAIDS 2006) that resulted in a strong recommendation for the development of research guidelines for community stakeholder engagement.

Following a series of regional consultations that defined the key elements needed for creating effective partnerships for biomedical HIV prevention trials, these were published as the GPP-HIV guidelines in 2007 (UNAIDS and AVAC 2007). GPP was first cited in the methodology of a scientific publication of trial results in 2010 (Grant et al. 2010). A revised edition of the GPP-HIV guidelines (UNAIDS and AVAC 2011) followed GPP application in different settings and formal consultations with stakeholder groups in Africa, the Americas, Asia, and Europe. Subsequently, the GPP guidelines were adapted for tuberculosis drug trials in 2012 (Critical Path to

TB Drug Regimens 2012) and the WHO published GPP guidelines in 2016 for trials of emerging (and re-emerging) pathogens that are likely to cause major outbreaks in the future and for which few or no interventions exist (WHO 2016). Web-based training curricula, tools, application examples, and other materials are currently available to assist stakeholders to understand, implement, and monitor GPP (AVAC 2018). Further, a Centre of Excellence is under development (Coalition to Accelerate and Support Prevention Research 2017) with an implementation framework that further advances GPP practitioners' understanding, ability, and technical skills to carry out GPP effectively.

Determining who the relevant stakeholders are for specific implementation research study is an essential step. Broadly, stakeholders are all individuals or groups of individuals who can influence or are affected by the conduct or outcome of the research, i.e. all those who have a stake. They include national research organizations and scientific councils, Schools of Medicine and Nursing, and ministries of Health, Education, Science, and Justice as well as others that may be involved in the eventual implementation of the interventions. Importantly, an important stakeholder group is the population to be recruited, study participants, families of study participants, people resident in the area where the research is to be conducted, local support groups or networks, religious leaders, opinion leaders, media, local healthcare authorities, healthcare providers, traditional healers, private sector partners, non-governmental organizations (NGOs), and community-based organizations (CBOs) including women and youth groups.

Local contextual analysis may be needed to define such stakeholders, the optimal level of engagement needed, and the types of structures to be created to facilitate stakeholder engagement. Funding for such contextual analyses should be anticipated, with time frames determined in part by the urgency of the issue to be evaluated. Formative research, qualitative studies, geospatial mapping, key informant interviews, and other methodologies assist in defining relevant stakeholders, including community, for implementation research. In a health emergency, the formative research phase should be nimble and rapid, aiming at stakeholder engagement to identify key voices to participate from the outset in discussions of trial design (Holzer et al. 2014). In less urgent situations, communities can be engaged, for example, in the randomization process itself, as was done in

the HPTN071 (PopART) Study, a cluster-randomized trial conducted in South Africa and Zambia (Bond et al. 2016).

It should be noted that stakeholders may have competing interests and perspectives. Disagreements during implementation research may arise from power imbalances, differences in social organization, hierarchies, gender issues, relative socioeconomic disparities, and other factors. A stakeholder engagement plan, developed together by research teams and relevant stakeholders, can define the range of stakeholders to be engaged; detail the frequency and types of engagement methods to be used, such as public meetings, workshops, joint decision-making models, or delegated decision-making; and clarify the type of engagement appropriate for each stakeholder, such as being informed, consulted, collaborated with, or emancipated to make decisions (Ray and Miller 2017). Such a plan can also delineate the nature of support for initial and ongoing activities aimed at building stakeholder capacity to understand concepts, purposes, practices, and limitations of implementation research so that the ability of stakeholders to provide meaningful input to the research process will be enhanced. Importantly, it clearly describes the methods to be used for resolution of differences and lays out key principles that underpin partnerships among stakeholders, including respect, fairness, integrity, transparency, accountability, and autonomy (UNAIDS and AVAC 2011).

Effective stakeholder and community engagement in implementation research is not only an ethical imperative, it enhances the conduct of research, contributes to robust research outcomes, and creates both impetus and support for scale-up following proven results.

Future directions

Implementation science is an important research approach that offers the potential to bring to practice many of the discoveries that exist in scientific journals. Its focus on advancing the learning of implementation and of scale-up is directly linked to the collective ability to advance the health and well-being of populations around the world. Yet, much remains to be done to achieve broad knowledge of this approach among researchers, funders, and policymakers. The diversity of frameworks and methods put forth for implementation science also highlights the need for a maturation of the field and for consensus building among experts in implementation science.

Additionally, efforts are needed to overcome some of the myths that relate to this research approach including highlighting its importance and the rigorous manner in which implementation research can be designed and conducted.

Creating a new generation of researchers who are knowledgeable of implementation science requires an investment in specific education and training opportunities for young scholars, as well as investigators at various career stages. These training programmes should emphasize a multidisciplinary approach including how to synthesize information across disciplines, test and adjust approaches in collaboration with stakeholders, use data to inform policies, and ways to formulate scalable interventions (Means et al. 2016). Such training would allow for broad health systems perspective and understanding of multisectoral ancillary effects of health interventions and policies. Core competencies for these programmes continue to be refined and need to be foundational and responsive to advances in the field (Gonzales et al. 2012; Chambers et al. 2017). Mentorship is also a critical component of advancing this science and should accompany and supplement any formal training programmes. An evaluation of a 2-year mentorship programme focused on dissemination and implementation research for cancer prevention and control resulted in improved dissemination and implementation research competency, high satisfaction for emotional support, career planning, and sharing resources (Padek et al. 2018).

With the adoption of implementation science as a means to help translate evidence-based-interventions related to health into positive outcomes for patients and communities, other fields have also turned to this approach as relevant for closing the know-do gap. Examples include initiatives related to special education (Cook and Odom 2013), conservation biology (Toomey et al. 2017) and environmental sustainability (Hering 2016). Thus, it is important that rather than each field developing separate frameworks and methodologies for guiding research and practice, efforts should be focused on aligning approaches and learnings across disparate fields and settings, with the ultimate goal of identifying a uniform scientific approach to close the know-do gap that is independent of setting and application.

It is also important for global health researchers to learn from programmes focused on the translation of innovations into commercial successes. In the US, the National Science Foundation’s Innovations Corps

(I-Corps) initiative is a public-private partnership aimed at equipping researchers in science and engineering with the skills and tools needed to translate their work from the laboratory to the marketplace (Robinson 2012). Akin to Innovation Science, the goal of this programme is to create an effective infrastructure for moving from basic discovery to a successful product.

Finally, an important goal of fostering research in implementation science and translational public health, is the full engagement of the academic enterprise in solving the pressing, practical issues of the world (El-Sadr et al. 2014). By actively seeking out collaborations with other fields within and beyond academia, implementation science can inspire an expansion of the core mission of academic institutions to include in addition to the generation of new knowledge, an emphasis on the need for engagement in taking knowledge to practice.

REFERENCES

Allotey, P., Reidpath, D.D., Ghalib, H., Pagnoni, F., and Skelly, W.C. (2008). Efficacious, effective, and embedded interventions: implementation research in infectious disease control. *BMC Public Health*, 8, 343.

AVAC. (2018). *Good Participatory Practice Tools* [Online]. Available at: https://www.avac.org/gpp-tools

Balas, E.A. and Boren, S.A. (2000). Managing clinical knowledge for health care improvement. *Yearbook of Medical Informatics*, 1, 65–70.

Barker, P.M., Reid, A., and Schall, M.W. (2016). A framework for scaling up health interventions: lessons from large-scale improvement initiatives in Africa. *Implementation Science*, 11, 12.

Bauer, M.S., Damschroder, L., Hagedorn, H., Smith, J., and Kilbourne, A.M. (2015). An introduction to implementation science for the non-specialist. *BMC Psychology*, 3, 32.

Bhattacharyya, O., Reeves, S., and Zwarenstein, M. (2009). What is implementation research?: rationale, concepts, and practices. *Research on Social Work Practice*, 19, 491–502.

Birken, S.A., et al. (2017). Criteria for selecting implementation science theories and frameworks: results from an international survey. *Implementation Science*, 12, 124.

Blanchard, J.F. and Aral, S.O. (2011). Program science: an initiative to improve the planning, implementation and evaluation of HIV/sexually transmitted infection prevention programmes. *Sexually Transmitted Infections*, 87, 2–3.

Bond, V., et al. (2016). 'The difference that makes a difference': highlighting the role of variable contexts within an HIV Prevention Community Randomised Trial (Hptn 071/PopArt) in 21 study communities in Zambia and South Africa. *AIDS Care*, 28 Suppl–3, 99–107.

Brown, C.H., et al. (2017). An overview of research and evaluation designs for dissemination and implementation. *Annual Review of Public Health*, 38, 1–22.

Canadian Institutes of Health Research. *About Knowledge Translation*. [Online] Available at: http://www.cihr-irsc.gc.ca/e/29418.html

Cane, J., O'Connor, D., and Michie, S. (2012). Validation of the theoretical domains framework for use in behaviour change and implementation research. *Implementation Science*, 7, 37.

Chambers, D.A., Proctor, E.K., Brownson, R.C., and Straus, S.E. (2017). Mapping training needs for dissemination and implementation research: lessons from a synthesis of existing D&I research training programs. *Translational Behavioral Medicine*, 7, 593–601.

Chen, L., et al. (2004). Human resources for health: overcoming the crisis. *The Lancet*, 364, 1984–90.

Coalition to Accelerate and Support Prevention Research (2017). *HIV Vaccine and Biomedical Prevention Research*. [Online]. Available at: https://www.usaid.gov/what-we-do/global-health/hiv-and-aids/technical-areas/hiv-vaccine-research-development

Cook, B.G. and Odom, S.L. (2013). Evidence-based practices and implementation science in special education. *Exceptional Children*, 79(2), 135–44.

Critical Path to TB Drug Regimens (2012). *Good Participatory Practice Guidelines for TB Drug Trials*. [Online] Available at: https://www.avac.org/sites/default/files/resource-files/GPP-TB%20Oct1%202012.pdf

Curran, G.M., Bauer, M., Mittman, B., Pyne, J.M., and Stetler, C. (2012). Effectiveness-implementation hybrid designs: combining elements of clinical effectiveness and implementation research to enhance public health impact. *Medical Care*, 50, 217–26.

Damschroder, L.J., Aron, D.C., Keith, R.E., Kirsh, S.R., Alexander, J.A., and Lowery, J.C. (2009). Fostering implementation of health services research findings into practice: a consolidated framework for advancing implementation science. *Implementation Science*, 4, 50.

Eccles, M. and Mittman, B.S. (2006). Welcome to implementation science. *Implementation Science*, 1, 1.

Eccles, M.P., et al. (2009). An implementation research agenda. *Implementation Science*, 4, 18.

El-Sadr, W.M., Philip, N.M., and Justman, J. (2014). Letting HIV transform academia--embracing implementation science. *New England Journal of Medicine*, 370(18), 1679–81.

Elzarrad, M.K., Eckstein, E.T., and Glasgow, R.E. (2013). Applying chronic illness care, implementation science, and self-management support to HIV. *American Journal of Preventive Medicine*, 44, S99–107.

Glasgow, R.E., Vinson, C., Chambers, D., Khoury, M.J., Kaplan, R.M., and Hunter, C. (2012). National Institutes of Health approaches to dissemination and implementation science: current and future directions. *American Journal of Public Health*, 102, 1274–81.

Glasgow, R.E., Vogt, T.M., and Boles, S.M. (1999). Evaluating the public health impact of health promotion interventions: the RE-AIM framework. *American Journal of Public Health*, 89, 1322–7.

Global Campaign for Microbicides (2009a). *Preventing Prevention Trial Failures: A Case Study and Lessons Learned for Future Trials from the 2004 Tenofovir Trial in Cambodia*. [Online] Available at: http://www.global-campaign.org/clientfiles/Cambodia.pdf

Global Campaign for Microbicides (2009b). *Research Rashomon: Lessons from the Cameroon Pre-Exposure Prophylaxis Trial Site*. [Online] Available at: https://www.avac.org/resource/research-rashomon-lessons-cameroon-pre-exposure-prophylaxis-trial-site

Gonzales, R., Handley, M.A., Ackerman, S., and O'Sullivan, P.S. (2012). A framework for training health professionals in implementation and dissemination science. *Academic medicine: journal of the Association of American Medical Colleges*, 87, 271–8.

Grant, R.M., et al. (2010). Preexposure chemoprophylaxis for HIV prevention in men who have sex with men. *The New England Journal of Medicine*, 363, 2587–99.

Hering, J.G. (2016). Do we need 'more research' or better implementation through knowledge brokering?. *Sustainability Science*, 11(2), 363–9.

Hirschhorn, L.R., Ojikutu, B., and Rodriguez, W. (2007). Research for change: using implementation research to strengthen HIV care and treatment scale-up in resource-limited settings. *Journal of Infectious Diseases*, *196* Suppl 3, S516–22.

Holzer, J.K., Ellis, L., and Merritt, M.W. (2014). Why we need community engagement in medical research. *Journal of Investigative Medicine*, 62, 851–5.
Kemp, C.G., et al. (2018). Implementation science for integration of HIV and non-communicable disease services in sub-Saharan Africa: a systematic review. *AIDS*, 32 Suppl 1, S93–s105.
Kim, J. (2018). *Remarks as Prepared for Delivery: World Bank Group President Jim Yong Kim at the Annual Meeting Plenary Session*. Washington, DC: The World Bank.
Knapp, H. and Anaya, H.D. (2013). Implementation science in the real world: a case study of HIV rapid testing. *International Journal of STD & AIDS*, 24, 5–11.
Kruk, M.E., et al. (2016). Transforming Global Health by Improving the Science of Scale-Up. *PLoS Biology*, 14, e1002360.
Logan, J., et al. (2010). The Ottawa model of research use. In: Rycroft-Malone, J. and Bucknall, T. (eds.) *Models and Frameworks for Implementing Evidence-Based Practice: Linking Evidence to Action*. Hoboken/Oxford: Wiley-Blackwell.
Madon, T., Hofman, K.J., Kupfer, L., and Glass, R.I. (2007). Public health. Implementation science. *Science*, 318, 1728–9.
Means, A.R., et al. (2016). The role of implementation science training in global health: from the perspective of graduates of the field's first dedicated doctoral program. *Glob Health Action*, 9, 31899.
Meyers, D.C., Durlak, J.A., and Wandersman, A. (2012). The quality implementation framework: a synthesis of critical steps in the implementation process. *American Journal of Community Psychology*, 50, 462–80.
Mills, E., Rachlis, B., Wu, P., Wong, E., Wilson, K., and Singh, S. (2005). Media reporting of tenofovir trials in Cambodia and Cameroon. *BMC International Health and Human Rights*, 5, 6.
National Implementation Research Network (2015). *Implementation Science Defined.* [Online] Available at: https://nirn.fpg.unc.edu/learn-implemenation-science-defined
Neta, G., et al. (2015). Implementation science in cancer prevention and control: a decade of grant funding by the National Cancer Institute and future directions. *Implement Sci*, 10, 4.
Nilsen, P. (2015). Making sense of implementation theories, models and frameworks. *Implementation Science*, 10, 53.
Odeny, T.A., et al. (2015). Definitions of implementation science in HIV/AIDS. *Lancet HIV*, 2, e178–80.
Padek, M., et al. (2018). Training scholars in dissemination and implementation research for cancer prevention and control: a mentored approach. *Implementation Science*, 13, 18.
Padian, N.S., Holmes, C.B., Mccoy, S.I., Lyerla, R., Bouey, P.D., and Goosby, E.P. 2011. Implementation science for the US President's emergency plan for aids relief (PEPFAR). *Journal of Acquired Immune Deficiency Syndromes*, 56, 199–203.
Palinkas, L.A., Aarons, G.A., Horwitz, S., Chamberlain, P., Hurlburt, M., and Landsverk, J. (2011). Mixed method designs in implementation research. *Administration and Policy in Mental Health and Mental Health Services*, 38, 44–53.
Peters, D.H., Adam, T., Alonge, O., Agyepong, I.A., and Tran, N. (2013). Implementation research: what it is and how to do it. *British Medical Journal*, 347, f6753.
Peters, D. and Adam, T. (2013). *Implementation Research in Health: A Practical Guide*. Geneva: Alliance for Health Policy and Systems Research, World Health Organization.
Proctor, E., et al. (2011). Outcomes for implementation research: conceptual distinctions, measurement challenges, and research agenda. *Administration and Policy in Mental Health and Mental Health Services*, 38, 65–76.
Rabin, B.A., Brownson, R.C., Haire-Joshu, D., Kreuter, M.W., and Weaver, N.L. (2008). A glossary for dissemination and implementation research in health. *Journal of Public Health Management Practice*, 14, 117–23.

Ray, K.N. and Miller, E. (2017). Strengthening stakeholder-engaged research and research on stakeholder engagement. *Journal of Comparative Effectiveness Research*, 6, 375–89.
Robinson, L. (2012). I-Corps and the business of great science. *Journal of Management*, 64(10), 1132.
RE-AIM. [Online] Available at: https://www.re-aim.org/
Salloum, R.G., Shenkman, E.A., Louviere, J.J., and Chambers, D.A. (2017). Application of discrete choice experiments to enhance stakeholder engagement as a strategy for advancing implementation: a systematic review. *Implementation Science*, 12, 140.
Sanders, D. and Haines, A. (2006). Implementation research is needed to achieve international health goals. *PLoS Medicine*, 3, e186.
Singh, J.A. and Mills, E.J. (2005). The abandoned trials of pre-exposure prophylaxis for HIV: what went wrong? *PLoS Medicine*, 2, e234.
Stetler, C. (2010). The Stetler model. In: Rycroft-Malone, J. and Bucknall, T. (eds.) *Models and Frameworks for Implementing Evidence-Based Practice: Linking Evidence to Action*. Hoboken/Oxford: Wiley-Blackwell.
Tabak, R.G., Khoong, E.C., Chambers, D.A., and Brownson, R.C. (2012). Bridging research and practice: models for dissemination and implementation research. *The American Journal of Preventive Medicine*, 43, 337–50.
The Global Fund to Fight AIDS Tuberculosis and Malaria USAID World Health Organization (2008). *Special Program for Research and Training in Tropical Diseases, UNAIDS, The World Bank. Framework for Operations and Implementation Research in Health and Disease Control Programs*. Geneva: WHO.
The Health Foundation (2011). *Evidence Scan: Improvement Science*. [Online] Available at: https://www.health.org.uk/sites/default/files/ImprovementScience.pdf
Toomey, A.H., Knight, A.T., and Barlow, J. (2017). Navigating the space between research and implementation in conservation. *Conservation Letters*, 10(5), 619–25.
UNAIDS (2006). Creating Effective Partnerships for HIV prevention trials: report of a UNAIDS consultation. *AIDS*, 20(6), W1–11.
UNAIDS and AVAC (2007). *Good Participatory Practice Guidelines for Biomedical HIV Prevention Trials*. [Online] Available at: https://www.avac.org/good-participatory-practice
UNAIDS and AVAC (2011). *Good Participatory Practice Guidelines for Biomedical HIV Prevention Trials*. [Online] Available at: https://www.avac.org/good-participatory-practice
Wilson, K.M., Brady, T.J., and Lesesne, C. (2011). An organizing framework for translation in public health: the Knowledge to Action Framework. *Preventing Chronic Disease*, 8, A46.
World Health Organization (WHO) (2016). Good participatory practice guidelines for trials of emerging (and re-emerging) pathogens that are likely to cause severe outbreaks in the near future and for which few or no medical countermeasures exist. [Online] Available at: https://www.who.int/blueprint/what/norms-standards/GPP-EPP-December2016.pdf?ua=1
Yamey, G. (2011). Scaling up global health interventions: a proposed framework for success. *PLoS Medicine*, 8, e1001049.
Zerhouni, E. (2003). Medicine. The NIH roadmap. *Science*, 302, 63–72.

SECTION 7

Environmental and occupational health sciences

7.1 Environmental health methods
Chien-Jen Chen and San-Lin You

7.2 Radiation and public health
7.3 Leeka Kheifets, Adele Green, and Richard Wakeford Occupational health
David Koh and Wee Hoe Gan

7.4 Toxicology and environmental risk analysis
David Koh and Ro-Ting Lin

7.5 Risk perception and communication
7.6 Baruch Fischhoff and Tamar Krishnamurti Urbanization and health
Jason Corburn

7.1

Environmental health methods

Chien-Jen Chen and San-Lin You

Environment and health: historical perspective

Human beings live in a complex environment. The importance of environmental factors in human health has been recognized since antiquity (Chen 2015). Hippocrates, in his classical writing *On Airs, Waters and Places*, emphasized the relevance of the environment to human health. He pointed out the influence of seasons and weather, location of residence, and nature of water on human health, and the occurrence of epidemics. Since ancient times, the seasonality of infectious diseases and the geographical clustering of endemic diseases have been documented worldwide. Several hypotheses had been advanced to explain the unique distribution of various diseases in time and place. The contagion theory proposed that both direct and indirect contact with sick people or animals might transmit diseases and hence both isolation and quarantine had been suggested to prevent the spread of diseases. Although they might have reduced the risk of spread of infectious diseases by ships, the strategies evolved from the contagion theory had not stopped the outbreak of many infectious diseases.

The miasma theory proposed that dirty living environments, unclean drinking water, and poor ventilation due to overcrowding might cause various diseases. This led to environmental sanitary reforms aimed at removing miasma by contributing to a cleaner environment in Europe. The hygienic movement emphasized the importance of implementing a safe drinking water supply and sewage disposal system. The *Public Health Act* enacted in Britain in 1848 was considered the first example of governmental involvement in public health affairs. However, a large outbreak of cholera still occurred in London despite its well-constructed sewage disposal system and much improved environmental sanitation in the 1850s. Through a

carefully designed survey, John Snow published his famous epidemiological study on cholera in 1855 while the germ theory was still in its infancy. He identified the consumption of contaminated drinking water as the cause of the cholera outbreak. The removal of the handle which pumped contaminated water ‘magically’ prevented the spread of the disease. The relocation of sources of drinking water successfully controlled the cholera outbreak. This was a pioneering epidemiological study in the identification of environmental factors in a disease outbreak.

In late nineteenth century, the development of germ theory led to the discovery of many specific infectious agents for human diseases. It also led to the development of several drugs for treatment of diseases, and improvements in sanitation and disinfection to control outbreaks of diseases. Robert Koch’s postulates, which emphasized a one-to-one relationship between cause and disease, were well accepted as criteria for the identification of causal agents of infectious diseases. Effective control of several infectious diseases through identification of specific agents, discovery of chemical and biological drugs, and invention of preventive and therapeutic vaccines had improved life quality and longevity of human beings. The discovery of animal reservoirs and vectors of infectious agents led to the control of vectors through environmental sanitation and disinfection. More and more biological and mechanical vectors have been identified for various viral, bacterial, and parasitic diseases since then. Control of vectors in the environment using pesticides became an important task for the prevention of infectious diseases. For example, malaria was eradicated in Taiwan through the detection and treatment of infected cases and use of the insecticide DDT. However, the widely used pesticides polluted the environment and became hazardous to wild animals and humans.

Life expectancy in most countries increased significantly after the 1950s, when conventional infectious diseases had been controlled through the widespread use of various vaccines and antibiotics. The major diseases that threaten human life have since gradually shifted from infectious to chronic non-communicable diseases. Environmental factors exposed through lifestyle habits, dietary intakes, and environmental pollution play important roles in the development of chronic diseases such as cancers, cardiovascular diseases, and neurological disorders. There are multiple risk factors in the complex multistage pathogenesis of chronic diseases. For examples,

environmental risk factors for liver cancer included hepatitis B virus, hepatitis C virus, aflatoxins, alcohol, vinyl chloride, thorium dioxide, anabolic steroids, and fine particle pollutants (Chen et al. 1997; Chen and Chen 2002; Pan et al. 2016). Some environmental pollutants may cause several chronic diseases in multiple organ systems. For example, arsenic in drinking water has been documented to induce cancers of the skin, bladder, kidney, and lung, as well as Blackfoot disease, ischaemic heart disease, stroke, hypertension, diabetes, erectile dysfunction, mental retardation, neurological disorder, cataract, pterygium, fungal infection, chronic kidney disease, and hepatitis and cirrhosis (Chen 2014; Hsu et al. 2016a, 2016b, 2017). As most chronic diseases have a multistage pathogenesis with a multifactorial aetiology without a unique one-to-one with a specific agent, the classical Robert Koch's postulates for judging the cause of disease are less applicable to chronic diseases. Bradford Hill's causation criteria including the strength, dose–response relation, temporality, specificity, and biological plausibility of the association between cause and disease have been widely adopted for the identification of causal agents for various diseases.

Endemic diseases are often characterized by their unique geographical clustering. This clustering may be associated with biological or chemical agents prevailing in local environments. Some parasitic diseases such as schistosomiasis remain endemic in areas where the ecological system maintains the life cycle of the parasites. The interruption of the life cycle through the identification and treatment of affected patients, eradication of vectors, and improvement of environmental sanitation may decrease the prevalence of these endemic parasitic diseases. Some endemic diseases such as goitre and Blackfoot disease are caused by the deficient or excessive intake of certain chemicals in the environment. The replacement of high-arsenic well water with low-arsenic surface water had decreased the occurrence of Blackfoot diseases in southwestern Taiwan. The endemic goitre caused by low dietary intake of iodine was effectively prevented by the national salt iodization programme in Taiwan.

There has been a long-standing recognition that the workplace environment can induce occupational diseases. Bernardino Ramazzini published his seminal treatise on diseases of workers. Scrotum cancer had long been considered a disease of chimney sweeps. The poor working environment in the nineteenth century sparked reform attempts in Europe,

but specific regulations to protect workers against occupational diseases were mostly implemented in the twentieth century.

Occupational hazards may spread from the workplace to its adjacent environments. For example, a major gas and chemical leak incident in a pesticide plant in Bhopal, India in 1984, and the 1986 accident at the Chernobyl nuclear power plant in Ukraine resulted in environmental disasters. In addition to the immediate casualties, both events also increased the risk of a number of chronic diseases in residents living in the vicinity. The health hazards induced by long-term exposure to industrial wastes are usually more insidious than those resulting from acute environmental disasters. Environmental pollution by industrial waste has been well documented to cause Itai-itai disease and Minamata disease in Japan. Chronic poisoning by heavy metals in industrial waste has become an important issue in environmental health. Soil and water contamination by widely used agricultural chemicals have brought significant threats to human health, wild animal life, and ecological sustainability. Air pollution has increased the risk of pulmonary diseases including asthma, chronic bronchitis, ischaemic heart disease, lung cancer, and liver cancer (Pan et al. 2016).

Natural disasters including snow-storms, hurricanes, floods, sand-storms, volcano eruptions, and forest fires have had serious consequences for public health. Other environmental disasters including earthquakes, tsunamis, and landslides are also detrimental to human health. The explosions that occurred at the Fukushima nuclear power plants in the aftermath of the major earthquake and tsunami in Fukuoka, Japan in 2011, resulted in a multiple environmental disaster. In addition to the heavy immediate casualties, natural environmental disasters threaten the function and efficiency of food, water, and electricity supply systems, ambulance and healthcare systems, sewage and waste disposal, and the control of infectious diseases. Protecting public health through proper environmental management is becoming a lasting challenge.

Causes of human diseases: host–agent interaction in the environment

Both host and environmental factors are involved in the development of human diseases, and the relative importance of environmental to host factors varies as a continuous spectrum for different diseases. Environmental factors seem to play a small role in the development of genetic diseases compared with infectious diseases or accidents. However, exposure to environmental factors may trigger the clinical manifestation of genetic diseases such as phenylketonuria and glucose-6-phosphate dehydrogenase deficiency. Environmental influences on human health are exerted through exposures to physical, chemical, and biological risk factors, and through related changes in human behavioural responses to those factors. The World Health Organization has released a country-by-country analysis of the impact of environmental factors on health (Pruss-Ustun et al. 2016). These data show that environmental risks accounted for a large fraction (23%) of the global burden of deaths and diseases. Environmental impacts on health were uneven across life course and gender. Low- and middle-income countries had the greatest share of environmental disease burden. Total environmental deaths remained unchanged from 2002 to 2012 with a strong sift to non-communicable diseases, and the evidence on quantitative links between health and environment has increased. Global health could be improved by reducing environmental risks including pollution, hazards in the work environment, ultraviolet radiation, noise, agricultural risks, climate, and ecosystem change. Proper environmental management is the key to prevent the proportion of all illnesses which are directly caused by environmental factors.

Human beings live in environments with many causes of diseases as shown in Table 7.1.1. Environmental causes for human diseases include physical, chemical, and biological agents. The physical component includes non-ionizing and ionizing radiation, noise, vibration, pressure, humidity, temperature, earthquakes, landslides, typhoons and hurricanes, and snow-storms and sand-storms. Ionizing radiation includes alpha and beta particles, gamma rays, and X-rays. Ionizing radiation may induce spontaneous abortion, congenital malformation, cancers, and haematopoietic disorders through its effects causing DNA damage and chromosomal aberrations. Non-ionizing radiation includes ultraviolet rays, visible light, infrared rays, microwaves, and electromagnetic fields. Ultraviolet rays may induce skin cancer and cataract; while electromagnetic fields may induce some cancers. The chemical component includes heavy metals, organic solvents,

agricultural chemicals, polycyclic hydrocarbons, and chlorinated organic compounds. Their health effects may be classified as acute, subacute, and chronic toxicity, carcinogenicity, mutagenicity, and teratogenicity. Some chemicals may persist in the environment and result in bioaccumulation. The biological component includes viruses, bacteria, fungi, parasites, allergens, arthropods, as well as plant and animal toxins. A number of infectious diseases of various organ systems are caused by biological agents. In addition to the acute and subacute symptoms and signs, these infectious agents may also induce chronic diseases such as cancers of the nasopharynx, stomach, liver, bladder, cervix, uteri, and lymphoid and soft tissue (Chen et al. 2014), pulmonary diseases, cardiovascular diseases, as well as neurological disorders.

Table 7.1.1 Environmental components of causes for human diseases

Component	Category
Physical	Ionizing radiation (α-particles, β-particles, γ-rays, X-rays)
	Non-ionizing radiation (ultraviolet rays, visible light, infrared rays, microwaves, electromagnetic fields)
	Noise
	Vibration
	Pressure
	Humidity and temperature
	Earthquakes and landslides
	Typhoons and hurricanes
	Snow-storms and sand-storms
Chemical	Heavy metals
	Organic solvents
	Agricultural chemicals
	Polycyclic hydrocarbons
	Chlorinated organic compounds
Biological	Viruses
	Bacteria
	Fungi
	Parasites
	Allergens
	Arthropods
	Animal and plant toxins

Host components in response to environmental components are also important in the determination of a number of human diseases as shown in Table 7.1.2. The genetic component includes a single gene for autosomal, X-

linked, Y-linked, and mitochondrial diseases, multiple genes for common diseases such as cancers and cardiovascular diseases which run in the family, and chromosomal aberrations. There also exist gene–gene interaction and gene–environment interaction in the development of human diseases. Genetic effects may be modified by environmental factors, and environmental factors may cause genetic damages and mutations. The behavioural component includes substance use, dietary intake, personal hygiene including hand-washing and safe sex, health promotion and disease prevention practice, transportation safety practice, occupational protection practice, sports and exercise, personal communication and networking, and response to stressful life events. All these behavioural components play very important roles in the prevention of environmental health hazards. For example, using cleaner fuel such as gas or electricity, using better cooking devices, improving ventilation, or keeping children away from smoke could reduce environmental risk of respiratory infections and diseases among women and children. Enforcement of helmet and seat belt laws may significantly decrease the mortality from motor vehicle accidents. The social component of human diseases includes built environments, occupational settings, agricultural methods and irrigation schemes, transportation systems, water supply systems, sewage and waste disposal, schooling systems, healthcare systems, social welfare and criminal prevention systems, anthropogenic climate changes, and ecosystem degradation. The improvement in environmental hygiene and public health systems has increased the global health status and life expectancy significantly in the last decades.

Table 7.1.2 Host components of causes for human diseases

Component	Category
Genetic Behavioural	Single gene (autosomal, X-linked, Y-linked, mitochondrial dominant/recessive)
	Chromosomal aberration
	Multiple genes (gene–gene and gene–environment interactions)
	Substance abuse (alcohol, tobacco, betel, drugs)
	Dietary intake (malnutrition, overnutrition)
	Personal hygiene and safe sex Health promotion and disease prevention practice Transportation safety practice (helmet and seat belt)
	Occupational protection practice
	Sports and exercise
	Personal communication and networking
	Stressful life events
Social	Housing and home safety
	Water supply systems
	Sewage and waste disposal
	Transportation systems
	Agricultural and irrigational methods
	Working environment
	Industrial hygiene and safety
	Schooling systems
	Healthcare systems
	Anthropogenic pollution and climate changes
	Ecosystem degradation

Several causal models have been proposed to describe the interaction between host and environment components in human health (Chen 2015).

The traditional epidemiological triangle model emphasizes the equal importance of host, agent, and environment, which are located in three angles of the epidemiological triangle. This causal model emphasizes the health effects of a specific agent, especially infectious agents, may be modified by both host and environment factors. Host factors may include the biological, sociodemographic, and behavioural characteristics including individual susceptibility and immunity, age, gender, race, socioeconomic status, occupation, genetic composition, body mass index, dietary intake, lifestyle habits, personal hygiene, and social activities. Environmental factors include living and working environments such as family, school, workplace, recreation, and healthcare place, where both agent and host interact. The disease agent may spread and transmit in the environment and affect susceptible individuals. Various strategies to prevent the occurrence of the disease caused by the agent may be developed through the understanding of the epidemiological triangle model. The agent may be controlled through disinfection and sterilization, antibiotic or antiviral treatment, intervention of drug resistance, and environmental sanitation. Active and passive immunization, improvement in nutritional status, optimal rest and exercise, immunomodulation, chemoprevention, and other host factors may be strengthened to reduce the risk of disease occurrence. The environment condition may be improved to prevent the occurrence of diseases through clean housing and workplaces, safe water and food supply, hygienic sewage and waste disposal, and others.

The epidemiological triangle model emphasizes the unique causal agent of a specific disease. This model is most appropriate for infectious diseases because the infectious agent is the necessary cause of a given infectious disease, even though the infectious agent may not be sufficient to cause the given disease. For many infectious agents, not all infected individuals will develop the disease. The one-to-one causal relation between agent and disease may not be observed under this circumstance. The triangle model is not appropriate for most chronic non-communicable diseases. Most chronic diseases have multiple causes. A single agent is neither necessary nor sufficient to cause the disease. The ecological wheel model was proposed to emphasize the host–environment interaction in the determination of human health. The ecological wheel model includes an axle (host component) surrounded by a tyre (environmental component). A disease may be caused by many environmental and host factors which are neither necessary nor

sufficient causes. They are considered as contributory causes of the disease. No unique single agent is specified for a disease in this model. The environmental component included various physicochemical, biological, and social environments just described. The host component included the genetic composition located in the core surrounded with individual immunity, lifestyle, personality, dietary intake, and social networking. The ecological wheel model emphasizes the importance of ecological balance in the maintenance of the health status of a host. The relative importance of host and various environment factors varies by diseases. For example, genetic factors are most important for the highly inheritable diseases; the biological environment for various infectious diseases; and the social environment for accidents. This casual model also stresses the interactions among host and various environments.

Neither the epidemiological triangle nor the ecological wheel takes the time dimension of disease development into consideration. The natural history of both acute and chronic diseases follows a specific temporal sequence. This usually starts from the contact of causal agent(s) with target cells, through the structural and functional changes in affected cells, the proliferation of causal agents or transformed cells, the development of lesions in target tissues or organs, the onset of clinical symptoms and signs, to the development of systemic illness. There are various agents and environment factors involved at different stages of the pathological progression. For example, the multistage process of carcinogenesis may be classified into initiation, promotion, progression, invasion, and metastasis. The development of atherosclerotic diseases includes stages of initial lesion, fatty streak, intermediate lesion, atheroma, fibroatheroma, complicated lesions, and clinical manifestation. An evolutional spiral model has been proposed to describe the disease causation through host–environment interactions over the entire natural history of the disease (Chen 1999). At the initial stage, both cause and molecular or cellular change are simple and microscopic, and the causal relationship tends to be one-to-one. With disease progression, however, the affected lesions keep expanding like a spiral with more and more complicated causal interactions between host and environment components at different stages.

The evolutional spiral model is most appropriate for the description of multistage pathogenesis of a disease with a multifactorial aetiology. As shown in Figure 7.1.1, the development of hepatocellular carcinoma (HCC)

is a multistage process with the involvement of multiple risk factors. There are many causes of HCC including hepatitis B virus (HBV), hepatitis C virus (HCV), aflatoxins, alcohol, tobacco, non-alcoholic fatty liver disease, diabetes mellitus, haemochromatosis, alpha 1-antitrypsin deficiency, and others. The exposure to causal agents may lead from mutations or epigenetic alterations of affected hepatocytes, through liver inflammation, chronic hepatitis, liver fibrosis, and cirrhosis, to HCC. During the hepatocarcinogenic process, the affected target starts from molecules and expands to cells, tissues, and organs like a spiral with more and more contributory causes involved to drive the progression of the disease. In addition to exposure to various chemical carcinogens, the inadequate intake of micronutrients including carotenoids and selenium is also associated with an increased risk of HCC. Host factors including humoral and cellular immunity, xenobiotic metabolism, hormones, and susceptibility genotypes can modify the risk of HCC development. Healthcare factors including immunization, antiviral therapy, liver disease screening, and other clinical interventions may be used to stop or slow the occurrence or recurrence of HCC.

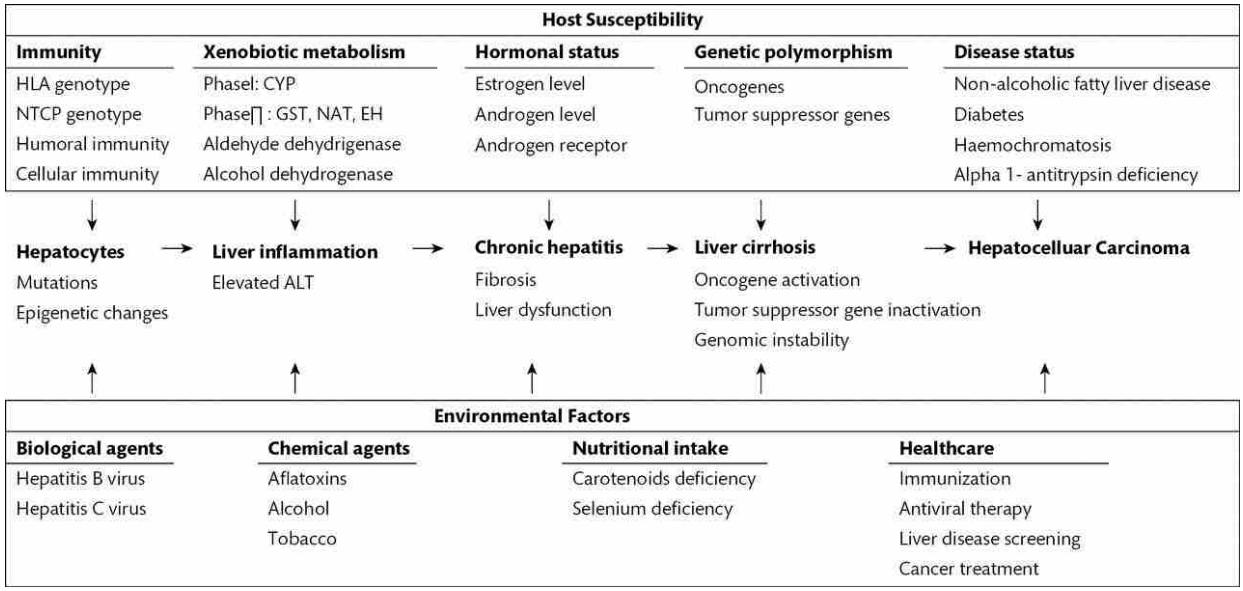


Fig. 7.1.1 Multistage hepatocarcinogenesis with a multifactorial aetiology.

Ecological studies: identification of environmental health hazards at aggregate level

Geographical clustering of endemic diseases in a small area is the most important characteristic of environmental health hazards. An extremely high prevalence of an endemic disease in a confined area should always lead to a search for unique environmental risk factors. Ecological studies are often used to compare the morbidity or mortality of environmental disease between endemic and non-endemic areas. In ecological studies, aggregate attributes such as the morbidity of residents in exposed and unexposed areas, rather than individual attributes such as the health status of exposed and unexposed subjects, are analysed and compared. Ecological studies tend to compare health indices in areas with and without the environmental exposure, or to examine quantitatively the correlation between health indices and the magnitude of environmental exposure in many areas. Some environmental characteristics or comorbidity rates may be used as surrogate indicators of the causal agents to elucidate the relation between the disease risk and the causal agent. The interpretation of ecological studies may be limited by the potential existence of ecological fallacy, when the agent–disease correlation observed at the aggregate level may not consistently be found at the individual level. However, ecological studies are frequently used to explore the environmental component of causes for human diseases.

As an illustrative example, Blackfoot disease is an endemic peripheral vascular disorder confined to a limited area on the southwest coast of Taiwan. Clinically the disease starts with numbness or coldness of one or more extremities and intermittent claudication, which progresses to black discoloration, ulcer, and gangrene. In end stages of the disease, spontaneous amputation of the distal parts of the affected extremities is common. The disease was hyperendemic in four neighbouring townships in southwestern Taiwan. As the water from shallow wells (6–8 metres in depth) in this area had high salinity, residents in some villages had started to use water from artesian wells (100–200 metres deep) since the 1920s. The use of artesian well water was found to be associated with the occurrence of the Blackfoot disease, and water samples from artesian wells of the endemic area were found to have high arsenic concentrations, which ranged from 0.35 to 1.14 mg/L.

Based on the analysis of mortality data, residents in the endemic area of Blackfoot disease had significantly elevated mortality from cancers of the skin, bladder, kidney, lung, and liver as shown in Table 7.1.3. Compared with the general population in Taiwan, the age-standardized mortality ratios for cancer of the bladder, kidney, skin, and lung were threefold or greater among male and female residents in the endemic area (Chen et al. 1985), suggesting arsenic in artesian water may cause these cancers as well as Blackfoot disease. As the four townships in this endemic area had different prevalences of Blackfoot disease, cancer mortality was further compared among townships. There was a trend that the higher the Blackfoot disease prevalence of a township, the greater the age-standardized mortality ratios for cancers of the skin, bladder, kidney, and lung.

Table 7.1.3 Age-standardized mortality ratios[1] of skin and internal cancers in four townships of arseniasis-endemic area in southwestern Taiwan by Blackfoot disease prevalence

Cancer	Gender	Blackfoot disease endemic area	Township by Blackfoot disease endemicity[2]			
			Very high	High	Medium	Low
Skin	Male	5.3	8.4	8.3	5.3	1.3
	Female	6.5	6.8	15.7	3.6	1.9
Bladder	Male	11.0	13.8	15.6	10.7	6.1
	Female	20.1	38.0	34.6	16.1	5.6
Kidney	Male	7.7	14.0	12.4	8.2	4.8
	Female	11.2	14.7	22.5	8.5	3.2
Lung	Male	3.2	3.8	5.3	2.6	1.7
	Female	4.1	6.3	6.8	3.3	1.9
Liver	Male	1.7	1.6	1.9	1.7	1.6
	Female	2.3	1.8	2.3	2.5	2.4

[1] Mortality rates of general population in Taiwan were used as the standard rates to calculate standardized mortality ratio (Taiwan = 1.0).

[2] Four townships including Peimen, Hsuehchia, Putai and Ichu with Blackfoot disease prevalence of 5.67, 3.87, 2.02 and 0.64 per 1,000, respectively. Age-standardized mortality ratios were calculated for males and females, respectively.

Source: data from Chen CJ et al., Malignant neoplasms among residents of a Blackfoot disease-endemic area in Taiwan: High-arsenic artesian well water and cancers, *Cancer Research*, Volume 45, pp. 5895–5899, Copyright © 1985 American Association for Cancer Research.

As 84 villages in these four townships also had different prevalences of Blackfoot disease and used different types of well water, more refined analyses of cancer mortality at the village level were carried out. A significant dose–response relation between the prevalence of Blackfoot disease and the age–sex-standardized mortality ratios for cancers of the skin, bladder, kidney, lung, and liver was observed at the village level as shown in Table 7.1.4. Furthermore, biological gradients of cancer mortality were also found by the type of wells used as drinking water source with the highest

cancer mortality in villages where only artesian wells were used and the lowest in villages which used only shallow wells. As both Blackfoot disease and artesian well use were significantly associated with the arsenic in drinking water, they were used as the surrogate environmental variable for arsenic exposure. It was thus implied that arsenic in drinking water might increase the risk of skin and internal cancers. However, the best way to assess the cancer risk associated with arsenic in drinking water is to obtain the arsenic levels in drinking water from wells.

Table 7.1.4 Age-sex-standardized mortality ratios[1] of skin and internal cancers in 84 villages of arseniasis-endemic area in southwestern Taiwan by Blackfoot disease prevalence and drinking water source

Cancer	Village by Blackfoot disease endemicity[2]			Village by types of wells as major drinking water source[3]		
	High	Medium	Low	Artesian only	Both	Shallow only
Skin	11.2	6.5	2.2	10.9	6.5	2.2
Bladder	40.3	17.2	3.8	26.3	11.2	4.5
Kidney	14.9	7.7	3.2	9.2	7.5	3.3
Lung	6.2	3.1	1.8	4.6	2.8	1.8
Liver	3.0	1.5	1.3	2.0	1.8	1.4

[1] Mortality rates of general population in Taiwan were used as the standard rates to calculate standardized mortality ratio (Taiwan = 1.0).

[2] A total of 84 villages were classified into three groups by Blackfoot disease prevalence: 0, 0.1–5.0 and >5.0 per 1,000, respectively.

[3] A total of 84 villages were classified into three groups by type of wells used as major drinking water source: artesian wells only, both artesian and shallow wells, and shallow wells only.

Reproduced with permission from Chen CJ et al., Malignant neoplasms among residents of a Blackfoot disease-endemic area in Taiwan: High-arsenic artesian well water and cancers, *Cancer Research*, Volume 45, pp. 5895–9, Copyright © 1985 American Association for Cancer Research.

The arsenic levels in water of shallow and artesian wells in 42 villages of the Blackfoot disease endemic area were available for further analysis of the association between arsenic in drinking water and cancer mortality. The age-adjusted morality for various cancer sites and the average concentration of

arsenic in water of wells in each village were derived. As shown in Table 7.1.5, there was a significant dose–response relation between arsenic level in well water and age-adjusted mortality for cancers of the bladder, kidney, lung, skin, and liver in both males and females, and prostate cancer in males at the village level (Chen et al. 1988a). In these 42 villages, a significant dose–response relation was also observed between level of arsenic in well water and age-adjusted mortality for cardiovascular diseases and peripheral vascular diseases at the village level for both males and females (Wu et al. 1989).

Table 7.1.5 Age-standardized mortality (per 100,000) from skin and internal cancers in 42 villages of arseniasis-endemic area in southwestern Taiwan by arsenic level in well water

Cancer	Gender	Village by arsenic level in well water			General population in Taiwan
		0.60+	0.30–0.59	<0.30	
Skin	Male	28.0	10.7	1.6	0.8
	Female	15.1	10.0	1.6	0.8
Bladder	Male	89.1	37.8	15.7	3.1
	Female	91.5	35.1	16.7	1.4
Kidney	Male	21.6	13.1	5.4	1.1
	Female	33.3	12.5	3.6	0.9
Lung	Male	87.9	64.7	35.1	19.4
	Female	83.8	40.9	26.5	9.5
Liver	Male	68.8	42.7	32.6	28.0
	Female	31.8	18.8	14.2	8.9
Prostate	Male	8.4	5.8	0.5	

Reproduced from *The Lancet*, Volume 1, Number 8582, Chen CJ et al., Arsenic and cancers (letter), pp. 414–15, Copyright © 1988, with permission from Elsevier, http://www.sciencedirect.com/science/journal/01406736.

In another large-scale ecological correlation study on age-adjusted cancer mortality of 314 precincts and townships in Taiwan (Chen and Wang 1990), the arsenic level in water of 83,656 wells was tested to derive the average arsenic level in well water in each precinct or township. Both cancer deaths and mid-year population in all study precincts and townships were obtained from the national death certification and household registration system. Based on the weighted multiple regression analysis, there were significant correlations between the average arsenic level in well water and the age-adjusted mortality rates of cancers of the liver, nasal cavity, lung, skin, bladder, kidney, and prostate at the precinct/township level. The increase in age-adjusted mortality per 100,000 person-years for every 0.1 ppm increase in arsenic level of well water for cancers of the liver, nasal cavity, lung, skin, bladder, and kidney, respectively, was 6.8, 0.7, 5.3, 0.9, 3.9, and 1.1 for males and 2.0, 0.4, 5.3, 1.0, 4.2, and 1.7 for females, after adjustment for indices of industrialization and urbanization. The increase in prostate cancer mortality for every 0.1 ppm increase in arsenic level in well water was 0.5 per 100,000 person-years. In these ecological correlation studies, both disease risk and environmental exposure were measured at a group (village or precinct/township) level. There exist potential ecological fallacies c and confounding effects of extraneous factors such as prevalence of cigarette smoking, alcohol drinking, environmental pollution, viral hepatitis, and others. The findings of ecological correlation studies have to be further validated by studies at the individual level.

Ecological correlation studies have also been carried out to imply possible causal associations between infectious agents and cancer (Chen et al. 2014). The similarity in geographical variation in prevalence of oncogenic infectious agent and incidence of specific cancer was observed for HBV/HCV and HCC, herpesvirus and Kaposi’s sarcoma, human T-cell leukaemia virus and leukaemia, liver fluke and cholangiocarcinoma, as well as *Helicobacter pylori* and stomach cancer. For example, the incidence of liver cancer and the prevalence of HBV and HCV infection are extremely high in the East, Southeast and Central Asia, and the sub-Saharan Africa. But the etiological proportion of liver cancer attributable to HBV and HCV varies in different countries and regions. HCV is more prevalent in Japan and Central Asia, and HBV is more prevalent in Taiwan, Southeast Asia, and China. Among townships in Taiwan, a significant ecological correlation between the HCC mortality and the prevalence of chronic HBV infection

was also observed. These ecological study findings suggest the importance of chronic HBV and HCV infection in the development of HCC. As around 75% of global liver cancer cases are caused by HBV and HCV. Thus, control of viral hepatitis may significantly reduce the burden of HCC globally (Chen 2018).

Cross-sectional surveys and longitudinal studies: assessment of environmental health hazards at the individual level

Several different study designs have been applied to assess the environmental health hazards at the individual level as shown in Table 7.1.6. Based on the time of the collection of disease risk (outcome) and environmental exposure (cause) data, they may be classified into cross-sectional and longitudinal studies. Cross-sectional studies collect the data on disease and exposure at the same time. The most common type of cross-sectional studies is the environmental health survey, in which participants are enrolled from exposed and unexposed areas to collect the information on health status, exposure to environmental factors, and potential confounding factors. Morbidity estimates from cross-sectional surveys represent disease prevalence rather than incidence. As the disease and exposure data are collected simultaneously, the causal temporality of disease and exposure needs further validation. If the environmental exposure is quite consistent over a long period of time before the onset of the disease, the causal temporality may be considered correct. Longitudinal studies may be further classified into case–control and cohort studies according to the enrolment of study participants. Conventional case–control studies recruit incident cases and matched controls from healthcare institutions or community, and collect the history of exposure to environmental factors through questionnaire interviews or biomarker tests. While the causal temporality is accounted for, the information on environmental exposure history may be subject to recall loss or bias and biomarkers may not reflect a long-term exposure before the disease onset. If there were detailed records of exposure history or biomarkers of long-term exposure, the biases may be reduced significantly. Conventional cohort studies recruit exposed and unexposed healthy participants from community or workplace, and follow the cohort until an adequate proportion of participants develop the disease under study. While

there is no bias of the exposure information, it may be limited by participants being lost to follow-up and resulting in the bias of disease ascertainment. If databases of disease registration, death certification, or health insurance were available, the loss to follow-up may be solved through computerized linkage with these databases.

Table 7.1.6 Major study designs of investigation on environmental health hazards at individual level

Temporality between disease and exposure	Collection of disease and exposure data	Study subjects selection	Limitations	Association estimates
Cross-sectional	Simultaneous	Random sample of population	Temporal correctness	Odds ratio
Longitudinal	Retrospective (retrieval of previous exposure)	Affected cases and unaffected controls	Recall bias	Odds ratio
Longitudinal	Prospective (follow-up of future disease)	Exposed and unexposed cohorts	Loss of follow-up	Hazard ratio

Returning to Blackfoot disease as an illustrative example, in a community-based cross-sectional survey of 40,421 residents in 37 villages in the endemic area in southwestern Taiwan, the relation between status of skin cancer and Blackfoot disease in participants, and arsenic in well water was examined (Tseng 1977). There was a significant dose–response relation between the arsenic concentration in the well water and the prevalence of skin cancer and Blackfoot disease for most age and sex groups. There was also a significant coexistence of both skin cancer and Blackfoot disease. In a community-based case–control study, a total of 353 pairs of Blackfoot disease cases and matched controls were interviewed (Chi and Blackwell 1968). The length of residing in the endemic villages was significantly associated with the risk of developing Blackfoot disease in a dose–response relation. While all 353 (100%) Blackfoot disease cases had ever used artesian well water as the principal source of drinking water during the 15 years before disease onset, only 233 (66%) matched controls had consumed artesian well water ($p < 0.001$). Blackfoot disease cases had lower

educational level and socioeconomic status than matched controls. In another case–control study on 241 Blackfoot disease patients and 759 matched healthy controls, questionnaire interviews were carried out to obtain information on the history of consuming artesian well water, diet, alcohol intake, and cigarette smoking, as well as the presence of arsenic-induced skin keratosis and cancer, and Blackfoot disease in first-degree relatives (Chen et al. 1988b). There was a significant dose–response relation between the risk of developing Blackfoot disease and the duration of consumption of artesian well water. In addition, arsenic-induced skin lesions, family history of Blackfoot disease, and undernourishment were also associated with the development of Blackfoot disease independent of the duration of consuming artesian well water.

In a cohort study of cause-specific mortality of 789 Blackfoot disease patients followed for 15 years (Chen et al. 1988b), age–sex-standardized mortality ratios were derived for Blackfoot disease patients using the mortality rates of the general population in Taiwan and the Blackfoot disease endemic area as the standard rates, respectively. As shown in Table 7.1.7, patients of Blackfoot disease had increased mortality from cardiovascular disease and cancers of the skin, bladder, lung, kidney, liver, and prostate compared to two reference populations. In another case–control study on cancers of bladder, lung, and liver in the endemic area of Blackfoot disease (Chen et al. 1986), the duration of consuming artesian well water was significantly associated with three cancers in a dose–response relation after adjustment for age and sex.

Table 7.1.7 Age-standardized mortality ratios for cancers and vascular diseases for Blackfoot disease patients in comparison with two referent populations

Disease	Standardized mortality ratio in Blackfoot disease patients	
	Taiwan population as reference[1]	Endemic area population as reference[2]
Skin cancer	28.5**	4.5*
Bladder cancer	38.8***	2.6**
Kidney cancer	19.5	1.6
Lung cancer	10.5***	2.8**
Liver cancer	4.7***	2.5**
Prostate cancer	17.3	2.7
Peripheral vascular disease	12.4***	3.5***
Cardiovascular disease	2.1***	1.6**
Cerebrovasular disease	1.2	1.1

[1] Mortality rates of general population in Taiwan were used as the standard rates (Taiwan = 1.0)

[2] Mortality rates of population in Blackfoot disease endemic area were used as the standard rates (endemic area = 1.0)

* p <0.5,

** p <0.01,

*** p <0.001

Reproduced with permission from Chen CJ et al., Atherogenicity and carcinogenicity of high-arsenic artesian well water: Multiple risk factors and related malignant neoplasms of Blackfoot disease, *Arteriosclerosis, Thrombosis, and Vascular Biology*, Volume 8, pp. 452–60, Copyright © 1988 American Heart Association, Inc. All rights reserved. DOI: 10.1161/01. ATV.8.5.452.

A community-based health survey study was carried out in three Blackfoot disease-hyperendemic villages to examine the associations between arsenic in drinking water and various chronic diseases. A total of 1,571 residents aged over 30 years were interviewed by public health nurses, and 1,071 of them had health examinations. Residents in these villages started using artesian wells as the principal drinking source in early 1910s,

and shifted to the public water supply system using surface water from a distant reservoir in early 1970s. The arsenic level in water of artesian wells of the Blackfoot disease-hyperendemic villages was surveyed in the early 1960s. As residents in a given village shared a few wells together, the median arsenic level of the water from the shared wells was used as the level of arsenic in drinking water for residents of the village. As a participant might migrate from one village to another, a detailed life history of residence and consumption of artesian well water was obtained through a structured questionnaire interview. The cumulative arsenic exposure was derived by the formula $\Sigma (C_i \times D_i)$; where C_i was the median arsenic level in water of shared wells of a village in which the participant lived, and D_i was the duration of drinking artesian well water in the village during the consecutive period i. The average arsenic exposure level in drinking water of a participant was derived by $\Sigma (C_i \times D_i)/\Sigma (D_i)$. Residents in hyperendemic villages of Blackfoot disease had a significantly higher age-specific prevalence of diabetes mellitus (Lai et al. 1994) and hypertension (Chen et al. 1995) than residents in non-endemic areas. Furthermore, there was a significant dose–response relation between the cumulative arsenic exposure and prevalence of diabetes mellitus (Lai et al. 1994), hypertension (Chen et al. 1995), peripheral vascular disease based on Doppler ultrasonography (Tseng et al. 1996), carotid atherosclerosis based on duplex ultrasonography (Wang et al. 2002), electrocardiogram-based ischaemic heart disease (Tseng et al. 2003), cataract (See et al. 2007), and pterygium (Lin et al. 2008) as shown in Table 7.1.8. Cohort follow-up studies also found a significant biological gradient of incidence of skin cancer (Hsueh et al. 1997) and lethal ischaemic heart disease (Chen et al. 1996a) with increasing cumulative arsenic exposure.

There was another survey of well water arsenic and health status enrolling a total of 8,102 residents in four townships located in northeast Taiwan. Residents in these townships had their own tube wells in their backyards. The arsenic levels in well water of these households were tested in early 1990s. The public water supply system using surface water was introduced in this area in early 1990s, and its coverage was as high as 95% in the year 2000. There was a dose–response relation between the cumulative exposure of arsenic in well water and prevalence of cerebrovascular disease, especially cerebral infarction, which was identified from home-visit personal interviews and ascertained through the review of hospital medical records

(Chiou et al. 1997). A significant dose–response association with arsenic level in drinking water was also observed for the prevalence of chronic hepatitis or cirrhosis in participants who were seronegative for hepatitis B surface antigens (HBsAg) and antibodies against HCV (Hsu et al. 2016a, b), as well as chronic kidney disease (Hsu et al. 2017).

Table 7.1.8 Dose–response relationship between cumulative arsenic exposure and risk of various diseases. Figures are relative risks (95% confidence intervals)

Disease/cause of death	Cumulative arsenic exposure			
	Low	Medium	Medium high	High
Skin cancer[1]	1.0 (referent)	2.8 (0.3–31.9)	2.6 (0.3–22.9)	7.6 (1.0–60.3)
Hypertension[2]	1.0 (referent)	0.9 (0.2–3.3)	2.4 (0.8–6.9)	3.6 (1.4–9.6)
Ischaemic heart disease death[3]	1.0 (referent)	2.5 (0.5–11.4)	4.0 (1.0–15.6)	6.5 (1.9–22.2)
Cataract (posterior subcapsular opacity)[4]	1.0 (referent)	2.2 (0.4–12.1)	4.8 (1.0–22.2)	5.7 (1.2–26.3)
Hepatitis and cirrhosis (without chronic viral hepatitis)[5]	1.0 (referent)	1.4 (0.6–3.7)	2.2 (0.8–6.4)	3.3 (1.2–9.1)
Chronic kidney disease[6]	1.0 (referent)	1.2 (0.9–1.5)	1.4 (1.1–1.8)	1.4 (1.0–1.8)
	Low	**Medium**	**High**	
Diabetes mellitus[7]	1.0 (referent)	6.6 (0.9–51.0)	10.1 (1.3–77.9)	
Electrocardiogram-based ischaemic heart disease[8]	1.0 (referent)	1.6 (0.5–5.3)	3.6 (1.1–11.7)	
Doppler ultrasonography-based peripheral vascular disease[9]	1.0 (referent)	2.8 (0.9–9.1)	4.3 (1.3–14.5)	
Duplex ultrasonography-based carotid atherosclerosis[9]	1.0 (referent)	1.8 (0.8–3.8)	3.1 (1.3–7.4)	
Cerebral infarction[10]	1.0 (referent)	2.7 (1.2–5.8)	3.4 (1.4–8.1)	
Pterygium[11]	1.0 (referent)	3.1 (1.8–5.6)	4.1 (2.2–7.4)	

[1] Follow-up study in southwestern Taiwan, arsenic level stratified as 0, 0.1–10.6, 10.7–17.7, >17.7 mg/L-years.

[2] Survey in southwestern Taiwan, arsenic level stratified as 0, 0.1–6.3, 6.4–10.8, 10.9–14.7 mg/L-years (two groups >14.7 mg/L-years are not shown here).

[3] Follow-up study in southwestern Taiwan, arsenic level stratified as 0, 0.1–9.9, 10.0–19.9, >19.9 mg/L-years.

[4] Survey in southwestern Taiwan, arsenic level stratified as 0, 0.1–12.0, 12.1–20.0, >20.0 mg/L-years.

[5] Survey in northeastern Taiwan, arsenic level stratified as ≤0.010, 0.010–0.049, 0.100–0.299, >0.300 mg/L.

[6] Survey in northeastern Taiwan, arsenic level stratified as ≤0.010, 0.010–0.049, 0.050–0.149, >0.150 mg/L.

[7] Survey in southwestern Taiwan, arsenic level stratified as 0, 0.1–15.0, >15.0 mg/L-years.

[8] Survey in southwestern Taiwan, arsenic level stratified as 0, 0.1–14.9, >14.9 mg/L-years.

[9] Survey in southwestern Taiwan, arsenic level stratified as 0, 0.1–19.9, >19.9 mg/L-years.

[10] Survey in northeastern Taiwan, arsenic level stratified as <0.1, 0.1–4.9, >4.9 mg/L-years.

[11] Survey in southwestern Taiwan, arsenic level stratified as 0, 0.1–15.0, >15.0 mg/L-years.

Source data from: skin cancer (Hsueh et al. 1997), hypertension (Chen et al. 1995), ischaemic heart disease death (Chen et al. 1996a), cataract (See et al. 2007), diabetes mellitus (Lai et al. 1994), electrocardiogram-based ischaemic heart disease (Tseng et al. 2003), Doppler ultrasonography-based peripheral vascular disease (Tseng et al. 1996), duplex ultrasonography-based carotid atherosclerosis (Wang et al. 2002), cerebral infarction (Chiou et al. 1997), pterygium (Lin et al. 2008), cirrhosis and hepatitis (Hsu et al. 2016), and chronic kidney disease (Hsu et al. 2017).

The occurrence of cancers of cohort participants enrolled from the endemic areas of arseniasis in southwestern and northeastern Taiwan was identified and ascertained through computerized data linkage with national cancer registry and death certification profiles. A dose–response relation was observed between long-term arsenic exposure from drinking well water and incidence of urothelial carcinoma (Chiou et al. 2001) and lung cancer (Chen et al. 2004). There was a synergistic effect on the development of lung cancer for cigarette smoking and exposure to arsenic in drinking water. Based on follow-up data from northeastern Taiwan, the cumulative lifetime (35–85 years) risk of lung cancer and urothelial carcinoma was found to increase with the arsenic concentration in well water (Yang et al. 2013). The lifetime lung cancer risk for arsenic levels in drinking water of <10, 10–99, and 100+ μg/L was 7.42%, 8.99%, and 17.09%, respectively; and the corresponding lifetime risk of urothelial carcinoma was 0.29%, 1.07%, and 3.43%, respectively. It is essential to ensure accuracy in the assessment of exposure to environmental factors in all cross-sectional surveys, case–control studies, and cohort studies in order to identify and characterize the environmental health hazards efficiently and effectively. It is also important

to be aware of the possibility of temporal variations in concentration of the causal agent in various environmental media, individual variations in frequency and quantity of contact with the contaminated environmental media, and variations in detection limit of various exposure assessment methods.

As one agent may cause many diseases like arsenic in drinking water, one disease may be caused by many agents. The multifactorial aetiology of HCC have been explored by many case–control and cohort studies (Chen et al. 1997; Chen 2018). The causes of HCC include HBV and HCV. In the 1970s, coexistence of HCC and HBsAg seropositivity, the biomarker of chronic HBV infection, was observed in several cross-sectional case–control studies. In the natural history of chronic HBV infection, spontaneous seroclearance of important milestone biomarkers including hepatitis B e-antigen (HBeAg), HBV DNA and HBsAg may occur (Chen and Yang 2011). HBeAg seroclearance occurs after a period of elevation of serum level of alanine aminotransferase (ALT), a seromarker of liver inflammation. The serum HBV DNA level (viral load) may remain high or gradually decline in HBeAg-seronegative patients. Seroclearance of HBsAg may occur after the serum HBV DNA level becomes undetectable. The seroclearance of these HBV biomarkers may occur before the time of HCC diagnosis. As HBeAg, HBV DNA, and HBsAg may be serocleared gradually with increasing age, their seroprevalence at the onset of HCC is much lower than that in childhood or young adulthood. It should thus be noted that cross-sectional studies might underestimate the relative HCC risk associated with the seropositivity of HBsAg, HBV DNA, and HBeAg if HCC cases and controls were not well matched on age. This may explain the wide range of relative HCC risk for HBsAg seropositivity in various cross-sectional case–control studies.

Longitudinal cohort studies may better elucidate the HCC risk associated with these risk predictors. The causality between chronic HBV infection and HCC was first confirmed by the GECC cohort study in Taiwan (Beasley et al. 1981), the first prospective follow-up study on virus-caused human cancer. The REVEAL-HBV cohort study has documented the increased risk of HCC and cirrhosis associated with HBeAg serostatus and increasing serum HBV DNA level in HBsAg-seropositive patients (Yang et al. 2002; Chen et al. 2006; Iloeje et al. 2006). Repeated measurements of serial biomarkers during the follow-up period may help examine the causal

temporality more accurately. For example, the long-term trajectory rather than baseline levels of HBV viral load has been documented to predict HCC risk more precisely (Chen et al. 2011). There was a significant dose–response relation between HBV viral load at study entry and risk of HCC newly developed during the follow-up. However, the HCC risk significantly varied among chronic hepatitis B patients with high viral loads at study entry. The HCC risk was highest for those maintaining a high viral load during the follow-up, and lowest for those whose viral load declined significantly to low levels. In other words, the follow-up biomarkers were very useful to stratify patients into more refined HCC risk groups. It has also been well documented that seroclearance of HBV biomarkers lead to a significant decrease in HCC risk (Liu et al. 2014). These observational studies suggest the most effective and efficient way to reduce the HCC risk is to eliminate chronic viral hepatitis through hepatitis B vaccination and antiviral therapy of chronic hepatitis B and C.

Elevated serum level of aflatoxin B_1-albumin adducts at study entry was found to increase the risk of cirrhosis and HCC in patients affected with chronic hepatitis B or C in case–control studies nested in the REVEAL-HBV cohort (Chu et al. 2017; Chu et al. 2018). It is important to note that the aflatoxin exposure at study entry may predict risk of end-stage liver diseases occurring within 8 or 9 years, but not those occurring afterward. In other words, aflatoxin may be a risk predictor only at the late stage of chronic liver disease development and/or aflatoxin exposure at study entry may not reflect the exposure in the long period of follow-up. This further supports the importance of serial measurements of exposure to risk predictor in cohort studies in order to elucidate the time-dependent effect of the disease predictors.

Intervention studies: validation of environmental health hazards through preventive intervention

It is unethical to carry out experimental trials on human beings to validate the suspected effects of environmental health hazards identified from observational studies discussed earlier. However, intervention trials or programmes may be used to validate possible effects of environmental agents on disease risk by removing the exposure to agents in the

environment. Intervention trials may be classified as individual or cluster trials depending on whether individual participants or clusters such as households, schools, or communities are allocated to the intervention. Intervention trials may also be classified as controlled and uncontrolled trials depending on the inclusion of a comparable control group or not. In controlled trials, the effect of the intervention on environmental health hazards is evaluated by the comparison of disease occurrence between experimental and control groups. In uncontrolled trials, the effect of intervention is assessed by the comparison of disease occurrence of the experimental group before and after the intervention.

The implementation of a public water supply system, which used uncontaminated surface water from distant reservoirs, in the Blackfoot disease endemic area of southwestern Taiwan was started in the early 1960s and completed in the 1970s. The arsenic-induced health hazards identified from observational studies could thus be validated through the comparison of the secular changes in morbidity and mortality of arsenic-induced diseases in the endemic and non-endemic areas. In a series of studies, cause-specific standardized mortality ratios of residents in the Blackfoot disease endemic from 1971 to 2003 were calculated using the general population in Taiwan as the standard population. Cumulative sum techniques were used to detect the occurrence of changes in the standardized mortality ratios. A significant decline in mortality from ischaemic heart disease (Chang et al. 2004), renal disease (Chiu and Yang 2005), peripheral vascular disease (Yang 2006), and cancers of the lung (Chiu et al. 2004b), kidney (Yang et al. 2004), and bladder (Yang et al. 2005) was observed for both males and females in the southwestern endemic area of Blackfoot disease. A significant decline in mortality from liver cancer (Chiu et al. 2004a) and diabetes mellitus (Chiu et al. 2006) was observed for females but not males. Based on the reversibility criterion, it was concluded that associations between arsenic exposure and these diseases were very likely to be causal. Although the findings are quite consistent with observational studies mentioned earlier, there are several issues which need further clarification: (1) Classification of underlying causes of death may vary by area and time. It is necessary to examine whether the variations in the classification of underlying causes in death certificates are comparable in different areas over the study period. (2) Competing causes of death may be quite different between the endemic area and Taiwan as a whole. Arsenic has a pleiotropic health effect that induces

various cancers, circulatory diseases, diabetes mellitus, hypertension, renal disease, and so forth. A person may die with several arsenic-induced diseases. It is thus difficult to select an underlying cause of death for an arsenic-exposed person. (3) Mortality associated with a disease is a function of the incidence and fatality of the disease. High mortality may imply an increased incidence and/or an elevated fatality rate. It is better to analyse incidence rather than mortality to clarify the causal association between agent and disease. (4) Risk factors other than arsenic may be important confounding factors, such as habits of cigarette smoking and alcohol drinking, chronic viral hepatitis, and obesity. They were not taken into consideration in the mortality analysis. It is thus more convincing to compare the secular changes in the incidence of arsenic-induced diseases after the implementation of public water supply system. Furthermore, age-cohort-period analysis may also help the identification of susceptible ages at exposure to arsenic and the assessment of reversibility effect of intervention.

A nationwide vaccination programme aimed at eradicating HBV infection was launched in 1984 in Taiwan (Chien et al. 2014). It was the first universal hepatitis B vaccination programme for infants in the world. During the first 2 years (July 1984 to June 1986) of the programme, only infants born to high-risk (HBsAg-positive) mothers were vaccinated. Since July 1986, however, all new-borns were vaccinated. The programme was further extended to the preschool children who had not received vaccination as neonates since 1987. There has been a dramatic decrease in the HBsAg carrier rate and HCC incidence among children and adolescents born after 1984 demonstrating the programme not only successfully prevented the transmission of HBV but also reduced the risk of HCC (Chang et al. 1997), which is the first study to demonstrate the prevention of human cancer through vaccination. Even more striking has been the effectiveness of the national hepatitis B vaccination programme in reducing long-term risk of infant fulminant hepatitis, chronic liver disease, and HCC 30 years after its launch (Chiang et al. 2013). Through the strong evidence of a striking decline in HCC incidence of the vaccinated infants, the causal relation of chronic HBV infection with HCC is well confirmed. However, the elucidation of the failure of hepatitis B vaccination to eliminate HCC in the vaccinated birth cohort requires further study at the individual rather than national level. In a 20-year cohort study of 3.8 million vaccinees, incomplete hepatitis B immunization and maternal HBV carrier status are associated

with the vaccine failure and resulting risk of acute and chronic liver diseases (Chien et al. 2014).

Several antivirals have been approved for chronic hepatitis B therapy, Lamivudine was first approved in 1998 which was found to decrease the HCC risk of treated patients significantly with a large proportion developed antiviral-resistant YMDD mutants (Liaw et al. 2004). Newly developed antivirals for hepatitis B therapy have higher genetic barriers to induce antiviral-resistant HBV. The interferon-based therapy was the standard for chronic hepatitis C treatment before direct acting agents were launched in 2013. Several recent studies have shown the decrease in incidence of HCC in chronic hepatitis B patients treated with oral nucleoside/nucleotide analogues and in chronic hepatitis C patients who cleared HCV with direct acting agents (Chen 2018). A national programme of antiviral therapy on high-risk CHB and CHC patients was launched in October 2003 in Taiwan. A significant reduction in the burden of end-stage liver diseases through this programme has recently been documented. A total of 157,570 CHB patients and 61.823 CHC patients were treated with reimbursed antiviral therapy from 2004 to 2011. There was 22% reduction in mortality from chronic liver diseases and cirrhosis, 24% reduction in HCC mortality and 14% reduction in HCC incidence in 2008–2011 compared to the 4-year period before the programme launch, i.e. 2000–2003 (Chiang et al. 2015). The study confirmed HBV and HCV are the most important aetiology of cirrhosis and HCC in Taiwan. Recent international studies has shown the decrease in the incidence of HCC in CHB patients treated with pegylated interferon or any nucleoside/nucleotide analogues (Lin et al. 2016) or treated with entecavir or tenofovir (Papatheodoridis et al. 2017), and in CHC patients who cleared HCV with pegylated interferon with ribavirin (Huang et al. 2017) or DAAs (Ioannou et al. 2017). These studies have further confirmed HBV and HCV are the most important aetiology of cirrhosis and HCC in the world.

Before the Second World War, endemic goitre was highly prevalent in mountainous areas in Taiwan. A national survey showed a prevalence as high as 70% in some aboriginal townships in mountainous areas. The prevalence of goitre in pigs was found to be significantly higher in the goitre-endemic area than non-endemic area. It was debated whether dietary deficiency in iodine or goitrogenic foodstuff intake was the major cause of the endemic goitre. The use of iodine tablets to supplement dietary iodine intake was found to be effective in lowering the prevalence of endemic

goitre in school children, but it was considered costly and inconvenient. In order to identify an effective and efficient prevention strategy at the community level, a controlled community intervention trial on iodized salt was carried out in two townships in northern Taiwan (Chen et al. 1976). All the dietary salt used in experimental and control townships was strictly provided by the research team. Before the implementation of the community trial, the endemic goitre prevalence among schoolchildren was similarly high. One year after the trial, the prevalence of endemic goitre in the experimental township was significantly lower than in the control township. While the study did not rule out the possible effects of the intake of goitrogenic foodstuff, iodine deficiency was confirmed to be the most important risk factor for endemic goitre in Taiwan. A nationwide salt iodization programme supported by UNICEF was implemented, and the goitre prevalence has been drastically reduced since then. As there may be changes in dietary intake of iodine other than iodized salt, it is essential to monitor and adjust the level of iodine in the salt to prevent thyroid toxicity due to excessively high intakes of iodine.

Individual susceptibility to environmental health hazards: identification and application of molecular and panomic biomarkers

The risk of developing an environmental disease may vary in different persons exposed to the same environment. In other words, there exists a host variation in the susceptibility to diseases caused by environmental factors. The individual susceptibility may come from genetic or behavioural components. For example, poor nutritional status and arsenic methylation capability may modify the risk associated with arsenic-induced health hazards including cancers and cardiovascular diseases. Multiple risk factors including cigarette smoking, aflatoxin exposure, antioxidant deficiency, and serum androgen levels are all important risk factors to HBV-induced HCC. Some environmental co-factors are difficult to detect and quantify unless biomarkers are used. Many studies have been carried out in recent decades on changes in the structures and functions of molecules, cells, tissues, and organ systems in response to exposures to environmental risk factors, to ascertain if the preclinical lesions of environmental diseases could be

detected at an early stage which would allow more timely intervention. More recently, the dramatic advances in genomics, proteomics, transcriptomics, proteomics, glycomics, lipidomics, and metabolomics have enabled more extensive investigations into the genetic polymorphisms and mutations as well as epigenetic changes of host to elucidate the effects of gene–gene and gene–environment interactions more precisely.

Various biomarkers of environmental exposure, health effect, and individual susceptibility of human diseases have been identified and applied in studies on environmental health hazards. The biomarkers for the exposure to environmental factors include molecular dosimetry of internal dose and biologically effective dose. The biomarkers for the health effect include early molecular and cellular changes, altered structures and functions of target tissues and organs, and preclinical lesions. The individual susceptibility biomarkers include both genetic and acquired susceptibility. As an illustration, since the development of chronic arsenic poisoning is a multistage pathogenesis, a series of biomarkers of arsenic-induced health hazards have been developed and applied as shown in Table 7.1.9 (Chen et al. 2005). There are several biomarkers of short-term internal dose for ingested arsenic including levels of arsenic in blood, urine, hair, and finger or toenails. Arsenic in urine, hair, and nails are better biomarkers for short-term exposure than arsenic in blood. Skin hyperpigmentation and palmoplantar hyperkeratosis, characteristic dermatological lesions induced by long-term exposure to arsenic, are excellent clinical biomarkers for long-term exposure to ingested arsenic. Patients with characteristic arsenic-induced skin lesions were found to have an increased risk of Blackfoot disease, lung cancer, and urothelial carcinoma. Inorganic arsenic is methylated into monomethylarsonic acid (MMA) and then further methylated into dimethylarsinic acid (DMA) in the human body. Monocarbon metabolism is involved in the methylation of inorganic arsenic. Genetic polymorphisms of arsenic (+3 oxidation state)-methyltransferase, purine nucleoside phosphorylase, and glutathione S-transferase (GST) omega 2 are associated with the arsenic methylation capability in humans. As MMA is more toxic than DMA, the percentage of MMA in the total urinary concentration of inorganic metabolites is considered as the biomarker for biologically effective dose of arsenic exposure. It has been well-documented that the higher the MMA percentage in the total urinary level of inorganic arsenic metabolites, the higher the risk of arsenic-induced

skin cancer, bladder cancer, lung cancer, peripheral vascular disease, carotid atherosclerosis, and hypertension.

The biomarkers of molecular changes induced by ingested arsenic include plasma levels of reactive oxidants and inflammatory molecules such as chemokine C-C motif ligand 2/monocyte chemotactic protein-1. The arsenic-induced cellular changes include sister chromatid exchanges, micronuclei, and chromosomal aberrations in peripheral lymphocytes and urothelial cells; as well as chromosomal loss and gain detected by comparative genomic hybridization and loss of heterozygosity in urothelial cells. Biomarkers of subclinical changes include carotid atherosclerosis, QT prolongation, and increased dispersion detected by electrocardiogram, retarded peripheral neural conduction, and retarded neurobehavioural function. The biomarkers of individual susceptibility to arsenic-induced health hazards include genetic polymorphisms of enzymes involved in arsenic and xenobiotic metabolism, arsenic methylation, oxidative stress, and DNA repair, as well as serum levels of carotenes, homocysteine and folate, body mass index, and cigarette smoking.

Table 7.1.9 Molecular and genomic biomarkers of exposure, effect, and susceptibility of arsenic-induced health hazards

Category	Group	Biomarkers
Exposure	Internal dose	
	Short term	Arsenic in urine, hair, and nail
	Long term	Skin hyperpigmentation and hyperkeratosis
	Biologically effective dose	Monomethylarsonic acid in urine
Effect	Molecular changes	Reactive oxidants in blood
		Inflammatory molecules in blood
	Cellular changes	Sister chromatid exchanges
		Micronuclei
		Chromosomal aberrations in target cells
		Chromosomal loss/gain and loss of heterozygosity in target cells
	Subclinical changes	QT abnormality in electrocardiogram
		Carotid atherosclerosis
		Retarded peripheral neural conduction
		Retarded neurobehavioural function
Susceptibility	Genetic susceptibility	Arsenic and xenobiotic metabolism enzymes
		Arsenic methylation enzymes
		DNA repair enzymes
		Oxidative stress-related enzymes
	Acquired susceptibility	Serum carotenes level Serum folate level Serum homocysteine level Body mass index Cigarette smoking

In the multistage hepatocarcinogenesis of chronic hepatitis B, there are many other risk factors that modify the risk of developing HBV-induced HCC (Chen and Chen 2002). Biomarkers associated with HBV-induced HCC are shown in Table 7.1.10. HBV infection markers include HBsAg, HBeAg, antibodies against hepatitis B core antigen (anti-HBc), antibodies against HBsAg (anti-HBs), as well as serum levels of HBV DNA (viral load), HBsAg, and anti-HBc. Different HBV infection markers have different associations with the development of HCC. In addition to the seropositivity of HBsAg and HBeAg, serum HBV DNA level is associated with an increasing risk of cirrhosis and HCC in a dose–response relation (Chen et al. 2006; Iloeje et al. 2006). Different genetic characteristics of HBV are also associated with different risk of liver cirrhosis and HCC. The genotype C, basal core promoter A1762T/G1764A mutant, and pre-S mutant of HBV are significantly associated with an increased risk of liver diseases; while the precore stop codon G1896A mutant is associated with a decreased risk (Yang et al. 2008). Both quantitative and qualitative characteristics of HBV are important in the development of HCC.

Table 7.1.10 Molecular and genomic biomarkers associated with hepatitis B virus-induced hepatocellular carcinoma

Category	Group	Biomarkers
Exposure	Hepatitis B virus	HBsAg/HBeAg seropositivity Serum level of HBV DNA (viral load) and HBsAg
		HBV Genotype
		HBV Mutants
		Serum level of antibody against HBV core antigen (anti-HBc)
	Aflatoxins	Urinary level of aflatoxin metabolites and guanine adducts
		Serum level of aflatoxin albumin adducts
	Tobacco	DNA adducts of 4-aminobiphenyl and polyaromatic hydrocarbons in hepatocytes
Effect	Asymptomatic carrier	Elevated serum level of alanine aminotransferase
	Chronic hepatitis	Serum ALT and AST level Fibrosis detected by biopsy/Fibroscan
	Cirrhosis	Child-Pugh score (albumin, bilirubin, prothrombin time, ascites, encephalopathy)
	HCC	Serum alpha-fetoprotein level Histopathological confirmation by biopsy Imaging by ultrasound, CT scan, and MRI
Susceptibility	Immunity	Anti-HBs-seropositivity
	Hormonal status	Serum levels of androgen and oestrogen
	Genetic polymorphisms	HBV receptor (sodium taurocholate co-transporting polypeptide) Human leukocyte antigens Xenobiotic metabolism enzymes
		DNA repair enzymes
		Hormone metabolism enzymes and receptors
	Nutritional intake	Serum levels of carotenoids and selenium

Dietary exposure to aflatoxins and habits of cigarette smoking and alcohol drinking have been found to increase the risk of hepatitis B-induced HCC. Due to the difficulties in measuring dietary intake of trace amount of aflatoxins and environmental exposures to tobacco smoke, several biomarkers are used for the molecular dosimetry of aflatoxin and tobacco smoke exposures. Aflatoxin metabolites in urine is used as a biomarkers for internal dose, and macromolecular adducts of aflatoxin as biomarkers for biologically effective dose. The hepatic DNA adducts of 4-amino-biphenyls and polyaromatic hydrocarbons are used to measure the biologically effective dose of exposures to tobacco smoke, while DNA and albumin adducts of aflatoxin B_1 were used as biomarkers of biologically effective dose. There is a dose–response relation between exposure to hepatotoxins and HCC risk. There are significant synergistic effects on HCC between chronic hepatitis B and environmental hepatotoxins. The hepatocarcinogenesis progresses from asymptomatic carrier status, chronic hepatitis, cirrhosis, to HCC. There are several biomarkers which may be used for the detection of various precancerous lesions.

The effect of environmental hepatotoxins on the HBV-induced HCC is modified by genetic polymorphisms of enzymes related to xenobiotic metabolism. Genetic polymorphisms of cytochrome P450 (CYP) enzymes 1A1 and 2E1, GST M1 and T1, N-acetyltransferase 2 were found to modify HBV-induced HCC risk associated with aflatoxin and tobacco exposure. For example, a significant dose–response relation between HCC risk and serum level of aflatoxin B_1 albumin adducts is observed in chronic HBV carriers with null genotype of GST M1 or T1, but no dose–response relation is observed for carriers with wild genotypes (Chen et al. 1996b). The dose–response relation between elevated serum testosterone level and HCC is much more obvious for chronic hepatitis B patients with the high-risk genotype of androgen receptor (Yu et al. 2000). Furthermore, the combination of putative high-risk genotypes of androgen receptor, CYP17, and steroid 5α-reductase is associated with highly elevated risk of hepatocellular carcinoma suggesting a significant additive effect of gene–gene interaction in the human host (Yu et al. 2001). The associations between HBV-induced HCC and low serum levels of micronutrients including carotenoids and selenium were modified by genetic polymorphisms of xenobiotic metabolism enzymes. The interactive effects

on human health between biological and chemical agents in the environment are complicated and deserve scrutiny.

Risk assessment of environmental human diseases: standards of arsenic in drinking water and risk calculators for hepatocellular carcinoma in viral hepatitis patients

In order to assure the safety of drinking water to protect public health, norms for drinking-water quality were initially recommended by World Health Organization in 1958. The *International Standards for Drinking-Water* established 0.20 mg/L as an allowable concentration for arsenic. The standard was re-evaluated and reduced to 0.05 mg/L in 1963. It was maintained as the 'Guideline Value' of World Health Organization in 1984, and many countries kept this as the national standard or an interim target. In the United States, congress directed Environmental Protection Agency to revise the standard in 1986. According to the 1993 edition of the *WHO Guidelines for Drinking-Water Quality*, inorganic arsenic was classified as a documented human carcinogen and 0.01 mg/L was established as a provisional guideline value for arsenic. The mortality from cancers of the liver, lung, and urinary bladder from exposure to arsenic in the drinking water of 42 villages in southwestern Taiwan was intensively analysed (Morales et al. 2000). The original standard of arsenic in drinking water (0.05 mg/L) was found to be associated with a substantial increased risk of cancer and is not sufficiently protective of public health. Based on the risk assessment and cost-effectiveness analysis, a new maximal contamination level of 0.01 mg/L became effective in February 2002, and was complied with in 2006 in the United States. Several hundred million people in the world are exposed to ground water containing arsenic level higher than 0.01 mg/L. Strategies for the mitigation of chronic arsenic poisoning from drinking water may include the elimination of wells with high-arsenic water, implementation of clean water supply system to areas where the arsenic concentration in drinking water is high, monitoring of arsenic levels in agricultural products irrigated by high-arsenic water, identification, and health surveillance of exposed population, as well as detection and management of preclinical and clinical outcomes of poisoned victims.

The prediction of disease risk is important for the planning and evaluation of public health programmes aimed at controlling the disease. Risk calculators of many diseases, especially cardiovascular diseases, have been developed and widely applied. As there are many risk predictors for most diseases, it is necessary to integrate several predictors into one measure of absolute risk using a regression model. Uncertainty about clinical interpretation of a single abnormal biomarker may be improved by the development of risk calculators based on the combination of multiple biomarkers. The risk calculator also allows for appropriate recognition of clinically important risk in persons with several but seemingly marginal risk factors that individually may otherwise not raise clinical concerns.

As described earlier, there are several risk predictors of HCC in patients affected with chronic hepatitis B. Based on the long-term follow-up findings of REVEAL-HBV study, long-term risk of cirrhosis and HCC have been developed and validated (Yang et al. 2010; Lee et al. 2013; Chen et al. 2015). For example, a nomogram of predicted HCC risk for chronic hepatitis B patients is shown in Figure 7.1.2. Risk predictors included in this nomogram were gender, age, family history of HCC, alcohol intake, serum ALT level, HBeAg serostatus, and serum HBV DNA level. Integer risk scores were assigned to different categories of various risk predictors. The nomogram showing 5- and 10-year cumulative risk of HCC by sum of risk score. A 64-year-old man with a family history of hepatocellular carcinoma, a habit of alcohol intake, a serum ALT level > 45 IU/L, and an HBeAg-seropositive status has a risk score of 2 + 6 + 2 + 2 + 2+ 6 = 20. His cumulative risk of developing HCC after 5 and 10 years is estimated at around 90% and 100%, respectively. The REACH-B risk calculator of HCC was derived from community-based REVEAL-HBV cohort and internationally validated by hospital cohorts of chronic hepatitis B patients from Hong Kong and South Korea (Yang et al. 2011). This HCC risk calculator was found to have a high sensitivity, specificity, and discriminating capability. These HCC risk calculators may be used to assess the efficacy of antiviral therapy for chronic hepatitis B. It has recently been documented that antiviral therapy may decrease the HCC risk in chronic hepatitis B patients with normal to minimally elevated ALT and no cirrhosis (Hoang et al. 2016). Risk calculators for many other environmental diseases may also be developed if there were good large-scale long-term cohort studies on the diseases.

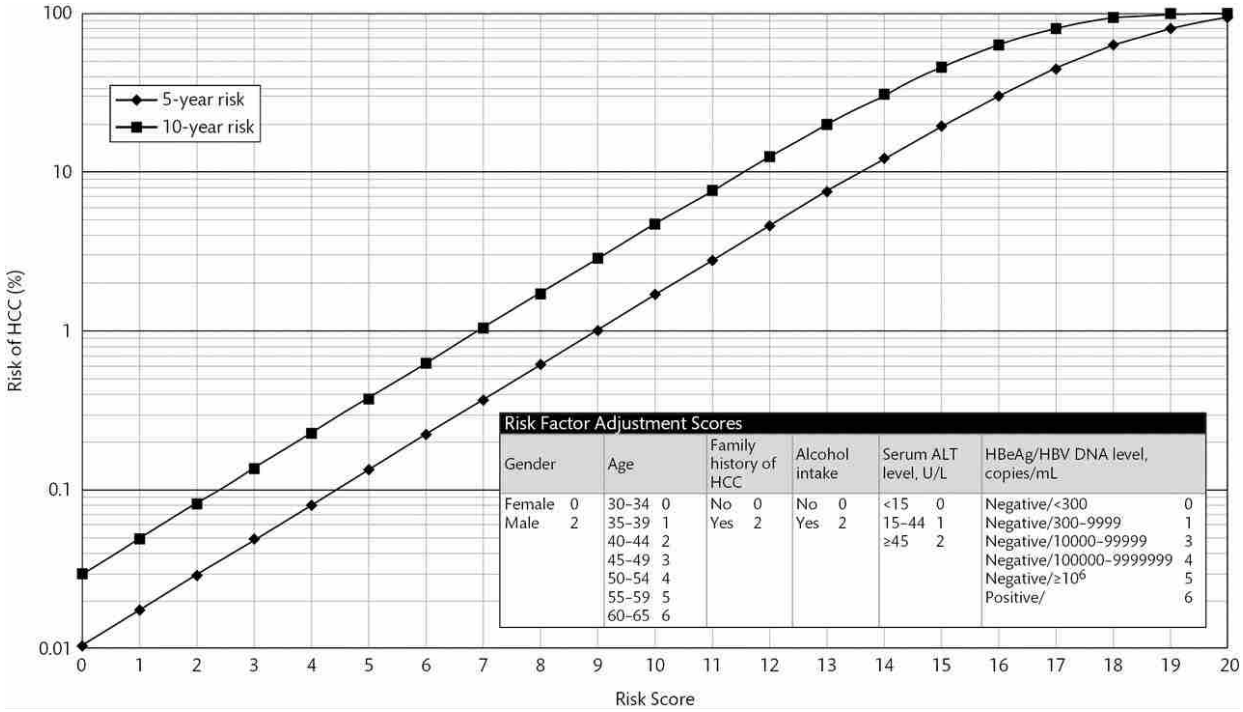


Fig. 7.1.2 Nomogram for the predicted risk of hepatocellular carcinoma caused by chronic hepatitis B. From Yang HI et al., Nomograms for risk of hepatocellular carcinoma in patients with chronic hepatitis B virus infection. *Journal of Clinical Oncology*, Volume 28, Issue 14, pp. 2437–744, Copyright © 2010 by the American Society of Clinical Oncology.

Promotion of global health through environmental health intervention

With increasing globalization, environmental pollution, or infectious disease outbreaks in one country may result in environmental disasters and health hazards in both adjacent and remote countries. Public health can be promoted through environmental health interventions including the prevention of air, water, and soil pollution with chemicals or biological agents; the reduction in exposures to ultraviolet, ionizing radiation, noise, and electromagnetic fields; improvement of built environments including housing, land use patterns, and roads; minimization of carbon dioxide emission, combating global warming leading to severe climate change, and ecosystem deterioration; optimization of agricultural methods and irrigation schemes; and promotion of occupational safety and security, and so forth.

Many environmental health interventions are more cost-effective compared with other kinds of health-sector efforts. These interventions often yield long-term benefits in addition to immediate health improvements. Reduction in the burden of environmental diseases may help the eradication of extreme poverty and hunger, the reduction in child mortality, the improvement in maternal health, the maintenance of environmental sustainability, and the control of epidemics including AIDS/HIV, malaria, tuberculosis, and other infectious diseases. Global partnerships need to be strengthened to achieve interrelated goals of health, environmental sustainability, and development.

REFERENCES

Beasley, R.P., et al. (1981). Hepatocellular carcinoma and hepatitis B virus: a prospective study of 22,707 men in Taiwan. *The Lancet*, 2, 1129–33.

Chang, C.C., Ho, S.C., Tsai, S.S., and Yang, C.Y. (2004). Ischemic heart disease mortality reduction in an arseniasis-endemic area in southwestern Taiwan after a switch in the tap-water supply system. *Journal of Toxicology and Environmental Health, Part A*, 67, 1353–61.

Chang, M.H., et al. (1997). Nationwide hepatitis B vaccination and the incidence of hepatocellular carcinoma in children in Taiwan. *The New England Journal of Medicine*, 336, 1855–9.

Chen, C.F., et al. (2011). Changes in serum levels of HBV DNA and alanine aminotransferase determine risk for hepatocellular carcinoma. *Gastroenterology*, 141, 1240–8.

Chen, C.J. (1999). *Epidemiology: Principles and Methods*. Taipei: Lian-Chin.

Chen, C.J. (2014). Health hazards and mitigation of chronic poisoning from arsenic in drinking water: Taiwan experiences. *Reviews on Environmental Health*, 29, 13–19.

Chen, C.J. (2015). Environmental health issues in public health. In: Detels, R., Gulliford, M., Karim, Q.A., Tan, C.C. (eds.) *Oxford Textbook of Global Public Health, Sixth Edition*. Oxford: Oxford University Press, pp. 623–839.

Chen, C.J. (2018). Global elimination of viral hepatitis and hepatocellular carcinoma: opportunities and challenges. *Gut*, 67, 595–8.

Chen, C.J. and Chen, D.S. (2002). Interaction of hepatitis B virus, chemical carcinogen and genetic susceptibility: multistage hepatocarcinogenesis with multifactorial etiology (Editorial). *Hepatology*, 36, 1046–9.

Chen, C.J., Chiou, H.Y., and Chiang, M.H. (1996a). Dose–response relationship between ischemic heart disease mortality and long-term arsenic exposure. *Arteriosclerosis, Thrombosis, and Vascular Biology*, 16, 504–10.

Chen, C.J., Chuang, Y.C., Lin, T.M., and Wu, H.Y. (1985). Malignant neoplasms among residents of a Blackfoot disease-endemic area in Taiwan: high-arsenic artesian well water and cancers. *Cancer Research*, 45, 5895–9.

Chen, C.J., Chuang, Y.C., You, S.L., Lin, T.M., and Wu, H.Y. (1986). A retrospective study on malignant neoplasms of bladder, lung and liver in Blackfoot disease endemic area in Taiwan. *British Journal of Cancer*, 53, 399–405.

Chen, C.J., et al. (2005). Biomarkers of exposure, effect and susceptibility of arsenic-induced health hazards in Taiwan. *Toxicology and Applied Pharmacology*, 206, 198–206.

Chen, C.J., et al. (2014). Epidemiology of virus infection and human cancer. *Recent Results in Cancer Research*, 193, 11–32.

Chen, C.J., Hsueh, Y.M., Lai, M.S., Hsu, M.P., Wu, M.M., and Tai, T.Y. (1995). Increased prevalence of hypertension and long-term arsenic exposure. *Hypertension*, 25, 53–60.

Chen, C.J., Kuo, T.L., and Wu, M.M. (1988a). Arsenic and cancers (letter). *The Lancet*, 1, 414–15.

Chen, C.J., Lee, M.H., Liu, J., and Yang, H.I. (2015). Hepatocellular carcinoma risk scores: ready to use in 2014? *Hepatological Oncology*, 2, 1–4.

Chen, C.J. and Wang, C.J. (1990). Ecological correlation between arsenic level in well water and age-adjusted mortality from malignant neoplasms. *Cancer Research*, 50, 5470–4.

Chen, C.J., Wu, M.M., Lee, S.S., Wang, J.D., Cheng, S.H., and Wu, H.Y. (1988b). Atherogenicity and carcinogenicity of high-arsenic artesian well water: multiple risk factors and related malignant neoplasms of Blackfoot disease. *Arteriosclerosis, Thrombosis, and Vascular Biology*, 8, 452–60.

Chen, C.J. and Yang, H.I. (2011). Natural history of chronic hepatitis B revealed. *Journal of Gastroenterology and Hepatology*, 26, 628–38.

Chen, C.J., et al. (2006). Risk of hepatocellular carcinoma across a biological gradient of serum hepatitis B virus DNA level. *Journal of the American Medical Association*, 295, 65–73.

Chen, C.J., Yu, M.W., and Liaw, Y.F. (1997). Epidemiology and multifactorial etiology of hepatocellular carcinoma. *Journal of Gastroenterology and Hepatology*, 12, S294–308.

Chen, C.J., et al. (1996b). Chronic hepatitis B carriers with null genotypes of glutathione S-transferase M1 and T1 polymorphisms who are exposed to aflatoxins are at increased risk of hepatocellular carcinoma. *American Journal of Human Genetics*, 59, 128–34.

Chen, C.L., et al. (2004). Ingested arsenic, cigarette smoking and lung cancer risk: a follow-up study in arseniasis-endemic areas in Taiwan. *Journal of the American Medical Association*, 292, 2984–90.

Chen, K.P., et al. (1976). Studies on the effect of salt iodization on endemic goiter, Taiwan. I. Mass survey on goiter of school children. *Journal of the Formosan Medical Association*, 75, 471–82.

Chi, I.C. and Blackwell, R.Q. (1968). A controlled retrospective study of Blackfoot disease, an endemic peripheral gangrene disease in Taiwan. *American Journal of Epidemiology*, 88, 7–24.

Chiang, C.J., Yang, Y.W., You, S.L., You, S.L., Lai, M.S., and Chen, C.J. (2013). Thirty-year outcomes of the national hepatitis B immunization program in Taiwan. *Journal of the American Medical Association*, 310, 974–6.

Chiang, C.J., et al. (2015). Significant reduction in end-stage liver diseases burden through national viral hepatitis therapy program in Taiwan. *Hepatology*, 6, 1154–62.

Chien, Y.C., Jan, C.F., Chiang, C.J., Kuo, H.S., You, S.L., and Chen, C.J. (2014). Incomplete hepatitis B immunization, maternal carrier status and increased risk of liver diseases: a 20-year cohort study of 3.8 million vaccinees. *Hepatology*, 60, 125–32.

Chiou, H.Y., et al. (2001). Incidence of transitional cell carcinoma and arsenic in drinking water: a follow-up study of 8,102 residents in an arseniasis-endemic area in northeastern Taiwan. *American Journal of Epidemiology*, 153, 411–18.

Chiou, H.Y., Huang, W.I., Su, C.L., Chang, S.F., Hsu, Y.H., and Chen, C.J. (1997). Dose–response relationship between prevalence of cerebrovascular disease and ingested inorganic arsenic. *Stroke*, 28, 1717–23.

Chiu, H.F., Chang, C.C., Tsai, S.S., and Yang, C.Y. (2006). Does arsenic exposure increase the risk for diabetes mellitus? *Journal of Occupational and Environmental Medicine*, 48, 63–7.

Chiu, H.F., Ho, S.C., Wang, L.Y., Wu, T.N., and Yang, C.Y. (2004a). Does arsenic exposure increase the risk for liver cancer? *Journal of Toxicology and Environmental Health, Part A*, 67, 1491–500.

Chiu, H.F., Ho, S.C., and Yang, C.Y. (2004b). Lung cancer mortality reduction after installation of tap-water supply system in an arsenic-endemic area in Southwestern Taiwan. *Lung Cancer*, 46, 265–70.

Chiu, H.F. and Yang, C.Y. (2005). Decreasing trend in renal disease mortality after cessation from arsenic exposure in a previous arseniasis-endemic area in southwestern Taiwan. *Journal of*

Toxicology and Environmental Health, Part A, 68, 319–27.
Chu, Y.J., et al. (2017). Elevated serum levels of aflatoxin B1-albumin adducts increase risk of cirrhosis and cirrhotic hepatocellular carcinoma among chronic hepatitis B virus carriers in Taiwan. *International Journal of Cancer*, 141, 711–20.
Chu, Y.J., et al. (2018). Aflatoxin B1 exposure increases the risk of hepatocellular carcinoma associated with hepatitis C virus infection or alcohol consumption. *European Journal of Cancer*, 94, 37–46.
Hoang, J.K., et al. (2016). Lower incidence of hepatocellular carcinoma with antiviral therapy in chronic hepatitis B patients with normal to minimally elevated ALT and no cirrhosis. *Medicine*, 95, e4433.
Hsu, L.I., et al. (2016a). Effects of arsenic in drinking water on risk of hepatitis or cirrhosis in persons with and without chronic viral hepatitis. *Clinical Gastroenterology and Hepatology*, 14, 1347–55.
Hsu, L.I., et al. (2016b). Cumulative arsenic exposure is associated with the development of fungal infections: two cohort studies based on southwestern and northeastern basins in Taiwan. *Environment International*, 96, 173–9.
Hsu, L.I., et al. (2017). Arsenic exposure from drinking water and the incidence of CKD in low to moderate exposed areas of Taiwan: a 14-year prospective study. *American Journal of Kidney Disease*, 70, 787–97.
Hsueh, Y.M., et al. (1997). Serum beta-carotene level, arsenic methylation capability and incidence of arsenic-induced skin cancer. *Cancer Epidemiology, Biomarkers & Prevention*, 6, 589–96.
Huang, C.F., et al. (2017). Risk of hepatitis C virus related hepatocellular carcinoma between subjects with spontaneous and treatment-induced viral clearance. *Oncotarget*, 8, 43925–33.
Iloeje, U.H., Yang, H.I., Su, J., Jen, C.L., You, S.L., and Chen, C.J. (2006). Predicting cirrhosis risk based on the level of circulating hepatitis B viral load. *Gastroenterology*, 130, 678–86.
Ioannou, G.N., Green, P.K., and Berry, K. (2017). HCV eradication induced by direct-acting antiviral agents reduces the risk of hepatocellular carcinoma. *Journal of Hepatology*, doi: http://dx.doi.org/10.1016/j.jhep.2017.08.030
Lai, M.S., et al. (1994). Ingested inorganic arsenic and prevalence of diabetes mellitus. *American Journal of Epidemiology*, 139, 484–92.
Lee, M.H., et al. (2013). Models predicting long-term risk of liver cirrhosis and hepatocellular carcinoma in chronic hepatitis B patients: risk scores integrating characteristics of host and hepatitis B virus. *Hepatology*, 58, 546–54.
Liaw, Y.F., et al. (2004). Lamivudine for patients with chronic hepatitis B and advanced liver disease. *New England Journal of Medicine*, 351, 1521–31.
Lin, D., et al. (2016). Reduction of chronic hepatitis B hepatocellular carcinoma with antiviral therapy, including low risk patients. *Alimentary Pharmacology and Therapeutics*, 44, 846–55.
Lin, W., et al. (2008). Associations between arsenic in drinking water and pterygium in southwestern Taiwan. *Environmental Health Perspectives*, 116, 952–5.
Liu, J., et al. (2014). Spontaneous seroclearance of hepatitis B seromarkers and subsequent risk of hepatocellular carcinoma. *Gut*, 63, 1648–57.
Morales, K.H., Ryan, L., Kuo, T.L., Wu, M.M., and Chen, C.J. (2000). Risk of internal cancers from arsenic in drinking water. *Environmental Health Perspectives*, 108, 655–61.
Pan, W.C., et al. (2016). Fine particle pollution, alanine transaminase and liver cancer: a Taiwanese prospective cohort study (REVEAL-HBV). *Journal of National Cancer Institute* 108, djv341.
Papatheodoridis, G.V., et al. (2017). The risk of hepatocellular carcinoma decreases after first 5 years of entecavir or tenofovir in Caucasians with chronic hepatitis B. *Hepatology*, 66, 1444–53.
Pruss-Ustun, A., Wolf, J., Corvalan, C., and Neira, M. (2016). *Preventing Disease through Healthy Environments: A Global Assessment of the Burden of Disease from Environmental Risks*. Geneva: World Health Organization.

See, L.C., et al. (2007). Dose–response relationship between ingested arsenic and cataract among residents in southwestern Taiwan. *Journal of Environmental Science and Health, Part A*, 42, 1843–51.

Tseng, C.H., Chong, C.K., Chen, C.J., and Tai, T.Y. (1996). Dose–response relationship between peripheral vascular disease and ingested inorganic arsenic among residents in Blackfoot disease endemic villages in Taiwan. *Atherosclerosis*, 120, 125–33.

Tseng, C.H., et al. (2003). Long-term arsenic exposure and ischemic heart disease in arseniasis-hyperendemic villages in Taiwan. *Toxicology Letters*, 137, 15–21.

Tseng, W.P. (1977). Effects and dose–response relationships of skin cancer and Blackfoot disease with arsenic. *Environmental Health Perspectives*, 19, 109–19.

Wang, C.H., et al. (2002). Biological gradient between long-term arsenic exposure and carotid atherosclerosis. *Circulation*, 105, 1804–9.

Wu, M.M., Kuo, T.L., Hwang, Y.H., and Chen, C.J. (1989). Dose–response relation between arsenic concentration in well water and mortality from cancers and vascular diseases. *American Journal of Epidemiology*, 130, 1123–32.

Yang, C.Y. (2006). Does arsenic exposure increase the risk of development of peripheral vascular diseases in humans? *Journal of Toxicology and Environmental Health, Part A*, 69, 1797–804.

Yang, C.Y., Chiu, H.F., Chang, C.C., Ho, S.C., and Wu, T.N. (2005). Bladder cancer mortality reduction after installation of a tap-water supply system in an arsenious-endemic area in southwestern Taiwan. *Environmental Research*, 98, 127–32.

Yang, C.Y., Chiu, H.F., Wu, T.N., Chuang, H.Y., and Ho, S.C. (2004). Reduction in kidney cancer mortality following installation of a tap water supply system in an arsenic-endemic area of Taiwan. *Archives of Environmental Health*, 59, 484–8.

Yang, H.I., et al. (2002). Hepatitis B e antigens and the risk of hepatocellular carcinoma. *The New England Journal of Medicine*, 347, 168–74.

Yang, H.I., et al. (2010). Nomograms for risk of hepatocellular carcinoma in patients with chronic hepatitis B virus infection. *Journal of Clinical Oncology*, 28, 2437–44.

Yang, H.I., et al. (2008). Association between Hepatitis B virus genotype and mutants and the risk of hepatocellular carcinoma. *Journal of National Cancer Institute*, 100, 1134–43.

Yang, H.I., et al. (2011). Risk estimation for hepatocellular carcinoma in chronic hepatitis B (REACH-B): development and validation of a predictive score. *Lancet Oncology*, 12, 568–74.

Yang, T.Y., et al. (2013). Lifetime risk of urothelial carcinoma and lung cancer in the arseniasis-endemic area of northeastern Taiwan. *Journal of Asian Earth Sciences*, 77, 332–7.

Yu, M.W., et al. (2000). Androgen-receptor CAG repeat, plasma testosterone levels, and risk of hepatitis B-related hepatocellular carcinoma. *Journal of the National Cancer Institute*, 92, 2023–8.

Yu, M.W., et al. (2001). Hormonal markers and hepatitis B virus-related hepatocellular carcinoma risk: a nested case-control study among men. *Journal of the National Cancer Institute*, 93, 1644–51.

7.2

Radiation and public health

Leeka Kheifets, Adele Green, and Richard Wakeford

Introduction to radiation and public health

The electromagnetic spectrum encompasses frequencies that range from above approximately 10^{20} hertz (Hz) for ionizing radiation at the high end of the spectrum to static fields and power frequencies of 50–60 Hz at the low end. In order of decreasing frequency, the spectrum comprises gamma-rays, X-rays, ultraviolet radiation, visible light, infrared radiation, microwaves, radio waves and extremely low frequency electric and magnetic fields (ELF EMF) (Fig. 7.2.1). This chapter reviews sources and health effects of human exposure to radiation within the electromagnetic spectrum, and for ionizing radiation exposure to subatomic particles, and reviews policies for limiting human exposure where appropriate.

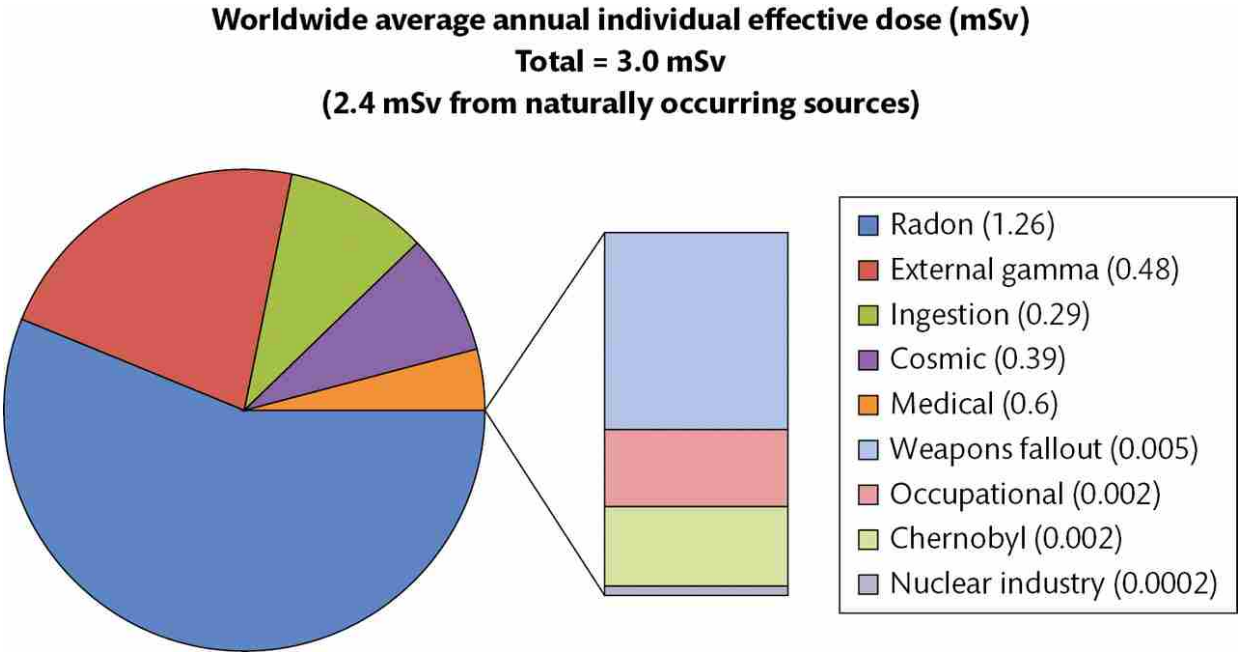


Fig. 7.2.1 Major sources and electromagnetic spectrum.

The focus of research and our knowledge of the health effects of these forms of energy have been driven, at least partly, by episodes of exposure of the general public (e.g. for ionizing radiation, the extensive studies of the Japanese survivors of the atomic-bombings of Hiroshima and Nagasaki), although valuable additional information has been derived from studies of occupational and medical exposures. Our understanding varies: for ionizing radiation, ultraviolet radiation, and extremely low frequency fields (ELF), numerous state-of-the art epidemiologic and toxicologic investigations encompassing a variety of endpoints have been undertaken; a lot of the recent research has been on radiofrequency fields (RF), but focused on m mobile phone use and tumours of the head and neck, with less attention to other sources and outcomes. Although numerous studies of static fields have been completed, here too, much of the key research is yet to be done; for infrared radiation and intermediate frequencies only scant information is available. Technological developments involving exposure to electromagnetic fields bring social and economic benefits to large sections of society but the health consequences can be difficult to predict and manage. As countries increase their capacity to generate and distribute electricity and take advantage of the many new technologies—such as telecommunications—to improve lifestyle and work efficiency, exposures to electric and magnetic fields from 0 to 300 GHz (non-ionizing radiation) has been

increasing rapidly. Further, the use of ionizing radiation in medicine, such as computed tomography (CT) scanning and nuclear medicine, has led to increases in exposure, particularly in economically developed countries. In this chapter, our main emphasis is on exposures common in everyday life rather than occupational exposures. When relevant, we also consider separately exposures from medical applications, as these exposures are commonly higher, but carry a different risk/benefit ratio for exposed individuals.

Static fields

Sources and environmental levels (static fields)

For a steady current in a circuit corresponding to zero-frequency (static fields), the charge density at any point of the circuit is constant, and therefore, the electric fields are constant in time. Since magnetic fields are created by moving electric charges, magnetic fields are also constant for static fields. The unit of measurement for magnetic fields is tesla (T). Another commonly used unit in engineering sciences is gauss (G); one T is equivalent to 10,000 G. Environmental exposure levels are described in microtesla (μT), or 10^{-6} T.

There is a natural static magnetic field that originates from the electric current flowing in the upper layer of the Earth's core, as well as solar activity and atmospheric processes. There are significant local differences in the strength of this field: the vertical component of the field reaches a maximum of about 70 μT at the magnetic poles, and approaches zero at the magnetic equator; conversely the horizontal component is close to zero at the poles and is a maximum of just over 30 μT at the magnetic equator.

In addition to the Earth's magnetic field are man-made magnetic fields of significantly higher strength. Major sources of exposure to the general public are transportation and magnetic resonance imaging (MRI). In general, electrified railway systems produce some of the largest static field levels encountered by the general public. Static fields up to several tens of microtesla have been reported inside trams operating on direct current (European Commission 1996). Modern magnetic levitation (maglev) systems use very high fields (around 1 T) directly on the rails; inside the trains fields vary between 50 μT and 10 mT depending on the cabin design.

MRI systems have proliferated in recent years and there are currently many thousands worldwide. The static magnetic fields are in the range of 0.2–3 T for systems for routine clinical use. MRI also utilizes much smaller time-varying gradient magnetic fields and radio frequency radiation. Medical applications use fields from 1.5 to 10 T (Gowland 2005). Sources of exposure that are either relatively uncommon or small include steel construction materials, high voltage DC transmission lines, headphones and telephone speakers, and steel-belted radial tires. Use of magnetic plasters, blankets, and mattresses for therapeutic properties have surface magnetic flux densities of about 50 mT but decay quickly within a few millimetres and are not used widely (European Commission 1996). High occupational exposures to static fields are encountered in several industrial processes including aluminium and chlorine production, gas welding, and in accelerator research.

Health effects (static fields)

The three established biophysical mechanisms are magnetic induction, and magneto-mechanical and electronic interactions. Laboratory studies of animals show aversive responses and field avoidance at 4 T or higher, probably vestibular in nature. Thresholds for motion-induced vertigo are estimated to be 1 T per second for greater than 1 second.

Cancer (static fields)

The few epidemiological studies published to date leave a number of issues unresolved concerning the possibility of increased cancer risk from exposure to static magnetic fields. These studies have been carried out almost exclusively on workers exposed to static magnetic fields generated by equipment using large direct currents (DC). Assessment of exposure has been poor, and the number of participants in the studies has been very small. Most of the studies were conducted in aluminium or other smelter plants. These studies are limited in their ability to provide useful information: in fact they do not provide clear evidence for increased risk due to more established, carcinogens present in some of these work environments: aluminium reduction creates coal tar, pitch volatiles, fluoride fumes, sulphur oxides, and carbon dioxide. Although some welders are exposed to relatively high static fields in some processes, they are also exposed to fumes and ELF

and RF magnetic fields. Furthermore, most of the studies on welders do not provide enough information to determine the type of welding used, limiting the usefulness of these data in assessing potential health effects of static magnetic field exposure. Other work environments with a potential for high fields, for example as experienced by MRI technicians, have not been adequately evaluated, but have been identified as a high research priority by the World Health Organization (World Health Organization [WHO] 2006). In short: there is insufficient information for proper risk assessment of the impacts of static magnetic fields on cancer. Evidence from animal carcinogenesis studies is inconclusive, in the few studies that have been carried out (van Rongen et al. 2007).

Other outcomes (static fields)

Even less information is available on other outcomes such as reproductive outcomes. One study examined fertility and pregnancy outcomes in female MRI operators, where the potential for exposure to relatively large static fields of up to ~1 T may have existed. The risk of miscarriage for pregnancies during MRI work was slightly increased compared to work in other jobs and was considerably higher than the risk of homemakers (Evans et al. 1993), but the study was methodologically weak. No adverse effects have been demonstrated in animal studies, but there are few good studies, especially with exposures in excess of 1 T (WHO 2006). A priority for further research is to investigate any potential harmful effects from exposure to the foetus during pregnancy, particularly during the first trimester.

Policy and prevention (static fields)

Exposure to equipment generating large static electric fields is uncommon. Exposures to large static magnetic fields (up to several tens of mT) occur as a result of the industrial use of DC electric currents, for example in welding and electrolytic processes, and less commonly in high energy physics laboratories and experimental fusion reactors. Few countries have developed exposure guidance.

However, exposure may be becoming more common. With the advent of superconductor technology in the latter part of the twentieth century, it has been possible to develop large electromagnets capable of generating fields in excess of 1 T. Patient exposure to fields of 1–3 T during clinical MRI

diagnostic procedures is now routine, and up to 10 T is becoming more common. MRI is increasingly used in paediatric diagnostic imaging, and a cohort study into the effects of MRI exposure on children is recommended as a high priority by the Scientific Committee on Emerging and Newly Identified Health Risks (SCENIHR) (2015). Additionally, a priority for further research is to investigate any potential harmful effects from MRI exposure to the foetus during pregnancy, particularly during the first trimester (ICNIRP 2009). Most MRI use involves relatively short scans of less than 1 hour, and there are usually clear benefits to the person being scanned, hence the approach to guidance on exposure is different from that adopted for occupational and public exposure, because of the different risk/benefit ratio to those concerned.

The International Commission on Non-Ionizing Radiation Protection (ICNIRP) (2009) recommended a spatial peak magnetic field of 2 T for head and trunk, which can be raised to 8 T for limbs or in a controlled environment. For the general public, it is recommended that an exposure should not exceed 400 mT. Special consideration should be given to patients: their whole-body exposure should be limited to 4 T in normal mode, and go up to 8 T in a controlled mode. More caution is recommended above 8 T, which is considered experimental and should be accompanied by clinical monitoring. Further, to avoid vertigo and nausea, patients should move or be moved slowly during the MRI procedures and in the presence of high fields.

Extremely low frequency fields

Sources and environmental levels (ELF)

Extremely low frequency electromagnetic fields (ELF EMF) are associated with all aspects of the production, transmission, and use of electricity. The fields are imperceptible to humans and are ubiquitously present in modern societies. ELF EMF are composed of two separate components, electric fields and magnetic fields. Electric fields are created by electric charges and are measured in volts/metre (V/m). Typical residential exposure levels are under 10 V/m. In the immediate vicinity of electric appliances, exposure levels can reach as high as several hundred V/m, whereas exposure levels immediately under high-tension power lines can reach several thousand V/m.

Average magnetic field exposures in the workplace have been found to be higher in electrical occupations which include power and telephone line workers, electricians, and electrical engineers, among others, than in other occupations such as office workers. Exposures range from 0.4 to 0.6 μT for electricians and electrical engineers, to approximately 1.0 μT for power line workers and above 3 μT for welders, railway engine drivers, and textile workers. By contrast, typical residential exposure levels are around 0.1 μT. In the immediate vicinity of electric appliances that are in use, magnetic fields could be as high as several hundred μT but are usually only of short duration.

Health effects (ELF)

Since the late 1970s, numerous epidemiologic studies of varying quality have investigated possible health risks from residential and occupational exposure to both electric and magnetic fields.

Cancer (ELF)

An association between higher-than-average magnetic field exposure levels and leukaemia risk is reported in most of almost 40 epidemiologic studies investigating childhood leukaemia and ELF EMF. However, most of them are based on a small number of highly exposed cases and thus lacked precision. A robust biologic hypothesis to explain this association is lacking. In 2000, two independently conducted pooled analyses of previously published studies showed a statistically significant, approximately twofold increase in childhood leukaemia risk for average residential exposure levels above 0.3 or 0.4 μT, compared to the lowest exposure category of ≤ 0.1 μT (Ahlbom et al. 2000; Greenland et al. 2000). More recently, Kheifets et al. (2010a) identified 14 studies published since the two previous 2000 pooled analyses, of which seven met their inclusion criteria. The odds ratios for exposure above 0.3 μT compared to <0.1 μT was 1.44 (95% confidence interval (CI): 0.88–2.36). Without the most influential study from Brazil, which is suspected to be particularly prone to bias, the odds ratios increased and became similar to previous pooled analyses. Risk increases this small are notoriously hard to evaluate in epidemiology because it is usually difficult to achieve enough precision to distinguish a small risk from no risk. Such small effect estimates, compared to larger ones, are also more likely to

result from inadvertent error, or bias, that can occur in epidemiologic studies (Greenland and Kheifets 2006). Selection bias, which can arise in the process used to select or enrol study participants, might explain some of the observed association (Mezei and Kheifets 2006). Given (1) the small associations observed in studies of magnetic fields and childhood leukaemia; (2) a limited understanding of causal risk factors for childhood leukaemia; (3) methodological difficulties in an assessment of a non-memorable and highly variable exposure; and (4) the potential for selection bias, a conclusive interpretation of these findings remains a challenge. The lack of a robust biophysical mechanism that would explain how environmental magnetic fields could cause cancer and a lack of support from laboratory investigations also indicates that the findings might be a result of study design and measurement issues rather than being truly causal. However, the remarkable consistency of epidemiologic studies is difficult to dismiss, or entirely explain.

Kheifets et al. (2010b) performed a similar pooled analysis for childhood brain tumours. In contrast to childhood leukaemia, there was little indication of an elevation in risk in the highest exposure category, and much less of a monotonic exposure-response relationship (with risk ratios below one in intermediate exposure categories). Results for other cancers, such as adult leukaemia and brain cancer (occupational exposures mostly) are less consistent than those for childhood leukaemia (Kheifets et al. 2008). For breast cancer a biologic mechanism has been proposed but has not been confirmed in recent epidemiologic studies which strongly argue against an association (WHO 2007).

Other health outcomes (ELF)

A number of additional health outcomes, particularly neurodegenerative disease, cardiovascular disease, and reproductive outcomes have been investigated in epidemiologic studies.

Few EMF studies on multiple sclerosis and epilepsy exist to evaluate the evidence. There is no evidence linking EMF and Parkinson’s Disease and only very weak evidence to suggest that they affect Alzheimer’s disease. The evidence that people employed in electrical occupations have an increased risk of developing Amyotrophic lateral sclerosis (ALS) substantial, perhaps because such employees have an increased risk of experiencing an electric shock rather than any long-term exposure to the fields *per se* (Vergara et al.

2013). The epidemiologic results are thus inconsistent and lack supportive laboratory evidence.

There is some evidence for increased miscarriage risk associated with high residential EMF exposure, however methodologic problems make interpretation of these studies difficult (Lewis et al. 2016). In particular, there are: difficulties in studying EMF exposure levels similar to background levels, capturing early miscarriages, and reporting accurate information free from recall bias. There is little evidence of increased risks of adverse reproductive outcomes other than miscarriage, although good data are lacking (WHO 2007). The results of the studies on paternal ELF EMF exposure and adverse reproductive outcomes are unconvincing.

Epidemiologic investigation of cardiovascular diseases has been motivated by biologically based hypotheses which stemmed from two independent lines of evidence: (1) magnetic fields were reported to reduce the normal heart rate variability in human experiments, and (2) several prospective cohort studies have suggested that reductions in some components of heart rate variability are associated with increased risk for heart disease, overall mortality rate in survivors of myocardial infarction, and sudden cardiovascular death. Thus, it was hypothesized that occupational exposure to magnetic fields increases the risk for cardiac arrhythmia-related conditions and acute myocardial infarction, but not chronic cardiovascular disease. Although an initial study appeared to support this hypothesis, recent larger and more rigorous studies have found no effects and have failed to confirm earlier findings and suggest that ELF magnetic fields do not play a role in the development of cardiovascular diseases (Kheifets et al. 2007).

Policy and prevention (ELF)

During the past decade, a number of national and international expert panels, including ones assembled by the United States National Institute of Environmental Health Sciences (NIEHS) (1999), the International Agency for Research on Cancer (IARC) (2002) and the WHO (2007), have reviewed the evidence on the potential relationship between exposure to ELF EMF and various adverse health outcomes. Evaluations by these expert panels generally agree that short-term, adverse effects do not occur at exposures to magnetic fields below 100 μT. Current guidelines (recently revised) are

based on induced internal electric fields and use a threshold for an induction of magnetic phosphenes. For general public exposure, the reference levels for power frequency (50 Hz) electric and magnetic fields are of the order of 5 kV/m and 200 μT, respectively. These values are well above levels encountered in most environments.

Based on these intensive reviews IARC and WHO classified ELF magnetic (but not electric) fields as a 'possible human carcinogen', or a Group 2B carcinogen. This classification was based both on epidemiologic evidence showing a consistent association between exposure to ELF magnetic fields and childhood leukaemia even though laboratory studies in animals and cells, do not support an association between exposure to ELF magnetic fields and cancer.

Several risk characterization attempts have indicated that the public health impact of residential magnetic fields on childhood leukaemia is likely to be limited (Kheifets et al. 2006). The combination of widespread exposures, established biological effects from acute, high-level exposures, and the possibility of leukaemia in children from low-level, chronic exposures have made it necessary but difficult to develop consistent public health policies. In view of these uncertainties, it might be advisable to adopt general no- and low-cost measures to reduce exposure (WHO 2007). The choice of precautionary measures can be informed by a health-economics analysis conducted so far only in a few countries (Kandel et al. 2016).

Intermediate frequency

Intermediate frequency (IF) exposures occur in the range 300 Hz to 10 MHz, partially encompassing frequencies also assigned to very low frequency (VLF; 0.3–30 kHz) and high (3–30 MHz) ranges.

Because few biological studies of exposures to intermediate frequencies have been conducted, exposure guidelines have largely been extrapolated from the encompassing ELF (at the low end) and RF (at the high end) limits.

Sources and environmental levels (IF)

IF exposures from occupational appliances (e.g. induction heaters for heating metals prior to manipulation and plasma heaters) are common. Smaller exposures occur from induction heating elements in residential

cooking appliances. Television, radio, and other communications transmitters can emit IF, but they are discussed in the section on RF, which constitutes the much more common wavelengths of exposure from such sources. Indirect exposures from medical equipment (e.g. MRI machines and magnetic bone stimulators where pulses may drop to IF levels) may occur rarely. Direct exposures to IF may also occur from surgical equipment used for tissue cutting and cauterizing. Common exposures to IF occur in proximity to devices that read magnetic cards or buttons for identification purposes (Polichetti and Vacchia 1998), and similarly, from antitheft devices that read magnetic identification tags. Very small exposures are obtained from television screens (and old computer monitors). Use of antitheft technologies, device tagging, and identification is increasing rapidly, exposures immediately beside radio-frequency identification (RFID) reader antennae can be locally higher, and simultaneous exposures can come from a variety of IF sources, locations of which may be unknown (ICNIRP 2008).

Health effects (IF)

Since exposure level and proximity to the source together determine the IF levels in tissue, knowledge of both is important in determining the biological effects of IF exposure levels in humans. Effects can be either thermal, when frequency is sufficient to induce a heating response, or non-thermal (i.e. cell membrane excitation or electroporation), when frequency does not induce a heating response. Studies of the impact of IF on human health, particularly with exposure from new technologies, are currently lacking.

Policy and prevention (IF)

Because of the lack of data, currently there are no specific recommendations for limiting exposure to IF, other than those extrapolated from EMF at extremes of the IF range (ELF and RF). As use of IF fields and subsequent exposure to the population is increasing, there is an increased interest in characterizing exposure and health effects in the IF range (Generalised EMF Research using Novel Methods [GERoNiMO] n.d.).

Radiofrequency radiation

Sources and environmental levels (RF)

With rapid advances in EMF technologies and communications, people are increasingly exposed to frequencies in the radiofrequency (RF) range. RF fields are produced by radio and TV broadcasts, mobile phone base stations, and other communication infrastructure. The most relevant exposure is to mobile phones. This technology typically uses frequencies from 450 to 2,500 MHz, although new technology has broadened this band to GHz range.

Other sources of exposure to the general population are radio and television transmitters which operate at between 200 kHz and 900 MHz. Radio and TV signals are broadcast to a large area from comparatively few sites (Neubauer et al. 2007). Compared to radio and TV transmitters, mobile phone base stations cover a smaller area, and produce much lower emissions, but are vastly more common in many countries. Residential exposures also come from wireless monitors used in children's cribs, cordless phones, and Wi-Fi (wireless Internet connections) commonly used at home and in schools. Occupational exposures include RF PVC welding machines, plasma etchers, and military and civil radar systems. All operate at different frequencies.

Handheld mobile phones available since the late 1980s became widely used by the general population only in the late 1990s. Currently there are more than 5 billion mobile phone users worldwide, with a penetration in some countries reaching 120% (i.e. many people have more than one). Use of mobile phones has changed markedly over recent years, concomitant with the development of new technologies (e.g. 2G to 3G, power control, handover management, and novel uses of smartphones). Phones operating under 2G and 3G can have significantly different output power; 3G is thought to be around 1% of the power emitted by a phone operating under 2G. 5G technologies and networks use the largely untapped bandwidth of the millimetre wavelength, between 30 and 300 GHz on the radio spectrum, which uses smaller base stations than current wireless technology. As a result, wireless antennae may be placed densely throughout neighbourhoods on infrastructure such as lamp posts, utility poles, and buildings. The Internet of Things (IoT) and machine-to-machine (M2M) devices are also becoming a reality.

In addition, modes of mobile phone use are constantly changing as texting, web-surfing, playing games, and listening to music and video streaming on mobile phones are now common. Exposure from mobile

phones is concentrated closest to the handset and the antenna. Absorption of RF from mobiles is localized and depends on the position of the phone during use. This represents a very important determinant of exposure: only calls made with the mobile phone close to the head result in absorption of RF energy inside the head.

For whole-body exposure, mobile phone base stations can be the largest individual source of RF, but other sources such as radio or TV transmitters can result in comparable exposures depending on where the measurements are taken. For RF sources other than mobile phones, typical power densities outdoors would be 0.01–1 mW/m^2, but could be orders of magnitude higher (i.e. 100 mW/m^2 and above). Indoor levels are often lower than outdoor exposures by orders of magnitude; for example, in Europe, a median indoor power density of 0.005 mW m^{-2} has been reported. Recent study of European children reports median total personal RF-EMF exposure of 75.5 $\mu W/m^2$ (Birks et al. 2018). Downlink was the largest contributor to total exposure (median: 27.2 $\mu W/m^2$) followed by broadcast (9.9 W/m^2). Exposure from uplink (4.7 $\mu W/m^2$) was lower. Wi-Fi and digital enhanced cordless telecommunications (DECT) contributed very little to exposure levels in children. While exposures from mobile phone base stations are several orders of magnitude lower than from phones, it differs from that of mobile phones in other ways: base stations expose the whole body, and the exposure duration is considerably longer. Perhaps more importantly, base station exposure has been a subject of much concern to the public because it is not under the control of the public and its presence is not perceived to be of direct individual benefit.

Population exposures to RF fields are less completely characterized than exposures to ELF fields. This is due to: (1) technical challenges; (2) the rapid evolution of RF-related technology (frequency, coding schemes); and (3) changing patterns of use (duration of calls, text messaging, web-surfing, etc.).

Health effects (RF)

Since RF radiation induces heating in body tissues and imposes a heat load on the whole body, prevention of excessive heating serves as a basis for most international guidelines for human exposure. Studies of the interaction of RF with tissue in the range used for mobile phones have led to the proposal of

many different non-thermal mechanisms for RF interaction. Generally, it is thought that non-thermal interactions are unlikely to be biologically significant at the RF levels below guidance values, but much of the on-going research is directed towards non-thermal mechanisms.

Cancer (RF)

Epidemiological studies of health effects related to RF exposure from mobile phones are numerous and have primarily focused on cancer, especially brain tumours, although studies with long-term exposure with sufficient latency are still limited. Currently it is only possible to evaluate short- to medium-term effects of mobile phone exposure; while some studies report effects for heavy or long-term users (Coureau et al. 2014), the majority of studies have found no effects on either brain or parotid gland tumour risk (Swerdlow et al. 2011). Exposure assessment remains problematic: substantial random error has been shown for even short-term recall of mobile phone use; and information bias appears to affect at least the reporting of the side of the head where the phone is commonly used. Also, some studies may be compromised by a non-representative control group, caused by an increased participation of mobile phone users. Results for acoustic neuroma are more suggestive albeit inconsistent (Swerdlow et al. 2011). Recently, a few studies have examined other cancers, such as leukaemia, non-Hodgkin's lymphoma, and uveal and testicular cancers. Results are unremarkable, but subject to the same limitations as brain tumour studies. So far only one study has examined the possible association between brain tumours and use of mobile phones in children (Aydin et al. 2012). Small and imprecise risks were reported in the high exposure categories, which became more pronounced in a subgroup of about one-third of the subjects for whom objective operator data were available. However, due to methodologic limitations, some internal inconsistencies, and most importantly, lack of increases in the brain tumour rates for children in the registry, the authors considered their data to argue against causality. Clearly, more studies of children are needed.

A few studies have assessed cancer risk in relation to radio and TV transmitters (Ahlbom et al. 2004). Often driven by a previously identified cancer cluster, these analyses are based simply on distance from the source and often include an extremely small number of cases. It is therefore not surprising that such studies have been uninformative. Four recent case–control studies of cancer risk related to mobile phone base stations (Ha et al.

2007; Merzenich et al. 2008; Elliott et al. 2010; Li et al. 2012) have employed improved methods both in terms of design and exposure assessment. While reporting some positive associations for disease and exposure subgroups, overall these studies provide no consistent evidence of association between exposure from base stations and other transmitters and risk of childhood cancer. However, numerous methodologic limitations remain, including the inability to detect small increases in risk.

Although occupational studies have been performed over a longer time period, we are only beginning to measure and learn about RF exposures in various occupations, and the exposure may not always be relevant for an assessment of effects of mobile phone frequencies. Although some increased risks have been found in certain studies, there is no consistent evidence of risk increases for any cancer sites. The studies have several methodological weaknesses: (1) studies are not based on measurements of the actual RF exposure for the subjects included; (2) exposure classification has often been based on job title alone; and (3) control of confounding, if any, has been limited (Ahlbom et al. 2004).

All of the studies have reported null results for carcinogenicity in normal animals at exposure levels compatible with mobile phones, however, co-carcinogenicity studies have been suggestive (IARC 2012). As a result, IARC classified RF as possibly carcinogenic to humans based on limited evidence in humans (from studies of glioma and acoustic neuroma in relation to mobile phones) and limited evidence in animals (based on co-carcinogenicity studies). Of note is the most comprehensive animal study conducted by NIEHS investigating whether mobile phone radiation increases cancer risk in rats and mice. The NIEHS study has not yet been published in peer-reviewed literature, but has been reviewed extensively by a panel of outside reviewers (National Toxicology Program 2018). The panel concluded that for most frequencies/modulations/species the evidence of carcinogenic activity was equivocal, while for several combinations of frequencies/modulations/species some evidence of carcinogenic activity was found, raising to clear evidence of carcinogenic activity for male rats based on incidences of malignant schwannoma in the heart and code-division multiple access (CDMA)-modulated mobile phone radio frequency radiation (RFR) exposure at 900 MHz. In addition, while GSM- or CDMA-modulated cell phone RFR at 1,900 MHz did not increase the incidence of any non-neoplastic lesions in male or female mice, increases in non-neoplastic

lesions in several organs in both males and females exposed to GSM and CDMA cell phone RFR at 900 MHz was reported. The final interpretation must await the report and peer- reviewed publications, which are expected in the latter part of 2018.

Other outcomes (RF)

It is well established in animal studies that hyperthermia during pregnancy can cause embryonic death, abortion, growth retardation, and developmental defects; development of the central nervous system is especially susceptible. Numerous studies have evaluated developmental effects of RF fields on mammals, birds, and other non-mammalian species (Heynick and Merritt 2003; Independent Advisory Group on Non-Ionizing Radiation 2003). These studies have shown that RF fields are teratogenic at exposure levels that are high enough to cause significant increases in temperature. There is no consistent evidence of effects at non-thermal exposure levels, although a few studies have evaluated possible effects on postnatal development using sensitive endpoints, such as behavioural effects. Serious health effects of hyperthermia in humans, however, are associated only with greatly elevated body temperatures (>40°C), and such temperature rises are well above those generated by the maximum allowable level for public RF exposure.

Several studies of occupational RF exposure, primarily in physiotherapists, have reported an increased risk of congenital malformations, but no specific type of malformation has been consistently reported, and there is a potential for recall bias in these studies (Larsen 1991). Exposure to RF during sensitive periods of development in early life may lead to lasting effects on health (Kheifets et al. 2005). No association was found between mobile phone use during pregnancy and early neurodevelopment in very young children in two studies (Vrijheid et al. 2010; Divan et al. 2011). A Danish study has raised the hypothesis that pregnancy and childhood exposure to mobile phones may result in common childhood behavioural problems (Divan et al. 2008). Prospective evaluations of this association were confirmed in the same cohort (Sudan et al. 2016). Behavioural problems related to mobile phone use in children stands out as the only association independently confirmed in several studies (Abramson et al. 2009; Divan et al. 2012; Thomas et al. 2010), including in pooling of individual participant data from five cohorts found that maternal mobile phone use during pregnancy may be associated with an increased risk for

behavioural problems, particularly hyperactivity/inattention problems, in the offspring (Birks et al. 2017).

Possible health effects based in part on anecdotal reports of numerous symptoms such as headaches and sleep disturbance from continuous whole-body RF exposure from base stations is an area of major public concern. Because of numerous methodologic shortcomings, data regarding effects of such RF exposure on symptoms are inadequate for assessment at present.

Policy and prevention (RF)

Guidance on exposure for the general public is intended to restrict local tissue temperature rises to acceptable levels and currently is set to 0.08 Wkg^{-1}, for the whole body, and 2 Wkg^{-1}, for the head. Mobile phones have only been in widespread use for a relatively short time and therefore, the possibility remains that long-term RF exposures from mobile telephony can have adverse health effects. More importantly, many of the health outcomes are yet to be studied and research on potential detrimental effects on children is particularly sparse.

The need for public health policy coupled with the paucity of data, particularly for children regarding the long-term health effects of mobile phone use, suggests that low-cost precautionary measures are appropriate. Mobile phone use is increasingly common among school children, and teenagers may be among the heaviest mobile phone users today with some exposures close to guideline limits (Kheifets et al. 2010c). Exposure can be reduced by users restricting the length of their calls, or by using ‘hands-free’ devices to keep mobile phones away from the head and body.

Infrared radiation

Sources and environmental levels (IR)

Wavelengths between 780 nm and 1 mm are classified as infrared radiation (IR). The IR spectrum is conventionally broken down into three separate categories: IR ‘A’ (780 nm –1.4 μm); IR ‘B’ (1.4–3 μm); and far-IR ‘C’ (3 μm–1 mm). Besides solar radiation and fire as common natural sources, artificial sources of IR include a variety of therapeutic and domestic applications such as heating devices, lights, ‘infrared saunas’, and industrial

sources related to heat production such as in steel/iron industries. Lasers comprise a special group that emits IR radiation over one or more extremely narrow wavelength bands, not covered in detail here.

Health effects (IR)

IR penetrates human skin and ocular tissues, namely the cornea, lens, and retina, to varying depths depending on wavelength and tissue type. For example, far-IR can penetrate to almost 4 cm beneath the skin surface (Vatansever and Hamblin 2012). Humans have protective aversion responses to thermal pain and to the bright light/visible radiation that is often also present, so that potentially hazardous exposure can be avoided. When harmful effects of IR in humans do occur, most studies suggest that the main mechanism of action is thermal injury, with little evidence of photochemical or other biological interactions. The principal chromophore of far-IR wavelengths is water and when far-IR is used therapeutically (see next), it is again believed to be mostly by thermal mechanisms causing for example increased nitric oxide production with beneficial effects on vascular endothelial function, but there may also be other less understood biological effects at energy levels too low to cause heating (Vatansever and Hamblin 2012; Shui et al. 2015).

Harmful effects (IR)

Harmful effects of any episode of hyperthermia of the skin depend on the intensity and the duration of the IR exposure (Moritz and Henriques 1947). When the temperature of the skin is held at 44°C, around 6 hours is required before irreversible skin damage is sustained, while at surface temperatures of 70°C or higher, less than 1 second is required to cause epidermal necrosis (Moritz and Henriques 1947). In industrial settings, whole-body heat stress tends to limit the duration of exposure to IR radiation, keeping it under the threshold for thermal damage to the skin, and so, only very brief exposures to very high irradiances pose a thermal hazard. Chronic IR exposure of the skin without burning (such as repeatedly sitting with exposed skin close to open fires—common in the past) can cause a distinctive red-brown mottling of the skin known as erythema ab igne. IR is not considered relevant in the initiation of skin cancer although it has been hypothesized that IR might

accelerate UV-induced skin carcinogenesis (Edwards et al. 1999), possibly by reducing DNA repair ability (Dewhirst et al. 2003), or by triggering damage by reactive oxygen species, but supportive evidence is scant.

The main detrimental effect of high IR exposure is to the eye due to the sensitivity of the cornea, iris, lens, and retina to thermal damage, though damage thresholds for these tissues vary greatly depending on the precise wavelength and exposure duration. For example, with laser irradiation at 1.315 μm, the cornea is the most sensitive of the tissues to continuous wave exposure. The retina, lens, and iris are sensitive to short exposure durations of ~0.1–1.0 seconds, and the retina is the eye tissue most sensitive to damage by submillisecond laser pulses (Zuclich et al. 2007). When incident IR radiation is absorbed by the cornea it is converted into heat which is conducted to the lens with marked decreases in lens proteins whose aggregation is thought to lead to the lens opacity of cataract (Aly and Mohamed 2011). There is considerable evidence of IR-induced cataract in industries with extremely high exposures, e.g. glass workers (Lydahl and Philipson 1984) and iron and steel workers (Wallace et al. 1971). Most non-industrial exposures to IR appear insufficient to cause cataract or other eye damage, however.

Workers exposed to high levels of IR contend with whole-body hyperthermic effects (rise in body temperature causing heat stress) which are similar to those in any industry involving physical activity in warm environments. The same is likely true of risks of excessive exposure to far-IR saunas or warming cabins. Thus, the exposure limits in these situations will rely mainly on ensuring that safe body temperatures are maintained.

In summary, there have been relatively few studies of the health effects of IR, but current evidence suggests that harmful effects are few and due to acute high intensity exposure or to chronic exposure, causing ocular damage and heat-related illness in certain situations.

Beneficial effects (IR)

Far-IR (around 10 μm wavelength) saunas operate at lower temperatures than traditional saunas and promote greater sweating since skin penetration by IR is greater than by hot air. Their use is widely believed to bring general salubrious effects as well as improved health in patients with cardiovascular disease, diabetes mellitus, and chronic kidney disease though supportive

evidence from randomized controlled trials, non-randomized trials and cohort studies is limited (Beever 2009). Other therapeutic applications include far-IR ray devices used for a range of conditions, from allergic rhinitis to arteriovenous fistula malfunction in haemodialysis patients, and far-IR-emitting ceramics and fabrics that are worn for example for relief of pain (e.g. arthritis, dysmenorrhoea) (Vatansever and Hamblin 2012).

Policy and prevention (IR)

There are various international standards for limiting exposure to IR that are provided separately for IR-A, B, and sometimes C (ICNIRP 2006). These limits, based on combinations of wavelength and time of exposure, are most appropriate for industrial applications where exposures are extreme and measurable. The heat-generating effects of IR tend to be avoided when humans react to minimize their exposure before thermal injury to the skin occurs. This is not true for exposures causing cataract, so specific recommendations exist for wearing protective eye coverings and limiting exposure to IR. Again this applies largely to occupational settings with extreme exposures and assumes that any exposure of the eye to IR is potentially harmful.

There are currently no recommendations for limiting or avoiding exposures to IR in saunas and other recreational settings other than to limit exposure to avoid heat-related illness.

Ultraviolet radiation

Sources and environmental levels (UV)

Ultraviolet (UV) radiation is categorized as UVA (400–320 nm), UVB (320–280 nm), and UVC (280–250 nm), and the sun is the natural source. All UVC and much of the UVB emitted by the sun are absorbed by the Earth's atmosphere such that the midday sun comprises about 95% UVA and 5% UVB. Although in recent decades depletion of ozone in the stratosphere by man-made chlorofluorocarbons has resulted in greater levels of UVB on the Earth's surface in some temperate and arctic regions. Solar UVB levels increase with decreasing latitude and increasing altitude and are highest in the summer and in the middle hours of the day; levels of ambient UVA

mostly parallel UVB levels. Light cloud cover prevents about 50% of solar UV energy from reaching the Earth's surface, while very heavy cloud cover can virtually eliminate UV radiation (Diffey 1991). Reflectance from light-coloured surfaces such as concrete, snow, and to a lesser degree, water, increases terrestrial UVB.

There are many sources of artificial UV found in occupational and medical settings and these include mercury vapour lamps, arc-welding, commercial germicidal UV lamps, and dental polymerizing equipment. In addition, over the last two decades the sunbed industry has emerged as a prevalent source of community UV exposure, sought after by fair-skinned people, particularly adolescents and young adults, seeking to acquire a tan by artificial means (Olsen and Green 2012).

Health effects (UV)

The penetration of human tissues by UV radiation and the ensuing health effects are not only wavelength-dependent with longer UVA wavelengths penetrating human tissues more deeply than UVB, but also age-dependent. The structure of juvenile skin and eyes may result in increased UV dosage and damage to target cells compared with adults resulting in childhood susceptibility to later, long-term degenerative or carcinogenic effects (Green et al. 2011a). For example, in the juvenile eye, the crystalline lens allows UV transmission to the retina and choroid, whereas after around 25 years of age the adult lens and cornea filter almost all UV (Lerman 1984).

Harmful effects (UV)

Acute exposure to UVB causes inflammation of the skin (sunburn) and eye (photokeratitis of the cornea) and immunosuppression (Young 2006). Chronic UV exposure of the skin causes photoaging, benign keratinocytic skin tumours (actinic keratoses), and skin cancer (basal cell and squamous cell carcinomas (BCC and SCC) and melanoma) (IARC 2012). Skin colour determines sensitivity to the acute and chronic effects of UV radiation on the skin, with white Caucasians bearing most of the burden of harmful effects compared with those with black or olive skin. Chronic UV exposure of the eye causes a range of degenerative effects, e.g. pterygium, cataract, macular

degeneration, as well as cancer (SCC and ocular melanoma) regardless of skin and eye colour.

The carcinogenic effects of solar UV radiation derive from the combined action of UVB and UVA: direct damage by UVB and to a less extent by UVA to DNA through cyclobutane pyrimidine dimer formation as well as the indirect oxidative damage induced by UVA, modulated by the efficiency of DNA repair pathways such as nucleotide excision repair (Ikehata and Ono 2007). UV radiation also suppresses the immune system by inducing immunosuppressive mediators and by interfering with immune responses of antigen-presenting cells (Halliday and Rana 2008).

A multitude of epidemiologic studies support the causation of cancers of the skin by solar UV radiation in white populations (IARC 2012), ranging from descriptive studies showing positive associations with birth and/or residence at low latitudes and rare occurrence at non-sun exposed body sites, to analytic studies showing that biologic indicators of excessive sun exposure, namely multiple sunburns (throughout life, but especially in childhood) and actinic keratoses, are significant risk factors (Whiteman et al. 2001; Fears et al. 2002; Chang et al. 2009). Further, there is now evidence from randomized controlled trials that regular application to the skin of a broad-spectrum sunscreen that filters at least 95% of UVB and a proportion of UVA (i.e. sunscreen with a Sun Protection Factor (SPF) of at least 15+), protects against actinic keratoses, SCC, and melanoma (Thompson et al. 1993; Green et al. 1999, 2011b). Convincing evidence is now also available from cohort studies and meta-analyses that the use of tanning beds increases the risks of all three major skin cancers, BCC, SCC, and melanoma, particularly if exposure occurs during the first three decades of life (Boniol et al. 2012; IARC (Working Group on Artificial Ultraviolet Light and Skin Cancer) 2007; Olsen and Green 2012; Wehner et al. 2012; Zhang et al. 2012).

Beneficial effects (UV)

Besides the modest photoprotection against further UV-induced skin damage provided by a suntan to pale skin, UV exposure enhances vitamin D synthesis which is essential to bone health. Beyond bone health, conclusive evidence of benefit of UV radiation mediated through vitamin D or otherwise, is not presently available, however (Institute of Medicine 2011;

Thacher and Clarke 2011; Allinson et al. 2012; LeBlanc et al. 2015). When a systematic review of vitamin D and cancer was performed by IARC in 2008, benefits were inconclusive and did not outweigh the harms (IARC 2008), and later systematic reviews have been consistent with this (Gandini et al. 2011).

Policy and prevention (UV)

Limiting exposure to solar UV

The Global Solar UV Index (UVI) is a unitless quantity estimated on a daily basis that describes the maximum intensity of solar UV radiation received on the Earth's surface at a nominated location (Allinson et al. 2012). Developed jointly by the WHO, the United Nations Environment Programme, the World Meteorological Organization, and the ICNIRP, the UVI serves to raise public awareness about the amount of harmful UV radiation present at a specified location (WHO 2002). Guidance about appropriate sun protection measures is also provided in line with standard recommendations of cancer control agencies worldwide on how best to avoid solar UV exposure for the primary prevention of skin cancer. These are: limit exposure during midday hours; seek shade; wear protective clothing including a broad-brimmed hat to protect the eyes, face, and neck; protect the eyes further with wrap-around sunglasses; use and reapply broad-spectrum sunscreen of at least SPF 15+ liberally; and protect babies and young children from intense sunlight (WHO 2002). The specific use of sunscreen is advocated as an adjunct to physical protection measures (shade; clothing cover) and sunscreen should be reapplied after sweating, swimming, and long periods in the sun. Despite much evidence from ecologic and observational studies about sun exposure and sunscreen use, vitamin D, and human health outcomes, currently there is an absence of evidence from randomized controlled trials about their quantitative relationships, and thus, there is as yet no established basis on which to include vitamin D considerations in sun protection recommendations (Allinson et al. 2012).

Limiting exposure to sunbeds

Many countries have enacted legislation to more tightly regulate the sunbed industry and limit people's exposure to this source of UV radiation. Sunbeds

are banned in Brazil and Australia, and legislation prohibiting use by those under 18 is in place in France, Spain, Portugal, Germany, Austria, Belgium, England, Wales, Northern Ireland, and Scotland, Canada, and the United States (Pawlak et al. 2012). These regulations need to be reinforced by warnings from health professionals and education about the risks of indoor tanning (WHO 2017), so that young people in particular are aware that their use of sunbeds for short-term cosmetic tanning materially raises their skin cancer risk (Olsen and Green 2012).

Limiting occupational exposures

ICNIRP recommends upper limits of exposure to UV in occupational settings, outdoor and indoor, to protect both the skin and the eyes of workers (ICNIRP 2004). It is recognized that the risks of UV exposure of the skin differ greatly depending on skin colour: skin protection must be emphasized for white-skinned people, while higher average threshold exposure is allowed for Asian and Black skin types. Eye protection against UV is emphasized for all skin types.

For the outdoor worker, exposure of the skin depends upon the work tasks and posture, exposure duration, the particular environment, time of day, and season, such that all of these factors need to be taken into account in establishing appropriate control measures. Under most situations, ocular UV exposure does not exceed the limits for even long exposure duration unless in environmental conditions where ground reflectance is unusually high (ICNIRP 2010). Standard protective clothing and headwear can be complemented by application of broad-spectrum sunscreen, while UV-protective goggles can effectively reduce ocular UV exposure from high-reflectance surfaces like snow. For indoor workers' exposure to artificial sources such as UVC-emitting lamps for sterilizing work areas in hospitals, the food industry, and laboratories, or UV lasers, safety precautions include engineering (structural) and administrative controls and Personal Protective Equipment and training. Educating individual workers about the risks of excessive UV exposure is also of paramount importance (ICNIRP 2010).

Ionizing radiation

Ionizing radiation consists of high frequency electromagnetic radiation—X-rays and gamma-rays—or of subatomic particles such as alpha-particles (^{4}He nuclei), beta-particles (electrons), neutrons, protons, pions, muons and atomic nuclei (such as ^{12}C ions). For radiation to be ionizing it must have high enough energy to eject electrons from atoms or molecules, thus creating electrically charged ions. Electrically charged particles (e.g. alpha- and beta-particles) are directly ionizing, whereas electrically neutral particles (e.g. gamma-rays and neutrons) are ionizing because they produce charged particles on interacting with matter. Ionizing radiation is capable of causing DNA strand breaks either by directly interacting with DNA or through the production of chemical free radicals. The energy deposited in matter by ionizing radiation is measured in gray (Gy), 1 Gy being 1 J/kg, which is the absorbed dose of radiation (National Research Council (NRC) 2006; ICRP 2007; United Nations Scientific Committee on the Effects of Atomic Radiation (UNSCEAR) 2008, 2010).

The amount of energy transferred by particles per unit track length when passing through matter is called the linear energy transfer (LET) of the radiation. For example, gamma- and X-rays, electrons, and muons are sparsely ionizing— low LET— radiations because relatively few ionization events occur along their paths. On the other hand, alpha-particles, energetic neutrons, and accelerated heavy charged nuclei (e.g. ^{12}C) generated by synchrotrons are termed high LET radiations because they transfer a relatively large amount of energy per unit track length, producing a high density of ionization. Measurement units commonly used are defined and listed in Table 7.2.1.

Table 7.2.1 Quantities and units of ionizing radiation (ICRP 2007)

Quantity dose	Unit	Definition
Absorbed dose	gray (Gy)	Energy deposited in matter (= 1 J/kg)
	rad (obsolete unit)	1 rad = 100 erg/g = 0.01 Gy = 1 cGy
Equivalent dose	sievert (Sv)	Absorbed dose multiplied by the radiation weighting factor, w_R, for the particular type of radiation. Defined in the context of radiological protection against low-level exposure
	rem (obsolete unit)	1 rem = 0.01 Sv = 1 cSv
Effective dose	sievert (Sv)	Sum of equivalent doses, each multiplied by the tissue weighting factor, w_T, for the particular organ/tissue irradiated. Defined in the context of radiological protection against low-level exposure
Collective effective dose	person-Sv	Sum of the individual effective doses for a population. Defined in the context of radiological protection against low-level exposure
Committed effective dose	sievert (Sv)	Cumulative effective dose to be received over a given period from a given intake of radioactive material. Defined in the context of radiological protection against low-level exposure
Radioactivity	becquerel (Bq)	The unit of activity, one radioactive transformation per second
	curie (Ci) (obsolete unit)	37 billion disintegrations per second (= 37 GBq)

Source: data from International Commission on Radiological Protection (ICRP), *The 2007 Recommendations of the International Commission on Radiological Protection*, ICRP Publication 103, Copyright © 2007.

For the purposes of radiological protection, in the context of low doses (<100 mGy of low LET radiation) or low dose rates (<0.1 mGy/min, as an hourly average, of low LET radiation), the absorbed dose is weighted by the different ionization densities of the various types of radiation and by the different radiosensitivities of the various human organ/tissues, reflecting the overall health detriment posed by such exposures. The equivalent dose is

given by the sum of the absorbed doses (averaged over a tissue) each multiplied by the radiation weighting factor, w_R, for the particular types of radiation involved, and the effective dose is the sum of the equivalent doses received by different tissues with each equivalent dose weighted by the tissue weighting factor, w_T—both equivalent and effective doses are measured in sievert (Sv). For low-level exposure, the effective dose gives the same overall risk of an adverse health effect as a uniform whole-body dose of gamma radiation. It is used in radiological protection when different organs/tissues receive different doses, possibly from radiations with different radiation weighting factors (ICRP 2007).

Sources and environmental levels (ionizing radiation)

Exposure to natural background radiation is ubiquitous. Additional exposure occurs from anthropogenic sources, mainly for the purposes of medical diagnosis. Overall, the world-averaged individual effective dose from naturally occurring sources is about 2.4 mSv/year, with a typical range of 1.0 to 13 mSv/year (UNSCEAR 2010). This represents about 80% of the global average individual effective dose, the ~0.6 mSv/year received from artificial sources being dominated by medical exposures (UNSCEAR 2010). The different sources contribute the fractional amounts as summarized in Figure 7.2.2. The rapid advance in radiodiagnostic technology (such as CT scanning and nuclear medicine) has led to a comparatively large increase in doses from medical sources in developed countries (UNSCEAR 2010).

Naturally occurring sources

In addition to external sources of radiation, such as cosmic-rays originating from outside the Earth and terrestrial gamma-rays from naturally occurring radioactive materials in the environment, radionuclides are inhaled or ingested, irradiating the body from within, and the inhalation of the radioisotopes of radon and their radioactive decay products contributes on average more than 50% of the effective dose from natural sources. Radon-222 is the main source of radon exposure and is produced by the decay via alpha-particle emission of radium-226, which is a component of the radioactive decay chain of naturally occurring uranium-238; rarer radon-220 is part of the decay chain of naturally occurring thorium-232. When radon is inhaled the dose is largely from alpha-particles to the bronchial epithelium,

with a consequent risk of lung cancer. The ingestion of naturally occurring radionuclides is mainly through the presence of the decay products of uranium-238 and thorium-232, and of long-lived radionuclides such as potassium-40, in food and drink. When cosmic-rays strike the upper atmosphere radionuclides are formed, mainly tritium (hydrogen-3) and carbon-14, and these enter the body largely through the consumption of food and drink.

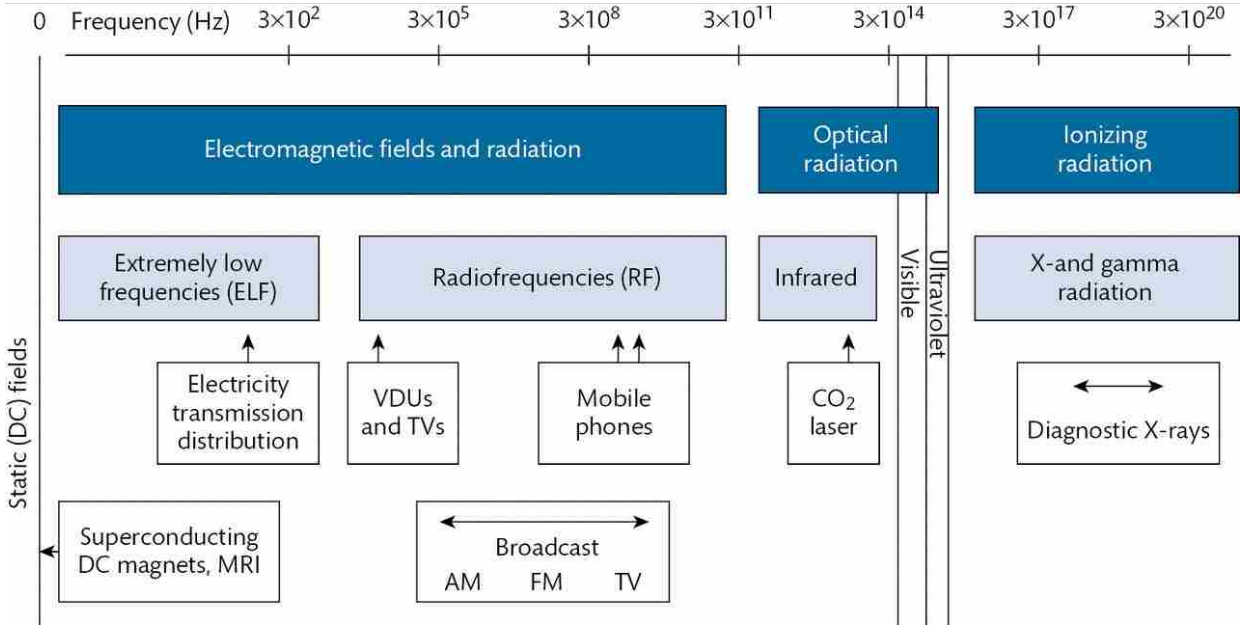


Fig. 7.2.2 Worldwide average annual individual effective dose (mSv) (UNSCEAR 2010).
Source: United Nations Scientific Committee on the Effects of Atomic Radiation 2010, 'Summary of low-dose radiation effects on health', in UNSCEAR 2008 Report to the General Assembly, with Scientific Annexes, United Nations, New York.

There are important variations in the natural background dose depending on factors such as the nature of rock and soils, building materials, and altitude. There exist on Earth high natural background radiation areas where annual effective doses can exceed 10 mSv and even 100 mSv (UNSCEAR 2010). These elevated levels may be due to gamma radiation, such as from the monazite sands of the coast of Kerala in India, but mainly they are due to gaseous radon entering buildings and producing high concentrations in air. Concentrations of radon and its decay products in some houses in the most affected areas of the world are sufficiently raised that measures are necessary

to reduce the consequent risk of lung cancer to acceptable levels. Cosmic radiation levels are increased at high altitudes due to decreased shielding by the atmosphere—air travel therefore raises the dose received from cosmic-rays, but with the exception of frequent travellers (and aircrew), this forms a relatively small increase in the overall dose from natural sources of radiation. Doses received occupationally and the numbers of workers involved worldwide are shown in Table 7.2.2.

Table 7.2.2 Global occupational radiation exposures (UNSCEAR 2010)

Source of exposure	Average annual effective dose (mSv)	Number of monitored workers worldwide (thousands)
Man-made sources [1]		
Nuclear fuel cycle (including uranium mining)	1.0	660
Industrial uses of radiation	0.3	869
Military activities	0.1	331
Medical uses of radiation	0.5	7,440
Miscellaneous	0.1	565
Total from man-made sources	0.4	9,865
Enhanced natural sources [2]		
Aircrew	3.0	300
Mining (other than coal)	3.0	4,600
Coal mining	2.4	6,900
Other workplaces	4.8	1,250
Total from natural sources	2.9	13,050

[1] 2000–2002
[2] 1995–2002

Data from (UNSCEAR 2010).

Man-made radiation sources

The main man-made global contribution to environmental exposure has come from radioactive fallout from the testing of nuclear weapons in the atmosphere, mainly during the late 1950s and early 1960s. The world average annual effective dose reached a peak of 0.11 mSv in 1963, and since then it has decreased to about 0.005 mSv in 2000, most of this dose is now coming from carbon-14 (UNSCEAR 2010). Other contributions are from the generation of electricity from nuclear energy and the production of nuclear weapons, and also from naturally occurring radionuclides in the discharges from fossil-fuelled power stations and other industrial plants, such as phosphate processing factories. Doses to members of the public from these exposures are now generally low at around a μSv. Doses to workers in the nuclear industry and in other occupations (e.g. industrial radiographers) have fallen over the years with improved radiological protection measures. The worldwide average annual dose in the nuclear industry is now about 1 mSv. Modern technological advances have led to other sources of exposure, such as smoke detectors and security scanners, although the doses received from these are almost always minute.

Some potential sources of exposure have raised public concern, such as lost medical or industrial radiation sources, deliberate contamination of items such as foodstuffs with radioactive materials, or ‘dirty bombs’. ‘Dirty bombs’ (a type of ‘radiological dispersion device’) spread radioactive materials by using conventional explosives (WHO 2003), but it is difficult to assess just how dangerous these terrorist weapons would be in terms of radiological effects, and it is likely that the greatest impact would be psychological as a result of fear and panic.

Exposure from medical procedures

Modern diagnostic radiography assures earlier, more precise diagnosis and monitoring of the progression of a large number of diseases. Screening procedures (such as mammography) are beneficial for specific populations who are at a higher risk from certain diseases (such as breast cancer). Diagnostic radiography using conventional X-ray examination, CT scanning, and positron emission tomography (PET) have significantly improved the staging and accuracy of diagnosis. In addition, a number of interventional radiologically-driven procedures (such as angioplasty) introduced in the last

10–20 years, contribute significantly to the effectiveness of treatment of serious diseases of the cardiovascular and central nervous systems and other organs. Such technological improvements can be critically important in effecting cures, but involve relatively high exposure, and therefore should be used with care.

In recent years the use of nuclear medicine and in particular CT scanning has increased to a marked extent in economically developed countries. This increase is illustrated by the average annual effective dose to members of the population of the United States from medical procedures (mainly from diagnostic exposures), which was 0.53 mSv in the early-1980s, but had risen to 3.0 mSv in 2006, a level that is comparable to the average annual effective dose received from natural sources. The doses received from medical procedures are dependent on economic development: world-averaged individual dose from medical exposures during 1997–2007 was 0.62 mSv/year, but in countries with advanced healthcare systems it was 1.92 mSv/year (UNSCEAR 2010). The number of procedures carried out in the United States during 2006 in each category of diagnostic and interventional procedures, together with the average effective dose received per procedure, is given in Table 7.2.3

Table 7.2.3 Estimated number of procedures, and the average effective dose received, in each category of diagnostic and interventional procedures carried out in the United States during 2006 (NCRP 2009)

Category	Number of procedures (million)	Average effective dose per procedure (mSv)	Collective dose (person-Sv)
Computed tomography (CT)	67	6.6	440,000
Conventional radiography and fluoroscopy	293	0.3	100,000
Interventional fluoroscopy	17	7.5	128,000
Nuclear medicine	18	12.8	231,000
Total	395	2.3	899,000

Radiation therapy is based on the ability of ionizing radiation to kill cells, and radiation can target a particular group of cells, such as cancerous cells. Radiotherapy saves many lives every year, and in the case of palliative treatment, substantially reduces suffering. There are also some non-malignant diseases whose treatment by radiation is a method of choice. Most of the side effects of conventional radiotherapy are predictable and expected, arising from the unintentional irradiation of cells outside the target tissues, and one of the aims of modern IMRT (intensity modulated radiation therapy) is to reduce the side effects to a minimum by sparing normal tissues (Hall 2006). However, radiotherapy inevitably leads to some degree of irradiation of healthy tissues and to consequent risk of adverse health effects. This is particularly relevant for deeply-seated tumours, tumours closely located to critical organs, and paediatric tumours. One newly-emerging method is the use of charged particles (e.g. protons and ^{12}C ions), with their positive physical and radiobiological properties, which allows highly conformal treatment of various kinds of tumours while delivering minimal doses to large volumes of surrounding healthy tissues. This means that they can cause severe damage to the DNA in cancer cells while sparing both traversed and deeper healthy tissue. This characteristic also enables them to be used for a more accurate irradiation of the tumour (Rosen 2001).

Recommendations for medical exposure

Medical exposures are intended to provide a direct benefit to the exposed individual. If the practice is justified and the protection optimized, the dose in the patient will be as low as is compatible with the medical purposes. Therefore, dose limits are generally not applied to medical exposures. However, there is considerable opportunity for dose reductions in diagnostic radiology using the techniques of optimization of protection, and this is important since the second largest exposure to typical individuals is from medical exposures, and this is increasing particularly in the developed countries.

The risk from diagnostic radiography is a particularly important consideration for children—they are still developing and are more sensitive to radiation-induced cancer than adults; in addition, they have a longer prospective life span, and hence, the period over which potential long-term side effects may arise is greater (Brenner et al. 2001; WHO 2005). Therefore, care should be taken to minimize the number of unnecessary X-ray examinations, including dental X-ray examinations, and in optimizing the CT exposure settings when this procedure is indicated.

Health effects (ionizing radiation)

Ionizing radiation causes two categories of adverse health effects: tissue reactions (also known as deterministic effects) and stochastic effects (ICRP 2007). Tissue reactions are caused by the ability of radiation to kill cells—if sufficient cells are killed then the tissue reaction occurs—and tissue reactions are characterized by a threshold dose below which the effect does not occur and above which the tissue reaction increases in severity with the dose received (as more cells are affected). Stochastic effects are caused by non-lethal cell modification, which may develop into a cancer if the affected cell is somatic or into a hereditary abnormality if the affected cell is a germ cell. Stochastic effects are understood to have no threshold dose, and at low doses or low dose rates the risk is assumed to be directly proportional to the received dose, although this is not universally accepted. The overall risk of stochastic effects is dominated by the risk of cancer in the exposed individual (ICRP 2007).

Ionizing radiation causes damage to DNA either through direct physical interaction with DNA or through the nearby creation of free radicals that

then chemically react with DNA. These interactions cause strand breaks in the DNA, and ionizing radiation is particularly efficient at producing clustered (i.e. localized) damage and consequently double strand breaks. DNA repair mechanisms attempt to deal with the strand breaks caused by ionizing radiation, but double strand breaks are especially difficult to repair accurately. Heavy damage to DNA from high acute doses of radiation often leads to a loss of DNA integrity that is sufficiently serious to lead to cell death, and when a particular proportion of cells in a tissue is killed then a tissue reaction occurs; the greater this proportion the more severe the tissue reaction. A cell may survive the DNA damage caused by radiation because the DNA is repaired correctly, but a cell may also survive with the DNA damage misrepaired, and this is understood to be the origin of the risk of cancer (in somatic cells) and hereditary effects (in germ cells).

The dose-response for stochastic effects is conventionally thought to be produced by a combination of the single traversal of the cell nucleus by a track of radiation and multiple traversals by independent tracks. The former occurs at low doses or low dose rates of sparsely ionizing radiation and produces a linear term in the dose-response, while the latter occurs at moderate and high doses delivered at a high dose-rate and leads to an upward curving quadratic term; this is the mechanistic origin of the Dose and Dose-Rate Effectiveness Factor (DDREF) of greater than 1 (conventionally, a value of 2), whereby the effect per unit dose (the slope of the dose-response) is greater at moderate-to-high acutely delivered doses. At high acute doses, cell killing becomes more likely, and the slope of the dose-response reduces as a consequence (since dead cells cannot contribute to the risk of stochastic effects). The linear no-threshold dose-response assumes that at low doses or low dose rates the risk of non-lethal DNA damage of relevance to stochastic effects is directly proportional to the dose, and therefore, to the overall number of tracks crossing cell nuclei.

For high LET radiations, the density of ionization is such that the passage of a single particle through a cell nucleus does sufficient damage to the DNA that only a linear term in the dose-response is apparent. Indeed, at the cellular level, the passage of a single alpha-particle track is a high dose event, with a reasonable probability of inducing cell death, so the passage of two alpha-particles is most likely to kill the cell. Thus, the dose-response for high LET radiations is characterized by linearity, with a slope greater than that for the same absorbed dose produced by low LET radiations (due to the

higher ionization density), followed by a reduction in slope at high doses due to cell killing.

However, this picture is a greatly simplified version of reality, the biological mechanisms involved being very complex. For example, it assumes that DNA repair following acute low doses or low dose rates of low LET radiation, producing just a few electron traversals of a cell nucleus, is not much more efficient (conventionally, twice as efficient—a DDREF of 2) than that following acute moderate doses, producing many electron traversals, and this assumption may not be correct. Further, there is evidence that epigenetic mechanisms may play a role in radiation-induced risk, in a way that is as yet unclear, and that these mechanisms may differ between high doses and low doses.

Some argue that since low doses produce a degree of DNA damage that is comparable to the overall level of damage that is constantly occurring through background biochemical processes, it is unlikely that low doses are harmful because the radiation-induced damage is ‘swept up’ by routine repair. Going even further, some argue that low doses of radiation stimulate the immune system, and so are beneficial because damage is repaired that would not have been otherwise (so-called radiation hormesis). Nonetheless, the ‘clustered’ nature of the damage caused by the passage of ionizing radiation through a cell nucleus must be taken into account in such a comparison. On the other hand, some mechanisms of damage may be more effective at low doses than at moderate doses, leading to a supralinear dose-response, and the bystander effect (whereby a cell hit by radiation adversely influences neighbouring cells that were not directly hit) is one example of such a mechanism because it affects more bystander cells at low doses than at moderate doses, where the effect is saturated because nearly all cells are hit. This serves to illustrate the complexity of radiobiology, and why risk estimates are still dominated by empirical epidemiological findings (ICRP 2007).

It has been known for many decades that brief exposure to a dose of several gray leads to death through the killing of sufficient numbers of cells in vital tissues such as the red bone marrow (i.e. death is caused by tissue reactions in radiosensitive tissues). An example from recent decades is the 28 emergency workers attending the Chernobyl nuclear reactor accident in Ukraine in 1986 who died within a few months as a result of the high doses they received over a short period. Another example is the deliberate

poisoning with a substantial dose of the alpha-emitter polonium-210 of Mr. Litvinenko in London in 2006. Other tissue reactions include skin burns, epilation, permanent and temporary sterility, depression of haematopoiesis, eye lens opacities and cataracts, and tissue necrosis.

Cancer (ionizing radiation)

Radiation has always been a natural part of our environment, and the types of radiation emitted from anthropogenic sources are fundamentally the same as those emitted from natural sources. At present it is not possible to distinguish between cancers resulting from radiation exposure (whether from man-made or naturally occurring sources) and those arising from other causes (NRC 2006; ICRP 2007; UNSCEAR 2008, 2010; IARC 2012). Consequently, in the absence of being able to determine the risk to human health from exposure to ionizing radiation from fundamental radiobiological principles, risk estimates must be derived empirically from epidemiological studies of suitably exposed groups of humans, augmented by an incomplete knowledge of radiobiological mechanisms. The evidence that has been constructed from radio-epidemiological studies is substantial, and with the possible exception of tobacco, no other exposure has been studied so extensively. These epidemiological studies fall into four broad categories: studies of the Japanese survivors of the atomic-bombings, those exposed for medical purposes either therapeutically or diagnostically, occupational exposure, and those exposed environmentally (NRC 2006; ICRP 2007; UNSCEAR 2008, 2010; IARC 2012).

The experience of the survivors of the atomic-bombings of Hiroshima and Nagasaki in 1945 provides the principal (but certainly not the only) epidemiological evidence on which radiation risk estimates are based. A large cohort study, the Life Span Study (LSS), was established from information gathered during the Japanese national census in October 1950, and analyses of data generated by the LSS continue today (Preston et al. 2007; Ozasa et al. 2012; Hsu et al. 2013). The LSS consists of about 93,500 survivors, around 86,500 of whom have individually estimated tissue-specific doses—considerable effort has been expended on the reconstruction of tissue-specific doses received during the bombings (mainly from γ-rays, but a component from neutrons), the latest database being the Dosimetry System 2002 (DS02), which has recently undergone a minor revision

(DS02R1) (Cullings et al. 2017). Of the LSS subjects, ~48,000 received non-trivial doses (>5 mSv), and of these ~18,000 received doses >100 mSv, and ~2,400 doses >1 Sv; almost 2/3 of the non-trivially exposed survivors received doses <100 mSv (i.e. low acute doses). At the end of the last mortality follow-up (2003), 58% of the LSS subjects had died, and 24% of the deaths were due to cancer; 527 solid cancer deaths (8% of such deaths among non-trivially exposed survivors) are estimated to be attributable to radiation exposure during the atomic-bombings (Ozasa et al. 2012). Approaching half of the 204 leukaemia deaths before 2001 among the LSS survivors receiving doses >5 mSv have been attributed to radiation (Preston et al. 2004), so overall, around 10% of cancer deaths in the atomic-bomb survivors are estimated to have been attributable to radiation from the bombings. As well as mortality, cancer incidence is ascertained in the LSS through two specialist registries, the most recent analysis of solid cancer incidence covering the period 1958–2009 (Grant et al. 2017), while the latest findings of the incidence of leukaemia, lymphoma, and multiple myeloma are from an analysis covering the period 1950–2001 (Hsu et al. 2013). The latest study of solid cancer incidence has revealed a difference between the sexes in the dose-response for all solid cancers combined: for males the dose-response is curvilinear upwards while for females it is linear (Preston et al. 2007). The reasons for this difference are being investigated. In a subset of the LSS, the Adult Health Study (AHS), survivors participate in regular medical examinations to record clinical parameters such as blood pressure. Other relevant cohorts are those who were exposed *in utero* and the survivors' offspring who were conceived after the bombings, the latter group providing information on the risks of the hereditable effects of exposure.

Early support for the initial findings of the studies of the Japanese atomic-bomb survivors came from investigations of long-term health effects among patients treated with radiotherapy, in particular British patients irradiated for ankylosing spondylitis and US women treated with radiation for cancer of the uterine cervix (McLean et al. 2017; UNSCEAR 2008). Now there is evidence deriving from a wide range of radio-therapeutic treatments such as various groups of cancer survivors, children irradiated for ringworm of the scalp or haemangioma of the skin or for an enlarged thymus, and patients injected with the alpha-emitter radium-224. Patients exposed to radiation for diagnostic purposes are providing increasingly useful information as the use of CT scans increases, but those X-rayed to monitor treatment for

tuberculosis or for scoliosis, and foetuses exposed during an abdominal X-ray examination of the pregnant mother have already provided valuable evidence on the risks of low-level exposures. A recent large cohort study of British children who received at least one CT scan found evidence of an increased risk of brain tumours and leukaemia that was related to the dose of radiation estimated to have been received from the scans (Pearce et al. 2012). An interesting group of patients is those injected with the radiographic contrast medium Thorotrast, which unfortunately was thorium-based and therefore radioactive and gave large doses to certain tissues, such as the liver and red bone marrow, leading to excess liver cancers and leukaemias.

Although the findings of studies of those exposed for medical purposes form a valuable complement to the evidence from studies of the atomic-bomb survivors, medical studies need to be interpreted with care since exposure occurred due to a known or suspected disease. For example, a CT scan may have taken place because of an undiagnosed cancer, so that the cancer caused the CT scan, not the reverse (so-called reverse causation). Similarly, a certain medical condition may predispose to cancer, but may also lead to other diseases that require a greater frequency of CT scans, leading to ‘confounding by indication’. These considerations have led to difficulties of interpretation of the findings of low-dose diagnostic exposures because the predicted risk of such exposures is small relative to the background risk of cancer so that subtle biases and confounding can lead to spurious associations with radiation. Radiotherapy usually involves localized high (cell killing) doses, which can affect incidentally irradiated healthy tissues and lead to a reduction of radiation-induced cancer risk in these tissues; concomitant chemotherapy can also be a difficulty of studies of those receiving radiotherapy.

Radiologists and radiographers have been subjects of study since the early uses of radiation in medicine, and US radiologists provided the first indication that radiation could cause leukaemia (March 1944). Of particular interest are groups of underground hard-rock miners (for example, uranium and tin miners) who inhale the radioactive gas radon and its decay products which irradiate the lungs (mainly with alpha-particles); and in the past, exposures have been high. Workers (mainly young women) who applied radium-based luminous paint to watch and instrument dials inadvertently ingested radium (an alpha-emitter) through the habit of licking the tips of

paintbrushes (‘tipping’), and some workers ingested large quantities of radium, which deposits in bones and led to an excess of bone cancers. Of increasing importance is the evidence obtained from the study of groups of nuclear industry workers, especially those who worked during the early years of the industry when doses were comparatively high relative to those received today. Large worker-studies are providing valuable information on the effects of protracted low-level exposure, and international collaborations have the potential to produce statistically powerful results. The International Nuclear Workers Study (INWORKS) is a recent example of such a collaboration, and has reported radiation-related increases in the risk of leukaemia and solid cancers (Leuraud et al. 2015; Richardson et al. 2015, 2018). Special cases are those workers involved in dealing with emergencies, such as the ‘liquidators’ who were brought in to deal with the aftermath of the Chernobyl accident in 1986 (Kashcheev et al. 2015), and those from the early years of operations at the Mayak nuclear complex in Russia who were particularly highly exposed to both external sources of radiation and plutonium, an alpha-emitter that was inhaled in relatively large amounts during the early weapons production programme of the former USSR (Sokolnikov et al. 2015; Kuznetsova et al. 2016).

Radon in homes has been a subject of substantial scientific attention because globally this is the source of the largest exposure to radiation for the great majority of people. Large case–control studies that take into account the important contribution of risk posed by cigarette smoking, have examined the risk of lung cancer arising from residential exposure to radon, and appropriately combined findings from these studies have produced radon-related risk estimates that are compatible with those obtained from underground hard-rock miners (ICRP 2010; McLean et al. 2017). Exposure to other sources of natural background radiation, in particular terrestrial gamma radiation, has also been studied, notably in those parts of the world where such exposures are materially greater than the global average. A study of cancer incidence in Kerala (India) has not found an excess cancer risk related to radiation exposure, although the power of this study is presently limited and the findings are compatible with those of the LSS (Nair et al. 2009). A large British case–control study concentrating on childhood leukaemia has found evidence of an excess risk from background gamma radiation that is consistent with risk models derived from the Japanese

atomic-bomb survivors (Kendall et al. 2013), but a similar study in France did not find an excess risk (Demoury et al. 2017).

Radioactive contamination from accidental and routine discharges from nuclear installations has been another area of study, radioactive releases from the accident at Chernobyl (UNSCEAR 2018) and substantial pollution of the Techa River by liquid effluent from the Mayak complex being notable examples (Krestinina et al. 2013; Davis et al. 2015). Radioiodine released during the Chernobyl accident in 1986 led to a large excess of thyroid cancer among those highly exposed (largely through the consumption of heavily contaminated milk) as children, and excess leukaemias and solid cancers have been found in the Techa River residents. The releases during the Fukushima nuclear reactors accident in 2011 provide another opportunity for study, although it would appear that doses are not sufficiently large for any effects to be detectable (WHO 2012a, 2012b). The Fukushima accident provides another illustration of the caution that is required in the interpretation of health data, in that the Japanese government established a health survey in Fukushima Prefecture that included a thyroid screening programme for all those at a young age at the time of the accident. An excess of thyroid cancer has been found by this screening programme, but it is most unlikely that this is as a result of radiation exposure during the accident, and probably due to screening using sensitive ultrasonography detecting cancers that would have been present but undiagnosed in the absence of the screening, an effect that has been found elsewhere (Ahn et al. 2014).

Fallout from atmospheric nuclear weapons testing, especially at its peak in the late-1950s and early-1960s, has provided another source of exposure for study, and those resident in the vicinity of weapons testing sites (such as the Semipalatinsk Testing Site in present-day Kazakhstan) are of particular interest. Global fallout has not influenced childhood leukaemia incidence much beyond the level predicted by standard models, showing that intakes of fallout radionuclides do not have the pronounced leukemogenic effect that has been suggested by some (Wakeford et al. 2010). A group of Taiwanese residents was irradiated by gamma-rays from ^{60}Co-contaminated steel accidentally used in building construction, and this group is being followed (Hsieh et al. 2017). There is some evidence of an excess of leukaemia and breast cancer.

Epidemiological studies have clearly established that acute exposure to moderate and high doses of ionizing radiation (>100 mGy of sparsely

ionizing radiation) increases the risk of cancer (NRC 2006; ICRP 2007; UNSCEAR 2008, 2010; IARC 2012; McLean et al. 2017) However, all types of cancer do not have equal sensitivity to radiation-induction—acute leukaemia, thyroid, and female breast cancer have a high sensitivity, whereas chronic lymphocytic leukaemia, Hodgkin lymphoma, and melanoma of the skin have a low sensitivity, with other cancers falling somewhere in between from higher (e.g. lung and colon cancers) to lower (e.g. pancreatic and prostate cancers) sensitivity. Leukaemia was the first cancer to be definitively linked to radiation exposure when a clear excess risk of the malignant disease was detected in the LSS of Japanese atomic-bomb survivors in the early-1950s (although unusually high numbers of cases were reported by alert clinicians in the late-1940s, before the LSS was established), and there had been earlier reports of elevated levels of leukaemia among radiologists (March 1944), which had been difficult to interpret at the time. As the follow-up of the atomic-bomb survivors progressed, raised mortality and incidence of other cancer sites were found to be related to exposure, and this evidence was supplemented by other exposed groups, such as an increased risk of a number of solid cancers among patients treated with radiotherapy for a variety of diseases, a marked excess of bone cancer among those groups with high intakes of radium, an excess risk of lung cancer among underground hard-rock miners exposed to radon, and an excess risk of liver cancer among the Thorotrast patients with high body burdens of thorium (McLean et al. 2017). Recently, the Chernobyl accident has produced a notable excess risk of thyroid cancer among those who as children inhaled and ingested high levels of radioactive iodine released during the accident, an excess that is compatible with that found among those exposed as children to external sources of radiation.

The most recent analysis of leukaemia mortality among the LSS cohort found that a linear-quadratic dose-response best described the full range of data, and that the excess relative risk was 3.1 (95% CI: 1.8–4.3) at 1 Gy red bone marrow dose and 0.15 (95% CI: –0.01–0.31) at 0.1 Gy, based on 318 deaths (Ozasa et al. 2012). About half of the leukaemia deaths among the non-trivially exposed (>5 mGy) survivors are attributed to irradiation during the bombings. However, modelling demonstrates a marked variation of risk with age-at-exposure and time-since-exposure: for young children, the excess relative risk (ERR, the proportional increase in risk) rises to a peak of ~50 some 7 years after exposure and then falls away to ~5 some 30 years

after exposure—a temporal wave of excess risk—while for exposed adults this wave is hardly apparent, with the ERR remaining below 5 at all times after exposure. From the latest data set for mortality from all solid cancers (malignant diseases other than leukaemia, lymphoma, and multiple myeloma) among the survivors, a linear dose–response fits the data best, with a slope of the ERR for solid cancer mortality for a survivor exposed at age 30 years and at an attained age of 70 years of 0.47 (95% CI: 0.38–0.56) Gy^{-1}, based upon almost 11,000 deaths. The ERR coefficients for leukaemia and solid cancer incidence are similar to those for mortality (Preston et al. 2007; Ozasa et al. 2012; Hsu et al. 2013), but the latest solid cancer incidence study has found a significant difference in dose-response between the sexes (Grant et al. 2017). As mentioned earlier, there is notable variation in the ERR/Gy for individual sites of cancer, the ERR for the incidence of breast, thyroid, and non-melanoma skin cancers being high while that for cancers of the rectum and uterus are low.

Risk models have been developed from the cancer data for the Japanese atomic-bomb survivors, supplemented by other data sources where these are found to make a useful contribution to evidence, such as for leukaemia and for thyroid and breast cancers following exposure for medical therapy. Lung cancer among radon-exposed underground hard-rock miners has provided a valuable source of information on the risk of radon-induced lung cancer, which is compatible with the risk of lung cancer among those exposed residentially to, in general, much lower levels of radon. The influence of tobacco smoke upon the risk of lung cancer is especially important to take into account in such studies, and it would appear that radiation and cigarette smoke interact to produce a combined risk of lung cancer that is more than additive, so that smokers are at a greater risk of radon-induced lung cancer than non-smokers.

In general, the findings of the studies of the Japanese atomic-bomb survivors are compatible with those from other exposed groups, although comparisons are often not straightforward. For example, doses from radiotherapy are usually localized and high, with the objective of killing the target cells, but the unintentional irradiation of healthy cells can also lead to their being killed and thereby to a reduction in the risk per unit dose of cancer. Doses away from the target are received from scattered radiation and are difficult to compute accurately, although modern dose reconstruction methods are improving dose estimates. Of growing importance are studies of

acute low doses or protracted low dose rates, because these exposure circumstances directly test the assumptions required for everyday radiological protection. Preliminary findings of such studies, e.g. of patients receiving CT scans, those living in areas affected by radioactive contamination, or nuclear industry workers, indicate that there is an increased risk from low doses/dose-rates (at least, for some types of cancer, notably leukaemia), but that this risk is not greatly different from the predictions of conventional risk models.

Other outcomes (ionizing radiation)

Epidemiological studies have failed to produce definitive evidence that exposure to ionizing radiation increases the risk of heritable genetic effects in the descendants of those exposed (ICRP 2007). Even the follow-up of more than 30,000 offspring of the Japanese atomic-bomb survivors who were conceived after the bombings has not provided clear evidence of an inherited effect from parental exposure. Nonetheless, large experiments involving laboratory animals, mainly mice, have produced unambiguous evidence that irradiation of parents does raise the risk of hereditary effects in offspring. The risk of radiation-induced hereditary effects in humans derived from this evidence suggests that this risk is much lower than the risk of cancer in the exposed individual.

The lens of the eye is sensitive to radiation-induced cataracts, and recent evidence indicates that the threshold for the production of cataracts, at around 500 mGy of low LET radiation, is much lower than the 5 Gy previously considered as a threshold dose (ICRP 2012). This evidence comes from the Japanese atomic-bomb survivors, the Chernobyl ‘liquidators’, and medical staff.

The embryo and foetus are sensitive to radiation-induced developmental (teratogenic) effects, such as, early in pregnancy, congenital malformations (although death may occur before birth defects become likely) (ICRP 2007). Severe mental retardation was found among Japanese atomic-bomb survivors exposed *in utero* to high doses during 8–15 weeks of gestation, and less so at 16–25 weeks, and lower doses may depress IQ.

An issue requiring resolution is whether low doses/dose-rates are capable of raising the risk of non-cancer somatic diseases, in particular diseases of the blood circulatory system, such as heart attack and stroke (ICRP 2012). It has been known for some time that high doses can lead to an excess risk of

these diseases, and this was attributed to tissue damage produced during such exposure, but epidemiological evidence is mounting for an effect at lower doses. However, a mechanism has yet to be identified, and the confounding effects of major (non-radiation) risk factors cannot yet be reliably eliminated as responsible for the association. Nonetheless, the potential risk of circulatory disease following low-level exposure to radiation is considered sufficiently important that significant research effort is being devoted to its investigation.

Policy and prevention (ionizing radiation)

The International Commission on Radiological Protection (ICRP), in its latest assessment of evidence provides detriment-adjusted nominal risk coefficients for all cancers combined: 5.49% Sv^{-1} for the whole population and 4.07% Sv^{-1} for the working population, for low doses/dose-rates of radiation (ICRP 2007). The detriment-adjusted nominal risk coefficients for hereditary effects are: 0.2% Sv^{-1} for the general population and 0.12% Sv^{-1} for a working population. Consequently, the aggregated detriment-adjusted nominal risk coefficients for stochastic effects following low-level exposure to radiation are 5.7% Sv^{-1} for a general population and 4.2% Sv^{-1} for a working population. These risk coefficients form the technical basis of the current framework of radiological protection recommended by the ICRP, and give rise to dose limits: for workers, the limit is an effective dose of 20 mSv per year averaged over a defined period of five years (with a maximum of 50 mSv in any one year), and for the general public the limit is an effective dose of 1 mSv in a year (although in special circumstances a higher dose could be received in any one year provided the annual average is 1 mSv over a 5-year period). These limits do not apply to medical or naturally occurring sources of exposure, although residential exposure to radon and its radioactive decay products in some areas of the world are assessed to produce a sufficiently high risk of lung cancer that remediation measures are required to reduce the exposure.

The three key principles of radiological protection remain: (1) the justification of activities that could cause or affect radiation exposures; (2) the optimization of protection in order to keep doses as low as reasonably achievable; and (3) the application of dose limits (ICRP 2007). The recommendations state:

No practice involving exposures to radiation should be adopted unless it produces sufficient benefit—to the exposed individuals or to society—to offset the detriment it causes.

In relation to any particular source within a practice, the magnitude of individual doses, the number of people exposed, and the likelihood of incurring exposures (where these are not certain to be received) should all be kept as low as reasonably achievable, economic and societal factors being taken into account.

The exposure of individuals resulting from the combination of all the relevant practices should be subject to dose limits, or to some control of risk in the case of potential exposures; hence, ensuring that no individual is exposed to radiation risks that are judged to be unacceptable from these practices in any normal circumstances. Dose (or risk) constraints are used to restrict to upper bounds the dose received (or potentially received) from a specific source.

Restrictions on effective dose sufficient to ensure the avoidance of deterministic effects in all body organs and tissues, and that limit the risk of stochastic effects to acceptable levels, are shown in Table 7.2.4.

Table 7.2.4 Recommended dose limits (ICRP 2007)

Type of limit	Occupational	Public
Effective dose	20 mSv per year, averaged over defined periods of 5 years, with a maximum of 50 mSv in any one year*	1 mSv in a year
Annual equivalent dose in		
Lens of the eye	20 mSv per year, averaged over defined periods of 5 years, with a maximum of 50 mSv in any one year	15 mSv
Skin	500 mSv	50 mSv
Hands and feet	500 mSv	-

* The basis for the control of the occupational exposure of women who are not pregnant is the same as that for men. However, once pregnancy has been declared, the embryo/foetus is protected—the effective dose limit for the embryo/foetus is 1 mSv.

Source: data from International Commission on Radiological Protection (ICRP), *The 2007 Recommendations of the International Commission on Radiological Protection*, ICRP Publication 103, Copyright © 2007.

Conclusions

The nature, frequency, and severity of adverse effects on human health caused by exposure to ionizing and non-ionizing radiation varies depending on the intensity and other exposure parameters, and is highly frequency-dependent. Potential effects range from rapidly fatal injuries to birth and hereditary defects, cancer, and other chronic diseases. Many effects occur only at high levels and are controlled or reduced by observing applicable exposure standards. In this chapter, our main emphasis is on exposures common in the everyday life and on long-term effects, which are typically of most concern to the public.

The IARC classified ionizing radiation and ultraviolet radiation as ‘known’ or Group 1 carcinogens, while extremely low frequency magnetic fields and radiofrequency fields have been classified as a ‘possible human carcinogen’, or a Group 2B, and static fields as well as extremely low frequency electric fields have been assigned Group 3 (not classifiable as to

carcinogenicity to humans). Intermediate frequencies and infrared radiation have not yet been classified as to their carcinogenic potential.

As it is not possible to eliminate exposure to ionizing radiation, these exposures are kept to a practical minimum. Most recommend avoiding high levels of solar UV exposure in order to facilitate the prevention of skin cancers. Because of the beneficial effects of solar UV exposure it is necessary to adopt a balanced set of recommendations regarding this exposure. For ELF magnetic fields, given exposure prevalence, considerable scientific uncertainty, and limited public health impact, no or very low-cost precautionary measures are justified. For RF, potential risk could affect a large number of people because of the ubiquitous nature of the exposure. Thus, given the important gaps in knowledge and availability of no- to low-cost measures, such as use of hands-free devices to reduce exposures from mobile phones, such measures can and should be adopted.

Rigorous studies capable of addressing knowledge gaps are urgently needed, particularly for rapidly growing technologies such as mobile phones, MRI exposures, IF antitheft, and device tagging, and modern magnetic levitation (maglev) systems.

ACKNOWLEDGEMENTS

We are grateful to Andrew Nguyen for his help in editing this chapter.

REFERENCES

Abramson, M.J., et al. (2009). Mobile telephone use is associated with changes in cognitive function in young adolescents. *Bioelectromagnetics*, 30(8), 678–86.

Ahlbom, A., et al. (2000). A pooled analysis of magnetic fields and childhood leukaemia. *British Journal of Cancer*, 83(5), 692–8.

Ahlbom, A., et al. (2004). Epidemiology of health effects of radiofrequency exposure. *Environmental Health Perspectives*, 112(17), 1741–54.

Ahn, H.S., Kim, H.J., and Welch, H.G. (2014). Korea's thyroid-cancer 'epidemic'—screening and overdiagnosis. *New England Journal of Medicine*, 371(19), 1765–7.

Allinson, S., et al. (2012). Validity and use of the UV index: report from the UVI Working Group, Schloss Hohenkammer, Germany, 5–7 December 2011. *Health Physics*, 103(3), 301–6.

Aly, E.M. and Mohamed, E.S. (2011). Effect of infrared radiation on the lens. *Indian Journal of Ophthalmology*, 59(2), 97–101.

Aydin, D., et al. (2012). Childhood brain tumours and use of mobile phones: comparison of a case–control study with incidence data. *Environmental Health*, 11, 35.

Beever, R. (2009). Far-infrared saunas for treatment of cardiovascular risk factors. *Canadian Family Physician*, 55(7), 691–6.

Birks, L., et al. (2017). Maternal cell phone use during pregnancy and child behavioral problems in five birth cohorts. *Environment International*, 104, 122–31.

Birks, L.E., et al. (2018). Spatial and temporal variability of personal environmental exposure to radio frequency electromagnetic fields in children in Europe. *Environmental International*, 117, 204–14.

Boniol, M., Autier, P., Boyle, P., and Gandini, S. (2012). Cutaneous melanoma attributable to sunbed use: systematic review and meta-analysis. *British Medical Journal*, 345, e4757.

Brenner, D.J., Elliston, C.D., Hall, E.J., and Berdon, W.E. (2001). Estimated risks of radiation-induced fatal cancer from pediatric CT. *American Journal of Roentgenology*, 176(2), 289–96.

Chang, Y.-M., et al. (2009). Sun exposure and melanoma risk at different latitudes: a pooled analysis of 5700 cases and 7216 controls. *International Journal of Epidemiology*, 38(3), 814–30.

Coureau, G., et al. (2014). Mobile phone use and brain tumours in the CERENAT case–control study. *Occupational and Environmental Medicine*, 71(7), 514–22.

Cullings, H.M., et al. (2017). DS02R1: improvements to atomic bomb survivors' input data and implementation of dosimetry system 2002 (DS02) and resulting changes in estimated doses. *Health Physics*, 112(1), 56–97.

Davis, F.G., Yu, K.L., Preston, D., Epifanova, S., Degteva, M., and Akleyev, A.V. (2015). Solid cancer incidence in the Techa river incidence cohort: 1956–2007. *Radiation Research*, 184(1), 56–65.

Demoury, C., et al. (2017). Residential exposure to natural background radiation and risk of childhood acute leukemia in France, 1990–2009. *Environmental Health Perspectives*, 125(4), 714–20.

Dewhirst, M.W., Viglianti, B.L., Lora-Michiels, M., Hanson, M., and Hoopes, P.J. (2003). Basic principles of thermal dosimetry and thermal thresholds for tissue damage from hyperthermia. *International Journal of Hyperthermia*, 19(3), 267–94.

Diffey, B.L. (1991). Solar ultraviolet radiation effects on biological systems. *Physics in Medicine and Biology*, 36(3), 299–328.

Divan, H.A., Kheifets, L., Obel, C., and Olsen, J. (2008). Prenatal and postnatal exposure to cell phone use and behavioral problems in children. *Epidemiology*, 19(4), 523–9.

Divan, HA, Kheifets, L, Obel, C and Olsen, J (2012). Cell phone use and behavioural problems in young children. *Journal of Epidemiology and Community Health*, 66(6), 524–9.

Divan, H.A., Kheifets, L., and Olsen, J. (2011). Prenatal cell phone use and developmental milestone delays among infants. *Scandinavian Journal of Work and Environmental Health*, 37(4), 341–8.

Edwards, C., Gaskell, S.A., Hill, S.A., Heggie, R., Pearse, A.D., and Marks, R. (1999). Effects on human epidermis of chronic suberythemal exposure to pure infrared radiation. *Archives of Dermatology*, 135(5), 608–9.

Elliott, P., et al. (2010). Mobile phone base stations and early childhood cancers: case–control study. *British Medical Journal*, 340, c3077.

European Commission (1996), *Non-Ionizing Radiation; Sources, Exposure and Health Effects*. B-1049 Brussels. [Online] Available at: https://op.europa.eu/en/publication-detail/-/publication/f675851e-b4a5-4676-843c-0d6aebb306df

Evans, J.A., Savitz, D.A., Kanal, E., and Gillen, J. (1993). Infertility and pregnancy outcome among magnetic resonance imaging workers. *Journal of Occupational Medicine*, 35(12), 1191–5.

Fears, T.R., et al. (2002). Average midrange ultraviolet radiation flux and time outdoors predict melanoma risk. *Cancer Research*, 62(14), 3992–6.

Gandini, S., et al. (2011). Meta-analysis of observational studies of serum 25-hydroxyvitamin D levels and colorectal, breast and prostate cancer and colorectal adenoma. *Int J Cancer*, 128(6), 1414–24.

Generalised EMF Research using Novel Methods [GERoNiMO] (n.d.). *Radiation Programme, ISGlobal*. [Online] Available at: http://radiation.isglobal.org/index.php/geronimo-home
Gowland, P.A. (2005). Present and future magnetic resonance sources of exposure to static fields. *Progress in Biophysics & Molecular Biology*, 87(2–3), 175–83.
Grant, E.J., et al. (2017). Solid cancer incidence among the life span study of atomic bomb survivors: 1958–2009. *Radiation Research*, 187(5), 513–37.
Green, A., et al. (1999). Daily sunscreen application and betacarotene supplementation in prevention of basal-cell and squamous-cell carcinomas of the skin: a randomised controlled trial. *The Lancet*, 354(9180), 723–29.
Green, A.C., Wallingford, S.C., and McBride, P. (2011a). Childhood exposure to ultraviolet radiation and harmful skin effects: epidemiological evidence. *Progress in Biophysics and Molecular Biology*, 107(3), 349–55.
Green, A.C., Williams, G.M., Logan, V., and Strutton, G.M. (2011b). Reduced melanoma after regular sunscreen use: randomized trial follow-up. *Journal of Clinical Oncology*, 29(3), 257–63.
Greenland, S. and Kheifets, L. (2006). Leukemia attributable to residential magnetic fields: results from analyses allowing for study biases. *Risk Analysis*, 26(2), 471–82.
Greenland, S., Sheppard, A.R., Kaune, W.T., Poole, C., and Kelsh, M.A. (2000). A pooled analysis of magnetic fields, wire codes, and childhood leukemia. Childhood Leukemia-EMF Study Group. *Epidemiology*, 11(6), 624–34.
Ha, M., et al. (2007). Radio-frequency radiation exposure from AM radio transmitters and childhood leukemia and brain cancer. *American Journal of Epidemiology*, 166(3), 270–9.
Hall, E.J. (2006). Intensity-modulated radiation therapy, protons, and the risk of second cancers. *International Journal of Radiation Oncology Biology Physics*, 65(1), 1–7.
Halliday, G.M. and Rana, S. (2008). Waveband and dose dependency of sunlight-induced immunomodulation and cellular changes. *Photochemistry and Photobiology*, 84(1), 35–46.
Heynick, L.N. and Merritt, J.H. (2003). Radiofrequency fields and teratogenesis. *Bioelectromagnetics*, (Suppl 6), S174–86.
Hsieh, W.H., Lin, I.F., Ho, J.C., and Chang, P.W. (2017). 30 years follow-up and increased risks of breast cancer and leukaemia after long-term low-dose-rate radiation exposure. *British Journal of Cancer*, 117(12), 1883–7.
Hsu, W-L, et al. (2013). The incidence of leukemia, lymphoma and multiple myeloma among atomic bomb survivors: 1950–2001. *Radiation Research*, 179(3), 361–82.
Ikehata, H. and Ono, T. (2007). Significance of CpG methylation for solar UV-induced mutagenesis and carcinogenesis in skin. *Photochemistry Photobiology*, 83(1), 196–204.
Independent Advisory Group on Non-Ionizing Radiation (2003). *Health Effects from Radiofrequency Electromagnetic Fields*. Chilton, UK: National Radiological Protection Board.
Institute of Medicine (2011). *Dietary Reference Intakes for Calcium and Vitamin D*. Washington, DC: The National Academies Press.
International Agency for Research on Cancer (IARC) (2002). *IARC Monographs on the Evaluation of Carcinogenic Risks to Humans. Non-Ionizing Radiation, Part 1: Static and Extremely Low-Frequency (ELF) Electric and Magnetic Fields.* Lyon: IARC.
International Agency for Research on Cancer (IARC) (2008). *Vitamin D and Cancer*. Lyon: IARC.
International Agency for Research on Cancer (IARC) (2012). *IARC Monographs on the Evaluation of Carcinogenic Risks to Humans*. Lyon: IARC.
International Agency for Research on Cancer (IARC) (Working Group on Artificial Ultraviolet Light and Skin Cancer) (2007). The association of use of sunbeds with cutaneous malignant melanoma and other skin cancers: a systematic review. *International Journal of Cancer*, 120(5), 1116–22.
International Commission on Non-Ionizing Radiation Protection (ICNIRP) (2004). Guidelines on limits of exposure to ultraviolet radiation of wavelengths between 180 nm and 400 nm (incoherent optical radiation). *Health Physics*, 87(2), 171–86.

International Commission on Non-Ionizing Radiation Protection (ICNIRP) (2006). ICNIRP statement on far infrared radiation exposure. *Health Physics*, 91(6), 630–45.
International Commission on Non-Ionizing Radiation Protection (ICNIRP) (2008). ICNIRP statement on emf-emitting new technologies. *Health Physics*, 94(4), 376–92.
International Commission on Non-Ionizing Radiation Protection (ICNIRP) (ICNIRP) (2009). Amendment to the ICNIRP 'Statement on Medical Magnetic Resonance (Mr) Procedures: Protection of Patients'. *Health Physics*, 97(3), 259–61.
International Commission on Non-Ionizing Radiation Protection (ICNIRP) (2010). ICNIRP statement—protection of workers against ultraviolet radiation. *Health Physics*, 99(1), 66–87.
International Commission on Radiological Protection (ICRP) (2007). *The 2007 Recommendations of the International Commission on Radiological Protection*. Oxford: Elsevier Science.
International Commission on Radiological Protection (ICRP) (2010). *Lung Cancer Risk from Radon and Progeny and Statement on Radon*. Oxford: ICRP.
International Commission on Radiological Protection (ICRP) (2012). *ICRP Statement on Tissue Reactions/Early and Late Effects of Radiation in Normal Tissues and Organs—Threshold Doses for Tissue Reactions in a Radiation Protection Context*. Oxford: ICRP.
Kandel, S., Swanson, J., and Kheifets, L. (2016). Health-economics analyses applied to ELF electric and magnetic fields. *Risk Analysis*, 36(6), 1277–86.
Kashcheev, V.V., et al. (2015). Incidence and mortality of solid cancer among emergency workers of the Chernobyl accident: assessment of radiation risks for the follow-up period of 1992–2009. *Radiation and Environmental Biophysics*, 54(1), 13–23.
Kendall, G.M., et al. (2013). A record-based case–control study of natural background radiation and the incidence of childhood leukaemia and other cancers in Great Britain during 1980–2006. *Leukemia*, 27(1), 3–9.
Kheifets, L., Afifi, A.A., and Shimkhada, R. (2006). Public health impact of extremely low-frequency electromagnetic fields. *Environmental Health Perspectives*, 114(10), 1532–37.
Kheifets, L., et al. (2010a). Pooled analysis of recent studies on magnetic fields and childhood leukaemia. *British Journal of Cancer*, 103(7), 1128–35.
Kheifets, L., et al. (2010b). A pooled analysis of extremely low-frequency magnetic fields and childhood brain tumors. *American Journal of Epidemiology*, 172(7), 752–61.
Kheifets, L., Ahlbom, A., Johansen, C., Feychting, M., Sahl, J., and Savitz, D. (2007). Extremely low-frequency magnetic fields and heart disease. *Scandinavian Journal of Work and Environmental Health*, 33(1), 5–12.
Kheifets, L., Monroe, J., Vergara, X., Mezei, G., and Afifi, A.A. (2008). Occupational electromagnetic fields and leukemia and brain cancer: an update to two meta-analyses. *Journal of Occupational and Environmental Medicine*, 50(6), 677–88.
Kheifets, L., Repacholi, M., Saunders, R., and Deventer, E. (2005). The Sensitivity of Children to Electromagnetic Fields. *Pediatrics*, 116(2), e303–e13.
Kheifets, L., Swanson, J., Kandel, S., and Malloy, T.F. (2010c). Risk governance for mobile phones, power lines, and other EMF technologies: risk governance for mobile phones and power lines. *Risk Analysis*, 30(10), 1481–94.
Krestinina, L.Y., et al. (2013). Leukaemia incidence in the Techa river cohort: 1953–2007. *British Journal of Cancer*, 109(11), 2886–93.
Kuznetsova, I.S., Labutina, E.V., and Hunter, N. (2016). Radiation risks of leukemia, lymphoma and multiple myeloma incidence in the Mayak cohort: 1948–2004. *PLoS One*, 11(9), e0162710.
Larsen, A.I. (1991). Congenital malformations and exposure to high-frequency electromagnetic radiation among Danish physiotherapists. *Scandinavian Journal of Work and Environmental Health*, 17(5), 318–23.
LeBlanc, E.S., Zakher, B., Daeges, M., Pappas, M., and Chou, R. (2015). Screening for vitamin D deficiency: a systematic review for the U.S. Preventive Services Task Force. *Annals of Internal Medicine*, 162(2), 109–22.

Lerman, S (1984). Biophysical aspects of corneal and lenticular transparency. *Current Eye Research*, 3(1), 3–14.
Leuraud, K., et al. (2015). Ionising radiation and risk of death from leukaemia and lymphoma in radiation-monitored workers (INWORKS): an international cohort study. *Lancet Haematology*, 2(7), e276–81.
Lewis, R.C., et al. (2016). Exposure to power-frequency magnetic fields and the risk of infertility and adverse pregnancy outcomes: update on the human evidence and recommendations for future study designs. *Journal of Toxicology and Environmental Health Part B: Critical Reviews*, 19(1), 29–45.
Li, C.Y., Liu, C.C., Chang, Y.H., Chou, L.P., and Ko, M.C. (2012). A population-based case–control study of radiofrequency exposure in relation to childhood neoplasm. *Science of the Total Environment*, 435–6, 472–8.
Lydahl, E. and Philipson, B. (1984). Infrared radiation and cataract II. Epidemiologic investigation of glass workers. *Acta Ophthalmologica (Copenhagen)*, 62(6), 976–92.
March, H.C. (1944). Leukemia in radiologists. *Radiology*, 43(3), 275–8.
McLean, A.R., et al. (2017). A restatement of the natural science evidence base concerning the health effects of low-level ionizing radiation. *Proceeding of the Royal Society B: Biological Sciences*, 284, 1862.
Merzenich, H., et al. (2008). Childhood leukemia in relation to radio frequency electromagnetic fields in the vicinity of TV and radio broadcast transmitters. *American Journal of Epidemiology*, 168(10), 1169–78.
Mezei, G. and Kheifets, L. (2006). Selection bias and its implications for case–control studies: a case study of magnetic field exposure and childhood leukaemia. *International Journal of Epidemiology*, 35(2), 397–406.
Moritz, A.R. and Henriques, F.C. (1947). Studies of thermal injury: II. The relative importance of time and surface temperature in the causation of cutaneous burns. *American Journal of Pathology*, 23(5), 695–720.
Nair, R.R., et al. (2009). Background radiation and cancer incidence in Kerala, India-Karanagappally cohort study. *Health Physics*, 96(1), 55–66.
National Council on Radiation Protection and Measurements (2009). *Ionizing Radiation Exposure of the Population of the United States*, no. 160. Bethesda, MD: National Council on Radiation Protection and Measurements.
National Institute of Environmental Health Sciences (1999). *NIEHS Report on Health Effects from Exposure to Power-Line Frequency Electric and Magnetic Fields*, NIH Publication, no. 99–4493. Durham, NC: National Institutes of Health.
National Research Council (2006). *Health Risks from Exposure to Low Levels of Ionizing Radiation: BEIR VII Phase 2*. Washington, DC: The National Academies Press.
National Toxicology Program (2018). *Actions from Peer Review of the Draft NTP Technical Reports on Cell Phone Radiofrequency Radiation*. Durham, NC: National Institute of Environmental Health Sciences Research Triangle Park.
Neubauer, G., et al. (2007). Feasibility of future epidemiological studies on possible health effects of mobile phone base stations. *Bioelectromagnetics*, 28(3), 224–30.
Olsen, C.M. and Green, A.C. (2012). More evidence of harms of sunbed use, particularly for young people. *British Medical Journal*, 345, e6101.
Ozasa, K., et al. (2012). Studies of the mortality of atomic bomb survivors, report 14, 1950–2003: an overview of cancer and noncancer diseases. *Radiation Research*, 177(3), 229–43.
Pawlak, M.T., Bui, M., Amir, M., Burkhardt, D.L., Chen, A.K., and Dellavalle, R.P. (2012). Legislation restricting access to indoor tanning throughout the world. *Arch Dermatol*, 148(9), 1006–12.
Pearce, M.S., et al. (2012). Radiation exposure from CT scans in childhood and subsequent risk of leukaemia and brain tumours: a retrospective cohort study. *Lancet*, 380(9840), 499–505.

Polichetti, A. and Vacchia, P. (1998). Exposure of the general public to low-and medium-frequency electromagnetic fields. In: *Proceedings of the 3rd COST 244bis Workshop on Intermediate Frequency Range.* Paris, pp. 21–30.

Preston, D.L., et al. (2004). Effect of recent changes in atomic bomb survivor dosimetry on cancer mortality risk estimates. *Radiation Research*, 162(4), 377–89.

Preston, D.L., et al. (2007). Solid cancer incidence in atomic bomb survivors: 1958–1998. *Radiation Research*, 168(1), 1–64.

Richardson, D.B., et al. (2018). Site-specific solid cancer mortality after exposure to ionizing radiation: a cohort study of workers (INWORKS). *Epidemiology*, 29(1), 31–40.

Richardson, D.B., et al. (2015). Risk of cancer from occupational exposure to ionising radiation: retrospective cohort study of workers in France, the United Kingdom, and the United States (INWORKS). *British Medical Journal*, 351, h5359.

Rosen, N.S. (2001). Taking care of children. *American Journal of Roentgenology*, 177(3), 715–17.

Scientific Committee on Emerging and Newly Identified Health Risks (2015). *Potential Health Effects of Exposure to Electromagnetic Fields (EMF)*. Luxembourg: European Commission.

Shui, S., Wang, X., Chiang, J.Y., and Zheng, L. (2015). Far-infrared therapy for cardiovascular, autoimmune, and other chronic health problems: a systematic review. *Experimental Biology and Medicine (Maywood)*, 240(10), 1257–65.

Sokolnikov, M., Preston, D., Gilbert, E., Schonfeld, S., and Koshurnikova, N. (2015). Radiation effects on mortality from solid cancers other than lung, liver, and bone cancer in the Mayak worker cohort: 1948–2008. *PLoS One*, 10(2), e0117784.

Sudan, M., Olsen, J., Arah, O.A., Obel, C., and Kheifets, L. (2016). Prospective cohort analysis of cellphone use and emotional and behavioural difficulties in children. *Journal of Epidemiology and Community Health*, 70(12), 1207–13.

Swerdlow, A.J., et al. (2011). Mobile phones, brain tumors, and the interphone study: where are we now?. *Environmental Health Perspectives*, 119(11), 1534–38.

Thacher, T.D. and Clarke, B.L. (2011). Vitamin D insufficiency. *Mayo Clinic Proceedings*, 86(1), 50–60.

Thomas, S., Heinrich, S., von Kries, R., and Radon, K. (2010). Exposure to radio-frequency electromagnetic fields and behavioural problems in Bavarian children and adolescents. *European Journal of Epidemiology*, 25(2), 135–41.

Thompson, S.C., Jolley, D., and Marks, R. (1993). Reduction of solar keratoses by regular sunscreen use. *New England Journal of Medicine*, 329(16), 1147–51.

United Nations Scientific Committee on the Effects of Atomic Radiation (2008). Summary of low-dose radiation effects on health. In: *UNSCEAR 2006 Report to the General Assembly, with Scientific Annexes*. New York: United Nations.

United Nations Scientific Committee on the Effects of Atomic Radiation (2010). Summary of low-dose radiation effects on health. In: *UNSCEAR 2008 Report to the General Assembly, with Scientific Annexes.* New York: United Nations.

United Nations Scientific Committee on the Effects of Atomic Radiation (2018). *Evaluation of Data on Thyroid Cancer in Regions Affected by the Chernobyl Accident*. New York: United Nations.

van Rongen, E., Saunders, R.D., van Deventer, E.T., and Repacholi, M.H. (2007). Static fields: biological effects and mechanisms relevant to exposure limits. *Health Physics*, 92(6), 584–90.

Vatansever, F. and Hamblin, M.R. (2012). Far infrared radiation (FIR): its biological effects and medical applications. *Photonics & Lasers in Medicine*, 4, 255–66.

Vergara, X., Kheifets, L., Greenland, S., Oksuzyan, S., Cho, Y.S., and Mezei, G. (2013). Occupational exposure to extremely low-frequency magnetic fields and neurodegenerative disease: a meta-analysis. *Journal of Occupational and Environmental Medicine*, 55(2), 135–46.

Vrijheid, M., et al. (2010). Prenatal exposure to cell phone use and neurodevelopment at 14 months. *Epidemiology*, 21(2), 259–62.

Wakeford, R., Darby, S.C., and Murphy, M.F. (2010). Temporal trends in childhood leukaemia incidence following exposure to radioactive fallout from atmospheric nuclear weapons testing. *Radiation and Environmental Biophysics*, 49(2), 213–27.

Wallace, J., Sweetnam, P.M., Warner, C.G., Graham, P.A., and Cochrane, A.L. (1971). An epidemiological study of lens opacities among steel workers. *British Journal of Industrial Medicine*, 28(3), 265–71.

Wehner, M.R., Shive, M.L., Chren, M.M., Han, J., Qureshi, A.A., and Linos, E. (2012). Indoor tanning and non-melanoma skin cancer: systematic review and meta-analysis. *British Medical Journal*, 345, e5909.

Whiteman, D.C., Whiteman, C.A., and Green, A.C. (2001). Childhood sun exposure as a risk factor for melanoma: a systematic review of epidemiologic studies. *Cancer Causes Control*, 12(1), 69–82.

World Health Organization (WHO) (2002). Global Solar UV Index: A Practical Guide. In: *A Joint Recommendation of World Health Organization, World Meteorological Organization, United Nations Environment Programme and the International Commission on Non-Ionizing Radiation Protection.* Geneva: WHO.

World Health Organization (WHO) (2003). *Radiological Dispersion Device (Dirty Bomb)*. Geneva: WHO.

World Health Organization (WHO) (2005). *The International Radon Project*. Geneva: WHO.

World Health Organization (WHO) (2006). *Static Fields*. Geneva: WHO.

World Health Organization (WHO) (2007). *Extremely Low Frequency (ELF) Fields. Environmental Health Criteria, 238.* Geneva: WHO.

World Health Organization (WHO) (2012a). *Health Risk Assessment from the Nuclear Accident After the 2011 Great East Japan Earthquake and Tsunami, Based on a Preliminary Dose Estimation*. Geneva: WHO.

World Health Organization (WHO) (2012b). *Preliminary Dose Estimation from the Nuclear Accident After the 2011 Great East Japan Earthquake and Tsunami*. Geneva: WHO.

World Health Organization (WHO) (2017). *Artificial Tanning Devices: Public Health Interventions to Manage Sunbeds*. Geneva: WHO.

Young, A.R. (2006). Acute effects of UVR on human eyes and skin. *Progress in Biophysics and Molecular Biology*, 92(1), 80–5.

Zhang, M., Qureshi, A.A., Geller, A.C., Frazier, L., Hunter, D.J., and Han, J. (2012). Use of tanning beds and incidence of skin cancer. *Journal of Clinical Oncology*, 30(14), 1588–93.

Zuclich, J.A., Lund, D.J., and Stuck, B.E. (2007). Wavelength dependence of ocular damage thresholds in the near-IR to far-IR transition region: proposed revisions to MPES. *Health Physics*, 92(1), 15–23.

7.3

Occupational health

David Koh and Wee Hoe Gan

Introduction

The World Bank Group (2017) estimated the global labour force in 2017 at 3.449 billion, about 45% of the world's population. Females comprised 39.3% of the total labour force. In addition to workers engaged in hazardous industries, child labour, migrant workers, and those in vulnerable employment continue to be groups of concern when it comes protecting workers' health. Rising life expectancies and declining birth rates have led to population ageing globally, which will impact the demographic profile of the workforce. The average age of the labour force is projected to increase from nearly 40 in 2017, to 41 in 2030 (ILO 2018). While this may be small from a global perspective, the proportion of aged workers will increase, especially in Europe and East Asia. This will put pressure on countries and industries to help older workers adapt physically and psychosocially to changes in the work environment.

Occupational health is therefore a cornerstone to meet the challenges of the workforce now and in the future. The term 'occupational health', as defined by a joint committee of the World Health Organization (WHO) and the International Labour Organization (ILO), involves the 'promotion and maintenance of the highest degree of physical, mental and social wellbeing of workers in all occupations' (Forsmann 1983). This definition emphasizes the term health rather than disease, and further implies a multidisciplinary responsibility as well as a mechanism for the provision of health services for the working population. Occupational health seeks to not only maintain and promote workers' health and working capacity, it also extends to the improvement of working environment and work to be conducive to safety and health (Coppee 2011).

Occupational health practice encompasses the health protection, optimization, and management at the levels of the working population and individuals (Harber et al. 2010). For the former, the main components are the formulation of occupational health legislations and policies to protect the health of the workforce, the systemic reduction of risks from exposure to workplace hazards, and health promotion for those at work. Occupational health interventions at the individual level include primary prevention (e.g. occupational health training for workers, immunizations for healthcare providers), secondary prevention (e.g. pre-placement and periodic medical surveillance), and tertiary prevention (e.g. return-to-work assessments and programmes). In tertiary healthcare institutions, specialized clinical activities such as multidisciplinary management (e.g. occupational dermatology clinic jointly managed by occupational physicians and dermatologists), disability assessments, and acute management of occupational exposures (e.g. excessive mercury exposure) are undertaken (Barron et al. 2005).

Occupational health must be practised taking the context of the worker's job roles and environment. Both work and non-work environmental conditions can affect the health of the worker and in turn, his ability to undertake work safely and effectively. From the traditional ambit of industrial hygiene and ergonomics, the scope of occupational health practice must necessarily extend to include environmental health considerations and the holistic management of the workforce. Examples include hazardous waste management and disposal in a chemical plant, prevention of outbreaks of occupational infections in a healthcare facility, emergency preparedness and response (American College of Occupational and Environmental Health (ACOEM) 2010).

History and evolution

The Italian physician Bernardino Ramazzini (1633–1714) is often described as the 'Father of Occupational Medicine'. His publication, *De Morbis Artificum Diatriba*, is the seminal text in occupational medicine. Ramazzini stated that according to Hippocrates' teaching, 'When you come to a patient's house, you should ask him what sort of pain he has, what caused them, how many days he has been ill, whether the bowels are working and

what sort of food he eats' (Ramazzini 1713/1964). Following this citation, Ramazzini wrote: 'I may venture to add one more question: *What occupation does he follow?*' Ramazzini described many occupational illnesses that are still seen today, and the principles for their control.

The Industrial Revolution in the eighteenth century influenced the development of occupational health profoundly. Engineering and scientific innovations in the Western world, related to newly introduced industrial processes and the setting up of factories, led to dramatic social changes. Previously, work was done mainly by craftsmen in rural cottage industries. The Industrial Revolution resulted in work being carried out in factories in urban centres.

Effects were seen both within the community, as well as in the individual worker. Family life was disrupted, with men leaving their families and moving to work in new industrial areas. In these new areas, health and social problems emerged, for example, poor housing and sanitation, alcoholism, prostitution, and poverty. Inside factories, individuals were exposed to long hours of work, uncontrolled occupational hazards, and the risk of accidents. Child labour and apprenticeship of young children were common, and there was an absence of labour legislation.

The World Wars in the first half of the twentieth century further propelled our understanding of hazardous occupational exposures and diseases. Workers were affected with tetrachloroethylene (TCE) poisoning from applying TCE to aircraft wings, and trinitrotoluene poisoning from filling ammunition shells with little or no personal protection. Some of the phenomena seen in the Industrial Revolution are being replicated in the rapidly industrializing developing nations. Even in developed nations, similar problems are still encountered by migrant workers and other deprived sectors of the working population.

Occupational health legislation

Occupational health legislation in many countries is one of the most important measures to protect workers' health. For example, in the United States (US), the Occupational Safety and Health Act was passed by Congress in 1970. The United Kingdom (UK) Health and Safety at Work Act 1974 provided a broad legislative framework for the protection of

workers. The European Union (EU) adopted a policy in 1989 on the 'Fundamental Social Rights of Workers', emphasizing the need for safety and health protection in the workplace, improvements in living and working conditions, and provision of social protection for workers.

Internationally, there is growing recognition that a collective ownership is necessary for occupational health and safety. At the Sixtieth World Health Assembly, the WHO endorsed the global plan of action on workers' health (WHO 2007). This calls for, among other goals, national policy frameworks on workers' health, the protection and promotion of health at the workplace, and the improvement of access to occupational health services.

Disease burden and epidemiology

The ILO reported that in 2015, there were an estimated 2.78 million cases of deaths from occupational causes (Hamalainen et al. 2017). The majority (2.40 million) were due to fatal work-related illnesses, while the rest resulted from work-related accidents. On average, over 7,500 work-related deaths occur every day. About 31% of work-related mortality were due to circulatory diseases, 26% due to malignant neoplasms, 17% due to respiratory diseases, 14% due to occupational injuries, and 9% due to communicable diseases. Respiratory diseases now constitute the third most common cause of work-related mortality as a result of the formal recognition of occupational exposures to a multitude of vapours, gases, dusts, and fumes as an important cause of chronic obstructive pulmonary disease (Collegium Ramazzini 2016).

Approximately 5–7% of global fatalities are thought to be attributable to work-related illnesses and occupational injuries (Lopez and Mathers 2006). The population attributable fraction (PAF), or proportion of a disease that would be avoided in the absence of the risk factor under consideration, is often used to estimate the burden of occupational diseases. While there are several methods to estimate PAF, a commonly used approach is to multiply the relative risk from occupational exposure with the proportion exposed in the population of interest (Steenland and Armstrong 2006). The PAF should be interpreted in context because assumptions may be applied—an example is the period of cancer latency, which affect the determination of the

appropriate occupational exposure period to the carcinogen of interest (Ruston 2017).

In terms of global morbidity, an estimated 37% of back pain, 16% of hearing loss, 13% of chronic obstructive pulmonary diseases, 11% of asthma, 10% of injuries, 10% of lung cancers, and 2% of leukaemia are related to work (Concha-Barrientos et al. 2004).

It is generally accepted that occupational and work-related illnesses are underreported. Occupational epidemiological studies are also often subjected to a selection bias known such as the healthy worker effect (HWE). HWE describes the better health status of workers relative to the general population and is often ascribed to the exclusion of unhealthy workers at employment and the study of active workers (Li and Sung 1999). Several methods have been described to minimize HWE, such as choosing active workers from another industry as the reference group instead of the general population. There is no standard solution and the approach needs to be customized for individual studies.

Over 100 occupational diseases have been classified according to the tenth revision of the International Classification of Diseases and Related Health Problems (ICD-10), and the list is expected to grow in ICD-11. To determine which occupational exposure leads the development of a particular occupational health condition may not be straightforward, given the multifactorial and dynamic nature of exposure-health interface in many workplaces. Recent occupational epidemiological research to map multiple occupational exposures with disease outcomes through novel models will allow better characterization and provide a clearer understanding of the multidimensional interactions between occupational hazards and ill-health (Bosson-Rieutort et al. 2018).

Estimates of costs and economic loss

The economic cost associated with work-related disease ranges from 1.8 to 6.0% of a country's gross domestic product (average 4%) (Takala et al. 2014). Occupational injuries and illnesses account for about US$250 billion annually to the overall national costs of all diseases, injuries, and deaths in the US (Leigh 2011). The UK estimated that the costs of workplace injuries and new cases of work-related ill-health from 2015 to 2016 to be £14.9 billion (Health and Safety Executive (HSE) 2017a).

The main contributors are direct costs for medical care, rehabilitation, and compensation; indirect costs related to lost opportunities for the injured worker and employer comprising salary costs, administrative costs, and productivity losses; and human costs as a result of the change in the quality of life of the worker and those around him (IRSST 2013).

Major types of occupational diseases and injury

Occupational illness can affect virtually every organ system. Occupational diseases of the lung and skin are common since these organs have substantial surface areas in direct contact with toxic substances. Noise-induced hearing loss and musculoskeletal disorders are among the most common disorders arising from physical factors in the workplace. Occupational cancer is a major concern because of its high mortality. More attention is now being channelled to the domains of psychosocial stress at work; occupational infections (especially with the emergence of multidrug-resistant organisms); and health hazards related to new agents such as nanomaterials; and maturing technologies such as the knowledge industry.

Occupational lung diseases

The respiratory system is an easily accessible target organ for airborne toxic agents. Major categories of occupational lung disease include the pneumoconioses (‘dust diseases’ of the lungs), lung cancer, occupational asthma, chronic bronchitis and other effects of respiratory irritants, and occupational pulmonary infections.

Silicosis is the most common form of pneumoconiosis worldwide. Exposure to silica occurs in a wide variety of occupations such as sandblasters, miners, millers, pottery workers, foundry workers, and workers using abrasives. Approximately 2.3 million workers in the US are exposed to respirable crystalline silica (RCS) at the workplace (Centers for Disease Control and Prevention 2017). The International Agency for Research on Cancer (IARC) has classified crystalline silica as a known human carcinogen (IARC 1997).

Asbestos is another important cause of lung diseases that include mesothelioma, lung cancer, and asbestosis (a form of progressive pulmonary fibrosis). There are different forms of this group of naturally

occurring fibrous minerals, with the main ones being chrysotile (white asbestos) and crocidolite (blue asbestos). All forms are carcinogenic. It is estimated that 125 million people in the world are exposed to asbestos at the workplace (WHO 2017a). Asbestos is responsible for over 100,000 deaths per year worldwide (Takala 2003), with an estimated 38,400 deaths annually due to mesothelioma (Odgerel et al. 2017). Given the long latency, the future burden of mortality resulting from asbestos will be substantial even if all future exposure were to be eliminated completely.

Work-related asthma (WRA) comprises occupational asthma and work-exacerbated asthma (WEA, i.e. asthma not caused by, but rather worsened by an occupational exposure). At least 15% of adult asthma is thought to be work-related (Tario and Lemiere 2014), and WEA may occur in up to 25% of working persons with asthma (Henneberger et al. 2011). WRA may be caused by occupational allergens (such as inhalation of hay dust in farmers) or irritants (cleaning agents is a common group of irritants). Unlike non-WRA, the optimal management of WRA is primary prevention by eliminating or reducing occupational exposures. If the socioeconomic situation precludes total avoidance, work re-entry programmes may enable sustained employment after diagnosis (Malo et al. 2015). Prevalence studies of occupational asthma usually underestimate the number of affected workers because workers tend to quit jobs where they suffer such symptoms, although their asthma symptoms and signs may continue even after leaving work.

Many vapours, gases, dusts, and fumes (VGDF), such as hydrogen chloride, organic solvents, and welding fumes, can cause irritant effects to the respiratory tract. Increasingly, occupational exposure to VGDF is also shown to be associated with chronic obstructive pulmonary disease (COPD). For example, the Swedish CArdioPulmonary BioImage Study (SCAPIS) demonstrated that the PAF for occupational exposure to VGDF was 0.37 for COPD and 0.23 for emphysema (Toren et al. 2017). It is important that the occupational risk of VGDF exposure be fully appreciated so that adequate preventive measures can be put in place for workers' protection.

Occupational cancers

Occupational carcinogens include chemical substances, for example, benzene and asbestos; physical hazards, for example, ionizing radiation;

and biological hazards, for example, viruses. It is estimated that approximately 16 million workers in the EU are exposed to carcinogens at work. The most common cancers due to these workplace exposures are cancers of the lung, bladder, skin, and liver.

Occupational cancer accounts for about 4–20% of all cancers in developed countries. Variability in the estimates arises from differences in data sets used and assumptions applied. In the UK, 8.2% and 2.3% of mortality from cancers in males and females, respectively were occupation-related, giving an overall PAF of 5.3% (Rushton et al. 2012). Due to the complex interactions between exposure and causation in occupational cancers, the probability of causation model has, to date, been successfully applied only to occupational cancers after ionizing radiation exposure (Niu and Deboodt 2010).

Occupational and work-related skin diseases

Skin disorders are among the most commonly reported occupational diseases. Similar to many occupational disease classifications, the main difference between work-related skin diseases (WRSD) and occupational skin diseases (OSD) is whether the occupational exposure is assessed to be partially or fully attributed to the development of the skin disorder. In Europe, the annual incidence of WRSD is estimated at 0.5–1 per 1,000 workers (Mahler 2017) with the most common condition being contact dermatitis. Although skin disorders are relatively easily diagnosed, OSD are believed to be under-reported, so that the actual rate is many times higher than officially reported (European Agency for Safety and Health at Work 2008).

WRSD/OSD are unevenly distributed between industries. Workers in agriculture, forestry, fishing, manufacturing e.g. metal workers, and some service industries, e.g. florists, hairdressers, and barbers, have a greater risk of developing WRSD/OSD. This is likely to be due to their exposure to a wider range of substances known to be irritants or allergens. With greater recognition and understanding of the role of occupational factors in skin diseases (e.g. occupational UV-induced skin cancer), there will be a revision in the Occupational Skin Diseases section of ICD-11 to provide a more comprehensive framework and facilitate precise definition and classification (Alfonso et al. 2017).

Occupational reproductive disorders

The overall contribution of occupational exposures to reproductive disorders is not well characterized because of limited research in this area. Few studies are available, especially on physical and biological agents that may affect fertility and pregnancy outcomes in humans.

Occupational exposure to chemical agents such as lead (Pb) and the pesticide dibromochloropropane (DBCP) have been documented to cause testicular effects with resultant reduction in sperm count. Pb can also cross the placenta in pregnant women to affect the developing foetus. Other substances associated with documented adverse reproductive outcomes include metal compounds (e.g. methyl mercury), solvents (e.g. carbon disulphide), oestrogenic agents, anaesthetic gases, antineoplastic drugs, carbon monoxide, ethylene oxide, ethylene glycol ethers, polychlorinated biphenyls, and physical agents (e.g. ionizing radiation).

Occupational exposures can cause a wide range of reproductive disorders in both males and females (Kumar 2004; Jensen et al. 2006). Effects in males include reduced sperm count, aberrant sperm shape, altered sperm function; and lowered hormone levels or libido. Exposures in females may cause menstrual cycle irregularities, infertility, chromosomal aberrations, reduced breast milk production, early onset of menopause, and suppressed libido.

Reproductive disorders also include adverse effects on the offspring of the exposed worker (Burdorf et al. 2011). Potential foetal and developmental effects from maternal exposures include preterm delivery, foetal loss, prenatal death, low birth weight, altered sex ratio of live births, congenital malformations, childhood malignancies, and neurobehavioral disorders in the offspring.

Occupational noise-induced hearing loss

Exposure to excessive noise is a major occupational hazard in many industries. In the US, about 22 million workers are exposed to hazardous noise levels at work (NIOSH 2018). An estimated 16% of disabling hearing loss in adults worldwide is attributed to occupational noise (Nelson et al. 2005). In most developing countries, noise-induced hearing loss (NIHL) is one of the most commonly reported occupational disorders.

Workers in manufacturing, mining, construction, transportation, agriculture, and the military are at high risk for NIHL. There is also an association between maternal occupational noise exposure and hearing dysfunction in the offspring (Selander et al. 2016). A Cochrane Review (Tikka et al. 2017) showed varying effectiveness of legislation, engineering control measures, and use of hearing protection devices in preventing NIHL. Nevertheless, the implementation of these measures has led to a systematic framework of noise hazard reduction for the at-risk industries.

Occupational infections

Healthcare is one of several sectors whose professionals are at risk of occupational infections. In addition to blood-borne diseases such as hepatitis B, hepatitis C, and HIV, airborne and droplet-borne infections such as measles, varicella, tuberculosis, and pertussis may also be transmitted from patients to healthcare workers (HCW). Many of these infections are preventable with immunizations. It has been estimated that 9–42 HCW per million die annually from occupational infections (Sepkowitz and Eisenberg 2005).

The incidence rate of sharps injuries in HCW is about 3.7 per 100 HCW per year (Elseviers 2014), though it is known to be underreported (Elder and Paterson 2006). Needlestick injuries accounted for about 40% of hepatitis B and hepatitis C infections, and 4.4% of HIV infections in HCW (Pruss-Ustun et al. 2005). However, with the stricter implementation of universal precautions, adherence to standard operating procedures, introduction of surgical and phlebotomy devices with safety features, and availability of effective post-exposure prophylaxis for occupational infections in high risk cases, a gradual reduction in the incidence of occupational transmission is achievable.

Transmission of *M. tuberculosis* is a recognized risk in healthcare facilities, and multidrug-resistant tuberculosis and extensively drug-resistant tuberculosis have re-emerged as a major occupational and public health problem. The risk of latent tuberculosis infection and tuberculosis disease is considered to be higher in HCW compared to the corresponding local communities (Baussano et al. 2011).

Emerging infectious diseases also pose a risk to HCW. Two examples are the Severe Acute Respiratory Syndrome (SARS) and the Middle East respiratory syndrome (MERS), both of which are caused by novel

coronaviruses. In the 2003 worldwide outbreak of SARS, 20% of patients were HCW (Koh et al. 2003). HCW have also been shown to be the propagating vehicles for healthcare facility outbreaks (Alfaraj et al. 2018).

Other workers besides HCW may also at risk for occupational infections. Some of the infections result directly from the work, while others are indirectly related to work. Examples include vector-borne diseases, such as malaria, dengue haemorrhagic fever, Lyme disease, and anthrax; water and food-borne diseases resulting from poor sanitation and inadequate potable water, such as cholera and typhoid; and zoonoses among veterinarians, farmers, agricultural and forestry workers, such as leptospirosis and rabies.

Occupational traumatic injuries

These injuries include such events as amputations, fractures, severe lacerations, eye injuries, and burns. In 2014, approximately 374 million workers sustained non-fatal occupational injuries that resulted in an absence of 4 days or more from work (Hamalainen P et al. 2017). This translates to an annual accident incidence of 11,096 per 100,000 workers. In the UK, statistics of non-fatal injuries showed that the top 5 causes are—slip, trip or fall on the same level (29%), lifting/handling (22%), struck by object (10%), fall from height (7%), and acts of violence (7%) (HSE 2017b).

Work-related diseases

Work-related diseases are diseases in which occurrence may be associated with workplace factors, but these need not be the only risk factor in each case. They occur more commonly than pure ‘occupational diseases’. The term is also used to refer to aggravation of a pre-existing medical condition by workplace factors. While prevention of occupational diseases is possible by the elimination of the workplace hazard, work-related diseases cannot be entirely prevented by only addressing occupational hazards. Common work-related diseases include hypertension, ischaemic heart disease, psychosomatic illnesses, musculoskeletal disorders, and non-specific respiratory disease. Two examples are given next.

Work-related musculoskeletal disorders

Acute and chronic work-related musculoskeletal disorders (WRMSD) are a major cause of morbidity affecting workers. The term encompasses a range of conditions from back pain, upper and lower limb disorders, neck and shoulder symptoms, hand–arm vibration syndrome (HAVS), carpal tunnel syndrome, repetitive strain injury, and other effects on the musculoskeletal system. There is substantive evidence to support the aetiologic importance of physical risk factors, including rapid work pace and repetitive motion, forceful exertions, non-neutral body postures, and vibration (Punnett 2014). Psychosocial factors such as high job demands, lack of job control, and monotonous work have additionally been shown to be associated with increased WRMSD risk.

In the UK, more than 0.5 million workers suffer from new or longstanding WRMSD between 2016 and 2017, resulting in a loss of 8.9 million working days (HSE 2017c). In the US, WRMSD accounted for 34% of all lost workdays, and one out of every three dollars spent on workers' compensation (Occupational Safety and Health Administration (OSHA) 2014). The implementation of ergonomic measures, and simple interventions such as stretching exercises may be effective in preventing WRMSD (Podniece et al. 2008; Aje et al. 2018).

A common WRMSD is low back pain (LBP). LBP from occupational ergonomic factors accounts for about one-third of all disabilities with occupational risk factors (Driscoll et al. 2014). Given the high prevalence of back pain in the general population and its episodic nature, it is not always easy to determine the work-relatedness of this condition. While there are many studies demonstrating the association between work-related factors and the development of LBP, a recent longitudinal study showed that early exposure to heavy physical workload is associated with more than 30% risk of local and radiating LBP after a follow-up of 21 years (Lallukka et al. 2017). This highlights the importance of primary prevention even in younger workers when they enter the labour market.

Stress-related ill-health

Work-related stress has been defined as the 'harmful physical and emotional responses that occur when the requirements of the job do not match the capabilities, resources or needs of the worker' (NIOSH 1999). Over the past half century, globalization and rapid technological advancements introduced significant changes to the workplace. New forms of work organization and

employment patterns (e.g. more temporary and independent contracting) bring about job flexibility, yet at the same time result in lower control and a heightened sense of job insecurity. The fast pace of technological renewals enhances work efficiency, but also requires workers to constantly keep up with changes. Greater connectivity through the internet has blurred the boundary between work and personal life. Past European Working Conditions Surveys gave a prevalence of between 11.8% and 27.0% of work-related stress in the UK working population (Houdmont et al. 2010).

The clinical manifestations of occupational stress are indistinguishable from non-occupational stress. It can present as emotional lability, anxiety, depression, insomnia, and even attempts at suicide. Adverse health outcomes of job stress are wide ranging, from increased risk of cardiovascular disease, musculoskeletal symptoms, impaired immune function, and gastrointestinal disorders (e.g. peptic ulcers and irritable bowel syndrome). The varying rates of coronary heart disease across occupations, and the link between work-related psychosocial factors and WRMSD support the role of work conditions in work-related stress (ILO 2016a). In extreme situations, work-related stress may result in *karoshi* (death by cerebrovascular and cardiovascular diseases due to overwork) and *karojisatsu* (suicide due to overwork) (Yamauchi et al. 2017).

Under-recognition of occupational ill-health

Many countries maintain occupational disease registries. However, the lack of internationally agreed reporting standards and, in some cases, tedious notification processes result in the underestimation of the prevalence of some of the occupational diseases under surveillance (Somayeh et al. 2017). It is generally accepted that occupational and work-related diseases are underdiagnosed. One explanation for this is the inherent difficulty in establishing cause-and-effect relationships for occupational diseases. The link between occupation and disease may sometimes be unclear, because most occupational diseases are not distinct clinically and pathologically from diseases associated with non-occupational aetiologies. For example, skin cancer caused by polycyclic aromatic hydrocarbons is similar in appearance to that caused by excessive exposure to sunlight. Similarly, solvent-induced encephalopathy may easily be confused with effects of advancing age. Only for a very limited number of diseases, for example,

mesothelioma (caused by asbestos), and angiosarcoma of the liver (caused by vinyl chloride monomer) are the causal association between occupational exposure and disease readily established based on strong epidemiological evidence and a good occupational history documenting exposure.

The long latency which typically elapses between occupational exposure and onset of illness is a second factor which may obscure the occupational aetiology. For example, occupational cancers rarely appear within 10 or even 20 years of first exposure to a carcinogen. Similarly, chronic solvent-induced encephalopathy may develop insidiously only after years of exposure.

Lack of a good understanding among health practitioners about the hazards found at work is a third cause of underestimation of occupational disease, reflecting the fact that most physicians are not adequately trained to suspect work as a cause of disease (Baillargeon et al. 2011). Technological evolutions and changes in employment patterns lead to new occupational diseases, further compounding this knowledge deficit. With competing demands for time in the medical curriculum from an increasing number of disciplines and subject areas, most medical schools now only allocate limited time to teaching medical students to take a proper occupational history, to recognize symptoms of common industrial toxins, or to recall known associations between occupational exposures and disease.

Compounding a lack of medical awareness is the limited ability of many workers to provide an accurate report of their exposures. Workers may have had multiple toxic exposures in a variety of jobs over a working lifetime. In many developing countries today, there is still no formal requirement to inform workers of the hazard of the materials with which they work. In many instances, a patient may not know or may not be able to remember all his or her past occupational exposures.

Finally, given the potential financial liability associated with the finding that a disease is of occupational origin, employers may be reluctant to recognize the work-relatedness of a disorder, especially in cases where personal habits or non-occupational pursuits are possible contributory factors. Since employers are often in the best position to recognize causal associations between workplace exposures and disease, this conflict of interest represents an obstacle to obtaining accurate estimates of the burden of occupational illness.

Many of the occupational hazards today are still not fully evaluated for their potential toxicities. While it is another reason for the under-recognition of occupational diseases, there is a concerted effort globally to improve this through better classification and communication of the risk to workers.

Special populations of workers

Workplace hazards affect some worker populations disproportionately, for example, child labourers, women employees, and impaired workers. These populations are especially vulnerable because of the interaction between their work roles and broader roles in society, as well as by their particular exposures in the workplace.

Child labour

The ILO estimates 218 million children between five and 17 years are in employment globally, of whom 152 million are victims of child labour. Of the latter group, 73 million are engaged in hazardous work, concentrated primarily in agriculture (e.g. fishing, forestry, livestock herding). About 22,000 children are killed at work every year (International Programme on the Elimination of Child Labour (IPEC) 2017).

Children are vulnerable to exploitation in the workplace and may be denied basic human rights, such as decent wages or access to education. Children in hazardous occupations are at greater risk of suffering ill effects than adult workers. They may have greater exposure to hazards than adult workers in the same occupation because children tend to be given the most menial jobs, which can involve higher exposures to toxic substances. Children are more susceptible to the same hazards faced by adult workers because they differ from adults in their physical, physiological, and psychological characteristics. Children using hand tools designed for adults run a higher risk of fatigue and injury. Personal protective equipment (PPE) designed for adults do not fit children, and therefore do not provide adequate protection. Furthermore, children may not be as aware as adults of workplace dangers, or as knowledgeable of precautions to be taken at work. Children are also more vulnerable to psychological and physical abuse than are adults and suffer deeper psychological damage when they are denigrated or bullied.

In a study of more than 42,000 child labourers in Bangladesh, the incidence of injury and death among child labourers was 24 per 100,000 children-years (Dalal et al. 2016). The top three causes of injury were cuts/open wounds, abrasions/lacerations, and sprains; and the corresponding top causes of death were suicide, road traffic injuries, and electrocution. Child labour not only deprives children of education and training for a better future, it also perpetuates poverty and impacts countries' economies through loss of competitiveness and productivity. The International Programme on the Elimination of Child Labour (IPEC), established under ILO, works with member States and partners to encourage the withdrawal of children from child labour by providing them with education, and assisting their families with skills training and employment.

Women workers

In 2015, nearly 1.3 billion women were in the workforce compared to almost 2 billion men. As a global average, the female employment-to-population ratio was 0.46, whereas the corresponding male ratio was 0.72 (ILO 2016b).

There is significant interplay between women's roles in society, socioeconomic condition, and occupation. Women's roles in traditional societies are defined in relation to childbearing and upbringing of children, and responsibility for the care and comfort of the family. Paid employment of women has increased in most countries, but this has led to a conflict between the necessity to be at work and women's traditional home and family responsibilities. In many societies, early marriage, repeated pregnancies, large families, low educational status, and poverty all disproportionately impact on women workers. Even today, international objectives for women's health, such as the Millennium Development Goals, are focussed on maternal-child health rather than advancing women's intrinsic health (Mohindra 2016).

Employment of women in most societies is characterized by occupational segregation, underemployment (often involved in doing seasonal and part-time work below their level of education), and barriers to advancement. Occupational segregation means that women tend to be clustered into a small number of occupations while being under-represented in most others. Women in developing countries tend to be employed in sectors such as

agriculture, textiles and clothing, food processing, and social services. Compared with men, women work for smaller industries or organizations, and have less opportunity for work control. Women are more likely to work in the informal sector, in specific types of informal work such as domestic work, street vending, and sex work, with their accompanying low social status and lack of legislative protection. Gender differences are also observed in the rates of occupational injuries and illnesses, including the higher rates of mental disorders in females based on studies of workers' compensation claims (Berecki-Gisolf et al. 2016).

Increasing recognition of the importance of gender differences towards risks at work has led to gender mainstreaming in occupational health and safety (OHS). This refers to the systematic consideration of the differences between the conditions, situations and needs of female and male workers, and integrating these into the design, implementation, monitoring, and evaluation of OHS (European Agency for Safety and Health at Work 2014).

Disabled workers

A large proportion of individuals with disabilities are in the working age group, and they are able to make constructive contributions in the workplace despite physical impairment. Legislation in some countries, such as UK's Equality Act 2010, prohibits employers from discriminating against individuals with disabilities, and encourages employers to make reasonable accommodation for those with a known impairment (HSE 2015). The Americans with Disabilities Act limits employers from asking job applicants to answer medical questions, take a medical examination, or identify a disability (US Equal Employment Opportunity Commission (EEOC) 2008).

France and Germany impose quotas on enterprises to employ a certain percentage of disabled people. India and Japan use employment promotion measures to ensure workplace accessibility and provide employment services in the form of job placement agencies. Singapore provides tax reductions as a financial incentive to compensate employers for any financial burden resulting from the employment of disabled people.

Reasonable accommodations are changes made to the work environment, job responsibilities, or conditions of work that provide opportunities for workers with special needs to perform essential job functions. The aim is to

ensure that, as far as reasonably practicable, a disabled person has the same access to everything that is involved in performing and keeping a job as a non-disabled person. These measures may include technical assistance devices; customization, including PPE and clothing; and changes to processes, location, or timing for essential job functions. Many studies have demonstrated benefits of hiring people with disabilities, including improvements in profitability (e.g. cost-effectiveness); staff retention, reliability, and punctuality; employee loyalty; company image; competitive advantage (e.g. diverse customer pool, innovation); and inclusive work culture (Lindsay et al. 2018).

Migrant workers

Migrant workers refer to all international migrants who are currently employed, or are unemployed and seeking employment in their present country of residence. The ILO estimates that there are 150.3 million migrant workers globally (ILO 2015). This forms 3.9% of the total global population and 4.4% of the global workforce. The majority of migrant workers are in the services (domestic work, food services, administrative work) sector. The other sectors significantly represented include manufacturing, construction, and agriculture.

Migrant workers are a particularly vulnerable population. Many of them perform what are known as 3-D (dirty, dangerous, and demanding) jobs (Moyce and Schenker 2018). They may be affected by poor nutrition and endemic diseases, often have lower educational backgrounds, and are inadequately trained to deal with potential work hazards. Many face additional obstacles because they do not speak the language of the host country adequately. They are often not familiar with local health and safety practices and regulations. The accommodation available for them can be temporary and often crowded with limited shared facilities, and with restricted access to medical care and other social services. Migrant workers may encounter racism, xenophobia, and exploitation because they do not have full legal status or awareness of their rights in the host country (Holmes 2006). A review of fatal occupational injuries in the US from 2003 to 2010 found a risk difference of 15% increase in relative hazard for immigrant workers compared to their native counterparts (Byler and Robinson 2018).

Shift workers

Shift work can take the form of fixed shifts, where workers constantly work the same shift period; or rotating shifts, where workers' shift periods vary over time. It is estimated that in the EU, 17% of workers are engaged in shift work (ILO 2011).

Apart from sleep deficits and suboptimal work performance, shift work has also been associated with adverse effects on psychological health (e.g. anxiety and depression), gastrointestinal disorders (e.g. peptic ulcer disease), metabolic disorders (e.g. obesity and impaired glucose tolerance), cardiovascular diseases, reproductive disorders (e.g. preterm birth), and cancer (especially breast cancer in female shift workers) (Costa 2010). The IARC classifies shift work as a Group 2A carcinogen, defining it as 'probably carcinogenic to humans'.

Shift work-related ill-health may also increase with time of 'exposure'. A study found that after the first five years of shift work, the risk of cardiovascular events increases by 7.1% for every additional five years of exposure (Torquati et al. 2018). Given the complex and multifaceted interactions between shift work and health, it is important to minimize downstream consequences for susceptible workers through health surveillance, education, and proper organization of shift schedules.

Total workplace safety and health

Work as a social determinant of health has been extensively researched. It is recognized that workload, duration, intensity, psychological stress, access to workplace health promotion resources, and other occupational-related factors have an impact on well-being. There is no unified definition of well-being. The concept of well-being encompasses the presence of positive emotions and moods, engaging activities and work, fulfilment, and positive functioning (CDC 2016). By integrating physical health and well-being into a multidimensional indicator, it provides a more holistic approach to the understanding of occupational and work-related diseases and, in turn, their prevention and management.

Well-being, alongside safety and physical health, is an important outcome in occupational health and contributes to the changing paradigm of a new total workplace safety and health model. This model recognizes that

in addition to addressing workplace safety issues, good workplace health (physical and psychosocial health) is equally important because it affects the worker's overall well-being and therefore, his work capacity, and the overall performance of the organization. A central concept in Total Workplace Safety and Health is the protection and promotion of the health, safety, and well-being of all workers and the sustainability of the workplace (Workplace Safety and Health Council 2017).

To this end, individual risk factors, such as worker's health and fitness, are taken into consideration in risk assessment. By the corollary, integrating programmes that control such risks at the workplace, including workplace health promotion and workers' chronic disease management, can address risks that impact individual performance, organizational productivity, workplace accidents, and sickness absenteeism.

There are ample justifications for this integrated approach. Epidemiological studies show the following key findings, namely that occupational exposures and risk factors for chronic diseases may have synergistic adverse health effects on workers; workers with high risk for hazardous workplace exposures often have more risk factors for chronic diseases; and integration of the management of occupational safety and health with health promotion may increase worker participation and benefit the overall work environment (NIOSH 2012).

An illustration of this integrated approach to occupational health is the Total Worker Health® (TWH®) launched by the US' National Institute for Occupational Safety and Health (NIOSH) in 2011. TWH® is defined as policies, programmes, and practices that integrate protection from work-related safety and health hazards with promotion of injury and illness prevention efforts to advance worker well-being (NIOSH 2017). There are many domains in TWH®, which include control of work hazards and exposures (e.g. risk assessment and management); work organization (e.g. shift work, fatigue and stress prevention); support for built environments (e.g. health air quality, access to healthy and affordable food options); employment patterns (e.g. contract work, financial and job security); work policies (e.g. promotion of productive ageing, return to work); compensation and benefits (e.g. chronic disease management, work injury compensation); and leadership (e.g. supportive leadership, employee engagement, provision of meaningful work). These domains complement and mutually reinforce to not only protect workers from hazards, but also to

advance their health and well-being by targeting the conditions of work. A systematic review of TWH® interventions (Anger et al. 2015) demonstrated improved risk factors for injuries (e.g. self-efficacy towards safety) and chronic diseases (e.g. exercise, dietary habits, and measurable indicators such as lipid profile, blood pressure, and body mass index).

Return on investment may be significant. The introduction of the Promoting Healthy Lifestyle: Alternative Models' Effects (PHLAME) wellness programme in fire departments resulted in significantly lower work compensation claims and medical costs compared to fire departments without similar programmes, yielding an average cost saving per firefighter of $4.60 for every $1 invested in the programme (Kuehl et al. 2013). Apart from cost considerations for employers, Total Workplace Safety and Health approaches also complement broader public health outcomes in reducing at-risk health behaviours and the burden of chronic diseases. Interventions from such integrated programmes have been shown to reduce sedentary behaviour and tobacco use in workers (Feltner et al. 2016).

Occupational risk assessment

In occupational health practice, an understanding of the difference between the terms 'hazard' and 'risk' is essential.

Hazard refers to a substance, agent, or physical situation with a potential for harm in terms of injury or ill-health, damage to property, damage to the environment, or a combination of these. Hazards can be physical, chemical, biological, ergonomic, or psychosocial in nature. Physical hazards include exposure to extremes of temperature, light, noise, vibration, electricity, and radiation. Chemical hazards cover a range of organic (aliphatic or aromatic) and inorganic chemicals. Biological hazards are those due to infectious organisms (e.g. viruses, bacteria, rickettsia), and also prions and proteins. Ergonomic and mechanical hazards result from poor design of workstations and disorganized systems of work, and psychosocial hazards are those that contribute to and cause workplace stress. Hazard identification is the process of recognizing that a hazard exists and defining its characteristics.

Risk relates to the likelihood of the harm or undesired event occurring, and the consequences of its occurrence. It is the probability that the substance or agent will cause adverse effects under the conditions of use

and/or exposure, and the possible extent of harm. It is thus a function of both exposure to the hazard and the likelihood of effects on health from the hazard. Extent of risk covers the population that might be affected by the risk, the numbers exposed, and the consequences.

Assessing risk

Risk assessment is the process of estimating the magnitude of risk and deciding if the risk is tolerable or acceptable. A tolerable risk may not always be acceptable. It merely refers to a willingness to live with a risk to secure certain benefits, and in the confidence that the risk is being properly controlled. The levels of tolerability of risk are different for different countries, and in different working populations and the general public. The views on tolerability depend to a large extent on the populations that might be affected by the risk.

Risk assessment and risk management must take into account both routine and non-routine activities and conditions, including foreseeable emergency situations. Hazards that are intrinsic to these situations or generated by such activities should be recognized. Exposed people should be identified, including non-employees and those who are susceptible and therefore at higher risk because of illness or other medical conditions. Existing control measures, if any, need to be evaluated.

The health risks from the hazards should be determined and assessed, and a decision made on whether the risk is acceptable or tolerable. Unacceptable risks have to be eliminated or reduced with new or improved control measures, and their effectiveness monitored. Such a process requires a team effort, involving the workers themselves as well as personnel with the relevant expertise. It is important to inform workers of the hazards, risks, and appropriate measures that have been and can be taken to protect them.

The steps for risk assessment for chemical, biological, ergonomic, and psychosocial hazards may differ. The assessments for chemical or physical exposures are generally more objective and precise than the assessment for psychological stressors. As an example, an initial assessment for a chemical exposure might include the following steps:

List substances in the area to be assessed.
Determine which substances are actually used, and in what quantities, and the frequency and duration of use.
Evaluate workers' concerns.
Assess the tasks of workers, their exposure, and methods of handling.
Obtain suppliers' safety data sheets.
Evaluate safety data sheets.
Inspect places where the substances are handled.
Evaluate method(s) of control.
Perform environmental monitoring for the chemical if required.
Decide on acceptability or tolerability of risk, and if further control measures are needed.

The assessment of psychosocial factors at work is more complex. It may include the evaluation of organizational dysfunction, work conditions, as well as a study of indicators such as sickness absence, staff turnover, and measurement of stress-related illness among employees. Possible work stressors are poor design of tasks, aggressive management style, inadequate interpersonal relationships, unclear work roles, limited career prospects, and adverse workplace environmental conditions. The extent of work stress can be assessed by use of validated questionnaires administered to staff; for example, the General Health Questionnaire (Jackson 2007) or the NIOSH Generic Job Stress Questionnaire (NIOSH 2008).

Assessing exposure

Environmental monitoring

Environmental or ambient monitoring in the workplace is undertaken to measure external exposure to harmful agents. The monitoring is to ensure that exposure is kept within 'permissible levels' so as to prevent the occurrence of disease. The concept of permissible levels assumes that for each substance, there is a level of exposure at or below which the exposed worker does not suffer any health impairment.

Permissible levels or occupational exposure limits are standards that have been set for physical, chemical, and biological hazards in workplaces. There are variations in permissible levels or exposure limits from different

countries and different agencies. Much depends on factors such as the philosophy and rationale for the standards, the process of standard setting, nomenclature, and applicability, and whether economic considerations and technical feasibility are taken into account. For example, threshold limit values (TLVs) for chemical substances and physical agents are set and reviewed annually by the American Conference of Governmental Industrial Hygienists (ACGIH). They are derived using information from industrial experience, as well as studies in both animal and human populations. These standards are adopted by many countries, especially those without their own national occupational exposure standard setting mechanism. The US also has recommended exposure limits (RELs) produced by NIOSH, and permissible exposure limits (PELs) from the OSHA. The UK has Workplace Exposure Limits (WELs), and in Germany there are two groups of OELs: (1) technical guidance concentrations (TRKs), and (2) maximum concentration for a chemical substance in the workplace (MAK).

Permissible levels have their limitations. For example, a ‘safe’ level of exposure can be difficult to determine for agents such as allergens and carcinogens. As such, every effort must be made to keep exposure levels as low as reasonably practicable and the permissible level is then a level above which exposure must not occur. This level may be mandated by legislation. Concerns regarding health risks have arisen from substances assumed previously to be safe, for example, glycol ethers in the electronic industry associated with the risk of spontaneous abortions. Standards that are set for an 8-hour working day would not be applicable for a 12-hour work shift. Furthermore, exposure to several hazards simultaneously does occur. In such situations, there can be possible synergistic or additive effects. This would then require more stringent control of each individual hazard.

The methods for environmental monitoring require choice of the correct collecting devices, sampling strategy, and analysis of the collected samples in accredited laboratories. Variations in age, gender, pre-existing disease, genetic makeup, and social habits, for example, smoking, influence individual susceptibility and have to be considered in applying the findings from exposure assessment to the exposed population. Despite these limitations, sensible use of environmental standards can often result in practical control of many common workplace hazards so that the majority of workers are protected.

Biological monitoring

Biological monitoring of workers refers to the measurement and assessment of workplace agents or their metabolites either in tissues, secreta, excreta, expired air, or any combinations of these to evaluate exposure and health risk. The specific chemical or its breakdown product can be measured to detect the total body burden of the substance. The method of measurement of these substances must be validated and there should be a means to interpret the results obtained in terms of the extent of exposure, and risk to health.

Biological monitoring and environmental monitoring complement each other in the assessment of health risk in the exposed worker. One major feature of biological, as compared to environmental, monitoring is that for a particular individual, it takes into account exposure from all routes of absorption. For example, for workplace exposure to organic solvents, skin absorption may be a significant route of entry of the solvent into the body, and ambient environmental air monitoring might be less useful as an indicator of exposure than biological monitoring. Furthermore, environmental monitoring at the workplace would not account for non-occupational or extraoccupational exposures. A person exposed to organic solvents at work may have additional exposure at home as a result of hobbies or other non-occupational activity.

While the term ‘biological monitoring’ has previously been used to also include clinical procedures (e.g. periodic X-rays, blood tests, symptom enquiry, etc.) for monitoring the health status of exposed workers, it is preferable to make a clear distinction between the terms ‘biological monitoring’ (used for assessing exposure), and ‘biological effects monitoring’ and ‘health surveillance’ (used for detecting effects).

Biological monitoring standards

The Biological Exposure Index (BEI) represents the levels of determinants that are most likely to be observed in specimens collected from healthy workers who have been exposed to chemicals to the same extent as workers with inhalation exposure at the TLV (ACGIH 2017). Exceptions would be made for chemicals for which TLVs are based on non-systemic effects, for example, irritation; and for chemicals with significant routes of entry via additional routes of entry (usually percutaneous absorption).

The ACGIH cautions that BEIs are not able to clearly distinguish hazardous from non-hazardous exposures. Biological variation in individuals may sometimes result in measurements in individuals to exceed recommended BEI levels without an increased health risk. BEIs are also not intended for use as a measure of adverse effect or diagnosis of occupational disease. However, if measurements of the individual or group of workers persistently exceed the BEIs, the cause of the excessive values should be investigated, and measures should be taken to reduce the exposure.

Biological effects monitoring

Biological effects monitoring refers to the measurement and assessment of early biological effects, of which the relationship to health impairment has not yet been established, in exposed workers to evaluate exposure and/or other health risk compared to an appropriate reference. Health surveillance of workers is a common modality of biological effects monitoring. Examples are audiometry for noise-exposed workers, and clinical examination for skin lesions in workers exposed to polycyclic aromatic hydrocarbon compounds in tar, pitch, and bitumen. Biological markers are also used where relevant, including the detection of alterations in enzyme levels (e.g. cholinesterase for workers exposed to organophosphorus or carbamate pesticides), or other biochemical changes such as delta-aminolaevulinic acid in urine of workers exposed to inorganic lead, or beta-2 microglobulin in the urine of cadmium exposed workers. These changes do not necessarily lead to any direct overt pathological damage in an individual, but may reflect effects that are reversible on removal of the worker from further exposure.

Biomarkers can be used to monitor exposure, effect, or susceptibility. A single biological specimen may contain several biomarkers, each of which with the potential to be used as in occupational and environmental health. For example, there have been occupational stress studies that measure levels of salivary cortisol (a stress hormone) and secretory Immunoglobulin A (a marker of effect) (Koh and Koh 2007). However, not every biomarker correlates well with the level of exposure, effect, or susceptibility. There are also other challenges, such as the compatibility of collection methods, stability of biomarkers, and other factors confounding the measured levels of biomarkers, such as age, diurnal variations, medications, and health behaviours (e.g. smoking).

The availability of genetic tests to identify susceptible workers raises issues of ethics, individual privacy, right to work, and the relevance of such tests. Given the limitations of individual molecular biomarkers in assessing health risk, and the multifactorial nature of environmental disease, more work needs to be done to develop our understanding of the exposome, which is the constellation of all exposure events (exogenous and endogenous, including lifestyle) from conception to death. The integrated knowledge of the exposome and our genome may allow us to subsequently address all the factors that affect the health of our workforce and better control those that are work-related (Schulte and Hauser 2012).

Risk management

Prevention and control of occupational disease can take place at various levels, such as at the national level, at the level of the workplace itself, or measures directed at the individual worker.

Primary prevention aims at reducing the occurrence of disease by eliminating the causal factors contributing to disease or reducing workplace exposure to safe levels. Examples are banning the use of asbestos and reduction of noise at its source to levels that do not cause hearing loss. Primary prevention with regard to chemicals requires either: (1) elimination of toxic materials and their replacement by less hazardous substitutes or (2) use of safe systems of work, and controls such as complete enclosure or local exhaust ventilation at the source of aerosol generation.

Secondary prevention aims to detect overexposure, or early reversible effects of disease in order to take corrective action. For example, regular monitoring of blood lead levels among exposed workers could indicate a potential for lead poisoning, or regular audiograms among noise-exposed workers may show temporary threshold shift (TTS). Removal of workers with TTS from further noise exposure, and subsequent containment of the source of the noise can allow recovery of hearing. Successful secondary prevention depends on the ability to identify work-related illness early and effectively through screening workers at high risk for occupational disease.

Tertiary prevention aims to minimize the consequences in people who already have established disease. This activity is primarily a curative and rehabilitative function and depends on availability of appropriate treatment.

However, even for acute poisonings, there are very few specific effective antidotes available. Hence, the focus should be on primary or secondary prevention.

Hierarchy of controls

Within the workplace, a hierarchy of control measures is often invoked for reducing occupational hazards. This consists of considering the following steps listed according to order of priority:

Hazard elimination or substitution.
Engineering controls.
Redesign of the workstation or work process.
Administrative controls and worker education.
Use of PPE.

Hazard Elimination or substitution

Elimination removes the health risk completely, and has been used for carcinogens like asbestos and benzene, or substances that can cause serious health effects such as heavy metals. For example, it is possible to produce new soldering materials that contain little or no cadmium at all, and the HSE in the UK has produced legislation for this elimination (Mason et al. 1999). Even for materials where total removal is not possible, elimination strategy through change-of-form approaches have been adopted. These include coating nanoparticles, or altering the physical state or particle size through pelletizing, or prilling to encapsulate nanoparticles (Schulte et al. 2016). These modified strategies reduce the hazardous exposure and may make it safer for the worker.

Substitution of the hazard with a less toxic alternative is another option. In the case of processes which use solvents, such as degreasing operations, a less toxic solvent such as 1,1,1 trichloroethane can be used, instead of the comparatively more toxic trichloroethylene or tetrachloroethylene. Selection of a less hazardous process or equipment also represents a meaningful control strategy. For example, substitution of a continuous process with an intermittent process usually results in a decrease in exposure. Instead of replacing an entire work process to reduce hazards,

equipment substitution may achieve the desired reduction in exposure. An example is use of a degreaser with a hoist instead of dipping components into solvents by hand.

Engineering controls

Automation, enclosure, or segregation of a work process, and the use of dampeners or mufflers to reduce vibration or noise have been successful engineering control measures used. Improved ventilation is also an effective and widely used control measure. Control of hazards by ventilation can be through local exhaust ventilation or general ventilation. The approach for implementing ventilation controls is firstly to evaluate sources and degree of exposure; secondly, to develop an engineering design; thirdly, to install a system based on the design; and lastly, to evaluate the completed system to ensure that the air contaminant has been effectively controlled. In healthcare, engineering measures such as negative pressure ventilation serve as an important airborne precaution strategy to minimize the transmission and propagation of infectious diseases such as SARS (Esswein et al. 2004).

Isolation is defined as the interposing of a barrier between a hazard and workers who might be injured or made ill by the hazard. Isolation may refer to storage of materials, such as flammable liquids, enclosure or removal of equipment to another area (such as noisy generators), or isolation of processes or of the workers themselves (e.g. by enclosing a sawmill worker in a soundproof ventilated booth to protect them from noise and wood dust). The petroleum industry, for example, uses automated remote processing in plants based on centralized computer control of process equipment. Workers are thus largely isolated from hazards except in maintenance operations and during process upsets.

Suppressing the substance by ‘wetting’ of dusty operations is another example of effective engineering control. Alteration of work practices can help to reduce exposure to hazards. An example is vacuuming cotton lint off spinning machines rather than blowing it off with compressed air, which creates airborne dust particles.

Redesign of workstation or work process

Workstation redesign to reduce unnecessary and repetitive bending or poor postures can minimize ergonomic hazards. Among computer operators, use of adjustable equipment, positioning of the workstation to reduce glare, and provision of appropriate work rest periods can prevent the occurrence of visual and musculoskeletal complaints.

Administrative controls and worker education

Administrative controls may be a viable alternative or an additional measure to reduce worker exposure to occupational hazards. This could take the form of job enlargement or job rotation, restriction of hours of work at a hazardous operation, or temporary job reassignment.

With administrative controls, the level of exposure to the hazard is not diminished; instead, the duration of exposure is reduced. For example, mandatory rest periods and shorter work hours for outdoor workers during extremely hot weather (as in desert environments) have helped to reduce the likelihood of heat stress. Another example is the reduction of overall noise exposure through rotation of tasks. In recent years, the allostatic concept of well-being and health premised upon the balance between work demands and restorative needs is getting renewed attention. It has resulted in more research into the benefits of alternations between physical and mental tasks within the same job to enhance the well-being of workers (Jahncke et al. 2017).

Training of workers to recognize work hazards, to work safely, and to know what to do in the event of an emergency is another important aspect of prevention. There may be considerable variability in exposure even among a group of workers supposedly performing the same work tasks. This variation can be explained by differences in hazard awareness, attitude, and practice of safe working techniques in different workers.

Programmes for encouraging personal hygiene constitute another approach to reducing exposure. The provision and use of showers and a change to clean clothes at the end of the working day is a practical personal hygiene measure. A subtle but potentially important route of exposure is ingestion of toxic agents through eating or smoking, at the workplace. To prevent such exposure, separate facilities for consumption of food away from production lines should be provided.

Personal protective equipment

The use of personal protective devices to prevent occupational ill-health should generally be considered as a last resort after all other methods for minimizing exposure to workplace hazards have been tried. PPE can never be as efficient a means of protection as engineering or process controls. Nevertheless, the use of PPE is often widely practised, perhaps because it is seen as a relatively low-cost measure that puts much of the onus on the worker for compliance with use. PPE can be useful for short-term (as for emergency rescue work) or infrequent exposure to occupational hazards. Respirators, gloves, protective clothing, ear plugs, and muffs are all common forms of PPE in use throughout industry. They can play an important role, provided that carefully designed PPE programmes are in place and that the devices are frequently and regularly checked.

PPE should be properly selected to be effective against specific hazards, for example, the choice of an appropriate glove for use with a particular solvent. Workers have to be trained to use the equipment correctly and to ensure that it is working effectively, such as respirator fit testing in the use of respirators. Worker compliance in the use of these devices has to be high, or its protective effects may be less than desired. Compliance can be an issue, because of discomfort, especially for workers in hot and humid climates. PPE also have to be properly maintained and replaced when necessary.

Secondary prevention

Pre-employment examinations

An often stated aim of pre-employment examination is to detect any mismatch between the job applicant's health and the intended job. It seeks to determine whether an individual is fit to perform a specified job without risk to himself or others. A decision can then be made by management as to whether or not the person should be hired. Cost considerations, concerns about inability to carry out the full range of work tasks, and possible harm to third parties are often given as reasons for rejection of an applicant.

There is little empirical evidence to support the usefulness of pre-employment examinations (WHO 2009). In practice, the detection rate for relevant clinical abnormalities is low (2% in prospective HCW in the UK)

(Whitaker and Aw 1995), such that the cost-effectiveness of the process is questionable. The low preventive effectiveness of pre-employment examinations (i.e. the effort to prevent one adverse outcome) implies that the medical assessment should consist of no more questions and tests than are required relevant to the specified job (de Kort and van Dijk 1997). There is an element of selection bias since individuals who are in poor health with existing clinical abnormalities are unlikely to be seeking employment.

Pre-employment examinations probably constitute one of the most frequent procedures performed by occupational health departments especially in developing countries. In many developed countries, there are laws prohibiting discrimination on the basis of health status. Individuals are hired primarily on the basis of their credentials and ability before being put through a *pre-placement examination* (Aw and Koh 2013). Subsequent adjustments to the job tasks or work environment can then be made if indicated for health reasons.

Where they are performed, the pre-employment examination or the pre-placement assessment can be used to obtain baseline data on health status or fitness, to inform and educate new workers about potential hazards in their work and the preventive measures that can be taken to safeguard their health. Immunization against diseases that may possibly be contracted on the job, and for which effective vaccines are available (for example hepatitis B, measles, mumps, and rubella), should also be given. Some genetic disorders which can be identified at the pre-employment stage may enable suitable advice to be given in regard to special care in handling chemicals at work, for example, people with glucose-6-phosphate dehydrogenase deficiency are at risk of haemolytic anaemia on exposure to naphthalene and fava beans. People with serum total alpha-1-antitrypsin deficiency may be susceptible to smoke, fumes, and other respiratory irritants.

Statutory periodic medical examinations

The aim of statutory medical examinations is to prevent special groups of ‘at-risk’ workers from developing serious occupational diseases. These may include workers exposed to heavy metals, silica, asbestos, benzene, radiation, and noise. Regular health examinations, which are specific for the type of hazard the worker is exposed to, are conducted. Workers found to

have signs of overexposure to the hazard or with early signs of disease can be removed from further exposure. They can be given alternative work until they are fit to return to their former jobs. Furthermore, if signs of overexposure are detected, additional control measures can be taken to reduce the exposure at source and prevent other workers from being similarly affected.

Special groups of workers are required to undergo periodic medical examinations for protecting the health of the public, for example, public transport operators, pilots, and food handlers.

Notification of occupational diseases

Most countries require the statutory notification of occupational diseases to the regulatory bodies. This allows investigation, confirmation, and follow-up of cases by the authorities, and an analysis of trends over time. A single notified case may indicate poor control of work hazards that can pose a risk to other workers in the same work environment. The responsibility for notification has involved doctors, employers, or employee representatives, and varies between countries. Many countries produce a list of notifiable occupational diseases. This can be different from a list of prescribed diseases for which individuals may obtain financial compensation.

Notification serves as an additional means of control of occupational diseases. It should initiate a chain of events, starting from investigation and confirmation of the index case, and active case finding of other affected people. Recommendations for specific preventive measures at the workplace are then prescribed. The authorities would follow-up by ensuring that the recommendations have been implemented. If necessary, further evaluation of the effectiveness of the preventive measures can be made.

Tertiary prevention

Tertiary prevention involves mainly early treatment and rehabilitation. For occupational ill-health, affected workers should be removed from further exposure, and the appropriate medical treatment given as necessary. Examples of appropriate treatment include prompt first aid after an injury, chelation for severe cases of heavy metal poisoning, atropine and aldoximes for organophosphate poisoning, antidotes for cyanide poisoning, use of

calcium gluconate for hydrofluoric acid burns, and hyperbaric treatment for compressed air illness. Treatment would also include prompt provision of clinical supportive measures.

Planning for emergency response

Occupational health physicians play a central role in emergency preparedness and response. The American College of Occupational and Environmental Medicine identifies this domain as one of ten core competencies of the occupational health practitioner (ACOEM 2010). Occupational physicians are equipped with both clinical knowledge as well as an in-depth understanding of the work environment, processes, and workforce. They are therefore in a unique position to support, and sometimes lead, in emergency preparedness and business continuity planning and response.

Rehabilitation

Rehabilitation of workers is another important aspect of occupational health practice. Management, co-workers, occupational health professionals, and the injured worker have to work together to ensure that suitable alternative duties are provided, and that any work restrictions or physical limitations are understood. There should be clear short- and long-term goals in rehabilitation, and alternative duties should be meaningful and contribute to production.

Return to work

Return to work (RTW) is an important therapeutic goal for the injured worker. Studies have shown that when the medical management of the injured or ill worker includes attention to occupational issues with a view to early RTW, they tend to have faster recoveries and a more sustainable time on the job. In contrast, after more than 12 weeks off work, there is a 50% probability that the worker will still be away from work at the 1-year mark (ACOEM 2017a).

Planning for RTW should commence early in the course of treatment between the occupational physician and the primary physician (where applicable). Important considerations include the impact of the medical

condition on functioning and work ability, identification of obstacles to RTW (e.g. addressing misconceptions about illness and work), and early engagement of the employer in discussions on work adjustments and accommodations. Success in RTW often extends beyond managing physical functioning, to the complexities related to beliefs, roles, and perceptions of many stakeholders (e.g. patients, employers, co-workers) (MacEachen et al. 2006). Other important determinants to maximize positive outcomes in RTW programmes include securing employers' and organizational commitment, having dedicated and well-trained RTW coordinators, focusing the RTW plan on work participation rather than restrictions, and incorporating a plan which does not disadvantage co-workers (Skivington et al. 2016).

The assessment of work ability centres around three parameters, namely risk (chance of harm to the patient, co-workers, or the general public); capacity (refers to concepts such as strength, flexibility, and endurance); and tolerance (the ability to tolerate sustained work or activity at a given level) (Talmage et al. 2011). Of these, only capacity may be measurable to some degree of precision. Risk assessment is often dependent on logical deduction or expert consensus, whereas the ability to tolerate sustained work is a psychophysiologic concept which relies heavily on the motivation of, and incentivization available to the patient.

Workplace accommodations are also important to reduce barriers to RTW. Studies have shown that workplace interventions are important in reducing further sickness absence in injured workers who returned to work, and for improving pain and functional status among those with musculoskeletal disorders (van Vilsteren et al. 2015).

Workers' compensation

In many countries, workers who are injured at work or fall ill from hazardous work exposures are eligible for compensation. Workers' compensation is designed to provide income support, medical payments, and rehabilitation payments to workers injured on the job, as well as to provide benefits to the family and dependants of fatally injured workers. Most industrialized countries have workers' compensation programmes. In some countries, certain categories of workers, for example, domestic helpers are excluded. In many countries, employers have social insurance to give protection to employment injury victims. The principle of social

insurance is that of sharing of risks and pooling financial resources. A social insurance scheme establishes a public channel through a government department or government supervised body, which oversees procedures of screening, determination of award and payment of benefits.

Benefits are payable for temporary incapacity or permanent incapacity, and include survivors' benefits for those killed at work. Guidelines for assessment of disability are available in most countries. The final assessments for disability are made when the worker's medical condition has stabilized, and not likely to improve or deteriorate further.

Future challenges in occupational health

The practice of occupational health has evolved fairly rapidly over the last half a century. From the prevention and management of 'conventional' occupational diseases, occupational health burdens have shifted in tandem with global changes to employment patterns, demographic trends of the workforce, and the introduction of new technologies. This is more apparent in developed countries because of the relative decline in centuries-old occupational diseases with improvement in hazard reduction and control measures. The advent of information and communication technologies in the workplace, and its integration into the work ecosystem, fundamentally changes work patterns, duration, and intensity. With this, work-related stress and other mental disorders are replacing traditional occupational diseases as a leading cause of occupational morbidity in many parts of the world.

Advances in material science and engineering technology have enabled many occupational hazards to be eliminated, substituted with safer alternatives, or controlled through limited exposures to at-risk workers. However, the increase in life expectancies also means that workers are working longer years into elderly age, thereby bringing non-communicable diseases burden into the occupational fitness equation. It is likely that the demand for, and utilization of occupational health resources will rise over the years, given the ageing workforce and possibly the increasing complexity of cases (ACOEM 2017b). In the years ahead, the focus of management of occupational diseases and injuries must not only be restricted to clinical treatment alone. There is strong evidence to support

vocational rehabilitation, with the emphasis on early RTW as a therapeutic goal. This holistic and integrated occupational health model of care will likely define the standard of care in the future. In addition, occupational health has also evolved to consideration of health issues beyond the ‘factory gate’, as the discipline has much in common with environmental health and encompasses the philosophy and principles of good public health practice.

To address these contemporary and future challenges, a strategic change in occupational health practice will be one that will become more public health-centric (Harrison and Dawson 2016). From mainly addressing occupational hazard reduction and treating acute toxicities and injuries of workers, there will be a paradigm shift to the preventive care approach of the worker’s holistic health. This means that health interventions will not only be limited to occupational and work-related disorders, they will be extended to the primary prevention and management of chronic diseases under a Total Workplace Safety and Health framework. Occupational health practice of the present and future will therefore be centred not only on a safe workplace, but also on the concept of a healthy workplace. The promotion and maintenance of health and well-being may also be implemented at a broader community-level, with the knowledge that the worker is an integral part of a larger ecosystem and that health-seeking behaviours often achieve greater success when implemented under conditions where there is group participation and peer support.

Conclusion

The importance of occupational health lies in its emphasis on prevention, and in this respect it shares common objectives and strategies with public health. Occupational health focuses mainly on a specific group in the community— workers. Notwithstanding, there is convincing evidence that many of the public health issues of the future, such as medical needs associated with ageing populations and the increasing prevalence of non-communicable diseases, will also have a significant impact on the occupational health landscape. The socioeconomic imperatives of the Industrial Revolution had propelled the development of occupational medicine in the last two centuries. It is almost certain that changing needs of future economies and demographic shifts in the global workforce will

also have a great impact on shaping and defining the evolution of tomorrow's occupational health.

REFERENCES

Aje, O.O., Smith-Campbell, B., and Bett, C. (2018). Preventing musculoskeletal disorders in factory workers: evaluating a new eight minute stretching program. *Workplace Health Safety*, 66(7), 343–7.

Alfaraj, S.H., Al-Tawfiq, J.A., Altuwaijri, T.A., Alanazi, M., Alzahrani, N., and Memish, Z.A. (2018). Middle East respiratory syndrome coronavirus transmission among health care workers: implication for infection control. *American Journal of Infection Control*, 46(2), 165–8.

Alfonso, J.H., et al (2017). Minimum standards on prevention, diagnosis and treatment of occupational and work-related skin diseases in Europe—position paper of the COST Action StanDerm (TD 1206). *Journal of the European Academy of Dermatology and Venereology*, 31 Suppl 4, 31–43.

American College of Occupational and Environmental Medicine (2010). *Role of Occupational and Environmental Medicine Physicians in Emergency Preparedness and Response. Position Statement*. Elk Grove Village, IL: ACOEM. [Online] Available at: http://www.acoem.org/OEMPhysicians_EmergencyPreparedness.aspx

American College of Occupational and Environmental Medicine (2017a). *The Personal Physician's Role in Helping Patients with Medical Conditions Stay at Work or Return to Work. Position Statement*. Elk Grove Village, IL: ACOEM. [Online] Available at: http://www.acoem.org/uploadedFiles/Public_Affairs/Policies_And_Position_Statements/Guidelines/Position_Statements/The_Personal_Physician_s_Role_in_Helping_Patients.21.pdf

American College of Occupational and Environmental Medicine (2017b). *Global Trends in Occupational Medicine. Guidance*. Elk Grove Village, IL: ACOEM. [Online] Available at: http://www.acoem.org/uploadedFiles/Public_Affairs/Policies_And_Position_Statements/Guidelines/Guidelines/GlobalTrendsinOccupationalMedicine.pdf

American Conference of Governmental Industrial Hygienists (2017). *TLVs and BEIs. Threshold Limit Values for Chemical Substances and Physical Agents*. *Biological Exposure Indices*. Cincinnati, OH: ACGIH.

Anger, W.K., et al. (2015). Effectiveness of total worker health interventions. *Journal of Occupational Health Psychology*, 20(2), 226–47.

Aw, T.C. and Koh, D. (2013). Health screening. In: Palmer, K.T., Brown, I., Hobson, J. (eds.) *Fitness for work* (5th ed.). Oxford: Oxford University Press, pp. 572–84.

Baillargeon, M., Maheux, B., and Gilbert, A. (2011). The challenge of teaching occupational medicine to medical students: the Universite de Montreal experience. *Journal of Occupational and Environmental Medicine*, 53(11), 1258–61.

Barron, B.A., Beckett, W.S., and Utell, M.J. (2005). Clinical activities in an academic hospital-based occupational medicine program. *Journal of Occupational and Environmental Medicine*, 47(6), 587–93.

Baussano, I., Nunn, P., Williams, B., Pivetta, E., Bugiani, M., and Scano, F. (2011). Tuberculosis among health care workers. *Emerging Infectious Disease*, 17(3), 488–94.

Berecki-Gisolf, J., Smith, P.M., Collie, A., and McClure, R.J. (2016). Gender differences in occupational injury incidence. *American Journal of Industrial Medicine*, 58(3), 299–307.

Bosson-Rieutort, D., de Gaudemaris, R., and Bicout, D.J. (2018). The spectrosome of occupational health problems. *PLoS One*, 13(1), e0190196.

Burdorf, A., Brand, T., Jaddoe V.W., Hofman, A., Mackenbach, J.P., and Steegers, E.A. (2011). The effects of work-related maternal risk factors on time to pregnancy, preterm birth and birth weight: the Generation R Study. *Occupational and Environmental Medicine*, 68(3), 197–204.

Byler, C.G. and Robinson, W.C. (2018). Differences in patterns of mortality between foreign-born and native-born workers due to fatal occupational injury in the USA from 2003 to 2010. *Journal of Immigrant and Minority Health*, 20(1), 26–32.

Centers for Disease Control and Prevention (CDC) (2016). *Well-Being Concepts*. Atlanta: CDC. [Online] Available at: https://www.cdc.gov/hrqol/wellbeing.htm

Centers for Disease Control and Prevention (CDC) (2017). *Preventing Silicosis.* Atlanta: CDC. [Online] Available at: https://www.cdc.gov/features/preventing-silicosis/index.html

Collegium Ramazzini (2016). *19th Statement: A New Approach to the Control of Chronic Obstructive Pulmonary Disease (COPD)*. Bologna: Collegium Ramazzini.

Concha-Barrientos, M., et al. (2004). Selected occupational risks. In: Ezzati, M., Lopez, A., Rodgers, A., and Murray, C.J.L. (eds.) *Comparative Quantification of Health Risks: Global and Regional Burden of Disease Attributable to Selected Major Risk Factors*, pp. 1652–801. Geneva: WHO.

Coppee, G.H. (2011). Occupational health services and practice. In: Stellman J.M. (ed.) *Encyclopaedia of Occupational Health and Safety* (4th ed.). Geneva: ILO.

Costa, G. (2010). Shift work and health: current problems and preventive actions. *Safety and Health at Work*, 1(2), 112–23.

Dalal, K., Rahman, F., Gifford, M., and Rahman, A. (2016). The magnitude of injury problems among child labourers in a rural community of Bangladesh: findings from an injury surveillance system. *International Health*, 8(1), 73–6.

de Kort, W. and van Dijk, F. (1997). Preventive effectiveness of pre-employment medical assessments. *Occupational and Environmental Medicine*, 54(1), 1–6.

Driscoll, T., et al. (2014). The global burden of occupationally related low back pain: estimates from the Global Burden of Disease 2010 study. *Annals of Rheumatic Diseases*, 73(6), 975–81.

Elder, A. and Paterson, C. (2006). Sharps injuries in UK health care: a review of injury rates, viral transmission and potential efficacy of safety devices. *Occupational Medicine (London)*, 56(8), 566–74.

Elseviers M.M., Arias-Guillen, M., Gorke, A., and Arens, H.J. (2014). Sharps injuries amongst healthcare workers: review of incidence, transmissions and costs. *Journal of Renal Care*, 40(3), 150–6.

Equal Employment Opportunity Commission (2008). *Disability Discrimination*. Washington, DC: US EEOC. [Online] Available at: https://www.eeoc.gov/laws/types/disability.cfm.

Esswein, E.J., et al. (2004). Environmental and occupational health response to SARS, Taiwan, 2003. *Emerging Infect Dis*, 10(7), 1187–94.

European Agency for Safety and Health at Work (2008). *Occupational Skin Diseases and Dermal Exposure in the European Union (EU-25): Policy and Practice Overview*. Luxembourg: Office for Official Publications of the European Communities.

European Agency for Safety and Health at Work (2014). *Mainstreaming Gender into Occupational Safety and Health Practice*. Luxembourg: Publications Office of the European Union.

Feltner, C., et al. (2016). The effectiveness of Total Worker Health interventions: a systematic review for a National Institutes of Health Pathways to Prevention Workshop. *Annals of Internal Medicine*, 165(4), 262–9.

Forsmann, S. (1983). Occupational health. In: Parmeggiani, L. (ed.) *Encyclopaedia of Occupational Health and Safety* (3rd ed.). Geneva: ILO, pp. 1491–3.

Hamalainen, P., Takala, J., and Tan, B.K. (2017). *Global Estimates of Occupational Injuries and Work-Related Diseases.* Singapore: Workplace Safety and Health Institute. [Online] Available

at: https://www.wsh-institute.sg/files/wshi/upload/cms/file/Global%20Estimate%20of%20Occupational%20Injuries%20and%20Work-Related%20Illnesses%202017%20(24Nov17).pdf
Harber, P., et al. (2010). Occupational medicine practice: one specialty or three? *Journal of Occupational and Environmental Medicine*, 52(7), 672–9.
Harrison, J. and Dawson, L. (2016). Occupational health: meeting the challenges of the next 20 years. *Safety and Health at Work*, 7(2), 143–9.
Health and Safety Executive (HSE) (2015). *Health and Safety for Disabled People—The Law.* London: HSE. [Online] Available at: http://www.hse.gov.uk/disability/law.htm
Health and Safety Executive (HSE) (2017a). *Costs to Great Britain of Workplace Injuries and New Cases of Work-Related Ill Health—2015/2016.* London: HSE. [Online] Available at: http://www.hse.gov.uk/statistics/cost.htm
Health and Safety Executive (HSE) (2017b). *Non-Fatal Injuries at Work in Great Britain—2016/2017*. London: HSE. [Online] Available at: http://www.hse.gov.uk/statistics/causinj/index.htm
Health and Safety Executive (HSE) (2017c). *Work-Related Musculoskeletal Disorders—2016/2017*. London: HSE. [Online] Available at: http://www.hse.gov.uk/statistics/causdis/musculoskeletal/index.htm
Henneberger, P.K., et al (2011). An official American thoracic society statement: work-exacerbated asthma. *American Journal of Respiratory and Critical Care Medicine*, 184(3), 368–78.
Holmes, S.M. (2006). An ethnographic study of the social context of migrant health in the United States. *PLoS Medicine*, 3(10), e448.
Houdmont, J., Cox, T., and Griffiths, A. (2010). Work-related stress case definitions and prevalence rates in national surveys. *Occupational Medicine (London)*, 60(8), 658–61.
Institut de Recherché Robert-Sauve en Sante et en Securite du Travail (IRSST) (2013). *The Costs of Occupational Injuries*. Montreal: IRSST. [Online] Available at: https://www.irsst.qc.ca/media/documents/PubIRSST/R-787.pdf
International Agency for Research on Cancer (1997). *IARC Monographs on the Evaluation of Carcinogenic Risks to Humans: Silica, Some Silicates, Coal Dust and Para-Aramid Fibrils*. Lyon: IARC.
International Labour Organization (2011). *Conditions of Work and Employment Programme: Working Time in the Twenty-First Century*. Geneva: ILO.
International Labour Organization (2015). *ILO Global Estimates on Migrant Workers*. Geneva: ILO.
International Labour Organization (2016a). *Workplace Stress—A Collective Challenge*. Geneva: ILO.
International Labour Organization (2016b). *Women at Work—Trends 2016*. Geneva: ILO.
International Labour Organization (2018). *World Employment Social Outlook—Trends 2018*. Geneva: ILO.
International Programme on the Elimination of Child Labour (2017). *Global Estimates of Child Labour: Results and Trends, 2012–2016*. Geneva: ILO.
Jackson, C. (2007). The General Health Questionnaire. *Occupational Medicine*, 57, 79.
Jahncke, H., Hygge, S., Mathiassen, S.E., Hallman, D., Mixter, S., and Lyskov, E. (2017). Variation at work: alternations between physically and mentally demanding tasks in blue-collar occupations. *Ergonomics*, 60(9), 1218–27.
Jensen, T.K., Bonde, J.P., and Joffe, M. (2006). The influence of occupational exposure on male reproductive function. *Occupational Medicine*, 56(8), 544–53.
Koh, D. and Koh, C.H.G. (2007). The use of salivary biomarkers in occupational and environmental medicine. *Occupational and Environmental Medicine*, 64(3), 202–10.

Koh, D., Lim, M.K., and Chia, S.E. (2003). SARS: health care work can be hazardous to health. *Occupational Medicine*, 53(4), 241–3.

Kuehl, K.S., Elliot, D.L., Goldberg, L., Moe, E.L., Perrier, E., and Smith, J. (2013). Economic benefit of the PHLAME wellness programme on firefighter injury. *Occupational Medicine (London)*, 63(3), 203–9.

Kumar, S. (2004). Occupational exposure associated with reproductive dysfunction. *Journal of Occupational Health*, 46, 1–19.

Lallukka, T., et al. (2017). Early work-related physical exposures and low back pain in midlife: the Cardiovascular Risk in Young Finns Study. *Occupational and Environmental Medicine*, 74(3), 163–8.

Leigh, J.P. (2011). Economic burden of occupational injury and illness in the United States. *Milbank Q*, 89(4), 728–72.

Li, C.Y. and Sung, F.C. (1999). A review of the healthy worker effect in occupational epidemiology. *Occupational Medicine (London)*, 49(4), 225–9.

Lindsay, S., Cagliostro, E., Albarico, M., Mortaji, N., Karon, L (2018). A systematic review of the benefits of hiring people with disabilities. *Journal of Occupational Rehabilitation*, 28(4), 634–55.

Lopez, A.D. and Mathers, C.D. (2006). Measuring the global burden of disease and epidemiological transitions: 2002–2030. *Annals of Tropical Medicine & Parasitology*, 100(5–6), 481–99.

MacEachen, E., Clarke, J., Franche, R.L., Irvin, E., and Workplace-based Return to Work Literature Review Group (2006). Systematic review of the qualitative literature on return to work after injury. *Scandinavian Journal of Work, Environment & Health*, 32(4), 257–69.

Mahler, V., et al. (2017). Occupational skin diseases: actual state analysis of patient management pathways in 28 European countries. *Journal of the European Academy of Dermatology and Venereology*, 31 Suppl 4, 12–30.

Malo, J.L., et al. (2015). An official American Thoracic Society Workshop Report: presentations and discussion of the fifth Jack Pepys Workshop on Asthma in the Workplace. Comparisons between asthma in the workplace and non-work-related asthma. *Annals of the American Thoracic Society*, 12(7), S99–S110.

Mason, H.J., et al. (1999). Follow-up of workers previously exposed to silver solder containing cadmium. *Occupational and Environmental Medicine*, 56, 553–8.

Mohindra, K.S. (2016). Promoting women's health in an era of globalisation: a South Asian perspective. *Global Health Promotion*, 25(2), 90–4.

Moyce, S.C. and Schenker, M. (2018). Migrant workers and their occupational health and safety. *Annual Reviews of Public Health*, 9, 351–65.

National Institute for Occupational Safety and Health (1999). *Stress at Work.* DHSS (NIOSH) Publication No. 99–101. Cincinnati, OH: NIOSH.

National Institute for Occupational Safety and Health (2008). *NIOSH Generic Job Stress Questionnaire (Organization of Work: Measurement Tools for Research and Practice)*. Cincinnati, OH: NIOSH. [Online] Available at: http://www.cdc.gov/niosh/.

National Institute for Occupational Safety and Health (2012). *Research Compendium: The NISOH Total Worker Health® Program: Seminal Research Papers*. Publication No. 2012–146. Cincinnati, OH: NIOSH.

National Institute for Occupational Safety and Health (2017). *Total Worker Health*. Cincinnati, OH: NIOSH. [Online] Available at: https://www.cdc.gov/niosh/twh/totalhealth.html

Nelson, D.I., Nelson, R.Y., Concha-Barrientos, M., and Fingerhut, M. (2005). The global burden of occupational noise-induced hearing loss. *American Journal of Industrial Medicine*, 48(6), 446–58.

Niu, S., Deboodt, P., and Zeeb H. (2010). *Approaches to Attribution of Detrimental Health Effects to Occupational Ionizing Radiation Exposure and Their Application in Compensation Programmes for Cancer*. *Occupational Safety and Health Series*, vol. 73. Geneva: WHO.

Occupational Safety and Health Administration (2014). *Prevention of Work-Related Musculoskeletal Disorders*. Washington, DC: US DOL. [Online] Available at: https://www.osha.gov/pls/oshaweb/owadisp.show_document?p_id=4481&p_table=UNIFIED_AGENDA

Odgerel, C.O., et al. (2017). Estimation of the global burden of mesothelioma deaths from incomplete national mortality data. *Occupational and Environmental Medicine*, 74(12), 851–8.

Podniece, Z., Heuvel, S. van den, and Blatter, B. (2008). *Work-Related Musculoskeletal Disorders: Prevention Report*. Bilbao: European Agency for Safety and Health at Work.

Pruss-Ustun, A., Rapiti, E., and Hutin, Y. (2005). Estimation of the global burden of disease attributable to contaminated sharps injuries among health-care workers. *American Journal of Industrial Medicine*, 38, 482–90.

Punnett, L. (2014). Musculoskeletal disorders and occupational exposures: how should we judge the evidence concerning the causal association? *Scandinavian Journal of Public Health*, 42(13 Suppl), 49–58.

Ramazzini, B. (1964). *De Moribs Artificum Diatriba* (Cave Wright, W., trans.). New York: Hafner Publishing Co. (Work originally published in 1713.)

Rushton, L, et al. (2012). Occupational cancer burden in Great Britain. *British Journal of Cancer*, 107(Suppl 1), S3–7.

Schulte, P.A. and Hauser, J.E. (2012). The use of biomarkers in occupational health research, practice, and policy. *Toxicology Letters*, 213(1), 91–9.

Schulte, P.A., et al. (2016). Taking stock of the occupational safety and health challenges of nanotechnology: 2000–2015. *Journal of Nanoparticle Research*, 18, 159.

Selander, J., Albin, M., Rosenhall, U., Rylander, L., Lewne, M., and Gustavsson, P. (2016). Maternal occupational exposure to noise during pregnancy and hearing dysfunction in children: a nationwide prospective cohort study in Sweden. *Environmental Health Perspectives*, 124(6), 855–60.

Sepkowitz, K.A. and Eisenberg, L. (2005). Occupational deaths among healthcare workers. *Emerging Infectious Diseases*, 11(7), 1003–8.

Skivington, K., Lifshen, M., and Mustard, C. (2016). Implementing a collaborative return-to-work program: lessons from a qualitative study in a large Canadian healthcare organization. *Work*, 55(3), 613–24.

Somayeh, D., et al. (2017). Occupational disease registries—characteristics and experiences. *Acta Informatica Medica*, 25(2), 136–40.

Steenland, K. and Armstrong, B. (2006). An overview of methods for calculating the burden of disease due to specific risk factors. *Epidemiology*, 17(5), 512–19.

Takala, J. (2003). ILO and asbestos. In: *Asbestos Conference*, Dresden, 3 September. [Online] Available at: http://www.ilo.org/safework

Takala, J., et al. (2014). Global estimates of the burden of injury and illness at work in 2012. *Journal of Occupational and Environmental Hygiene*, 11(5), 326–37.

Talmage, J.B., Melhorn, J.M., and Hyman, M. (2011). How to think about work ability and work restrictions: risk, capacity, and tolerance. In: Talmage, J.B., Melhorn, J.M., and Hyman, M. (eds.) *AMA Guides to the Evaluation of Work Ability and Return to Work* (2nd ed.). Chicago, IL: AMA.

Tikka, C., Verbeek, J.H., Kateman E., Morata, T.C., Dreschler, W.A., and Ferrite, S. (2017). Interventions to prevent occupational noise-induced hearing loss. *Cochrane Database of Systematic Reviews*, 7, CD006396.

Toren, K., Vikgren, J., Olin, A.C., Rosengren, A., Bergstrom, G., and Brandberg, J. (2017). Occupational exposure to vapour, gas, dust, or fumes and chronic airflow limitation, COPD, and emphysema: the Swedish CArdioPulmonary BioImage Study (SCAPIS pilot). *International Journal of Chronic Obstructive Pulmonary Disease*, 12, 3407–13.

Torquati, L., Mielke, G.I., Brown, W.J., and Kolbe-Alexander, T. (2018). Shift work and the risk of cardiovascular disease. A systematic review and meta-analysis including dose-response relationship. *Scandinavian Journal of Work and Environmental Health*, 44(3), 229–38.

van Vilsteren, M., van Oostrom, S.H., de Vet, H.C., Franche, R.L., Boot, C.R., and Anema, J.R. (2015). Workplace interventions to prevent work disability in workers on sick leave. *Cochrane Database of Systematic Reviews*, 10, CD006955.

Whitaker, S. and Aw, T.C. (1995) Audit of pre-employment assessments by occupational health departments in the National Health Service. *Occupational Medicine*, 45(2), 75–80.

Workplace Safety and Health Council (2017). *Total Workplace Safety and Health*. Singapore: WSHC.

World Bank Group (2017). *Labour Force, Total*. Washington, DC: World Bank Group. [Online] Available at: https://data.worldbank.org/indicator/SL.TLF.TOTL.IN/

World Health Organization (WHO) (2007). *Workers' Health: Global Plan of Action*. Geneva: WHO.

World Health Organization (WHO) (2009). Evidence base for pre-employment medical screening. *Bulletin of the World Health Organization*, 87, 529–34.

World Health Organization (WHO) (2011). *5 Keys to Healthy Workplaces: WHO Global Model for Action*. Geneva: WHO.

World Health Organization (WHO) (2017a). *Asbestos: Elimination of Asbestos-Related Diseases*. Geneva: WHO.

World Health Organization (WHO) (2017b). *Protecting Workers' Health*. Geneva: WHO.

Yamauchi, T., et al. (2017). Overwork-related disorders in Japan: recent trends and development of a national policy to promote preventive measures. *Industrial Health*, 55(3), 293–302.

7.4

Toxicology and environmental risk analysis

David Koh and Ro-Ting Lin

Introduction to toxicology and risk analysis

Toxicology is the study of poisons. In theory, any substance can cause harm to health. To paraphrase the physician-alchemist Paracelsus, ‘It is only the dose that distinguishes a poison from a remedy’. To that we can add that the route of entry into the body can also determine the extent of injury or ill health. In practice, most poisons are either chemicals or biological materials that cause severe effects in relatively small doses. Poisonous substances that contaminate the environment can lead to an increased risk of ill health in exposed populations. The level of risk is a function of the extent of exposure and the susceptibility of those exposed. An understanding of toxicology is essential to determining and managing environmental risk.

Risk analysis has been defined as ‘a process consisting of three components: risk assessment, risk management, and risk communication’ (Food and Agriculture Organization and World Health Organization 2009). Environmental risk analysis is a broad field, encompassing risks to ecosystems and materials as well as to human health. This chapter discusses human health risks only. The goal is to synthesize toxicology and risk analysis as a basis for evaluating risks to human health posed by hazardous agents in the environment. Other disciplines, such as epidemiology and exposure assessment, are also required to better understand risk, and for many specific agents these disciplines will provide the primary information needed for risk assessment, risk management, and risk communication.

Toxicology has these two important roles in environmental risk analysis:

To ascertain cause and effect relationships linking hazardous agents to adverse effects in humans or the general environment.
To develop techniques capable of preventing these problems.

Toxicologists usually approach questions of disease causation by starting with the chemical or physical agent and studying its effects in laboratory animals or in test-tube systems. Modern toxicology also develops tools, primarily through molecular biology and advanced computational techniques, capable of probing the extent to which a given disease in an individual is caused by an external agent. This reverse approach, starting with disease and moving toward determining cause, is enabled by the increasing ability of epidemiology to link subtle biological markers indicative of early effects to biological markers indicative of exposure.

Toxicology is also an important discipline in the primary and secondary prevention of human health effects. When toxicologists understand the mechanisms by which chemical or physical agents cause biological effects, they can develop tests to help prevent further spread of harmful chemicals or promote early detection of potential adverse effects.

General concepts of toxicology relevant to risk analysis

Knowledge about poisons extends back to the beginning of history, as humans became aware of the toxicity of natural food components. Toxicologists credit Paracelsus with the first law of toxicology, that the dose makes the poison. Two other laws underlie modern toxicology: that chemicals have specific biological effects, a maxim credited to Ambrose Pare (Goldstein and Gallo 2001); and that humans are members of the animal kingdom.

The ‘laws’ of toxicology

The following section discusses the ‘laws’ and general concepts of toxicology pertinent to understanding how a chemical or physical agent acts in a biological system. The focus here is on the biological response to an agent, rather than on the intrinsic properties of the agent. These are the three ‘laws’ of toxicology:

The dose makes the poison.
Chemical/physical agents have specific effects.
Humans are animals.

The first law: the dose makes the poison

Exploration of the relation between dose and response is central to toxicology. There are two main types of dose–response curves (Fig. 7.4.1). One is an S-shaped curve characterized by having no observed effect at the lowest doses, gradual development of an increasing response as the dose increases followed by a linear phase of increase in relation to dose, and, eventually, a dose level at which no further increase in response is observed. Of particular pertinence to environmental toxicology is that this curve presumes that there exists a threshold level below which no harm whatsoever is to be expected. There is ample scientific basis for the existence of thresholds for specific effects. For example, one drop of undiluted sulphuric acid splashed on the skin is capable of producing a severe burn. Yet one drop of pure sulphuric acid in a bathtub of water is sufficiently dilute to be without effect. Thresholds for an adverse effect will differ among individuals based upon a variety of circumstances, some genetically determined and others representing stages of life or other specific circumstances. In the example of sulphuric acid on the skin, there are genetically determined differences in susceptibility related to the protective presence of skin and hair. Babies will be more susceptible than adults; and skin that is already damaged will be at particular risk. The S-shaped dose–response curve is assumed to fit all toxic effects except those that are produced by direct reaction with genetic material.

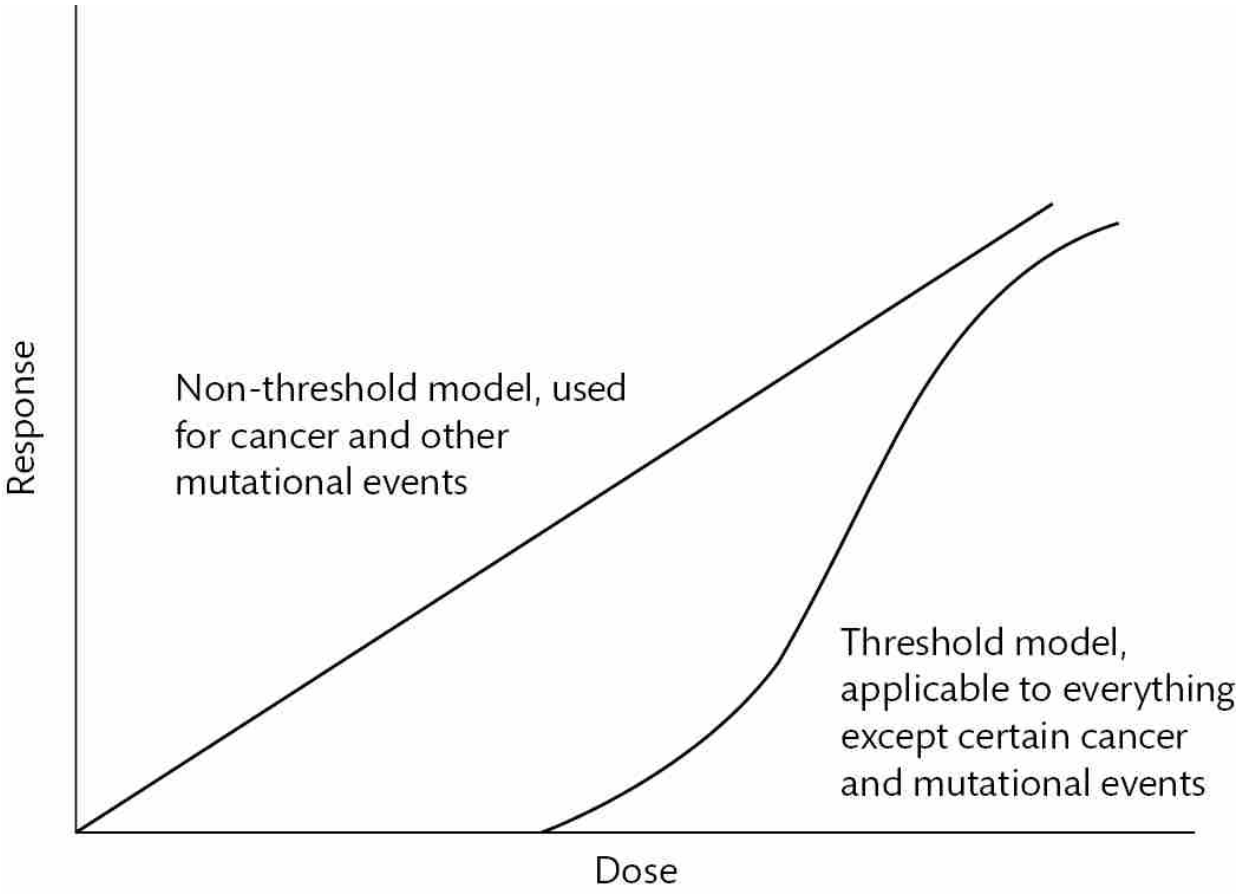


Fig. 7.4.1 Dose–response curve.

The second general type of dose–response curve covers endpoints caused by persistent changes in the genes. For example, in cancer, a somatic mutation occurring in a single cell results in a clone of cancer cell progeny. Another example is inherited mutations of the genetic components of cells involved in reproduction. It is believed that a single change in DNA can alter the genetic code in a way that can lead to a mutated cell. It follows that any single molecule of a carcinogenic chemical or packet of physical energy, such as ionizing radiation, that can alter DNA is theoretically capable of causing a persistent mutation. This presumption—that a single molecule or ionizing ray could change a normal cell to a cancerous cell—implies that there is no threshold below which the risk is zero. The resultant dose–response curve thus starts at a single molecule. As a further simplification, the shape of the curve can be linear in that the risk of two molecules of a DNA-altering chemical causing a mutation is conceivably twice that of one molecule, and so on, until a dose level results in dead cells.

Features of both S-shaped and linear dose–response curves occur in the human body's response to allergens. It is likely that a relatively large dose of an allergen is required to first sensitize an individual (i.e. the threshold dose for sensitization may vary with the agent and the individual). However, once sensitized, the affected individual may react to even molecules of the causative agent, with no apparent threshold for elicitation of a reaction.

There are also a few chemicals that have a U-shaped dose–response curve, where high levels of exposure cause harm and insufficient exposure leads to a deficiency which results in ill health. Examples are fat-soluble vitamins and essential trace minerals. Excess vitamin D intake causes hypervitaminosis D, and vitamin D deficiency leads to rickets and osteomalacia. There is, therefore, an ideal range for vitamin D levels which results neither in deficiency nor poisoning.

The second law: chemicals have specific effects

Understanding that chemical and physical agents have specific effects is in essence no different from recognizing that possession of a gun does not make one a murder suspect if the victim was stabbed to death. The law of specificity is well understood by the general public in terms of drugs: aspirin will help with your headache but is useless for constipation, while laxatives have the opposite effect. However, various surveys suggest that the selectivity of the effects of environmental chemicals is not widely understood and that many believe a chemical that can cause cancer in a particular organ can cause cancer and other diseases anywhere in the body. Some chemicals may cause malignancy in more than one body site (e.g. asbestos can cause lung cancer, peritoneal mesothelioma, and laryngeal cancer), but in general, chemical and physical agents are highly specific in their effects.

The specificity of effects is due to both chemistry and biology. Understanding the relationship between chemical structure and biological effect has been central to both pharmacology and toxicology. Structure activity relationship (SAR) is often used as a means to design a chemical with a specific effect that might be useful as a therapeutic agent. SAR is also used to predict whether a new chemical being readied for manufacture might be of potential harm. While SAR is a useful tool being further improved through modern computational approaches, its predictive value remains too

limited for use with a potentially toxic agent without recourse to additional testing. For example, only one simple methyl group separates toluene from benzene (only the latter known to cause bone marrow damage and leukaemia); ethanol from methanol (the latter causing metabolic acidosis, renal failure, and blindness); and *n*-hexane from either *n*-heptane or *n*-pentane (only *n*-hexane is responsible for peripheral nerve damage). These examples of specificity reflect both the formation of toxic metabolites, such as active species derived from the metabolism of benzene, and the interaction of a chemical or its metabolite within specific biological niches, such as the diketone metabolite (2,5-hexanedione) of *n*-hexane within neuronal axons.

Specificity of effects is also conferred by cellular processes that cause certain cells to be more susceptible to environmental agents. For example, red blood cells have the iron-containing protein haemoglobin that is responsible for the transport of oxygen. Toxicity through alteration of efficient oxygen delivery occurs through certain specific mechanisms. One is through the oxidation of the reduced ferrous form to the ferric form of iron, known as methaemoglobin, which can no longer carry oxygen. This occurs with a limited number of agents (e.g. nitrates and phenols) that, once within the red blood cell, are capable of oxidizing intracellular iron. Another specific mechanism of interference is exemplified by carbon monoxide. This otherwise relatively inert gas has a physical chemistry that sufficiently resembles oxygen so that it is able to tightly combine with the oxygen-combining site of haemoglobin, thereby displacing oxygen. There are many other examples in which a normal body process is disrupted by an exogenous chemical through a specific chemical alteration, such as oxidation or covalent addition, or by fitting into a niche designed through evolution to accommodate a necessary internal chemical which it superficially resembles.

The third law: humans are animals

The conceptual foundation for extrapolating from animals to humans is central to modern toxicology. The basic principles of cell function are common to all of biology: all cells must obtain energy, build structure, and release waste. Cell function in complex organisms, such as humans, is highly specialized, but there is still great similarity in cellular and organ function among mammals, which facilitates extrapolation from one species to

another. In general, a kidney poison in one species is likely to be a kidney poison in another, although there are certainly exceptions. As an example, arsenic is a human carcinogen with a lack of evidence for carcinogenicity among other animals.

However, despite similarities in effects, dose–response considerations often vary substantially between species, reflecting differences in absorption, distribution, metabolism, excretion, function, and target organ susceptibility. Understanding the factors responsible for interspecies differences greatly facilitates extrapolation from animals to humans. Once elucidated, the role of different absorption rates, metabolism, or other factors can be taken into account, often through a mathematical approximation that has come to be called physiologically based pharmacokinetics (or toxicokinetics).

The use of alternative methods to replace animal testing is becoming popular. For example, under the 2006 Registration, Evaluation, Authorisation, and Restriction of Chemicals (REACH) regulation in the European Union (EU), companies that manufacture or import chemical substances in quantities beyond a certain volume are required to submit reports of testing on vertebrate animals. To avoid unnecessary animal tests, the European Chemicals Agency (ECHA) introduced the concept of alternative methods such as read-across, SAR, weight of evidence, and *in vitro* methods (European Chemicals Agency 2017).

Pathways of exogenous agents within the body

Absorption, distribution, metabolism, and excretion of exogenous agents can vary greatly across individuals and within the same individual depending upon, for example, stage of life and state of health. These variations are among the major reasons for individual differences in susceptibility to risks from exposure to toxic agents. Increased understanding of how the human genotype and phenotype affect absorption, distribution, metabolism, and excretion of exogenous agents, including foods, is providing insight into the oldest human question about disease: ‘Why is this happening to me?’ (Omenn 2000).

Absorption

Absorption of a chemical into the body occurs through ingestion, inhalation, and skin absorption. Depending upon the specific chemical, the route of exposure can have major implications for the extent of absorption and the resultant toxicity. For example, almost 100% of inhaled lead-containing fine particles are absorbed into the body, while a much smaller percentage of ingested lead is absorbed. Internal factors also affect absorption, particularly in the gastrointestinal tract. In the case of lead absorption, both iron and calcium deficiencies increase the absorption of ingested lead. The matrix of the exposure agent also may have an effect. For example, the rate at which benzene in gasoline is absorbed through the skin is likely increased by oxygenated components of the gasoline mixture, and the absorption of dioxins from contaminated soil can vary enormously (Umbreit et al. 1986). The American Conference of Governmental Industrial Hygienists provides a 'skin' notation for chemicals that can be absorbed through the skin to cause systemic effects following exposure (American Conference of Governmental Industrial Hygienists 2013).

Often, a single route of absorption is dominant. But in many instances, more than one route may be important. For example, exposure to chlorinated disinfection products in drinking water systems, or to gasoline contamination of well water through a leaky underground storage tank, is usually thought of solely in terms of the ingestion of water. However, during showering, there is likely to be both inhalation and transdermal absorption, and if groundwater is contaminated, there can be off-gassing from the soil into the home. Epidemiological studies of the potential adverse consequences of water contamination need to take all of these exposure routes into account (Arbuckle et al. 2002).

Less common routes of absorption are intravenous, intramuscular, intraperitoneal, and sublingual, or via suppositories. These occur mainly in animal experiments or in human drug delivery. A novel route of absorption is the translocation of nanoparticles deposited in the nasal area via the olfactory nerves directly into the brain (Oberdorster et al. 2004). For example, inhaled airborne magnetite particulate matter (PM) has been found to cross the damaged olfactory unit, pass through the axons of the olfactory nerve, and enter the brain, which can potentially be related to neurodegenerative diseases such as Alzheimer's (Maher et al. 2016).

Distribution

Once inside the body, chemicals are transported by the blood. Distribution is determined by factors including the route of absorption, the chemical and physical properties of the substance, and the barriers that compartmentalize the body.

The route of absorption affects the different pathways through which chemical substances are distributed. Most compounds absorbed in the gastrointestinal tract go directly to the liver and may go no further. The liver has great capacity to detoxify many absorbed chemicals, especially when the chemical is absorbed by the digestive system and enters the hepatic portal system into the liver before it reaches the rest of the body. Enzymes in the gastrointestinal bacteria, lumen, gut wall, and liver contribute to this first-pass effect, which may greatly reduce the level at which a particular agent remains bioavailable in systemic circulation. On the other hand, inhaled agents go first to the lungs or other parts of the respiratory tract and then to general circulation, without being filtered through this ‘first-pass’ effect.

The specific chemical and physical properties of a substance and their affinity for certain organs or tissues also affect distribution of a chemical substance. Small inhaled particles tend to be distributed deep within the respiratory tract, while larger particles are filtered out in the nose or are trapped and cleared by the bronchiolar mucociliary escalator. Chemicals that are poorly soluble in water, such as oils, usually distribute within fatty tissues. Organ-specific factors are also important. For example, a specific pump located in the thyroid gland facilitates uptake of iodine and makes the thyroid particularly vulnerable to the adverse impact of radioactive iodine.

Some tissues, such as the blood-brain barrier and the placental barrier, limit the distribution of chemical substances. The blood-brain barrier, with its capillary endothelial cells, can protect the central nervous system. Only fat-soluble substances, such as alcohols, can cross the blood-brain barrier. The placental barrier prevents many toxicants from reaching the foetus but allows oxygen and nutrients to pass from mother to foetus (Burton et al. 2016).

Metabolism

Metabolism, in the narrowest sense of the term, refers to alteration of chemicals by the body. The major metabolic function of the body is to alter food into energy or structural materials. Metabolism of xenobiotics is often

protective, converting unwanted absorbed materials into chemical forms that are readily excretable. Thus, a fat-soluble agent can often be converted into water-soluble agents capable of being excreted in the urine. However, for certain classes of chemicals, metabolism results in conversion of relatively inactive compounds into harmful agents. For example, some carcinogens, including polycyclic aromatic hydrocarbon components of soot and benzene, require metabolic activation to change from pro-carcinogens to ultimate carcinogens.

All organs appear to have metabolic capability, often related both to organ function and to susceptibility to toxic agents. Understanding the specific enzymes and enzyme families responsible for metabolism within cell types is important to understanding why chemicals have specific effects in specific organs.

In the case of inhaled benzene, about 50% of the body burden is exhaled unchanged as the parent compound, and the rest is metabolized to potentially toxic metabolites. Slowing down benzene metabolism leads to an increase in the relative amount that is exhaled rather than metabolized, and thus a decrease in bone marrow toxicity. Individuals with what appear to be genetically determined increases in benzene metabolism to toxic metabolites, or decreases in the detoxification of these metabolites experience increased haematological risk. The effects of having both polymorphisms together appear to be at least additive and perhaps multiplicative in increasing the risk (Rothman et al. 1997; Kim et al. 2007). The application of genomics and proteomics to metabolism is often known as metabolomics.

Excretion

Excretion of chemicals from the body occurs through a variety of routes, including the gastrointestinal tract for unabsorbed compounds and for compounds dissolved in bile, and the renal tract (e.g. in urine) for water-soluble agents of appropriate molecular weight and charge. Significant excretion of volatile compounds can occur through the respiratory tract. Other routes of excretion include sweating and lactation, the latter which may put the breastfed infant at risk.

Multiple chemical interactions

Toxicological studies of absorption, distribution, metabolism, and excretion are often performed for single chemicals. Yet in real life, multiple and combined exposures are the norm rather than the exception. For such situations, the possibility of additive or synergistic effects, or potentiation or antagonism of effects, has to be considered. For example, exposure to two organophosphate insecticides may result in a simple additive adverse effect. In the case of exposure to carbon tetrachloride and ethanol, or asbestos exposure and cigarette smoking, the combined effects are synergistic. Antagonistic interactions take various forms. Functional antagonism is exemplified by the opposing functional effects of combined exposure to barbiturate and norepinephrine. An example of chemical antagonism is adding BAL (British anti-Lewisite) to inactivate metal ions through chelation. Dispositional antagonism is made use of when treating poisoning with charcoal, or when using an emetic or a diuretic. Receptor antagonism is seen in the case of naloxone as treatment for morphine overdose.

Risk assessment

Risk assessment follows two separate streams of toxicological reasoning: (1) for toxic agents implicitly or explicitly assumed to have a threshold, and (2) for carcinogens. Chemical safety assessment developed from simplified approaches such as studies on laboratory animals to determine the dose (LD_{50}, expressed usually as mg/kg body weight) or the concentration (LC_{50}, expressed as parts per million of the substance in air or water or other medium) capable of killing 50% of the test animals. The observed LD_{50} or LC_{50} was used as a basis for extrapolating to permissible levels in humans, often using several tenfold safety factors. Such protective factors were based on the concern that humans as a species could be more sensitive than the laboratory animals; that there would be greater variability in sensitivity among humans than among genetically inbred laboratory animals all raised in a similar environment; and that there could be adverse non-lethal effects that should be avoided. A presumptive tenfold safety factor has been added specifically to protect children, in recognition of their greater risk to certain chemicals compared to adults (National Research Council 1993).

The inherent assumption in the safety factor approach is that there is a threshold dose level below which there are no adverse effects. However, in

carcinogenesis, a single mutation could be the basis for the entire cancer process. As each molecule of a carcinogen, at least theoretically, could cause this mutation, a threshold cannot be assumed, and thus, in a simplified expression, there is no level of exposure that is without risk. However, almost all DNA damage is repaired by efficient cellular processes. Some unrepaired mutations are lethal to the cell, and as dead cells do not reproduce, these mutations cannot be the basis for cancer or inherited abnormalities. The majority of mutations are silent and have no discernible effects. Accordingly, the risk of any one molecule actually causing cancer is infinitesimally small—literally trillions of molecules of carcinogens are inhaled with every cigarette, yet not all cigarette smokers develop cancer. However, the assumption that there is no zero risk in regards to chemical and physical carcinogens has a major impact on communicating to the public about cancer risk.

There are known circumstances in which cancer causation does depend upon exceeding a threshold level of a chemical (e.g. the mechanism by which saccharin causes bladder cancer in laboratory animals appears to proceed through the precipitation of saccharin in the bladder, which requires a dose sufficient to exceed the physicochemical processes determining saccharin solubility). However, prudent management of cancer risk usually assumes that a carcinogen is ‘guilty until proven innocent’ of having no risk-free level. In essence, the burden of proof is on industry to demonstrate scientifically that a cancer-causing chemical does indeed have a threshold.

The four major components of risk assessment are listed in Box 7.4.1.

Box 7.4.1 Components of risk assessment

Hazard identification: Determining whether a specific chemical or physical agent is causally linked to a specific endpoint of concern; that is, chemicals have specific effects, or the second law of toxicology.

Dose–response evaluation: Determining the relation between the magnitude of exposure and the probability of occurrence of the specific endpoint of concern; that is, the dose makes the poison, or the first law of toxicology.

Exposure evaluation: Determining who and how many people will be exposed; the routes of exposure; and the magnitude, duration, and timing of exposure.

Risk characterization: Describing the nature and often the magnitude of the human risk, including attendant uncertainty.

Hazard identification

Hazard is the potential for harm. It is an intrinsic property of a substance or situation. Risk is the likelihood of harm occurring (and the likely severity of the resulting effect). It is a function of both hazard and exposure: for example, asbestos fibres sealed in a ceiling of an office have the potential to cause health effects (hazard), but the likelihood of that occurring is low if the ceiling is left intact (risk). However, if the ceiling is demolished and dust containing the asbestos fibres is released, the risk to health can be considerable. The risk has changed because of the circumstances of exposure. Identifying the presence of a hazard is just one step in risk assessment; subsequent steps, such as exposure evaluation, are key to determining what preventive action needs to be taken to reduce the risk.

Regulatory and quasi-scientific bodies often use a weight-of-evidence approach to identify a hazard. In essence, a panel of scientists judges whether sufficient evidence exists to identify an agent or condition as posing a risk of a specific effect in humans, or in some other target such as an ecosystem. In the United States, the approach to permitting the marketing of a new chemical is to have an internal Environmental Protection Agency scientific group review the chemical structure and other data submitted by industry. Similarly, the US Food and Drug Administration relies heavily on an advisory committee process to review evidence of efficacy and toxicity before approving a new pharmaceutical agent or medical device.

Formal weight-of-evidence approaches have been particularly useful in evaluating potential human carcinogens. The International Agency for Research on Cancer (IARC) of the World Health Organization uses this approach, convening expert panels to evaluate the evidence for carcinogenicity of specific chemical compounds or defined mixtures (e.g. diesel fuel, wood dust). The focus is on the weight of the evidence for carcinogenicity based upon carefully framed criteria—considering animal toxicology, epidemiology, mechanistic information, and exposure data—but not on the potency of the compound as a cancer-causing agent. When assigning a score for carcinogenicity, IARC now includes increased weight on understanding toxicological mechanisms (Cogliano 2004). Many countries use the information produced by IARC to determine governmental regulatory approaches. In the United States, the IARC listing of carcinogens has no official status, but it carries much weight with regulators, as does a

similar process used by the National Toxicology Program for its Report on Carcinogens (National Toxicology Program 2015).

Relatively few chemicals have been reported capable of causing cancer in humans. Of the perhaps 70,000 to 100,000 chemicals in commerce, about 100 are known to be human carcinogens. To an extent, this represents the success of environmental health science in providing tools that guide chemical manufacturers away from new chemicals that are potentially carcinogenic. During the chemical development process, early application of simple test batteries that evaluate the potential for mutagenesis or other predictors of cancer causation provides a responsible chemical industry with the means to avoid producing carcinogens or other potentially harmful products—and the means to avoid the regulatory and toxic tort consequences of harming the public. The value of this primary preventive approach depends upon the availability of effective toxicological test batteries. Such tests depend upon a basic understanding of the chemical and biological processes underlying toxic effects. Unfortunately, the investment in using standardized test batteries for high-production-volume chemicals, and the major increase in such investments due to new requirements such as the EU REACH legislation (see section 5.5), has not been accompanied by recognition of the need to develop better and more effective tests to protect the public. Advances in molecular toxicology provide many opportunities to improve these test batteries (National Research Council 2007).

Dose–response evaluation

The key issues in dose–response evaluation involve how to extrapolate from the high doses at which an effect is observed in an animal or epidemiological study to the usually much lower levels of risk which are of public or policy concern. Crucial to extrapolation are assumptions about the shape of the dose–response curve). It must be emphasized that the levels of risk desired by society (e.g. in the range of less than 1 in 10,000 to less than 1 in a million over a lifetime) are usually too low to be scientifically verifiable. This is particularly true given that the endpoints of concern cannot be attributed solely to the environmental hazard under consideration, as illustrated in Box 7.4.2.

Box 7.4.2 Benzene and leukaemia in the United States

Based upon extrapolation from both epidemiological and animal studies, an individual's lifetime exposure to 1 microgram/m^3 benzene in air is estimated to result in an increase in the lifetime risk of leukaemia of 2.2–7.8 in a million (US Environmental Protection Agency 2002). In the United States, a reasonable average outdoor level of benzene is approximately 3 micrograms/m^3, which would predict a lifetime risk of 6.6–23.4 in a million caused by this benzene exposure. Regulatory approaches that decrease that outdoor background level by two-thirds to 1.0 micrograms/m^3 benzene nationwide would be estimated to decrease the risk of benzene-induced leukaemia by two-thirds. This would mean that nationwide, there would be 4.4–15.6 fewer cases of leukaemia per lifetime for every 1 million Americans, or a lifetime risk of approximately 10 in a million.

Assuming a 70-year lifetime, and 350 million Americans, one can estimate that there would be 50 fewer cases of leukaemia a year nationwide as a result of a two-thirds decrease in outdoor benzene levels. This is a very small percentage of National Cancer Institute's estimate of 62,130 new cases of leukaemia in the United States in 2017 (http://www.cancer.gov/cancertopics/types/leukemia). While preventing 50 leukaemia cases is socially desirable, there are no current epidemiological or animal toxicology methods that could scientifically validate these assumptions. A further complication is that for most of the US population, unregulated and highly variable indoor exposure to benzene, as well as to many other volatile organic compounds, far exceeds outdoor exposure. In fact, the major reason for a decrease in personal benzene exposure in the United States has been the decline in cigarette smoking, for smokers, and its restriction from public places, for non-smokers (Goldstein et al. 2011).

Exposure evaluation

Exposure evaluation determines the extent to which an individual is exposed to a toxic agent (i.e. determining the dose), and thus involves both monitoring of environmental and individual samples as well as modelling. This process is complicated because the selected indicators of exposure should be representative—reflecting the actual dose. Box 7.4.3 (A) demonstrates the complications of representative exposure. Note also that knowing the dose alone does not allow one to assess the risk posed by the toxic agent. The dose–response relationship may change under different exposure scenarios. Several factors related to exposure, such as timing (i.e. dose-time-response relationship), may mediate the dose–response relationship and should be considered during risk assessment, as illustrated in Box 7.4.3 (B).

Box 7.4.3 Asbestos and asbestos-related diseases

(A) Measuring asbestos exposure level and assessing mesothelioma risk

Asbestos has been classified as a Group 1 human carcinogen by the IARC. Exposure to asbestos can lead to cancer, such as lung cancer and mesothelioma. To estimate the risk of respiratory tract cancers in relation to asbestos exposure in occupational settings, airborne asbestos concentration measured by fibre per cubic centimetre (f/cc) or fibre per cubic metre (f/m^3) has been widely used as an indicator. However, asbestos also exists as a health threat in the general living environment, such as in older buildings and structures. Data on residential exposure are often lacking, which limits the estimation of cumulative exposure level, let alone lifetime mesothelioma risk. Researchers have applied an alternative method to calculate individual asbestos exposure level by analysing the asbestos in a lung sample. Based on a population-based study in the United Kingdom, the lifetime risk of mesothelioma is 0.02% per 1,000 amphibole fibres per gram of dry lung tissue (Gilham et al. 2016).

(B) Dose-time-response relationship between asbestos and asbestos-related diseases

A majority of mesothelioma cases are caused by occupational exposure to asbestos. The dose–response relationship between asbestos and mesothelioma may change with different age at first exposure. Epidemiological studies suggest that the lifetime mesothelioma risk increases with cumulative level of exposure, duration of exposure, and also the timing of first exposure. That is, early exposure to asbestos is likely to exacerbate the risk of mesothelioma and thus warrants long-term surveillance programmes that begin in childhood (Lacourt et al. 2012; Lacourt et al. 2017).

Exposure evaluation is central to managing environmental risks. Advances in the field of exposure science are particularly crucial to understanding aggregate and cumulative risk (International Life Science Institute 2003; US Environmental Protection Agency 2003). Aggregate risk takes into account the different pathways of exposure for the same chemical. Cumulative risk describes the multiple effects of different agents through different routes. In essence, this is an assessment of the impact of the mix of external synthetic and natural chemicals in which we all live. Cumulative risk assessment is particularly pertinent to environmental justice considerations.

The importance of exposure assessment to understanding risk is exemplified by investigations into potential adverse health consequences resulting from man-made and natural disasters. Notable examples include the attack on the World Trade Center in New York City in September 2001, and the earthquake and tsunami that destroyed the Fukushima Daiichi Nuclear Power Plant in March 2011. Careful evaluation of disease endpoints in relation to exposure will help unravel the highly litigious issue of whether responders or the general public have an increased incidence of disease from exposures related to an event or the clean-up process. New protocols and

tools to assess exposures resulting from man-made and natural disasters may help better evaluate the exposure–effect relationship (Lioy et al. 2006).

Risk characterization

The seemingly straightforward process of characterizing risk through hazard identification, dose–response evaluation, and exposure evaluation presents many challenges. To start with, those doing the characterization can put their own spin on the findings. For example, people are likely to respond differently to something characterized as 99% free of risk than to the numerically equivalent 1% likelihood of a serious consequence including death. Characterizing exactly who is at risk (the denominator) is also a challenge. Reporting risk in terms of the entire exposed general public can trivialize the risk experienced by a highly sensitive subpopulation, such as those with asthma.

Risk can be displaced from one country to another. For instance, in 2006, hazardous waste illegally sent from Europe and dumped in the Ivory Coast was reported to have caused 9,000 acute illnesses and 6 deaths in Abidjan (Greenpeace 2006, 2012). The Ivory Coast, lacking internal expertise, had to use its funds to hire a European company to retrieve, ship, and process the toxic waste. The European company that dumped the waste made a series of payments to the Ivorian government and to individual claimants without admitting liability. In July 2010, a Dutch court found the company guilty of illegally exporting toxic waste from Amsterdam and concealing the nature of the cargo, and imposed a fine on the company.

A long-standing debate exists on the extent to which numerical uncertainty, rather than a simple qualitative statement of the major sources of uncertainty, should be a routine part of risk characterization. Those in favour of routinely providing numerical boundaries that quantify the extent of uncertainty point out that the many estimates and default assumptions in a risk assessment provide wide ranges of uncertainty that should be presented to the risk manager and the general public.

Those in favour of more restricted use of quantitative uncertainty analysis point out that most risk assessments are scoping activities aimed at considering alternatives or developing priorities. Further, major societal decisions are often made on numerical estimates for which no uncertainty factors are given (e.g. gross domestic product, unemployment estimates).

Note that the qualitative issues underlying uncertainty in a risk analysis should always be transparent to the risk manager and affected stakeholders.

The future of risk assessment

Risk assessment as a formal process for evaluating environmental agents continues to evolve, and in the last few decades, more sophisticated approaches to assessing cancer risk and cumulative and aggregate risk have been developed. The use of molecular toxicology to replace standard default assumptions is particularly promising (National Research Council 2007). Just as in other natural sciences, advances in data handling and informatics provide opportunities to assess larger and more complex databases. Advances in epidemiologic methodology using biological indicators of exposure and effect based upon ecogenetics and other molecular biological techniques may be particularly fruitful (Omenn 2000). Conceptually, our genetic makeup loads the gun, but the environment pulls the trigger. Identification of subpopulations sensitive to environmental agents will challenge regulatory and legal interpretations of the many environmental health laws aimed at protecting susceptible populations.

Global harmonization of risk assessment has been underway for decades. This will particularly be needed to avoid the use of environmental health principles as a façade for trade barriers. Evidence-based approaches to toxicology, similar to those increasingly used in medicine, are a welcome development. It remains a challenge to use such approaches for risk assessments that depend heavily on extrapolation to levels of risk below the readily observable—that is, where evidence will be indirect.

The regulatory approach to protecting worker health from toxic chemicals is often based upon both a measurable workplace standard and a subtle measure of effect. Thus for benzene, there is a 1 part per million, 8-hour time-weighted average workplace air standard, as well as a requirement for routine blood counts. The latter informs the former, in terms of both whether unmeasured exposures may be occurring and whether a reconsideration of the allowable external standard is needed. Red cell, white cell, and platelet counts can be affected by a variety of common conditions (e.g. infections, iron-deficiency anaemia), thus surveillance findings must be carefully evaluated before any observations are ascribed to benzene exposure.

In contrast, environmental standards are almost always measures of external emissions or ambient levels of pollutants. Such standards are surrogates for the desired goal of avoiding adverse consequences to human health and the environment. Achieving a level of scientific knowledge that would permit direct evaluation of subtle biological precursors of adverse effects is a desired route to developing emission standards that are truly protective. There is much work in progress in this field of environmental health indicators.

The history of protecting human life from the consequences of environmental agents is the history of catching up to the adverse effects of otherwise beneficial new technology—starting with the human use of fire. A new technology with potential for both beneficial and adverse consequences is nanotechnology. Decreased particle size can result in unexpected new physicochemical properties, in part due to a very high surface-to-volume ratio (Helland et al. 2007). There is concern about what nanoparticles may do at the cellular level and whether current toxicological testing schemes and regulatory processes are adequate to protect against potential harm.

Multiwalled carbon nanotubes (MWCNTs) are a type of manufactured nanomaterial. As use of MWCNTs widens, increased exposure via raw materials and consumer products is expected. The MWCNT-7 version was classified as possibly carcinogenic to humans (Group 2B) by the IARC in 2014, following results from numerous *in vivo* studies (Grosse et al. 2014). In 2017, the IARC declared other MWCNTs, as well as single-walled carbon nanotubes, not classifiable for carcinogenicity to humans (Group 3) (International Agency for Research on Cancer 2017). However, toxicological principles are still relevant and central to the risk assessment process, as demonstrated in Box 7.4.4.

Box 7.4.4 Potential environmental risks of carbon nanomaterials

Scientists have identified concerns about the spread of nanomaterials in consumer products to the environment. Carbon nanomaterials (CNMs), for example, have been widely used in commercial products, such as rechargeable batteries, automotive parts, sporting goods, boat hulls, and water filters (De Volder et al. 2013). CNMs can be released into the environment via manufacturing facilities, landfills, wastewater treatment plants, or accidental release, further disrupting ecosystems (Nowack and Bucheli 2007). Several types of CNMs, such as C60 fullerene, multiwalled carbon nanotubes, and single-walled carbon nanotubes, could be adsorbed by and accumulate in plant roots where the soil or culture medium contains such CNMs.

CNMs have both positive and negative effects on plant growth. For example, CNMs may inhibit rice growth, as measured by shoot height and root length (Hao et al. 2018). However, CNMs have also been used as fertilizer, and carbon nanotubes are used in agricultural sciences to enhance plant growth, by positioning micro-tubes inside the plant to enhance the efficiency of water uptake (Vithanage et al. 2017). Although Bjorkland et al. found low bioaccumulation of carbon nanotubes in plants, invertebrates, and non-mammalian vertebrates (Bjorkland et al. 2017), the overall health risk of CNMs to humans via plant consumption, as well as the risk from environmental exposures, needs further assessment (Guinee et al. 2017).

Future developments in risk assessment are likely to include discovery of new risks from old agents in new forms of use, and estimation of the combined effects of low-dose exposure to mixed agents (Miller et al. 2017). With more scientific evidence showing health risks attributable to toxic agents will come more opportunities for scientists to provide timely and robust policy recommendations. To protect public and environmental health, we will need policies that promptly regulate the use and distribution of identified toxic agents. The challenge is to narrow the gaps between scientific evidence and policy application (Gross and Birnbaum 2017).

Regulatory system design drives the ways that the risk of chemicals can be controlled. Chemical agents are used in a wide range of industries to produce consumer products, and new technology may involve new use of such agents. However, the environmental risks of most chemicals have not been quantified: industrial production processes often grow and change too fast for adequate risk assessment to be conducted. In the mid-twentieth century, there was often insufficient toxicological information to estimate the environmental health risks posed by extensive consumption of all the new chemical substances. Some regulatory schemes, such as the Toxic Substances Control Act in the United States, enacted in 1976, were developed under the assumption that ‘a substance is safe, unless it can be proven unsafe’. This placed the burden of proof on the governmental agency conducting the risk assessment. Since the EU’s REACH passed in 2006, the regulatory paradigm has been shifting to a new assumption, that ‘a substance is unsafe, unless it can be proven safe’ (Krimsky 2017).

In addition, many of the established dose–response relationships address single toxic agents, as do the regulations that set exposure limits and standards. The cumulative impact of low-dose exposure to multiple toxic agents is often neglected. For example, according to the Environmental Impact Assessment Act in Taiwan, governmental approval of new industrial developments that exceed a certain production capacity requires a health risk

assessment. Yet the risk assessments performed have addressed only the incremental additional health risk associated with the new development, ignoring the cumulative health risk associated with industrial and other factors across the entire geographic area (Lin et al. 2016). If low-dose and long-term exposures to multiple toxic agents continue, people who are initially considered as a low-risk group can eventually become a high-risk group.

Although it is important to identify people at high risk in order to develop early prevention strategies (high-risk strategy), interventions for widespread environmental hazards should also cater to those at lower risk (population strategy) (Lanphear 2017); unless there is a compelling reason to exclude them. This is the concept of prevention paradox (see Box 7.4.5.)

Box 7.4.5 The prevention paradox

The prevention paradox, espoused by Geoffrey Rose in 1981, suggests that a large number of people exposed to low risk may generate more cases than a small number exposed to high risk. This is because when there is a small elevation in risk, the total burden of disease will increase more in the large group than in the small group. Conversely, an intervention that offers large benefits to the overall population might offer little benefit to the individual (Rose 1981). Two types of prevention strategy are the high-risk strategy and the population strategy. The high-risk strategy focuses on identifying high-risk individuals and providing individual protection. The population strategy focuses on controlling determinants of incidence and shifting the mean of hazard exposure to a lower level (Rose 2001).

An estimate of population health gains resulting from 22 new preventive interventions in the Netherlands from 1970 to 2010 suggested that three-quarters of the gains were attributed to the population strategy and one-quarter to the high-risk strategy (Mackenbach et al. 2013). It should be noted that these two strategies are not always in competition. Choices can be made between the two strategies when establishing risk prevention policies, but the primary concern should be controlling the causes of incidence (Rose 2001).

Risk assessments should not be one-off activities, but instead involve continued monitoring for potential or encountered adverse health effects through, for example, the development of new risk assessment tools or post-marketing surveillance. This is illustrated in the case study in Box 7.4.6.

Box 7.4.6 Case study: safety of genetically modified crops

Advances in genetically modified (GM) technology in foods, or genetically modified organisms 2.0, have brought on a rapid revolution of crops to meet the global demand for food both quantitatively and qualitatively. In the United States, common GM crops include canola, corn, cotton, soy, and sugar beet. Two types of health risks have arisen: direct risk from the intake of GM foods, and indirect exposure to herbicides used on GM crops.

Currently, the health risk from dietary intake of GM foods is believed to be low. Two National Academy of Sciences reviews, in 2000 and 2004, concluded that genetically modified organisms pose no unique hazards to health. However, the reviews also recommended the development of new risk assessment tools and post-marketing surveillance.

In contrast, the herbicides used on GM crops carry substantial health risks such as carcinogenicity. There has been a sharp increase of herbicides applied to GM crops in the past few decades. For example, the herbicide most widely used on GM crops is glyphosate, and its use in the United States increased by more than 250 times in the past 40 years (from 0.4 million kg used in 1974 to 113 million kg in 2014). The global use of this herbicide has increased by more than 10 times.

As glyphosate-resistant weeds became more widespread, *Enlist Duo*—a combined herbicide product containing glyphosate and 2,4-dichlorophenoxyacetic acid (2,4-D)—was introduced by a chemical manufacturing company. The United States Environmental Protection Agency (US EPA) first approved the use of *Enlist Duo* in 2014, and its use is projected to increase. The following year, in 2015, IARC classified glyphosate as a probable human carcinogen (Group 2A) and 2,4-D as a possible human carcinogen (Group 2B). In 2015, the US EPA revoked its approval of Enlist Duo, stating that the herbicide might be more toxic than previously thought. However, in 2017, the US EPA approved *Enlist Duo* once again, despite the lack of any significant change in its environmental and health assessments. These ever-changing developments raise new concerns about the safety of GM crops (Guyton et al. 2015; Landrigan and Benbrook 2015).

An important goal of evaluating human health risks is to reduce the population's exposure to toxic agents and improve environmental quality. While conventional risk assessments focus on the mechanism of toxic agents on human health, cumulative risk assessments should include both environmental sources of toxic agents and social determinants of exposure. An abundance of scientific literature shows that environmental health risks cause more burden in socioeconomically disadvantaged populations and communities—the same communities which have long been excluded from the decision-making processes resulting in environmental laws, regulations, and policies (Mohai et al. 2009; Kalnins and Dowell 2015).

Incorporating social determinants of health into risk assessment can promote environmental justice. While uneven exposure to environmental hazards is now well recognized and awareness of environmental justice is rising, the pollution gap between rich and poor communities remains and has even widened. Take the emission of toxic release inventory in the United States: people living in the poorest counties have seen a decrease in emissions of 18%, yet those in the richest counties had a 68% reduction in emissions (Kalnins and Dowell 2015).

Risk management and risk communication

Following risk assessment—which is evaluating the consequences of exposure to environmental agents—risk management becomes important. Risk management is a decision-making process of selecting and determining how to manage the identified environmental risks and how to implement the regulatory response (Commission on Risk Assessment and Risk Management 1997). This process can also help governments prioritize how they allocate resources to risks (Rocks et al. 2017).

However, in the real world, there is often a gap between risk levels, governmental priorities, and public interest (Commission on Risk Assessment and Risk Management 1997). Thus, understanding the risk perceived by the general public and properly communicating potential hazards to stakeholders can lead to more effective risk management (Commission on Risk Assessment and Risk Management 1997; Rocks et al. 2017). For environmental emergency scenarios, communication of risk to the general public should start as early as possible to ensure comprehensive understanding and prevent panic (Bol 2016).

Risk communication is the exchange of information about environmental risks among risk assessors, managers, news media, interested groups, and the general public (National Research Council 1989). An example is the concern over adverse health effects following the large-scale nuclear accident at the Fukushima Daiichi Nuclear Power Plant in Japan in March 2011. The first concern raised was about health risks to the local population due to radiation exposure. Although such risks were found to be limited (World Health Organization 2013), related information should have been delivered to residents in a timely and appropriate way to prevent anxiety and mistrust of the authorities.

In addition to reporting the level of toxic agents and the risk of developing disease, risk communication should also aim to prevent potential stigma from attaching to polluted areas. Development of a health literacy toolkit that translates scientific and medical jargon into terms understandable by the general population can help medical professionals answer people's questions and monitor their health (Murakami et al. 2017; Goto et al. 2018).

Another concern related to the Fukushima Daiichi event was food safety, both inside and outside the region, due to radioactive materials leaking from the nuclear plant. This also posed a challenge to international trade and public health in other countries. The general public have expressed concerns

about the indirect and long-term health risks from dietary exposure to contaminated food (Hasegawa et al. 2015; Cho et al. 2017). Right after the accident, the Japanese government set limits for radioactive elements and restricted the sale and distribution of food products from Fukushima and other areas with significant contamination. The immediate actions (e.g. announcing new guidelines) and timely communication (e.g. providing food safety education and disclosing information to the public) were key to relieving concerns about radiation residues in food products, promoting food safety, and securing food supply stability (Gilmour et al. 2016; Cho et al. 2017).

The precautionary principle

A key issue facing risk management is how to proceed in the presence of hazard data with limited or no evidence of human toxicity. Applying the precautionary principle while additional scientific information is being gathered is one approach. The precautionary principle was given impetus by the 1992 Rio Declaration on Environment and Development, which provided the following statement: 'In order to protect the environment, the precautionary approach shall be widely applied by States according to their capabilities. Where there are threats of serious or irreversible damage, lack of full scientific certainty shall not be used as a reason for postponing cost-effective measures to prevent environmental degradation' (United Nations Environmental Programme 1992). The Rio Declaration forces the consideration of precautionary intervention but does not necessarily require such intervention. It includes qualifying language such as 'according to their capabilities' and 'cost-effective measures'. Furthermore, the triple negative notion—that the *absence* of rigorous proof of danger does *not* justify *in*action—is viewed by some as rather weak (World Commission on the Ethics of Scientific Knowledge and Technology 2005).

There are many variants of the precautionary principle, and extensive literature devoted to developing a more rigorous definition. To some, the precautionary principle is merely a means to build more public health protection into quantitative risk analysis, with additional prudent defaults and safety factors to protect at-risk populations and a further focus on uncertainty. To others, the precautionary principle is a new way of addressing environmental risk and dealing with complexity and uncertainty.

Its use is more likely to provide timely and preventive interventions (Kriebel and Tickner 2001; Martuzzi 2007). Some of the important approaches advocated by the precautionary principle, such as transparency and involvement of stakeholders, have also been frequently advocated under the rubric of risk assessment and management.

The United States banned the production of polychlorinated biphenyls (PCBs) in 1976, despite industry opposition, then and now, on the grounds of uncertain science. The continual decline in body burden of PCBs and dioxins has thus been accomplished based on regulatory decisions. Yet advancing the science needed for public health decision-making must remain a major goal, despite the usefulness of the precautionary principle (Foster et al. 2000; Goldstein and Carruth 2003, 2004; Grandjean et al. 2004).

The working definition of the precautionary principle proposed by UNESCO (see Box 7.4.7; italics in the original) includes the firm statement that actions based upon the precautionary principle should be subject to continuous analysis and review (World Commission on the Ethics of Scientific Knowledge and Technology 2005). A number of major concerns have been raised about the precautionary principle, for example, what does it add to standard public health concepts? Is it true that complex scientific questions are unsolvable and, if so, is the precautionary principle needed to promote action in the face of scientific uncertainty? In view of its alleged use to justify trade barriers, is it still possible to advocate the precautionary principle as an antidote to biased decision-making?

Box 7.4.7 Working definition of the precautionary principle

'When human activities may lead to morally unacceptable harm that is scientifically plausible but uncertain, actions shall be taken to avoid or diminish that harm. *Morally unacceptable harm* refers to harm to humans or the environment that is

- threatening to human life or health, or
- serious and effectively irreversible, or
- inequitable to present or future generations, or
- imposed without adequate consideration of the human rights of those affected.

The judgement of *plausibility* should be grounded in scientific analysis. Analysis should be ongoing so that chosen actions are subject to review. *Uncertainty* may apply to, but need not be limited to, causality or the bounds of the possible harm. *Actions* are interventions that are undertaken before harm occurs that seek to avoid or diminish the harm. Actions should be chosen that are

proportional to the seriousness of the potential harm, with consideration of their positive and negative consequences, and with an assessment of the moral implications of both action and inaction. The choice of action should be the result of a participatory process.'

Reproduced with permission from COMEST (World Commission on the Ethics of Scientific Knowledge and Technology), *The Precautionary Principle*, Copyright © 2005 United Nations Educational, Scientific and Cultural Organization, available from http://unesdoc.unesco.org/images/0013/001395/139578e.pdf.

The precautionary principle is said to be pertinent in situations where the scientific evidence is uncertain. One example is the management of the health and environmental risks of endocrine disruptors. This is particularly challenging due to the need to consider the interactive effects of multiple chemicals with a wide range of additive, synergistic, and antagonistic interactions (Kortenkamp 2007). A case study is shown in Box 7.4.8.

Box 7.4.8 The precautionary principle and bisphenol A

Worldwide, over 3 million tonnes of bisphenol A (BPA) are produced annually. BPA can be found in many consumer products. These include reusable water bottles, baby bottles, implantable medical devices, and dental sealants. BPA is also a component of epoxy resins, which are used to overlay the inside of food and beverage cans, in cardboard and papers used for food wrapping, in some plastic water pipes, and in paints and cigarette filters. Of concern is the fact that BPA leaches out from these consumer products, even under normal use conditions. The leaching increases under high temperatures, and with exposure to acidic or basic solutions.

Animal studies have shown that BPA may alter male and female reproductive tract development, the brain, and the mammary glands. In large-scale human population studies in various parts of the world, BPA has been detected in urine, blood, saliva, breast milk, and tissues and fluids associated with pregnancy at levels similar to those that have been shown to cause damage in animals. In a study of over 5,000 people in Canada, over 90% had detectable levels of BPA in urine, with a geometric mean concentration of 1.16 ng/mL (Bushnik et al. 2010). Health Canada has stated that precaution is warranted, since it is not possible to dismiss the potential harmful effects of BPA (Health Canada 2010). With this statement, Canada became the first country to declare BPA a health hazard, but additional regulatory action has yet to be taken (Vandenberg 2011).

Practical insight into the debate about the precautionary principle versus risk assessment can be found in the way the EU and other countries evaluate the safety of chemicals. The EU promulgated the REACH regulations, after much debate, with its proponents focusing on the precautionary principle as a rationale for the new legislation. REACH requires industry to develop data, assess risk, and provide information about virtually all chemicals in use, including constituents of product mixtures. No distinction is made

between newly developed chemicals or those long available in commerce—a contrast with the US Toxic Substances Control Act, whose weakness in this regard led to the inadequate testing of compounds such as the gasoline additive methyl tert-butyl ether (MTBE) before its inappropriate release into the environment (Erdal and Goldstein 2000).

The cost of implementing REACH is estimated at US$3–6 billion for obtaining data and registering compounds over the first 11 years. Risk assessment, based on both toxicity and exposure, will be used extensively throughout the process, including for setting priorities for data needs and making decisions on regulatory approaches. Unfortunately, there appears to be no uniform coordinated effort to develop the research where toxicological data is scant.

Unfortunately, the precautionary principle has been tainted through disagreements over application. The United States, Canada, and other countries disagreed with the EU on the extent of risk to health from consumption of hormone-treated beef or genetically modified foods. The EU's use of the precautionary principle to establish a stringent aflatoxin standard, to the detriment of exports of sub-Saharan countries, has been criticized as the misuse of public health principles for trade protectionism (Goldstein 2007). The risk of exposure to electromagnetic fields from the use of mobile phones is another contentious and confused issue, resulting in advice based on a mix of the precautionary principle and the scientific evidence.

The primary missing ingredient in the approach to ever more complex environmental challenges, including such broader issues as global warming, is a systems-based approach incorporating the best science and focusing on the most important questions. We need to have a more systematic and coordinated effort to respond to the challenges of fragmented national and international approaches to environmental issues that produce piecemeal efforts which may lead us to fall further behind in protecting public health and the environment.

Conclusion

Understanding the web of environmental cause and effect relations is an increasing challenge in an ever more connected world. Advances in toxicology, filtered through an appropriate appreciation of the optimal

approaches to analysing and presenting risks to an involved public, are crucial to protecting public health and the environment.

REFERENCES

American Conference of Governmental Industrial Hygienists (2013). *Threshold Limit Values for Chemical and Physical Agents and Biological Exposure Indices*. Cincinnati, OH: ACGIH.

Arbuckle, T.E., et al. (2002). Assessing exposure in epidemiologic studies to disinfection by-products in drinking water: report from an international workshop. *Environmental Health Perspectives*, 110, 53–60.

Bjorkland, R., Tobias, D.A., and Petersen, E.J. (2017). Increasing evidence indicates low bioaccumulation of carbon nanotubes. *Environmental Science-Nano*, 4, 747–66.

Bol, G.F. (2016). Risk communication in times of crisis: pitfalls and challenges in ensuring preparedness instead of hysterics. *EMBO Reports*, 17, 1–9.

Burton, G.J., Fowden, A.L., and Thornburg, K.L. (2016). Placental origins of chronic disease. *Physiological Reviews*, 96, 1509–65.

Bushnik, T., et al. (2010). Lead and bisphenol A concentrations in the Canadian population. *Health Reports*, 21, 7–18.

Cho, T.J., et al. (2017). Development of an effective tool for risk communication about food safety issues after the Fukushima nuclear accident: What should be considered? *Food Control*, 79, 17–26.

Cogliano, V.J. (2004). Current criteria to establish human carcinogens. *Seminars in Cancer Biology*, 14, 407–12.

Commission on Risk Assessment and Risk Management (1997). *Framework for Environmental Health Risk Management*. Washington, DC: Presidential/Congressional Commission on Risk Assessment and Risk Management.

De Volder, M.F., Tawfick, S.H., Baughman, R.H., and Hart, A.J. (2013). Carbon nanotubes: Present and future commercial applications. *Science*, 339, 535–9.

Erdal, S., and Goldstein, B.D. (2000). Methyl tert-butyl ether as a gasoline oxygenate: Lessons for environmental public policy. *Annual Review of Energy and the Environment*, 25, 765–802.

European Chemicals Agency (2017). How to avoid unnecessary testing on animals. [Online] Available at: https://echa.europa.eu/support/registration/how-to-avoid-unnecessary-testing-on-animals

Food and Agriculture Organization, and World Health Organization (2009). *Principles and Methods for the Risk Assessment of Chemicals in Food. Environmental Health Criteria 240*. Geneva: World Health Organization.

Foster, K.R., Vecchia, P., and Repacholi, M.H. (2000). Science and the precautionary principle. *Science*, 288, 979–81.

Gilham, C., et al. (2016). Pleural mesothelioma and lung cancer risks in relation to occupational history and asbestos lung burden. *Occupational and Environmental Medicine*, 73, 290–9.

Gilmour, S., Miyagawa, S., Kasuga, F., and Shibuya, K. (2016). Current measures on radioactive contamination in Japan: a policy situation analysis. *PLoS One*, 11, 12.

Goldstein, B.D. (2007). Problems in applying the precautionary principle to public health. *Occupational and Environmental Medicine*, 64, 571–4.

Goldstein, B.D., and Carruth, R.S. (2003). Implications of the precautionary principle to environmental regulation in the United States: examples from the control of hazardous air pollutants in the 1990 Clean Air Act amendments. *Law and Contemporary Problems*, 66, 247–61.

Goldstein, B.D. and Carruth, R.S. (2004). Implications of the precautionary principle: Is it a threat to science? *International Journal of Occupational Medicine and Environmental Health*, 17, 153–61.
Goldstein, B.D. and Gallo, M.A. (2001). Pare's law: the second law of toxicology. *Toxicological Sciences*, 60, 194–5.
Goldstein, B.D., Liu, Y., Wu, F., and Lioy, P. (2011). Comparison of the effects of the US Clean Air Act and of smoking prevention and cessation efforts on the risk of acute myelogenous leukemia. *American Journal of Public Health*, 101, 2357–61.
Goto, A., et al. (2018). Collaborative processes of developing a health literacy toolkit: A case from Fukushima after the nuclear accident. *Journal of Health Communication*, 23(2), 200–6.
Grandjean, P., et al. (2004). Implications of the precautionary principle in research and policy-making. *American Journal of Industrial Medicine*, 45, 382–5.
Greenpeace (2006). Toxic waste in Abidjan. [Online] Available at: https://www.greenpeace.org/archive-international/en/news/features/toxic-waste-in-abidjan-green/
Greenpeace (2012). The toxic truth. [Online] Available at: https://www.greenpeace.org/international/publication/7245/the-toxic-truth/
Gross, L. and Birnbaum, L.S. (2017). Regulating toxic chemicals for public and environmental health. *PLoS Biology*, 15, e2004814.
Grosse, Y., et al. (2014). Carcinogenicity of fluoro-edenite, silicon carbide fibres and whiskers, and carbon nanotubes. *Lancet Oncology*, 15, 1427–8.
Guinee, J.B., Heijungs, R., Vijver, M.G., and Peijnenburg, W. (2017). Setting the stage for debating the roles of risk assessment and life-cycle assessment of engineered nanomaterials. *Nature Nanotechnology*, 12, 727–33.
Guyton, K.Z., et al. (2015). Carcinogenicity of tetrachlorvinphos, parathion, malathion, diazinon, and glyphosate. *Lancet Oncology*, 16, 490–1.
Hao, Y., et al. (2018). Carbon nanomaterials alter plant physiology and soil bacterial community composition in a rice-soil-bacterial ecosystem. *Environmental Pollution*, 232, 123–36.
Hasegawa, A., et al. (2015). Health effects of radiation and other health problems in the aftermath of nuclear accidents, with an emphasis on Fukushima. *The Lancet*, 386, 479–88.
Health Canada (2010). *Order adding a toxic substance to Schedule 1 to the Canadian Environmental Protection Act, 1999*. [Online] Available at: https://www.ec.gc.ca/lcpe-cepa/eng/orders/OrderTypeList.cfm?intCategory=27
Helland, A., Wick, P., Koehler, A., Schmid, K., and Som, C. (2007). Reviewing the environmental and human health knowledge base of carbon nanotubes. *Environmental Health Perspectives*, 115, 1125–31.
International Agency for Research on Cancer (2017). *Volume 111: Some Nanomaterials and Some Fibres*. Lyon: IARC. [Online] Available at: https://monographs.iarc.fr/ENG/Monographs/vol111/mono111-01.pdf
International Life Science Institute (2003). *A Framework for Cumulative Risk Assessment; Workshop Report*. Washington, DC: ILSI.
Kalnins, A. and Dowell, G. (2015). Community characteristics and changes in toxic chemical releases: does information disclosure affect environmental injustice? *Journal of Business Ethics*, 145, 277–92.
Kim, S., et al. (2007). Genetic polymorphisms and benzene metabolism in humans exposed to a wide range of air concentrations. *Pharmacogenetics and Genomics*, 17, 789–801.
Kortenkamp, A. (2007). Ten years of mixing cocktails: A review of combination effects of endocrine-disrupting chemicals. *Environmental Health Perspectives*, 115 Suppl 1, 98–105.
Kriebel, D. and Tickner, J. (2001). Reenergizing public health through precaution. *American Journal of Public Health*, 91, 1351–5.

Krimsky, S. (2017). The unsteady state and inertia of chemical regulation under the US Toxic Substances Control Act. *PLoS Biology*, 15, e2002404.
Lacourt, A., et al. (2012). Temporal patterns of occupational asbestos exposure and risk of pleural mesothelioma. *European Respiratory Journal*, 39, 1304–12.
Lacourt, A., et al. (2017). Dose-time-response association between occupational asbestos exposure and pleural mesothelioma. *Occupational and Environmental Medicine*, 74, 691–7.
Landrigan, P.J. and Benbrook, C. (2015). GMOs, herbicides, and public health. *New England Journal of Medicine*, 373, 693–5.
Lanphear, B.P. (2017). Low-level toxicity of chemicals: No acceptable levels? *PLoS Biology*, 15, e2003066.
Lin, R.T., et al. (2016). Increased risk of respiratory mortality associated with the high-tech manufacturing industry: A 26-year study. *International Journal of Environmental Research and Public Health*, 13, 12.
Lioy, P.J., Pellizzari, E., and Prezant, D. (2006). The World Trade Center aftermath and its effects on health: Understanding and learning through human-exposure science. *Environmental Science and Technology*, 40, 6876–85.
Mackenbach, J.P., Lingsma, H.F., van Ravesteyn, N.T., and Kamphuis, C.B. (2013). The population and high-risk approaches to prevention: quantitative estimates of their contribution to population health in the Netherlands, 1970–2010. *European Journal of Public Health*, 23, 909–15.
Maher, B.A., et al. (2016). Magnetite pollution nanoparticles in the human brain. *Proceedings of the National Academy of Sciences of the United States of America*, 113, 10797–801.
Martuzzi, M. (2007). The precautionary principle: in action for public health. *Occupational and Environmental Medicine*, 64, 569–70.
Miller, M.F., et al. (2017). Low-dose mixture hypothesis of carcinogenesis workshop: Scientific underpinnings and research recommendations. *Environmental Health Perspectives*, 125, 163–9.
Mohai, P., Pellow, D., and Roberts, J.T. (2009). Environmental justice. *Annual Review of Environment and Resources*, 34, 405–30.
Murakami, M., et al. (2017). Communicating with residents about risks following the Fukushima nuclear accident. *Asia-Pacific Journal of Public Health*, 29, 74S–89S.
National Research Council (1989). *Improving Risk Communication*. Washington, DC: National Academies Press.
National Research Council (1993). *Pesticides in the Diets of Infants and Children*. Washington, DC: National Academies Press.
National Research Council (2007). *Toxicity Testing in the 21st Century: A Vision and a Strategy*. Washington, DC: NRC.
National Toxicology Program (2015). *Handbook for Preparing Report on Carcinogens Monographs*. Research Triangle Park, NC: NTP.
Nowack, B., and Bucheli, T.D. (2007). Occurrence, behavior and effects of nanoparticles in the environment. *Environmental Pollution*, 150, 5–22.
Oberdorster, G., et al. (2004). Translocation of inhaled ultrafine particles to the brain. *Inhalation Toxicology*, 16, 437–45.
Omenn, G.S. (2000). Public health genetics: An emerging interdisciplinary field for the post-genomic era. *Annual Review of Public Health*, 21, 1–13.
Rocks, S.A., et al. (2017). Engaging with comparative risk appraisals: Public views on policy priorities for environmental risk governance. *Risk Analysis*, 37, 1683–92.
Rose, G. (1981). Strategy of prevention: Lessons from cardiovascular disease. *British Medical Journal*, 282, 1847–51.
Rose, G. (2001). Sick individuals and sick populations. *International Journal of Epidemiology*, 30, 427–32.

Rothman, N., et al. (1997). Benzene poisoning, a risk factor for hematological malignancy, is associated with the NQ01 C-609->T mutation and rapid fractional excretion of chlorzoxazone. *Cancer Research*, 57, 2839–42.

Umbreit, T.H., Hesse, E.J., and Gallo, M.A. (1986). Bioavailability of dioxin in soil from a 2,4,5-T manufacturing site. *Science*, 232, 497–9.

United Nations Environmental Programme (1992). *Rio Declaration on Environment and Development, Principle 15*. [Online] Available at: http://www.un.org/documents/ga/conf151/aconf15126-1annex1.htm

US Environmental Protection Agency (2002). *Integrated Risk Information System: Benzene CASRN 71-43-2*. Washington, DC: US EPA.

US Environmental Protection Agency (2003). *Framework for Cumulative Risk Assessment*. Washington, DC: US EPA.

Vandenberg, L.N. (2011). Exposure to bisphenol A in Canada: invoking the precautionary principle. *Canadian Medical Association Journal*, 183, 1265–70.

Vithanage, M., Seneviratne, M., Ahmad, M., Sarkar, B., and Ok, Y.S. (2017). Contrasting effects of engineered carbon nanotubes on plants: a review. *Environmental Geochemistry and Health*, 39, 1421–39.

World Commission on the Ethics of Scientific Knowledge and Technology (2005). *The Precautionary Principle*. Paris, France: United Nations Educational, Scientific, and Cultural Organization.

World Health Organization (2013). *Health Risk Assessment from the Nuclear Accident after the 2011 Great East Japan Earthquake and Tsunami, Based on a Preliminary Dose Estimation*. Geneva, Switzerland: WHO.

7.5

Risk perception and communication

Baruch Fischhoff and Tamar Krishnamurti

Public health depends on laypeople's ability to understand the health-related choices that they and their societies face. The study of risk perception examines that ability. The study of risk communication examines the processes that determine how well communications help (or hinder) their recipients. Although focused on decisions involving risk, such research necessarily considers potential benefits as well, if only the benefits that come from reducing risks (e.g. through medical treatments, lifestyle changes, or improved air quality). Communication is a two-way process. Without listening to people, it is impossible to understand what they know and value, as a foundation for providing relevant information in a comprehensible form. Here, we draw on basic research in decision science (behavioural decision research) to inform the specific challenges faced by public health researchers and practitioners.

The decision science approach is conceptually straightforward. First, analyse the decisions that people face, in sufficiently precise terms to identify the information most critical to their outcomes. Second, describe decision-makers' existing beliefs in terms comparable to the analysis. Third, develop communications that bridge the critical gaps between what people know and what they need to know. Next, evaluate the success of those communications. Repeat as necessary. These steps are interdependent. Descriptive research can reveal goals, obstacles, and capabilities that the analyses missed; communication failures can prompt additional descriptive research into decision-making processes. Evaluation is essential. One cannot expect to get it right the first time.

Effective communication requires four kinds of expertise: (a) subject matter specialists, for ensuring accuracy; (b) risk and decision analysts, for characterizing choices and identifying critical information; (c) social and behavioural scientists, for describing existing beliefs, values and capabilities,

and designing communications; and (d) communication practitioners and implementation scientists, for executing and evaluating sustainable programmes. The success of the enterprise depends on continuing two-way communication with audience members, as depicted in Figure 7.5.1, so that professionals understand their needs, they understand what the professionals are doing, and the process creates the personal ties needed for mutual trust.

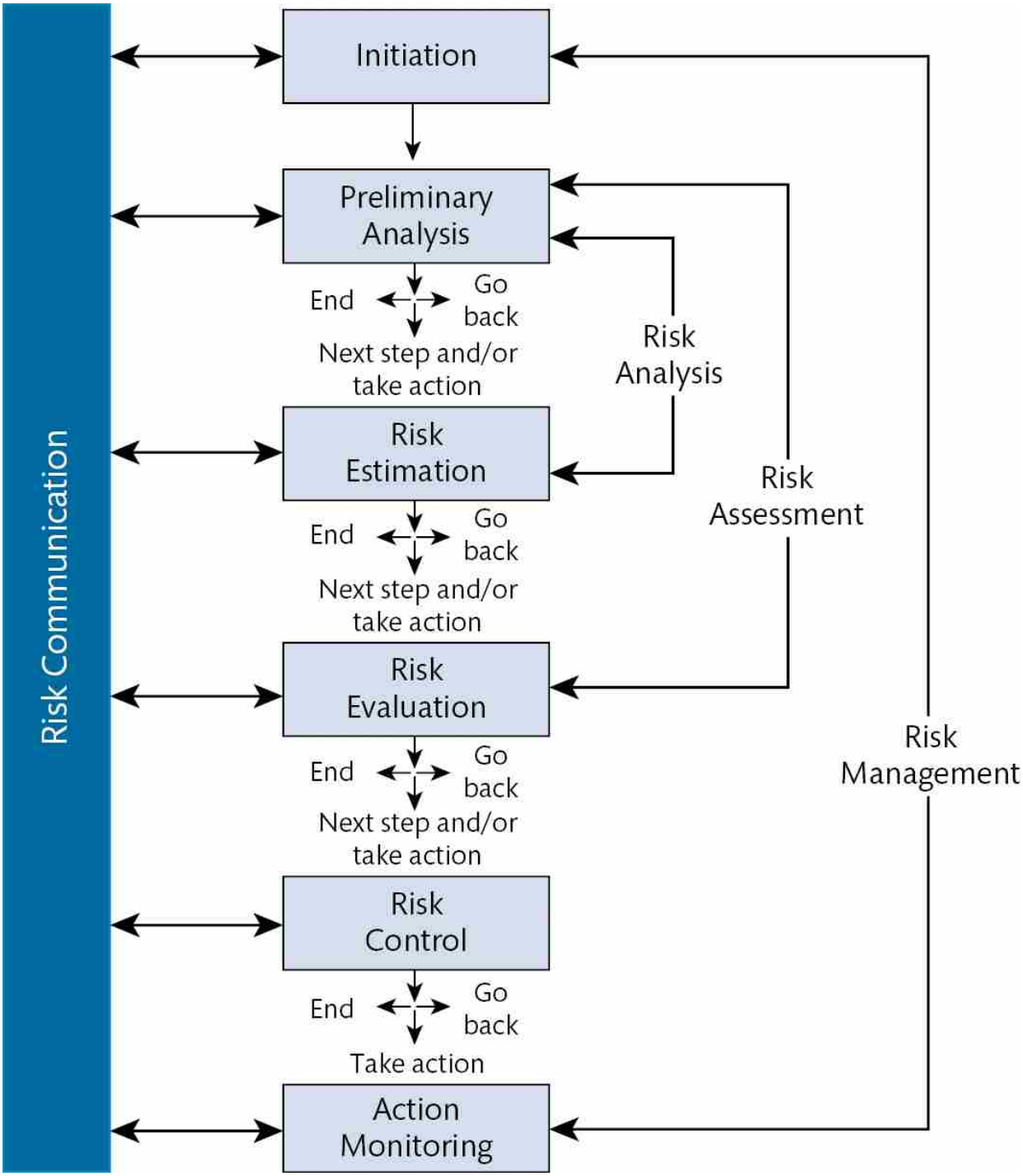


Fig. 7.5.1 Two-way communication.

Decision science provides guidance on two topics central to this endeavour. One is how to identify threats to risk-related decision-making and ways to overcome them. The second is how to measure communications' success. Without such evaluation, not only may

communications fail, but experts may come to doubt their audience's competence. If laypeople are held responsible for failing to learn from flawed communications, then they may be needlessly denied the opportunity for active participation in health decisions. One should no more release untested communications than untested pharmaceuticals. The chapter seeks to help experts help laypeople make wise health choices.

Introduction to risk perception and communication

Sound health risk decisions require understanding the risks and benefits of possible actions. Some choices are personal. They include whether to wear bicycle helmets and seat belts, whether to read and follow safety warnings, whether to buy and use condoms, and how to select and cook food. Other choices are public. They include whether to protest the siting of hazardous waste incinerators and halfway houses, whether to support fluoridation and 'green' candidates, and what topics to include in sex education (e.g. relationships as well as risks).

Sometimes, single choices have large effects (e.g. buying a safe car, taking a dangerous job, having a child). Sometimes, small effects accumulate over multiple choices (e.g. exercising, avoiding trans-fats, wearing seatbelts, using escort services, getting tattoos). Sometimes, health-related choices focus on health; sometimes, they do not (e.g. purchasing homes that require long commutes, choosing friends who exercise regularly, joining groups opposed to vaccination).

This chapter reviews the research base for assessing and improving individuals' understanding of the risks and possible benefits of health-related choices. Following convention, these pursuits are called *risk perception* and *risk communication*, respectively. Psychologists sometimes reserve the term 'perception' for direct physiological responses, using 'judgement' for their translation into deliberate actions.

Inaccurate judgements about risks can harm people. So can inaccurate beliefs about those judgements. If their decision-making ability is overestimated, people may face impossibly hard choices (e.g. among unfamiliar medical options, without adequate counselling). If their ability is underestimated, people may be needlessly denied the right to choose. As a result, the chapter assumes: (1) that descriptive statements about how people make decisions require supporting evidence and (2) that evaluative

statements about how well people make decisions require rigorous analysis of what they need to know. To these ends, the chapter emphasizes methodological safeguards against misguided assessments.

The next two sections present methods for assessing people's estimates of the size of risks and what they believe about the processes that create and control those risks. They are followed by sections on how to create and evaluate communications that address critical gaps in understanding. The concluding section considers the strategic importance of risk communication in public health.

The present chapter focuses on how to help people to be as informed as possible when making personal and public health decisions. Access to research on complementary social and emotional processes might begin with Breakwell (2014), and Slovic (2010), Our methodological focus recognizes that basic science has been disproportionately conducted in developed, Western countries—meaning that studies of what people believe and want should be repeated to capture local conditions (Medin et al. 2017). (Results related to how people process information might be more robust.) However, decision science methods should apply broadly, using experimental, survey, interview, and observational approaches to determine breadth and population distribution of beliefs, desires, and needs.

Quantitative assessment

Estimating risk magnitude

A common complaint among experts is that 'the public doesn't realize how small (or large) Risk X is'. There is empirical evidence demonstrating such biases (Slovic 2001). However, that evidence has often been collected in experimental settings designed to reveal biases, so that researchers can see just how they work. Such studies say little about how large or common biases are in everyday life. As a guide to interpreting and conducting studies that assess lay beliefs, we examine one widely cited study.

Participants

Lichtenstein et al. (1978) asked members of a civic group in Eugene, Oregon, to estimate the annual number of deaths in the United States from

30 causes. They chose causes that were familiar, well-defined, and for which public health statistics were available (e.g. botulism, tornadoes, motor vehicle accidents). They also elicited judgements in two ways that allowed comparing responses with those statistics. One of these *response modes* asked people to pick the more frequent of two paired causes of death (e.g. asthma, botulism) and then estimate the ratio of their frequencies. The second response mode asked people to estimate the number of deaths, after telling them the value for one cause (either electrocution or motor vehicle accidents) in order to give respondents a feeling for annual death rates—after pre-tests found that many people had little idea about what numbers to give. The results were typical of such studies.

Results

Judgements of the *relative* risk from different causes were similar however the question was asked, as though people have a sense of relative risk, which they express consistently even with such unfamiliar tasks.

Judgements of *absolute* risk, however, were affected by the procedure. People told about the 50,000 annual deaths from auto accidents gave higher estimates than did those told about the 1,000 annual deaths from electrocution. Thus, people have less feeling for absolute frequency, rendering them sensitive to implicit cues given by how questions are posed (Poulton 1989; Schwarz 1999).

Absolute risk judgements varied less than the corresponding statistical estimates. A substantive interpretation is that people do not realize just how large the differences are between life's big and small risks. A methodological interpretation is that these judgements were *anchored* in the initial values presented to participants (for electrocution or motor vehicle accidents). Because these values were in the middle of the range, participants overestimated small frequencies and underestimated large ones. That pattern might change with other anchors. For example, a lower anchor (e.g. botulism) should reduce (or eliminate) overestimation of small frequencies, while increasing the underestimation of large ones.

Both relative and absolute risk judgements seemed to reflect *availability bias*, whereby easily imagined causes of death (e.g. flood, homicide, tornadoes) seem disproportionately likely, whereas the opposite is true for 'quiet killers' (e.g. stroke, stomach cancer). Telling people about the possibility of availability bias did not help them to overcome it, as tracking frequency of events is so automatic that people do not realize how observations shape their perceptions.

Thus, some response patterns were affected by the procedure (e.g. absolute estimates) and some were not (e.g. relative risk judgements). A century-plus of psychophysics research (Poulton 1989) has identified many procedural details that can affect quantitative judgements, when people don't know quite what to say. Awareness of that research can be a valuable antidote to taking quantitative estimates too literally. Eliciting judgements in several ways can establish how big such context effects are, and how much they could affect decisions.

Evaluative standards

Risk judgements can be evaluated in terms of *consistency* or *accuracy*. Evaluating consistency means asking logically related questions and comparing the answers (e.g. do risk estimates increase with increases in the number of exposures or time period?). Evaluating accuracy means asking questions that can be compared to risk estimates. This requires greater precision than is common to many surveys.

The alternative to asking people about their beliefs is trying to infer them from their actions. For example, first-time gun sales often spike after mass shootings in the United States (Studdert et al. 2017). That common reaction could reflect fear of gun violence or of gun control. After the 9/11 attacks, some observers claimed that some Americans had increased their risk level by choosing to drive, rather than fly, citing historical risk statistics. However, at the time, no one knew if the historical statistics still applied or if those travellers assumed any additional risk voluntarily, weighing the cost and hassle of flying (once that became possible again). It is often hard to tell, by observation alone, what beliefs and preferences motivate decisions.

Probability judgements

Health decisions depend on the probability, and not just the possibility, of good or bad things happening (e.g. treatments working, side effects occurring). Although many experts are sceptical about how well people can use probabilities, the research shows a more complex picture (e.g. O'Hagan et al. 2006; Morgan 2014):

People interpret verbal quantifiers (e.g. likely) differently depending on the context (e.g. rare disease vs. rare sunny day), making it hard to know what they mean or infer (Budescu and Wallsten 1996).

People prefer to *receive* numeric probabilities, which provide more information, but to *provide* verbal ones, which involve less effort and accountability (Erev and Cohen 1990).

People can use well-designed numeric scales. For example, Woloshin et al. (1998) elicited very small probabilities with a scale that was linear from 1% to 100%, but had expanded 0–1% with a four-order log scale.

Numeric judgements often have good *construct validity*, in the sense of correlating sensibly with other variables. For example, in a nationally representative sample, teens giving higher probabilities of becoming pregnant also reported more sexual activity; teens giving higher probabilities of getting arrested also reported more neighbourhood violence.

Numeric judgements often have good predictive validity. For example, in that survey, teens who gave higher probabilities were more likely to get pregnant or arrested.

Biased numeric judgements may reflect misinformation, rather than difficulty with numbers, as when people underestimate their own probability of causing an accident because others' poor driving is more available to them or physicians fail to prescribe antidepressants as pregnancies progress (Yonkers et al. 2014), even though foetal risk dramatically diminishes after the first trimester.

Probability judgements can be deliberately biased, when people respond strategically. For example, Christensen-Szalanski and Bushyhead (1993) found that physicians overestimated the probability of pneumonia, fearing that unlikely cases might be neglected.

Transient emotions can affect judgements. For example, anger increases optimism, fear the opposite (Lerner and Keltner 2001), with effects large enough to tip close decisions.

Judgements of the probability of being correct are moderately correlated with how much people know. The result is *overconfidence* with hard questions (e.g. correct 55% of the time when 70% confident) and underconfidence with easy ones (e.g. correct 90% of the time when 80% confident).

Some numeric values are special. People like round numbers. When uncertain about their answer, they sometimes say 50 in the sense of 50–50,

rather than a numeric probability (Bruine de Bruin et al. 2000).
Probability judgement processes mature by middle adolescence. For example, teens are no more likely than adults to believe in their own invulnerability (Quadrel et al. 1993).
Some people use probabilities better than others; they also tend to be better at other judgement tasks and have better life outcomes (Mellers et al. 2015; Parker et al. 2018).
People who are low in numeracy and literacy can make probability judgements with carefully designed displays (e.g. icon arrays, pictographics; Hawley et al. 2008).
Clear, prompt feedback can improve how well people use probabilities (Lichtenstein and Fischhoff 1980; Mellers et al. 2015).
Experts are people, too, when forced to go beyond established knowledge and judge probabilities of uncertain events (O'Hagan et al. 2006).

Unless people provide numeric judgements, it is impossible to know how accurate their beliefs are and to tell, for example, whether public health communications have exaggerated fears of smoking (Viscusi 1992) and breast cancer (Black et al. 1995).

Defining risk

Studies like Lichtenstein et al. (1978) measure 'risk' perceptions, only if 'risk' means 'chance of death'. However, 'risk' has multiple meanings, even among experts (National Research Council 1996; Fischhoff and Kadvany 2011). It might include just deaths or also other outcomes, such as morbidity and trauma. Even if it includes just fatalities, it might be measured in probability of death, expected life years lost, deaths per person exposed, or per hour of exposure. Each definition entails an ethical position. For example, *probability of death* treats all deaths equally, whereas *life years lost* places extra weight on deaths of young people as they incur more lost years. In public health policy, that definition would direct resources to reducing injuries, which affect young people, relative to chronic illnesses, affecting older ones. Defining 'risk' to Include social and psychological costs would increase concern for alcohol and illegal drugs, which can ruin lives without ending them (Phillips et al. 2017).

Without a clear, shared definition of 'risk', people can unwittingly speak at cross purposes. One source of confusion is treating risk as a discrete

variable (deeming some activities as ‘risky’ and others as ‘safe’), rather than as a continuous one, requiring decision-makers to decide how much risk to tolerate. Calling something ‘risky’ could mean an unreasonable demand for zero risk, a shorthand for being ‘acceptable (given the benefits)’, or a call for invoking a *precautionary principle*, making it as too great to countenance (Löfstedt et al. 2002).

Dimensions of risk

One early risk perception study asked experts and laypeople to estimate the ‘risk of death’ from 30 activities and technologies (Slovic 1987), then correlated their judgements with statistical estimates of annual fatalities. That correlation was stronger for experts than for laypeople. However, when asked to estimate ‘fatalities in an average year’, experts and laypeople responded similarly. Comparing those judgements suggested that lay respondents interpreted ‘risk of death’ as including the catastrophic potential of non-average years. If so, then experts and laypeople agreed about routine risks (for which the science is often good), but disagreed about possible anomalies (for which the science is naturally weaker). Thus, when experts and laypeople disagree about risks, they might be seeing the facts differently or looking at different facts, relevant to their definition of ‘risk’.

Considering catastrophic potential could reflect one of (at least) two ethical concerns. One is caring more about lives lost at once, rather than lost individually. The second is caring more about uncertain risks, poorly enough understood that catastrophes are possible (Slovic 1987). Beginning with Starr (1969), many features, like uncertainty and catastrophic potential, have been proposed as affecting definitions of risk (Fischhoff and Kadvany 2011). One way to reduce that set to manageable size is to examine correlations between the features. Figure 7.5.2 shows results from a first study. It finds two *dimensions* of risk. The vertical dimension reflects the fact that hazards seen as new also tend to be seen as unknown, involuntary, and with delayed effects. The horizontal dimension reflects that fact that hazards with catastrophic potential also tend to evoke a feeling of dread. The factors labelled *unknown* and *dread*, respectively, might be seen as capturing cognitive and emotional aspects of people’s concern. Many studies following this ‘psychometric paradigm’ have found roughly similar dimensions, using differing elicitation modes, scaling techniques, items, and participants (Slovic 2001; Glassman-Fox and Weber 2016). Where hazards fall in the

space correlates with attitudes towards them, such as how stringently they should be regulated.

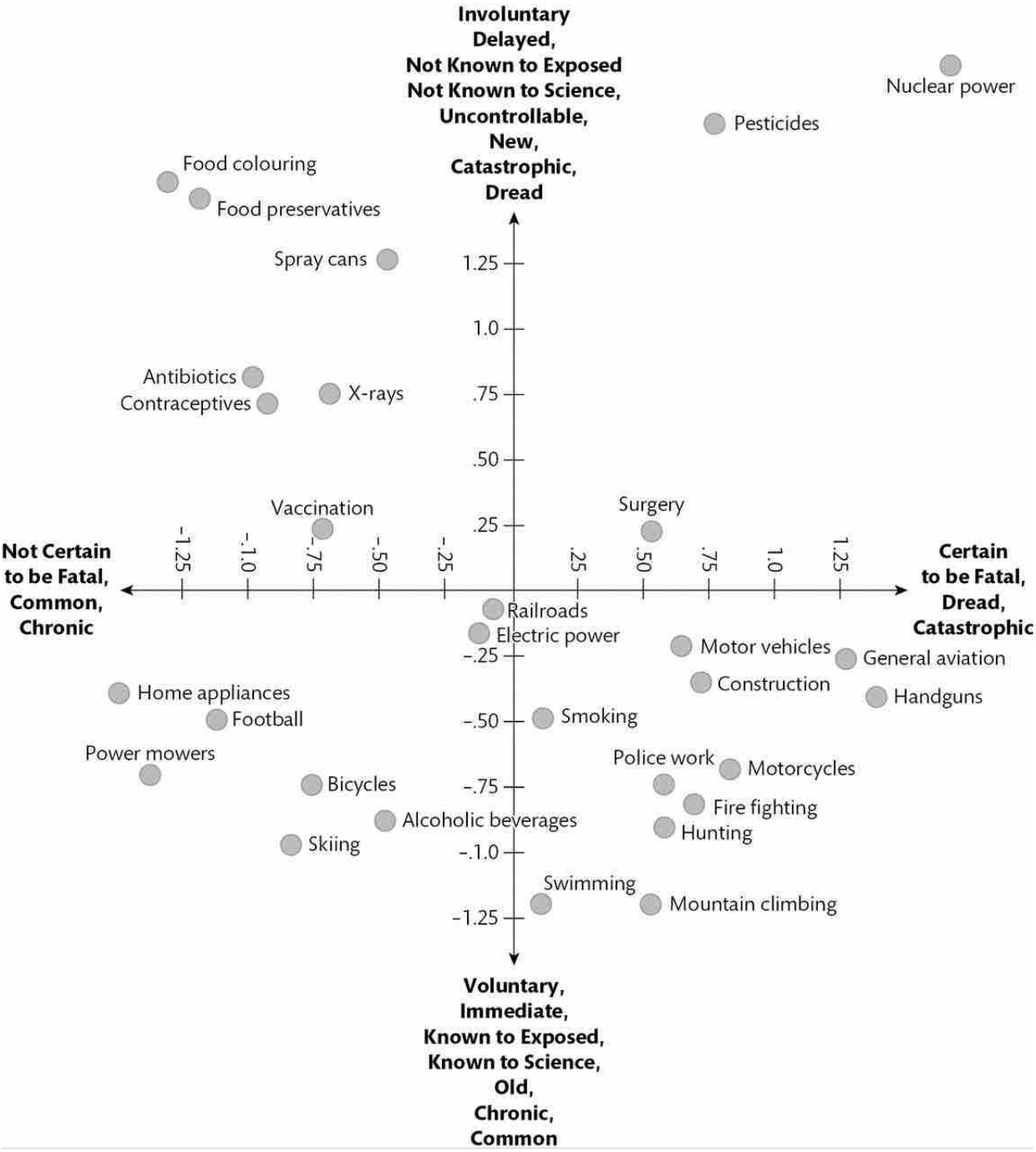


Fig. 7.5.2 Dimensions of risk.

Comparing risks

The multidimensionality of ‘risk’ belies simple comparisons, when people wish to determine which risks are most important or whether risks are being managed consistently. For example, concern about terrorism might be exaggerated given the number of fatalities, but not given its catastrophic potential or the sense of dread that it can induce. Thus, any comparison requires agreement on what ‘risk’ means.

One ambitious effort to develop socially acceptable definitions of risk was undertaken by the US Environmental Protection Agency (USEPA 1993). Following consultation with its staff and Scientific Advisory Board, it conducted or facilitated dozens of regional, state, and national ‘risk-ranking exercises’, in which participants identified the dimensions important to them, then deliberated priorities, supported by technical staff providing needed evidence. Letting participants choose the dimensions made these exercises more relevant to their needs, but reduced comparability across exercises. Florig et al. (2001) developed a method for standardizing such comparisons, based on the risk dimensions research. The UK government employed a variant (HM Treasury 2005).

The US Food and Drug Administration’s Voice of the Patient initiative (https://www.fda.gov/forindustry/userfees/prescriptiondruguserfee/ucm368342.htm) created a structured methodology for soliciting input from diverse stakeholders for determining the features that it should consider when evaluating the risks and benefits of proposed treatments. The US National Institute of Health’s PROMIS initiative (http://www.healthmeasures.net/explore-measurement-systems/promis) offers a standard set of (free) patient-reported outcomes measures that could be used to characterize risks. The related PROPr project has developed ways to weight the outcomes that decision-makers choose to consider (Dewitt et al. 2018).

Qualitative assessment

Event definitions

Once adequately defined, ‘risk’ can be estimated. With sexually transmitted infections (or pregnancy), that might mean specifying such details as

frequency and timing of intercourse, contraceptives used, and partners' physical state. Two experts with different definitions may see the same data and produce different estimates. So might laypeople asked for risk perceptions, but forced to guess exactly what is meant. For example, a prominent national survey asked: 'How likely do you think it is that a person will get the AIDS virus from sharing plates, forks, or glasses with someone who has AIDS?' After answering this question, US college students were asked what they had inferred about the kind and amount of sharing. Most agreed about the kind, with 82% selecting 'sharing during a meal' from a set of options. However, they disagreed about the frequency, with 39% selecting 'a single occasion', 20% 'several occasions', 28% 'routinely', and 12% uncertain (Fischhoff 1996). Respondents making different assumptions were, in effect, answering different questions, whose meaning researchers must guess, if they are to offer any conclusions about lay risk perceptions.

Laypeople are, similarly, left guessing when experts communicate about risks ambiguously. For example, McIntyre and West (1992) found that teens knew that 'safe sex' was important, but disagreed about what it entailed. Downs et al. (2004) found that teens interpret 'it can only take once' as meaning that they will get pregnant after having sex once. If they do not, some infer that they are infertile, encouraging unsafe sex. The best strategy for ensuring that people are talking about the same thing is empirical study, as discussed next.

Supplying details

The details that people infer, when given ambiguous and incomplete risk questions or messages, reveal their intuitive theories. For example, bleeding or fluid loss during pregnancy can have serious health implications. However, some spotting or light bleeding is common and normal. Women who experience spotting may worry unduly whereas women who have experienced spotting in the past with no ill-effects, may not realize the significance of heavier bleeding. They might attribute normal, light bleeding in pregnancy to menstruation, leading them to misattribute abnormal heavier bleeding to a routine reproductive process. Krishnamurti et al. (2017) explicitly addressed this incomplete mental model with a real-time symptom tracker designed to help women distinguish harmless from harmful bleeding.

Cumulative risk—a case in point

There is no full substitute for directly studying the beliefs that people bring to and take away from communications, especially when they come from backgrounds different than those of the communicators (Medin et al. 2017). However, the research literature provides a basis for anticipating those beliefs (Fischhoff et al. 2011). For example, unrealistic optimism is so widespread that one can assume that people see themselves as facing less risk than other people, whenever some personal control seems feasible (Shepherd et al. 2015).

Just as teens' insensitivity to the amount of sex, when judging STI risks, reflects a well-known insensitivity to how small risks accumulate over repeated exposures, people cannot be expected to infer the cumulative accident risk from driving without a seat belt (Slovic et al. 1978), or the pregnancy risk from having sex without effective contraceptives (Shaklee and Fischhoff 1990). A corollary is not realizing the cumulative impact of small differences in single-exposure risks (e.g. slightly better contraceptives, wearing a seat belt). For example, college students estimated the probability of male-to-female HIV transmission, as 10% in one case of protected sex and 25% in 100 cases. Given then-current estimates, the former was much too high and the latter fairly accurate. That inconsistency was only revealed by asking about both. The more relevant question for predicting behaviour depends on whether people think about single or multiple exposures (Fischhoff 1996). People similarly underestimate exponential growth, meaning that they may be able to estimate a disease's basic reproduction number (R0) without appreciating how quickly it will spread (Fischhoff et al. 2018).

Mental models of risk processes

The role of mental models

As mentioned, when people lack explicit information about the magnitude of a risk (or benefit), they must infer it. Judgement heuristics, like availability, provide one class of inferential rules for deriving specific estimates from general knowledge. A second class draws on individuals' *mental models* of the general processes that create and control risks. Mental models have a long history in psychology, having been studied for topics as diverse as how

people understand physical processes, international tensions, complex equipment, energy conservation, climate change, interpersonal relations, and drug effects (Ericsson and Simon 1993; Morgan et al. 2001). They are how people navigate their world.

Mental models can fail people when their premises are faulty or incomplete, or when inferential processes are flawed. For example, not realizing how quickly the risks of pregnancy and STIs accumulate could make other knowledge seem irrelevant. In a study of domestic radon, Bostrom et al. (1992) found that people often knew that it was a colourless, odourless, radioactive gas. However, they overestimated the risks and underestimated the opportunities for remediation, because they thought that radioactivity meant permanent contamination—not knowing that the radon by-products (or 'progeny') have short half-lives. As a result, once radon intrusions stop, the problem disappears. Yet, while it persists, rapid decay means rapid energy release. Homeowners unaware of these facts might reasonably decide not to test for radon—the problem doesn't seem urgent and there is nothing to do anyway if they find it.

Morgan et al. (2001) offer a general approach appropriate to studying mental models for complex, uncertain processes, like those of many public health risks. It begins by creating a formal (or 'expert') model, summarizing relevant scientific knowledge, with enough precision to compute quantitative predictions, were its data needs met (Fischhoff et al. 2006). A common formalism is the influence diagram, a directed graph with nodes, representing variables, and connecting arrows, representing relationships between them.

The research continues with open-ended, one-on-one interviews, structured around the model, eliciting lay beliefs in their intuitive formulation. Those 'mental model' interviews begin with general questions, asking respondents what they believe about the topic, then to elaborate on each issue that they raise. The tone is non-judgemental, seeking to understand, not evaluate respondents' perspectives. After exhausting responses to general questions, interviewers ask increasingly pointed ones, starting with general processes (e.g. exposure, effects, mitigation), and proceeding to specific issues (e.g. 'How does the amount of sex [or number of partners] affect HIV risk?'; 'What does "safe sex" mean?'). A variant has people think aloud while sorting photographs by their relevance, hoping for insights into topics that were otherwise missed. For example, seeing a

supermarket produce section prompted some respondents to say that radon might contaminate plants (Bostrom et al. 1992). The interviews continue until no new material emerges, achieving thematic saturation.

Once transcribed, interviews are coded into the formal model, adding new elements raised by respondents, marked as either misunderstandings or lay expertise (e.g. how equipment really works). The precision of the formal model typically allows reliable coding. Once mapped, lay beliefs can be analysed in terms of their accuracy, relevance, and specificity. For example, Bostrom et al. (1992) found that most respondents, drawn from civic groups, knew that radon is a gas (88%), which concentrates indoors (92%), is detectable with a test kit (96%), comes from underground (83%), and can cause cancer (63%). However, many believed erroneously that radon affects plants (58%), contaminates blood (38%), and causes breast cancer (29%). Few (8%) mentioned that radon decays. The interviews led to a structured survey suited to assessing the prevalence of beliefs in larger samples, with questions having ecological validity, in the sense of sampling the key topics in the formal model (Bruine de Bruin et al. 2007).

This overall approach has been successfully applied in numerous health contexts including STIs, sexual assault, emergency contraception use, Lyme disease, falls, breast cancer, vaccination, and infectious disease, as well as topics in many other domains, ranging from wireless technologies to nuclear energy sources in space (Fischhoff 1992; Morgan et al. 2001; Krishnamurti et al. 2008; Downs et al. 2008; Krishnamurti et al. 2012).

From risk beliefs to risk decisions

The adequacy of risk perceptions depends on the decisions that depend on them. Some decisions require precise estimates, others just a rough idea. For example, von Winterfeldt and Edwards (1986) showed that many decisions with continuous options (e.g. invest $X) are insensitive to the precise values assigned to the probabilities and utilities of possible outcomes. Dawes et al. (1989) showed that choices with discrete options (e.g. choosing graduate candidates) are often insensitive to exactly how predictors are weighted, meaning that simple linear (weighted-sum) models may do as well as more complicated ones. That result has good news for researchers hoping to predict behaviour. They should have some success with any model that includes the issues that matter to people or variables correlated with them. It

has bad news for researchers hoping to explain or change behaviour. Many models with different variables, implying different psychological processes, will have similar predictive success. As a result, one cannot tell which explains the predicted behaviour. The health-belief model is such weighted-sum model. It structures diverse topics in common terms, but must be supplemented by studies showing how people think about and respond to each part of their world. The greater the social distance between the experts and the public, the more important such studies become—lest experts miss options, uncertainties, or outcomes that occupy lay decision-makers, but would never occur to them, or vice versa.

As with any conversation, eliciting health beliefs and attitudes requires posing questions that people can answer in ways that are useful to the researcher. Because structured surveys offer limited feedback on differences on how respondents conceptualize a domain, there is no substitute for including open-ended procedures that allow people to express themselves in their own terms. For example, Beyth-Marom et al. (1993) found that teens think about the possible consequences of risky decisions (e.g. drinking and driving, smoking marijuana, skipping school) differently when framed in terms of doing and not doing them, even though the two are complementary actions. In that study, thinking about accepting risky options produced more consequences (suggesting that action is more evocative than inaction), a higher ratio of bad to good consequences (suggesting that action makes risks more available), and fewer references to social consequences (suggesting that social pressure is more salient when resisting temptation than when yielding to it). When participants were asked about making choices repeatedly, rather than just once, they often produced different consequences (e.g. repeatedly ‘accepting an offer to smoke marijuana at a party’ evoked more mentions of social reactions than did thinking about doing it once). Parents of the teens cited similar outcomes, except for more often mentioning long-term consequences (e.g. ruining career prospects). Thus, if parents and teens see these choices differently, it is not because they see different possible outcomes, but because they disagree about how likely and important those outcomes are. Such differences would be hidden with structured surveys that elicit ratings of fixed consequences, framed in just one way.

In a study imposing even less structure, teens chose difficult personal decisions, which they described in their own words (Fischhoff 1996). These

descriptions were coded by content (which choices are on teens' minds) and structure (how they are formulated). For example, these teens never mentioned decisions about drinking and driving, but often described decisions about drinking, suggesting that the latter decisions are a better focus for interventions. These teens typically focused on a single option (e.g. whether to attend a party with drinking), suggesting the value of helping teens to feel that they have options and to create attractive ones (e.g. socially acceptable ways to avoid excess alcohol). Seeing different ways to accept and reject risks should also reduce insensitivity to opportunity costs (foregone benefits) due to neglecting less visible options (Thaler 1991). For example, the direct risks of vaccinating children can loom much larger than the indirect risks of not vaccinating them (Ritov and Baron 1990).

Different methods for eliciting decision-makers' perspective have different, often complementary strengths and weaknesses (Ericsson and Simon 1993). Structured methods (e.g. surveys) allow efficient, readily analysed data collection, but increase the risk of miscommunication between researchers and respondents. Open-ended methods (e.g. mental models interviews) hear whatever is on respondents' minds, but increase the risk of interviewers sending non-verbal cues that influence what is said. Combined methods can provide rounded pictures, especially when grounded in shared formal analyses of the decision. One method that is common in commercial research, but rare in scientific studies, is focus groups. Indeed, their inventor, Robert Merton (1987) rejected them as sources of evidence (rather than just ideas), given the unnatural discourse of even well-moderated group. He preferred *focused interviews*, akin to mental models interviews, designed to hear individuals' views in depth.

Creating communications

Selecting information

Communication design begins by selecting content. The gold standard is a normative analysis, identifying the information most relevant to the choices that the communication is meant to inform (Fischhoff 2013). In practice, though, the content-selection process is often ad hoc, with experts intuiting 'what people ought to know'. That process can run afoul of the 'common knowledge effect', whereby people exaggerate how much goes without

saying, hence fail to provide essential details (Nickerson 1999). Poorly selected information can waste recipients' time, repeat relevant content, bury facts that people need to know among ones that are nice to know, and erode recipients' faith by showing insensitivity to their needs. It can also undermine experts' faith in their audience, if they fail to realize why their messages have missed the mark or asked recipients to do the impossible (Fischhoff 2011).

The logic of setting information priorities is straightforward: first provide the facts that will have the greatest impact, if properly understood. One place where information can overwhelm recipients is with informed consent procedures for medical treatments or trials. Merz et al. (1993) demonstrate a prioritization procedure using 'value of information analysis' (von Winterfeldt and Edwards 1986), finding that few of the many potential risks of carotid endarterectomy should matter to many patients. Krishnamurti and Argo (2016) demonstrated and validated another procedure, asking potential participants to identify the information that they most want to know. Although nothing should be hidden, communications should be sure to get the few key facts across.

At times, people face no specific choice, but just want to understand processes that could pose many decisions (e.g. new diseases, food-borne illnesses). The same logic of prioritization applies. Communications should focus on information that would be most useful for predicting the most important events (e.g. critical health signs, key factors in food safety). That information completes the mental models needed to monitor the environment, generate options, and follow discussions (Morgan et al. 2001). For example, Downs et al. (2004) found that most teens knew enough about HIV/AIDS to focus communications on a few critical gaps, such as how risks mount through repeated exposure and how hard it is for sexual partners to self-diagnose their own disease status.

An essential part of the content of any communication is the strength of the evidence supporting it (O'Hagan et al. 2006; Politi et al. 2007). Campbell (2011) shows ways to represent uncertainty graphically. Schwartz and Woloshin (2013) show how to describe the data quality (e.g. the length, size, and quality of clinical trials). Funtowicz and Ravetz (1990) show how to characterize the quality of the underlying science, including its *pedigree* (e.g. how well empirical patterns are supported by theory). Fortney (1988) found that oral contraceptives may increase a non-smoking woman's life

expectancy by up to 4 days and decrease it by up to 80 days, drawing on a research base so large that no conceivable study could materially change those bounds.

Formatting information

Once selected, information must be presented. There are many summaries of the relevant research (e.g. Wogalter 2006; Fischhoff et al. 2011). They note, for example, that comprehension improves when the text: (1) has a clear structure, corresponding to recipients' intuitive representation; (2) has a clear information hierarchy; and (3) provides *adjunct aids*, such as highlighting, advanced organizers (showing what to expect), and summaries.

However, even scientifically established design principles provide but a point of departure for communication design. Although better 'best guesses' than mere intuition, their success in any specific application is an empirical question. Standard usability testing procedures include seeing how long users take to find needed information, how often they reach the wrong information, and how often they realize that (Wogalter 2006). When there is variability in search patterns, designers need to decide which users' needs are most pressing. In a demonstration study, Riley et al. (2001) showed how users of dangerous consumer product (methylene chloride-based paint stripper) require widely varying efforts to reach safety information with different labels, then on the market. With some brands, they might not reach it at all.

Over the past 20 years, technology-based risk communications have become an integral part of public health (Eysenbach 2000). Increasingly, lay people and healthcare providers are turning to social media and digital health information technologies (e.g. patient and provider portals, weblogs, instant messaging platforms, social networks, and electronic consent) to monitor and share health risk information. There are currently over 300,000 commercially available mobile health apps for smartphones. These new platforms for health information exchange have great promise. Information can be presented in multiple non-text formats, making it more accessible to users with special needs and learning styles. Their potential for dynamic, interactive displays could allow tailoring information, targeting its content or format based on background knowledge, literacy, or numeracy. Increasingly ubiquitous access to the internet and smartphones makes it possible to reach

wider, more geographically diverse audiences. Predictive modelling of risk behaviour has the potential to further target communications. In a systematic review of social media use for health communication, Moorhead et al. (2013) also notes the data security risks with such communications.

Mohan et al. (2017) report a randomized controlled trial finding that online ‘serious game’ training conveyed risk information to emergency department physicians more effectively that conventional text-based training. To date, though, few of the many online applications have any published empirical evaluations. Each poses challenges of User-Interface (UI) and User-Experience (UX) design. The former focuses on the visual content and interactive elements; the latter focuses on the user’s enjoyment and engagement. Successful design can be informed by basic research on UI and UX. However, it still needs iterative testing with potential users, under conditions approaching actual use. One particular design risk is sacrificing effectiveness (and even accuracy) in pursuit of increased engagement. A second risk is that the increased the bandwidth demands of more engaging designs will disenfranchise those with limited access.

Evaluating communications

One should no more release an untested health communication than an untested drug. Indeed, communications are part of any medical product or procedure, affecting when it is chosen, how it is used, and whether problems are noticed in time to be remedied and reported. Arguably, evidence about such communications should be part of the submissions when products face regulatory approval. One standard for such evaluations to meet is that:

A communication is *adequate* if it:

Includes the information that recipients need, in order to make their decisions.
Affords them ready access to that information, given their normal search patterns.
Allows them to comprehend that information, with a reasonable effort.

Applying these tests requires evidence. Knowing what information people need requires learning their goals. Knowing whether people can find the information that is there requires observing how they search. Knowing how

much they comprehend requires seeing how well they have mastered the content.

Publication-quality evaluation requires resources and the training acquired by implementation researchers. However, simple forms of evaluation are within the reach of any communicator. In the US Food and Drug Administration *Communicating Risks and Benefits: An Evidence-Based User's Guide* (Fischhoff et al. 2011), each chapter ends with a section on how to conduct evaluations at no cost, small cost, or cost commensurate with the health, political, and economic stakes riding on effective communication. Central to any evaluation is listening, without presuming to know recipients' goals, beliefs, uncertainties, emotions, or modes of expression. In order to identify individuals' information needs, one can ask how they see the risks in their lives. In order to see how easily people can access information, one can watch as they search for it (e.g. online, in drafts; Downs et al. 2008). In order to assess a communication's comprehensibility, one can ask people to recall it, paraphrase it, make inferences from it, or create scenarios using it (Bruine de Bruin et al. 2007).

These are structured ways of talking about technical topics of mutual interest. As with any successful conversation, they must signal mutual respect. That means being able to hear unexpected responses. Even a few open-ended, one-on-one interviews might reveal incomprehensible or offensive material, or prompt questions about omissions. It also means framing those formative evaluations as testing the message, not its potential recipients. If they have not learned facts that matter to them, then the message must not have gotten that information across. Only when the resources of scientific communication have been exhausted should it be assumed that laypeople cannot learn what they need. The stakes riding on lay decision-making should justify that investment and humility. Amateurish communications can be worse than nothing, holding recipients responsible, or denying them the right, for decision-making, without giving them a fair chance.

The science of communication can guide *persuasive communications*, designed to induce behaviour that follows the communicators' wishes; *non-persuasive communications*, designed to help people make independent choices; and the *libertarian paternalist* hybrid where tasks are framed so that people voluntarily make decisions that someone has determined is 'for their own good' (Thaler and Sunstein 2009). Without studying people's goals,

however, one risks imposing experts' views on them. For example, Bostrom et al. (1992) found people who refused to follow communications that advocated testing for radon because they wanted to avoid creating evidence that could complicate selling their homes. Fischhoff (1992) describes the conflicting advice on reducing the risk of sexual assault, reflecting differences in experts' beliefs about women's (goals and the effectiveness of defence strategies). Reasonable individuals may 'defeat' safety measures by gaining more benefit from a product (e.g. driving faster with cars that handle better), frustrating experts concerned solely with safety.

Managing communication processes

Accomplishing these tasks requires four kinds of expertise:

Subject matter experts, who can identify the processes that create and control risks (and benefits).
Analysts, who can estimate the risks (and benefits) most pertinent to decision-makers.
Behavioural scientists, who can assess decision-makers' beliefs and goals, guide a communication's design, and evaluate its success.
Practitioners, who can create communication products and channels, for getting messages to audiences and obtaining feedback from them.

The work of these experts must be coordinated, so that they play appropriate roles. For example, behavioural scientists should not revise text, in an effort to improve its comprehensibility, without having subject matter experts check that the content is still accurate. Subject matter experts should not slant the facts according to their pet theories of how alarmed or calmed the public needs to be. Practitioners should not 'spin' health facts, the way that they might try to manipulate political agendas.

As noted, it is critical to involve target audience members in message and channel development. In the US, the Patient Centered Outcomes Research Institute (PCORI) was established to fund health research guided by the patients whom that research aims to serve (https://www.pcori.org/). It placed target audience involvement as central to its mission, starting from determining where better risk communication was critical, through to design and dissemination.

Figure 7.5.1 depicts the kind of two-communication needed to ensure that public health experts know how members of the public view their decisions and keep them informed of their actions. Creating such structured processes is a task for communication practitioners. They may need, for example, to interest members of the general public in topics far outside their current concerns, ascertain whether interest groups actually speak on behalf of their claimed constituencies, or empower those who might otherwise be neglected.

Conclusion

Effective risk communication is essential to health decision-making. Without it, individuals are denied the best possible chance of making sound choices —before, during, and after problems arise. As a result, they may suffer avoidable injury, while incurring the insult of feeling that they have failed to protect themselves or their loved ones. They may also feel anger at the experts have let them down, by not creating needed information and disseminating it in a timely, comprehensible way. As a result, communication must be central to how public health risks are studied and managed. It cannot be just an afterthought, letting the public know what the experts have learned or decided on its behalf. As depicted in Figure 7.5.1, it requires continuing, two-way communication.

This chapter has focused on general methods for assessing and addressing risk communication needs. It has avoided sweeping generalizations about what ‘people do’, ‘people believe’, and ‘people want’ That is because critical details vary across risk decisions and decision-makers. Responsive communication requires attending to the details of individuals’ lives and decisions. The basic science of risk perception and communication presented here could support research programmes dedicated to communicating the science presented in many chapters in the *Oxford Textbook of Public Health*, ensuring that the public gets full value from that science. Although the subject matter of each such research programme would be distinct, all could draw on the same basic research and methods. Given a well-characterized decision or risk, it is relatively straightforward, if technically demanding, to assess risk perceptions. If decision-makers’ risk (and benefit) perceptions have been measured well, then their choices can often be roughly predicted with simple linear models (Dawes et al. 1989). More precise prediction

requires more detailed understanding of the cognitive processes and mental models shaping these beliefs, and of the emotional, social, and economic factors impinging on the decision.

Meeting the challenge of effective risk communication requires coordinating four kinds of experts: subject matter specialists, analysts, behavioural scientists, and communication practitioners. Assembling such teams requires strategic leadership, which sees communication as essential to the public health mission. Communication research is inexpensive, relative to the stakes riding on sound decision-making, for individuals and for the public health organizations meant to serve them. Thus, there is no good reason not to assess risk perceptions and evaluate risk communications, drawing on the science summarized here. Such a scientific approach to risk communication would, for example, avoid the known pitfalls of ignoring the multidimensional character of 'risk', the problems with verbal quantifiers, or the need to help people see how risks mount up through repeated exposure. With the novel, complex, uncertain, fateful issues that can arise with health risk decisions, one cannot trust to intuition, when trying to meet decision-makers' needs.

Better risk communication should help recipients to make better decisions. However, it need not make their decisions easier, if it places hard choices in stark relief. Nor need it make communicators' lives easier, as recipients may discover bona fide disagreements. However, it should avoid needless misunderstanding, leading to fewer but better conflicts.

ACKNOWLEDGEMENT

The preparation of this chapter was supported by the Center for Climate and Energy Decision Making (SES-0949710) through a cooperative agreement between the National Science Foundation and Carnegie Mellon University and the Swedish Foundation for Humanities and Social Sciences and by an institutional training award from the NIH (KL2TR001856). The views expressed are the authors'.

REFERENCES

Beyth-Marom, R., et al. (1993). Perceived consequences of risky behaviors. *Developmental Psychology*, 29, 549–63.

Black, W.C., Nease, R.F., and Tosteson, A.N.A. (1995). Perceptions of breast cancer risk and screening effectiveness in women younger than 50 years of age. *Journal of the National Cancer Institute*, 8, 720–31.

Bostrom, A., Fischhoff, B., and Morgan, M.G. (1992). Characterizing mental models of hazardous processes: a methodology and an application to radon. *Journal of Social Issues*, 48(4), 85–100.

Breakwell, G.M. (2014). *The Psychology of Risk* (2nd ed.). Cambridge: Cambridge University Press.

Bruine de Bruin, W., Downs, J.S., Fischhoff, B., and Palmgren, C. (2007). Development and evaluation of an HIV/AIDS knowledge measure for adolescents focusing on misconceptions. *Journal of HIV/AIDS Prevention in Children and Youth*, 8(1), 35–57.

Bruine de Bruin, W., et al. (2000). Expressing epistemic uncertainty: it's a fifty-fifty chance. *Organizational Behavior and Human Decision Processes*, 81, 115–31.

Budescu, D.F. and Wallsten, T.S. (1996). Processing linguistic probabilities. In: Busemeyer, J.R., Hastie, R., and Medin, D.L. (eds.) *Decision Making from a Cognitive Perspective*. New York: Academic Press, pp. 275–318.

Campbell, P. (2011). Understanding the receivers and the reception of science's uncertain messages. *Philosophical Transactions of the Royal Society*, 369, 4891–912.

Christensen-Szalanski, J. and Bushyhead, J. (1993). Physicians' misunderstanding of medical findings. *Medical Decision Making*, 3, 169–75.

Dawes, R.M., Faust, D., and Meehl, P. (1989). Clinical versus actuarial judgment. *Science*, 243, 1668–74.

Dewitt, B., et al. (2018). Estimation of a preference-based summary score for the Patient-Reported Outcomes Measurement Information System: the PROMIS®-Preference (PROPr) Scoring System. *Medical Decision Making*, *38*(6), 683–98.

Downs, J.S., Bruine de Bruin, W., and Fischhoff, B. (2008). Patients' vaccination comprehension and decisions. *Vaccine*, 26, 1595–607.

Downs, J.S., et al. (2004). An interactive video program to reduce adolescent females' STD risk: a randomized controlled trial. *Social Science and Medicine*, 59, 1561–72.

Erev, I. and Cohen, B.L. (1990). Verbal versus numerical probabilities: efficiency, biases and the preference paradox. *Organizational Behavior and Human Decision Processes*, 45, 1–18.

Ericsson, K.A. and Simon, H.A. (1993). *Verbal Reports as Data*. Cambridge, MA: MIT Press.

Eysenbach, G. (2000). Consumer health informatics. *British Medical Journal*, 320(7251), 713–1716.

Fischhoff, B. (1992). Giving advice: decision theory perspectives on sexual assault. *American Psychologist*, 47, 577–88.

Fischhoff, B. (1996). The real world: what good is it? *Organizational Behavior and Human Decision Processes*, 65, 232–48.

Fischhoff, B. (2011). Communicating the risks of terrorism (and anything else). *American Psychologist*, 66, 520–31.

Fischhoff, B. (2013). The sciences of science communication. *Proceedings of the National Academy of Sciences*, 110(Supplement 3), 14033–9.

Fischhoff, B., Brewer, N.T., and Downs, J.S. (eds.) (2011). *Communicating Risks and Benefits: An Evidence-Based User's Guide*. Washington, DC: US Food and Drug Administration.

Fischhoff, B., et al. (2006). Analyzing disaster risks and plans: an avian flu example. *Journal of Risk and Uncertainty*, 33, 133–51.

Fischhoff, B., et al. (2018). Public understanding of Ebola risks: mastering an unfamiliar threat. *Risk Analysis*, 38, 71–83

Fischhoff, B. and Kadvany, J. (2011). *Risk: A Very Short Introduction*. Oxford: Oxford University Press.

Florig, H.K., et al. (2001). A deliberative method for ranking risks. *Risk Analysis*, 21, 913–22.

Fortney, J. (1988). Contraception: a life long perspective. In: *Dying for Love*. Washington, DC: National Council for International Health, pp. 33–8.

Funtowicz, S.O. and Ravetz, J. (1990). *Uncertainty and Quality in Science for Policy*. London: Kluwer.

Glassman-Fox, K.T. and Weber, E.U. (2016). What makes risks acceptable? Revisiting the 1978 psychological dimensions of risk. *Journal of Mathematical Psychology*, 75, 157–69.

Hawley, S.T., et al. (2008). The impact of the format of graphical presentation on health-related knowledge and treatment choices. *Patient Education and Counselling*, 73(3), 448–55.

HM Treasury (2005). *Managing Risks to the Public*. London: HM Treasury.

Krishnamurti T, Eggers SL, Fischhoff B. (2008). The impact of over-the-counter availability of Plan B on teens' contraceptive decision making. *Social Science &Medicine*, 67(4), 618–27.

Krishnamurti, T., et al. (2012). Preparing for smart grid technologies: a behavioral decision research approach to understanding consumer expectations about smart meters. *Energy Policy*, 41, 790–7.

Krishnamurti, T. and Argo, N. (2016). A patient-centered approach to informed consent: results from a survey and randomized trial. *Medical Decision Making*, 36(6), 726–40.

Krishnamurti, T., et al. (2017). Development and testing of the MyHealthyPregnancy app: a behavioral decision research-based tool for assessing and communicating pregnancy risk. *JMIR mHealth and uHealth.* 5(4), e42.

Lerner, J.S. and Keltner, D. (2001). Fear, anger, and risk. *Journal of Personality and Social Psychology*, 81, 146–59.

Lichtenstein, S. and Fischhoff, B. (1980). Training for calibration. *Organizational Behavior and Human Performance*, 26, 149–71.

Lichtenstein, S., et al. (1978). Judged frequency of lethal events. *Journal of Experimental Psychology: Human Learning and Memory*, 4, 551–78.

Löfstedt, R., Fischhoff, B., and Fischhoff, I. (2002). Precautionary principles: general definitions and specific applications to genetically modified organisms (GMOs). *Journal of Policy Analysis and Management*, 21, 381–407.

McIntyre, S. and West, P. (1992). What does the phrase 'safer sex' mean to you? Understanding among Glaswegian 18 year olds in 1990. *AIDS*, 7, 121–6.

Medin, D., Ojalehto, B., Marin, A., and Bang, M. (2017). Systems of (non-)diversity. *Nature Human Behaviour*, 1. doi:10.1038/s41562-017-0088.

Mellers, B., et al. (2015). Identifying and cultivating superforecasters as a method of improving probabilistic predictions. *Perspectives on Psychological Science*, 10(3), 267–81.

Merton, R.F. (1987). The focussed interview and focus groups. *Public Opinion Quarterly*, 51, 550–66.

Merz, J., et al. (1993). Decision-analytic approach to developing standards of disclosure for medical informed consent. *Journal of Toxics and Liability*, 15, 191–215.

Mohan, D., et al. (2017). Testing the efficacy of a video game vs. a traditional education program at improving physician decision making in trauma triage: a randomized controlled trial. *British Medical Journal*, 359, j5416.

Moorhead, S.A., et al. (2013). A new dimension of health care: systematic review of the uses, benefits, and limitations of social media for health communication. *Journal of Medical Internet Research*, 15(4), e85.

Morgan, M.G. (2014). Use (and abuse) of expert elicitation in support of decision making for public policy. *Proceedings of the National Academy of Sciences*, 111(20), 7176–84.

Morgan, M.G., et al. (2001). *Risk Communication: The Mental Models Approach*. New York: Cambridge University Press.

National Research Council (1996). *Understanding Risk: Informing Decisions in a Democratic Society*. Washington, DC: National Academy Press.

Nickerson, R.A. (1999). How we know—and sometimes misjudge—what others know: imputing our own knowledge to others. *Psychological Bulletin*, 125, 737–59.
O'Hagan, A., et al. (2006). *Uncertain Judgements: Eliciting Expert Probabilities*. Chichester: Wiley.
Parker, A.M., Bruine de Bruin, W., Fischhoff, B., and Weller, J. (2018). Robustness of decision-making competence: evidence from two measures and an 11-year longitudinal study. *Journal of Behavioral Decision Making*, 31, 380–91.
Phillips, J.K., Ford, M.A., and Bonnie, R.J. (eds.). (2017). *Pain Management and the Opioid Epidemic: Balancing Societal and Individual Benefits and Risks of Prescription Opioid Use*. Washington (DC): National Academies Press (US).
Politi, M.C., Han, P.K.J., and Col. N. (2007). Communicating the uncertainty of harms and benefits of medical procedures. *Medical Decision Making*, 27, 681–95.
Poulton, E.C. (1989). *Bias in quantifying judgment*. Hillsdale, NJ: Lawrence Erlbaum.
Quadrel, M.J., Fischhoff, B., and Davis, W. (1993). Adolescent (in)vulnerability. *American Psychologist*, 48, 102–16.
Riley, D.M., et al. (2001). Evaluating the effectiveness of risk-reduction strategies for consumer chemical products. *Risk Analysis*, 21, 357–69.
Ritov, I. and Baron, J. (1990). Status quo and omission bias. Reluctance to vaccinate. *Journal of Behavioral Decision Making*, 3, 263–77.
Schwartz, L.M. and Woloshin, S. (2013). The drug fact box: improving the communication of medical information. *PNAS*, 110, 14069–74.
Schwarz, N. (1999). Self reports. *American Psychologist*, 54, 93–105.
Shaklee, H. and Fischhoff, B. (1990). The psychology of contraceptive surprises: judging the cumulative risk of contraceptive failure. *Journal of Applied Psychology*, 20, 385–403.
Shepherd, J.A., Waters, E., Weinstein, N.D., and Klein, W.M.P. (2015). A primer on unrealistic optimism. *Current Directives in Psychological Science*, 24(3), 232–7.
Slovic, P. (1987). Perception of risk. *Science*, 236(17), 280–5.
Slovic, P. (ed.) (2001). *Perception of Risk*. London: Earthscan.
Slovic, P. (ed.). (2010). *The Feeling of Risk: New Perspectives on Risk Perception*. New York: Earthscan.
Slovic, P., Fischhoff, B., and Lichtenstein, S. (1978). Accident probabilities and seat-belt usage: a psychological perspective. *Accident Analysis and Prevention*, 10, 281–5.
Starr, C. (1969). Societal benefit versus technological risk. *Science*, 165, 1232–8.
Studdert, D.M., et al. (2017). Handgun acquisitions in California after two mass shootings. *Annals of Internal Medicine*, 166(10), 698–706.
Thaler, R. (1991). *Quasi-Rational Economics*. New York: Russell Sage Foundation.
Thaler, R. and Sunstein, C. (2009). *Nudge: Improving Decisions about Health, Wealth and Happiness*. New Haven, CT: Yale University Press.
USEPA (1993). *A Guidebook to Comparing Risks and Setting Environmental Priorities*. Washington, DC: USEPA.
Viscusi, K. (1992). *Smoking: Making the Risky Decision*. New York: Oxford University Press.
von Winterfeldt, D. and Edwards, W. (1986). *Decision Analysis and Behavioral Research*. New York: Cambridge University Press.
Wogalter, M., ed. (2006). *The Handbook of Warnings*. Hillsdale, NJ: Lawrence Erlbaum Associates.
Woloshin, S., et al. (1998). Scales for assessing perceptions of event probability: a validation study. *Medical Decision Making*, 14, 490–503.
Yonkers, K. A., Blackwell, K. A., and Forray, A. (2014). Antidepressant use in pregnant and postpartum women. *Annual Review of Clinical Psychology*, 10, 369–92.

7.6

Urbanization and health

Jason Corburn

Introduction

The majority of the world's population (4.2 billion) are now living in cities and municipal regions. According to the UN, 55% of the world was living in cities in 2018 and over 68% were expected to live in urban areas by 2050 (UN 2018). Urbanization is a dynamic and evolving physical, social and economic transformation that shapes the health and well-being of populations living in cities and around the world. City living can be healthy, since they can offer more population groups the health benefits of life-supporting infrastructure such as clean water and sanitation, education, and social services, as well as greater cultural, religious, and political expression and freedoms (Dye 2008). This holds true in both rich and poor cities of the global North and South. Thus, where we live, learn, work, and play, sometime referred to as our postal code, can matter as much or more than our genetic code for determining whether one will get sick, suffer unnecessarily, and even die prematurely.

Yet, not everyone living in cities can take advantage of the urban resources that can promote well-being. The increasing number of the urban poor and socially marginalized often experience health inequities, or differences in access to health-promoting resources that are unnecessary, avoidable, and unfair. As the UN Habitat and World Health Organization (WHO) stated in their 2010 report, *Hidden Cities: Unmasking and Overcoming Health Inequities in Urban Settings*:

> Health inequities are the result of the circumstances in which people grow, live, work and age, and the health systems they can access, which in turn are shaped by broader political, social and economic forces. They are not distributed randomly, but rather show a consistent pattern across the population, often by socioeconomic status or geographical location. No city—large or small, rich or poor, east or west, north or south—has been shown to be immune to the problem of health inequity. (WHO and UN Habitat 2010)

Importantly, the almost 1 billion now living in urban slums, informal settlements, or as 'pavement dwellers' in marginalized situations often have the worst health outcomes of all urban residents and their health is frequently worse than populations living in rural poverty (Lilford et al. 2017). The urban poor often experience inadequate housing and sanitation, lack of running water, poor or undernutrition, overcrowding, indoor and outdoor pollution, violence, limited access to healthcare, increased vulnerability to climatic events such as flooding and drought, and lack political power to influence policy decisions that can change their living conditions (Friel et al. 2011).

The links between global urbanization, inequities, and human well-being have been recognized by recent global agendas. The 2016 WHO/UN Habitat *Global Report on Urban Health* noted that national level wealth does not equate to health at the city level, conditions for health vary widely within the same city, and that cities must adopt a multisectoral response to promote health equity (WHO and UN Habitat 2016). The 2016 United Nations Conference on Housing and Sustainability (Habitat III) in Quito, Ecuador, also prioritized health and well-being through the release of the New Urban Agenda. Finally, the 2030 Sustainable Development Goals (SDGs) include a number of broad objectives that can only be met by addressing features of urbanization, such as poverty (SDG1), human health (SDG3), gender inequities (SDG5), water and sanitation (SDG6), decent work and economic growth (SDG8), reducing inequalities (SDG10) and climate action (SDG13). At least one goal, SDG 11, is dedicated to making cities inclusive, safe, sustainable, and resilient.

This chapter briefly reviews the historical debates around the connections between human health and urbanization and highlights some challenges for addressing twenty-first century urbanization. Twenty-first century urbanization presents new challenges for urban health (Galea et al. 2005). First, much rapid urbanization is happening in medium to small-size cities, not the mega-city growth the planet experienced through much of the twentieth century. These cities may not have the institutional structures or

national government support to plan for rapid growth and related health needs. Second, twenty-first century urbanization, particularly in sub-Saharan Africa, is not always accompanied by economic development, as was frequently the case in the nineteenth and twentieth centuries. Informal work and living conditions characterize much of planetary urbanization in the twenty-first century, and this can contribute to large portions of society being 'unseen' and 'unrecognized' by governments and international organizations interested in global sustainability and poverty alleviation. Third, and related to the second challenge, is that twenty-first century urbanization is characterized by rising and persistent social inequalities—poverty, discrimination, and spatial marginalization of minority ethnic groups, religions, migrants, refugees, and others. Cities in the twenty-first century are more diverse than their nineteenth century counterparts in terms of their populations, social characteristics, governance structures, and built environments. Further, the health outcomes that afflict urban dwellers are different than those of 150–100 years ago. New and drug resistant infectious diseases are combined with non-communicable diseases and issues of mental health and ageing in twenty-first century cities. Thus, we need new frameworks for understanding and addressing twenty-first century urban health.

This chapter explores how urbanization has and will continue to act as a fundamental determinant of human health (Galea et al. 2005). As I will show next, concerns with the well-being of urban populations has shaped global public from nineteenth century theories of miasma, to twentieth-century germ theory, to today's biomedical model, which attributes disease to unhealthy lifestyles, behaviours, and/or genetics (Porter 1999). For example, in the nineteenth century the urban Sanitary Era built water and sewer systems, required housing to have ventilation and cleaned streets. By the twentieth century, cities were chlorinating drinking water and pasteurizing milk to protect populations from disease and later, education and restrictions on 'unhealthy behaviours' such as tobacco use, shaped urban health. By the end of the twentieth century, epidemiologists began to re-emphasize that the social determinants of health (SDOH), or the institutional and structural forces including living condition, the distribution of wealth and power are the key drivers of population health. I will suggest that an emerging framework of population health that builds on the SDOH called 'toxic stress' reflects a twenty-first century evolution in urban health understanding and

action. *Toxic or chronic stress*, which I explain in more detail to follow, occurs when a group experiences adverse insults over a lifetime, often from in-utero through old age. These insults or 'stressors' have multiple sources and are often a combination of negative social—such as racism and segregation—economic, and physical exposures. These insults cause our body's immune system to 'over-react' through constant release of fight-or-flight hormones such as adrenaline and cortisol, and this wears away at our immune response, weakening our biologic response to infections and damaging our brain, heart, lungs, and other vital organs (McEwen and McEwen 2017). I will suggest here that twenty-first century urban health must take seriously the toxic stress hypothesis and this represent the *fourth wave* for urban health, with sanitation, germs, and genetics being the first three major public health movements.

Today, most urban health interventions are not focused on reducing multiple stressors in communities but rather on bringing either social services, primary care, economic opportunities, or physical improvements to urban residents or their neighbourhoods. Interventions tend to focus either on people or places, but rarely both at the same time. Interventions are generally conceptualized and led by experts, and focused on one disease, one risk factor, one hazardous exposure, one population group, or one suspected 'cause' of poor urban health, such as poverty (Corburn 2005). More care, more places offering care, more prevention, and more services tend to be the focus of many urban health strategies. This is what I call the *health in cities approach*, which focuses on healthy people but rarely on improving places, living conditions, or changing the institutions and policies that shape health opportunities or the presence of adverse stressors. In this chapter, I discuss twenty-first century urbanization and health by first reviewing some urban health history. I then highlight how twenty-first century cities, the physical, social, and political conditions, are linked to human health. I then explore the toxic stress framework and reveal its implications for twenty-first century framing of urban health. I close with some ideas for more healthy urban governance that embraces complexity science, uses processes of ecological adaptive management, and emphasizes the expertise of city residents as key participants and beneficiaries of urban health problem solving.

Modern urbanization and human health: a brief history

The notion that our urban environment—or the places where we live and work—influences human health is not new. Hippocrates, for some the Western father of medicine, noted in his *Airs, Waters, Places* (370 BC) that '[y]ou will find, as a general rule, that the constitutions and habits of a people follow the nature of the land where they live'. By the 1820s, French epidemiologist Louis René Villermé highlighted that the wealthier the Parisian neighbourhood, or Arrondissement, the lower the mortality rate and likelihood of illness. Edwin Chadwick's 'Report on the Sanitary Conditions of the Labouring Population in Great Britain in 1842', documented unsanitary living conditions afflicting the urban poor across Britain and contributed to the Royal Commission on the Health of Towns concluding that physical changes to the city were necessary to improve the well-being of the poor (Chadwick 1842, p. 87). Miasma—filth or dirty air—was the leading theory of disease causation, and sanitary commissions were created in European and American cities to clean up urban environments with the hope of arresting infectious disease epidemics (Duffy 1990). Sanitary engineers tended to address urban health issues by employing new technologies to remove waste by, for instance, piping it away from cities into rivers and oceans (Peterson 1979). Engineers laid claim to having the knowledge for designing healthy water delivery, street networks, and sewer systems—the 'arteries of the city'. When removing the miasma did not seem to reduce disease, the sick were removed from society. *Contagion*, the belief in the direct passage of poison from one person to another, led to large quarantines of immigrants and justified state-sponsored interventions in the economy (Porter 1999).

Early twentieth century urbanization and health: germ theory

By the turn of the new century, it was well-known in public health that both miasma and contagion failed to explain certain aspects of urban health, such as why, with ubiquitous filth, epidemics only occurred sometimes and in some places. By this time, the driving ideology in public health shifted to *germ theory*, which stated that microbes were the specific agents that caused infectious disease (Susser and Stein 2009). Medical treatment and disease management began to replace strategies of physically removing harms, and public health shifted toward laboratory research and interventions aimed at eliminating bacteria, such as vaccinations, pasteurization of milk, and

chlorination of municipal drinking water supplies (Leavitt 1982). Separate municipal bureaucracies emerged during this era to manage these urban improvement schemes, such as water service delivery infrastructure, sewage systems, public housing, parks and recreation, school-based health, and others, resulting in greater fragmentation and segmentation of work in urban health (Peterson 1979).

The model of laboratory science seen as producing definitive, apolitical knowledge also contributed to a view of city management during this time. City planners interested in improving health were often confronted with finding ways to pay for politically and economically controversial infrastructure projects that often wouldn't produce results for a decade or more (Rosen 1993). Planners and engineers turned to the prestige of laboratory science for a model of city management, since in the lab results were often unquestioned since researchers were viewed as neutral and detached from the social context where investigations were taking place, inputs into lab experiments were highly controlled and restricted, and analytic techniques were mechanical and standardized, so that results from a lab anywhere could be applied everywhere. City managers demanded political autonomy in the interest of the city's health, and used new 'efficiency' technologies such as benefit-cost analyses, land use zoning, and long-range comprehensive planning, to radically alter urban governance (Corburn 2013, 2017).

Social inequalities kept the fields of science, medicine, and urban management somewhat linked. In 1906, W.E.B. Du Bois published *The Health and Physique of the Negro American* and used data from census reports, vital statistics, and insurance company records to generate some of the first documentation of the poor health status of African Americans living in both southern and northern cities in comparison to white Americans. Du Bois questioned whether improvements in science, technology, housing, and other services were benefiting all Americans and concluded that racial health disparities were a result of social conditions—not Eugenic ideas of inherent racial inferiority that were commonly accepted in science and medicine at the time (Du Bois 1906).

Importantly, social movements organized to shape the priorities and direction of urban health. Unequal distributions of power in cities contributed to women reformers, such as Alice Hamilton, Florence Kelley, and Jane Addams, to organize workers and immigrants in the United States

to demand safer workplaces, social housing, and urban green spaces for play (Porter 1999). Importantly, urban public health was not just about science but shaped by power and politics of civil society and social movements (Szreter 2005).

Colonialism, urbanization, and health

Today's cities, and many of the challenges they face, can't be divorced from colonial-era strategies that blamed indigenous populations for disease and subsequently segregated them from life-supporting resources. Almost all migrants to colonial cities had to make or find their own housing. Laws of private property were almost non-existent for indigenous peoples in colonialized Asia, Africa, and the Caribbean in the nineteenth century. Doctors became all-purpose experts and authorities frequently put in charge of everything from 'native affairs' to town or city planning (Home 2013). Colonial medical authorities tended to blame the racial characteristics, living conditions and practices of the indigenous populations for diseases, and to make indigenous populations research subjects to help colonial powers control infectious diseases. Eugenics was the pseudo-science arguing that indigenous and slave populations were susceptible to disease due to their inherent racial inferiority (Curtain 1985). Francis Galton coined the term eugenics and the movement argued that the 'feeble minded' and destitute segments of urban society suffered from inherited genetic defects. One solution of eugenic societies that emerged in cities around the world was to control reproduction through forced sterilization. The advocates behind eugenics aimed to challenge the sanitarian ideas that reducing poverty, improving environmental and housing conditions would actually improve the health of the poor and were worth the expense (Krieger 2011). British sanitary engineers were sent to such colonial cities as Singapore, Calcutta, Pune, Lagos, and Cape Town, to help control infectious disease outbreaks. In practice, urban sanitarians used eugenic ideas of racial inferiority to justify subjecting indigenous practices to inspection, regulation, and disciplinary action—including razing homes and markets to 'ventilate' slums of their 'unhealthy' dark and dirty conditions (Garside 1988).

The biomedical model, housing and the fractured city: 1950s–1990s

By the mid-century, the driving theory in public health would shift again to the *biomedical model* of disease (Susser and Stein 2009). This model combined laboratory science with a focus on individuals and attributed morbidity and mortality to molecular-level pathogens brought about by lifestyles, behaviours, hereditary biology, or genetics. The biomedical model would provide public health with explanations for emerging chronic diseases such as heart disease, and shifted research and interventions to personal health 'risk factors' such as smoking, diet, and exercise (Krieger 2011). By the 1940s, the United States Housing Authority and New York City's Public Health Department blamed slums for crime and infant mortality and was advocating for removing slums and replacing them with modern, high-rise buildings (Corburn 2013; Hachmann et al. 2018). The American Public Health Association published a guide called 'Planning the Neighborhood' in 1948 that also recommended removing older slums in cities and suggested healthy housing was new, suburban-style development (APHA 1948). Slums and slum dwellers became synonymous with unhealthy places and populations. As Gilbert notes:

> The most worrying ingredient in most people's use of the word 'slum,' therefore, is the survival of these wholly negative connotations. Slum dwellers are not just people living in poor housing; they are considered by others to be people with personal defects. In Brazil, a favelado is not just someone who lives in a favela, he or she is thought to be someone who deserves to live there. (Gilbert 2007)

Slums often generated fear in the minds and actions of those who do not live there and the institutional response was largely to eliminate, not improve, slums. Slum had effectively become a pejorative and associated with unhealthiness, no matter the context or time period (Rao 2006).

For instance, American housing policies from this time through the 1950s subsidized new single-family suburban homes by offering mortgage insurance for new homes but not to older homes located within cities and urban ghettos (Corburn 2009). The US housing policies also mandated suburban design patterns that would set the stage for twentieth-century suburban sprawl. The US Housing Act of 1949 institutionalized urban renewal, where municipalities began razing 'slum' neighbourhoods and

displacing thousands of poor, largely African-American residents. Urban renewal was a programme and theory that aimed to remove downtown blight—still viewed as the cause of moral evil and the breeding ground for disease—and rebuild whole sections of the city using the best of modern technology and scientifically rational design. Yet urban renewal tended to only increase poverty for residents of poor neighbourhoods because their homes were replaced with either inadequate public housing or, as was more often the case, private real estate developers acquired the downtown land cheaply and opted not to build new housing but expensive high-rise office towers (Fullilove 2004). Not only were neighbourhoods physically fractured, but social and emotional ties, trust, and notions of collective efficacy were also severed by urban renewal, further diminishing the health-promoting resources available for African Americans. Shut out from most new suburbs, African Americans were denied other health benefits that can come with homeownership, such as capital accumulation, access to better-funded schools, and participation in the growing economy (Fullilove 2004).

Twenty-first century urban health equity

By the mid-1980s, academics, government agencies, and the World Health Organization (WHO) came together to create the Healthy Cities programme in order to re-invigorate the linkages between city planning and public health (Duhl and Sanchez 1999). Focused primarily in Europe, the Healthy Cities programme enrolled cities to draft 'healthy city plans' and share examples of the opportunities and barriers for planning more Healthy Cities. By the 1990s, public health researchers began to re-conceptualize explanations for the distribution of disease across populations in order to explain health disparities, energizing the field of social epidemiology (Krieger 2011). Urban place inequities—such as residential segregation, urban divestment, and environmental injustices—began to be seen by epidemiologists by the early twenty-first century as linked SDOH (Kjellstrom et al. 2007).

How urbanization influences human health

The physical and built environments

The physical environments of cities still matter for health, such as whether there is access to safe and affordable drinking water, sanitation, drainage, and garbage collection. Urban air and noise pollution remain critical health determinants in highly industrialized cities of Europe and Latin America, as well as developing cities in China and sub-Saharan Africa (Lilford et al. 2017). Urban air pollution is linked to up to 1 million premature deaths and 1 million pre-native deaths each year. Over 90% of air pollution in cities in these countries is attributed to vehicle emissions brought about by high number of older vehicles coupled with poor vehicle maintenance, inadequate infrastructure and low fuel quality (Friel et al. 2011). Urban particulate matter is associated with cardiovascular death and asthma. Noise pollution, a common urban nuisance, is associated with hearing impairment, hypertension, and ischaemic heart disease. Exposure to environmental pathogens in urban air and water can contribute to both infectious (i.e. parasitic diseases, diarrheal diseases, intestinal parasites, etc.) and non-infectious diseases.

Pedestrian conflicts with motor vehicles are one of the leading causes of injuries in urban areas. When a new development project includes new housing and commercial activity, pedestrian activity increases, and this can lead to an increase in injuries. However, greater pedestrian activity can promote physical activity that reduces heart disease, stroke, and mental illness, and increase functional status and the longevity of independence among the elderly. Creating new opportunities for pedestrian activity can also improve well-being by increasing the likelihood of social interactions that can reduce feelings of isolation. However, the construction of highways contributes to vehicular air pollution and suburban sprawl, while the lack of or inadequate transport inhibits access to employment and health-promoting services, especially for the urban poor. Highways and streets can limit green open space, which can act as a site for physical activity, social interactions, and an urban heat sink, reducing the likelihood of adverse human health impacts from climate change induced urban heat islands (Frumkin 2003).

The social environment and urban health

The physical and social environments of cities frequently interact and cannot be disassociated when trying to understand and improve urban health inequities. The urban social environment includes the institutions that shape

the structure and characteristics of relationships and opportunities among people and different population groups within a given community (Kjellstrom et al. 2007). Perhaps the most well researched aspect of the social environment that influences health is economic status or class. While debate continues whether absolute or relative poverty matter more for influencing health, there is agreement that being poor in any city increases one likelihood of a range of health risks across the life-course, from infant mortality and low birth weight, to stunted physical and cognitive development to early on-set chronic illnesses and higher rates of infections (Grant et al. 2017).

The economic 'environment' also influences well-being. Neighbourhoods with high concentrations of liquor stores also have high rates of addiction. However, local businesses can act as a source for employment and culturally-appropriate food and other services. Displacement of local businesses can adversely impact health by altering the availability and affordability of essential goods and services and the type of local employment possibilities. Business displacement can also contribute to physical blight—the tooth-gaped landscape all too common in poor neighbourhoods where widespread property abandonment has taken hold. Property abandonment can adversely influence health by increasing the likelihood of illegal dumping of garbage and hazardous wastes (Ellaway et al. 2012).

The social environment also influences health through a variety of other pathways, including the support of individual or group behaviours that affect health, buffering or enhancing the impact of stressors, and providing access to goods and services that influence health (for example, housing, food, informal healthcare) (Krieger 2011). Limited social supports may predispose persons to poorer coping and adverse health. High levels of social stressors such as social isolation and violence, are also known to adversely impact the health of urban residents. In cities, the greater spatial proximity of one's social networks may accentuate their role in shaping individual and population health (Friel et al. 2011).

Perhaps the most crucial social and political force in cities that influences health is spatial segregation. Many cities worldwide are highly segregated with discrimination against certain racial, ethnic, caste or tribal groups often acting as justification (Galea et al. 2005). Spatial segregation can have multiple effects, including the enforcement of homogeneity in resources and

social network ties and suppressing diversity that may benefit persons of lower socioeconomic status. Persons who live in segregated communities may have disproportionate exposure, susceptibility, and response to economic and social deprivation, toxic substances, and hazardous conditions (Corburn 2009).

Table 7.6.1 Urbanization and human health: resources and risks

Health resource	Urban physical and social influences on health (examples)
Environmental quality, including noise, air, soil, and water pollution	Vehicle emissions exacerbate respiratory disease and increase cardiopulmonary mortality, while indoor allergens exacerbate asthma Chronic noise exposure adversely harms sleep, temperament, hearing, and blood pressure, all of which can lead to developmental delays in children Trees and green space remove air pollution from the air and mitigate the urban heat island effect
Climate change resiliency, including protection from heat, draught, flooding, food supply disruptions, and other events	Heat islands, or elevated temperatures, are more frequent in dense, urban areas with limited green spaces Extreme weather events, such as draught and typhoons, exacerbate vulnerability of the urban poor already living on precarious land Climate change is contributing to an increase prevalence of vector-borne diseases in urban areas, such as dengue, malaria, and zika
Access to high-quality transit and safe roadways, sidewalks, and bicycle lanes	Vehicle/pedestrian injuries are most severe where sidewalks and crosswalks are non-existent Sidewalks and bicycle lanes facilitate physical activity, reducing heart disease, diabetes, obesity, blood pressure, osteoporosis, and symptoms of depression Public transit provides access to employment, education, parks, and healthcare services
Access to quality childcare, education, and healthcare facilities	Quality childcare can build disease immunities and increase likelihood of future educational attainment and earnings Education can enhance health literacy about preventative behaviours and services Timely access to primary health services prevents serious illness
Affordable, safe, stable, and socially integrated housing	Crowded and substandard housing conditions increase risks for infections, respiratory disease, fires, and stress Unaffordable rents or mortgages result in trade-offs between housing, food, and medical care Racial residential segregation limits economic and educational opportunities, concentrates disadvantage and increases social distance between racial/ethnic groups
Access to safe and quality open space, parks, cultural and recreational facilities	Clean and safe parks can increase the frequency of physical activity Cultural activities can promote cross-cultural understanding, decrease violence, and enhance social cohesion

Health resource	Urban physical and social influences on health (examples)
Employment providing meaningful, safe, and living wage jobs	Higher income is associated with better overall health, reduced mortality, and higher emotional stability Unemployment is a source of chronic stress, while job autonomy increases self-esteem
Access to affordable and quality goods and services	Neighbourhood grocery stores support nutritious diets Local financial institutions help families create and maintain wealth
Protection from crime and physical violence	Indirect effects of violence and crime include fear, stress, anxiety, and unhealthy coping behaviours, overeating, smoking, and alcohol/drug abuse Fear of crime can force children to stay indoors, increasingly exposure to toxic indoor air and allergens, and limiting physical activity outside
Social cohesion and political power	Physical and emotional support buffers stressful situations, prevents isolation, contributes to self-esteem, and reduces the risk of early death Stress from severed/lack of social ties/support can contribute to low birth weight, which increases risk of infant death, slow cognitive development, hyperactivity, breathing problems, overweight, and heart disease

Sources: Galea et al. 2005; WHO 2010.

Urbanization, governance, and health

The physical and social characteristics and dynamics in cities that influence health do not happen randomly or by accident. Cities and metropolitan regions more generally are not shaped by faceless forces of natural succession and competition. People and organized groups or coalitions actively accomplish cities and urban neighbourhoods, and the process is rarely the same from here to there. There are real winners and losers in the *political struggles* of place making, and static definitions of physical and social variables rarely captures this dynamic of places. The political processes of urban place-making and remaking are what I refer to as ‘urban governance’. (Burris et al. 2007) As Szreter (2005) notes (see Chapter 1.2), social movements were critical in reshaping public health priorities, whose knowledge is considered, and how civil society plays an important role in democratic decision-making.

The term governance is broadly understood as attentive to the relationships between the overlapping spheres of political, economic, and social life, as each aim to influence collective action. Governance is not government, but is inherently about the struggle and conflict between formal institutions and organizations and informal norms and practices. Urban governance includes a complex mix of different contexts, actors, arenas, and issues, where struggles over power can be manifested in public discourses or tacit day-to-day routines. An urban governance approach to urban health is attentive to both processes and outcomes, and must ask questions such as: what conditions lead to planners using or abusing power, responding to or even resisting market forces, working to empower some groups and dis-empower others, promoting multiparty consensual decision-making discourses or simply rationalizing decisions already made? (Corburn 2013). The governance approach views urban planning as a process that includes the shaping of public agendas and attention, available evidence, and norms of inquiry, inclusive or exclusive deliberations, and responses (or lack thereof) to bias, discrimination, inequality, and a recalcitrant state. An attention to governance moves health equity planning away from only focusing on vulnerable people or places to also include altering the political processes and institutions that work to maintain the social inequities driving urban health inequities (Friel et al. 2011).

Toward a twenty-first century science for urbanization and health: addressing toxic stress

Twenty first century urbanization must embrace the now well-documented idea that social stress can be toxic to the human body, especially when place-based stressors are chronic and cumulative across one's lifetime (McEwen and McEwen 2017). While stress can be life saving for most—think of the fight-or-flight mechanism—constant adversity is toxic, meaning that the prolonged activation of the stress response systems can disrupt the development of the brain architecture and other biologic systems. Some of the multiple, chronic stressors experienced by the poor and marginalized populations in cities appear in Figure 7.6.1 (Hajat et al. 2015). These stressors often begin in-utero, and individual behaviour changes alone cannot redress their impacts.

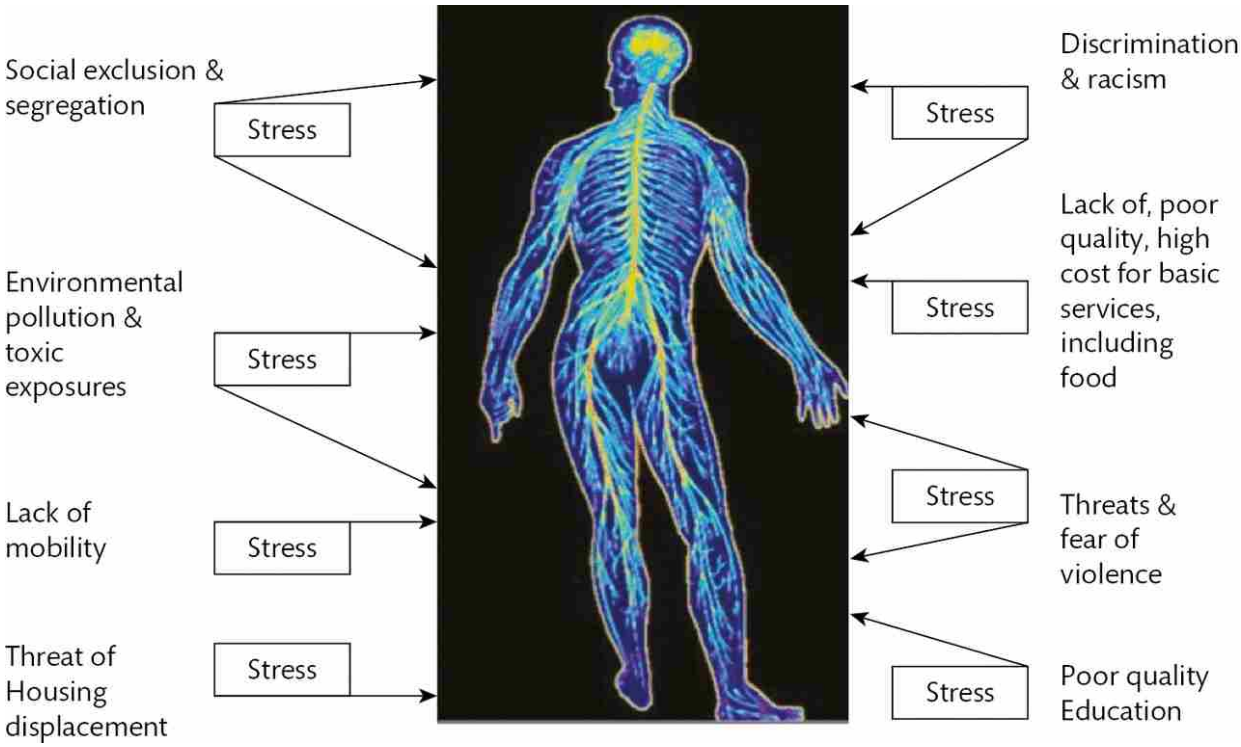


Fig. 7.6.1 Urban toxic stressors that adversely impact human health.
Corburn, J. 2013. Healthy City Planning: From Neighbourhood to National Health Equity. London: Routledge.

As shown in Figure 7.6.2, under ‘normal’ stressful situations the human body undergoes a range of physical and chemical responses, but primarily the hormones epinephrine (adrenaline) and cortisol are released to heighten one’s awareness, but these eventually ‘shut off’ when the threat is over and the body returns to homeostasis (Fig. 7.6.2, solid line). The body’s ability to maintain stability under stress has been called allostasis (McEwen and McEwen 2017). In toxic stress situations, stressors are constant and the ‘allostatic load’ continues to increase and the release of ‘fight-or-flight’ hormones does not properly regulate or shut-off (Fig. 7.6.2, dashed line). Increased allostatic load wears away at the immune system as it overworks to manage the hormonal releases and attempt to return to homeostasis. Under toxic stress circumstances, the oversecretion of cortisol and adrenaline trigger other biologic responses like poor immune response, poor glucose regulation, and constant feelings of hunger that can contribute to chronic diseases such as overweight and obesity, diabetes, hypertension,

cardiovascular disease, stroke, asthma, and mental health disorders (Velasquez-Manoff 2012). Some known toxic stressors include chronic poverty, racial, gender, and other forms of discrimination and marginalization, physical or emotional abuse, exposure to violence and housing instability (Berger and Sarnyai 2014). Allostaic load has been linked with changes in the length of telomeres, which are DNA-protein complexes capping the ends of chromosomes that protect them against damage. Telomere shortening is considered a marker of cellular ageing (Price et al. 2013).

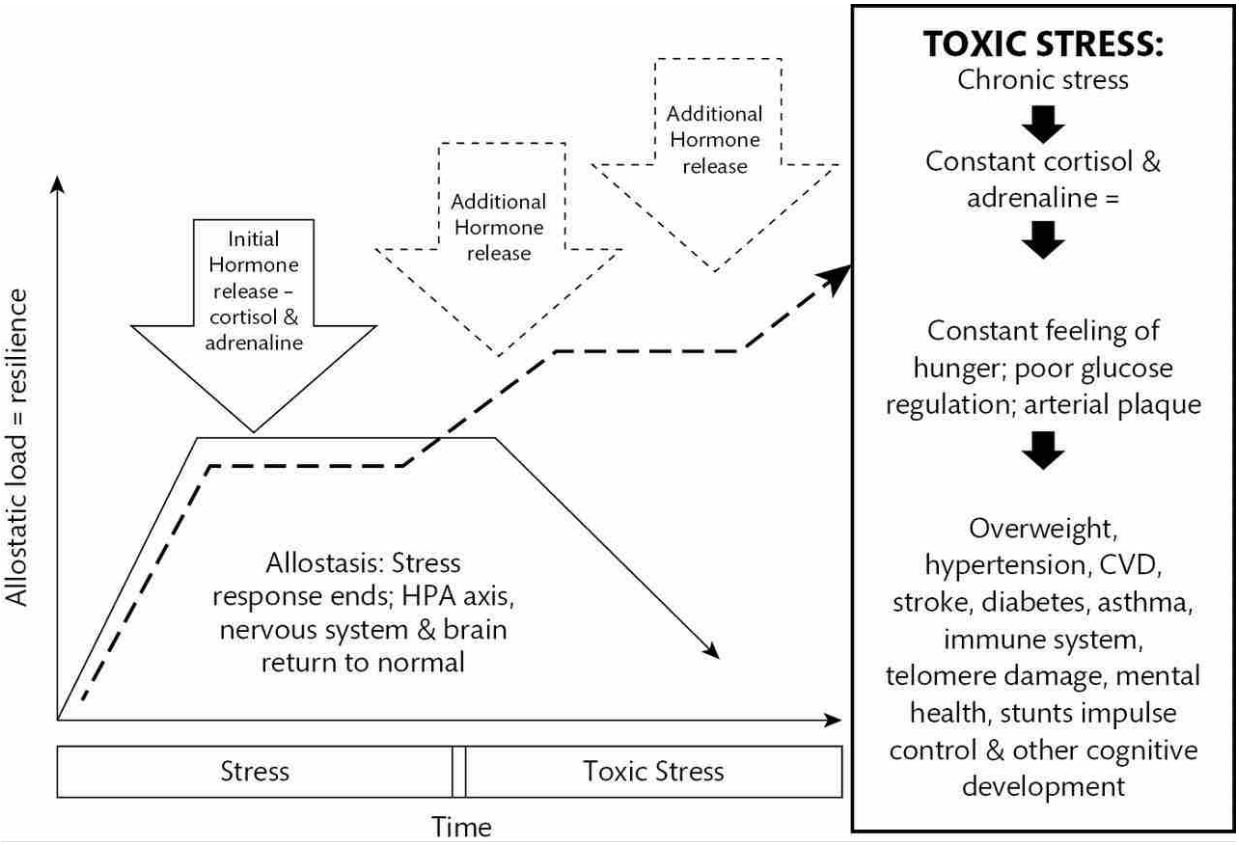


Fig. 7.6.2 Chronic stress and its biologic impacts.

Since populations in many deprived and impoverished areas of cities are likely experiencing multiple 'toxic stressors', twenty-first century urbanization and health must be attentive to the physical and social stressors in places, such as chronic poverty and discrimination, housing instability, and exposure to violence. The toxic stress idea suggests that a twenty-first

century science for the healthy city can no longer view a person as, for example, one day an African immigrant, another day born low birth weight, another day raised in a home bearing remnants of lead paint, another day subjected to racial discrimination at work, and still another day living in a racially segregated neighbourhood without a supermarket or access to transport. Just as the body does not neatly partition these experiences—all of which may serve to increase risk of uncontrolled hypertension and related morbidity and premature mortality—the twenty-first century urban health planner can no longer separate these spheres and sectors in research or practice (Corburn 2013).

Twenty-first century urbanization and health governance

Measuring and addressing urban toxic stress will demand new, interdisciplinary action-research into the context-specific factors that might be making particular urban residents unhealthy and more susceptible to disease, disability, and death. This work must privilege partnerships, participation, and power-sharing between scientists, the urban poor themselves, social movements, and state-level decision-makers. There is no one-size-fits all approach. I borrow a term from the Science Studies, economics, and development literatures to call this collaborative urban health approach *co-production* (Jasanoff 2004). In the co-production of urban health, science is understood as dependent on the natural world, as well as historical events, social practices, material resources, and institutions that contribute to the construction, dissemination, and use of scientific knowledge. Political decision-making, in the co-production framework, does not take ‘scientific knowledge’ as a given, but seeks to reveal how science is conducted, communicated, and used. The co-production model problematizes knowledge and notions of expertise, challenging hard distinctions between expert-lay ways of knowing. Acting to understand and improve urban health, under the co-production approach, requires a negotiation among the always partial and plural positions of professionals and lay people. The co-production model also destabilizes the dominant view in science policy-making that science can be uncritically accepted as ‘fact’ and ‘truth’. The destabilizing stories and emphasis on the need for ‘negotiating expertise’, suggest that a deliberative science is necessary for the co-production of slum health knowledge for action (Corburn 2005).

Adaptive ecosystem management offers a model of co-producing scientific research and practice for global urban health. Adaptive management acknowledges the failures of linear processes where narrow disciplinary scientists have aimed to develop complex models, predict long-term outcomes, and suggest one-time policy standards. Instead, adaptive management begins with an acknowledgement of the inherent complexity and uncertainty within systems, that this complexity demands an iterative, ongoing learning process among a range of expert stakeholders, and policy interventions must be adjusted to reflect newly acquired knowledge (Corburn 2013). Another difference between adaptive management and conventional science policy is that adaptive management does not postpone actions until definitive causality is known about a system, but rather emphasizes the importance of action in the face of uncertain science and couples these decisions tightly to rigorous monitoring.

Urban climate change adaptation

Perhaps the key twenty-first century urban health challenge is ensuring the poor and most vulnerable residents and communities are able to adapt to a changing climate. The urban poor are some of the most vulnerable to the human health impacts of global climate change since, for instance, low-income, urban neighbourhoods may have less green space, housing located on flood-prone land, and fewer economic and social resources to escape and recover from heat, weather, and other climate-related stressors. Urban climate justice demands co-produced climate change adaptation plans that reduce risk and improve the health status of the poor and already vulnerable communities. This will demand new investments in urban infrastructure, such as housing, water, and sanitation, social and healthcare services, in urban communities that already experience health inequities. As many local governments, particularly in the Global South, experience severe capacity and resource constraints, co-produced adaptation demands a leadership role for the urban poor and civil society actors. Using this approach, urban climate justice would aim to not just reduce existing risks but also build-upon local institutional strengths and civic capacities.

One such approach, used in Quito, Ecuador, resulted in the city's Climate Change Action Plan that articulated the needs of marginalized groups, facilitated a youth-led Convention on Climate Change, incorporated local

traditional and indigenous knowledge in the process, and defined ongoing civil society participation for climate action. Based on recommendations from Quito's plan, the municipality has invested in socially vulnerable groups by building new water and sewerage infrastructure, social housing that enabled the poor to avoid living in unsafe structures on steep hillsides, reforested areas of the municipality as an 'ecosystem service' that could reduce mudslides and cool temperatures, and community-and school-based organic farming (AGRUPAR, Agricultura Urbana Participativa) that improves nutrition and food security for the urban poor (Chu et al. 2016). Quito's food programme has multiple co-benefits; reducing malnutrition and vulnerability to climate-induced food supply disruptions for the urban poor, increasing employment and incomes, improving learning for children, enhancing water conservation and reuse, and reducing carbon emissions by limiting the travel distance along the food supply chain.

Health in all urban policies

Health in All Policies (HiAP) is an approach to decision-making that aims to put the adaptive and co-production approach into practice (Shankardass et al. 2018). HiAP recognizes most public policies have the potential to influence health and health equity, either positively or negatively, but policymakers outside of the health sector may not be routinely considering the health consequences of their choices and thereby missing opportunities to advance health. HiAP emerged, in part, from the WHO's Declaration of Alma-Ata in 1978 and reflects numerous international calls for policy action to address the SDOH (Freiler et al. 2013). The European Union endorsed HiAP in 2006 and the governments of Finland, Norway, Sweden, United Kingdom, Canada, South Australia, Malaysia, Sri Lanka, Iran, and Brazil have all adopted some form of HiAP within governmental practice. The Dutch government actively supports municipalities to draft and implement HiAP strategies, but very few have managed to complete and implement these strategies, noting the difficulty in getting government actors and institutions outside the health sector to consider health impacts in their work (Storm et al. 2014). Some cities are actively working to integrate HiAP into their land use and planning decision-making. For example, Washington, DC passed the Sustainable DC Transformation Order in 2013, creating a HiAP Task Force

and the Chicago Department of Public Health created the Healthy Chicago Interagency Council to integrate health across all city departments.

The City of Richmond, California, in the San Francisco Bay Area, became the first city in the US to pass a HiAP ordinance in 2014 (Corburn et al. 2014). Richmond's HiAP Ordinance is aligned with their General Plan and Budget, and includes intervention areas focused on how city policy, management, and service decisions can reduce the multiple toxic stressors residents and planners identified in Richmond. The HiAP Ordinance address six planning areas, including: (1) Governance and Leadership; (2) Economic Development and Education; (3) Full Service and Safe Communities; (4) Neighbourhood Built Environments; (5) Environmental Health and Justice; and (6) Quality and Accessible Health Homes and Social Services. In addition, HiAP includes a racial equity evaluation for all city projects, policies, and plans (Corburn et al. 2014).

Conclusions: pursuing greater urban health equity in the twenty-first century

Contemporary inquiry into global urbanization and health can learn from the past to better understand how current trends gained resonance and what alternative futures might be possible. Alternative paradigms can be both practical and socially just, as the examples here suggest. However, more work needs to be done to avoid overgeneralizations and investigate specific contexts, how culture and politics shapes urbanization and appropriate health interventions. As global efforts to prioritize urban health continue, such as through the SDGs, a historical perspective is necessary to critically re-engage with urbanization as a structural determinant of health (WHO and UN Habitat 2016).

As suggested here, re-orienting urbanization and the institutions behind it toward health will require moving toward integrated, multisector partnerships that target equity. Urban planners have a specific role to play by explicitly expanding the narrow framing of the 'built environment and health' to include social and political determinants of health, such as an attention to how racial discrimination contributes to place-based toxic stressors. This also demands that planners work with epidemiologists and others to break-down issue and sector fragmentation and the tendency in

medicine and public health to focus only on one disease, risk factor, or behavioural change at a time. Civil society involvement, such as through activist social movements and 'citizen scientists', is crucial to ensure twenty-first century urbanization serves the well-being of the most vulnerable and marginalized populations and places.

As discussed here, those concerned with urban health cannot wait until the science of health equity is definitive, but instead can rely on over a century of evidence that simultaneously addressing social, environmental, and economic inequalities can promote greater urban health equity.

REFERENCES

American Public Health Association (APHA) Committee on the Hygiene of Housing (1948). *Planning the Neighborhood: Standards for Healthful Housing*. Chicago, IL: Public Administration Service.

Berger, M. and Sarnyai, Z. (2014). More than skin deep: stress neurobiology and mental health consequences of racial discrimination. *Stress*, 18, 1–10.

Burris, S., Hancock, T., Lin, V., and Herzog, A. (2007). Emerging strategies for healthy urban governance. *Journal of Urban Health: Bulletin of the New York Academy of Medicine*, 84(1), 154–63.

Chadwick, E. (1842). *Report on the Sanitary Condition of the Labouring Population and on the Means of its Improvement, London, May 1842*. [Online] Available at: https://www.parliament.uk/about/living-heritage/transformingsociety/livinglearning/coll-9-health1/health-02/

Chu, E., Anguelovski, I., and Carmin, J. (2016). Inclusive approaches to urban climate adaptation planning and implementation in the Global South. *Climate Policy*, 16(3), 372–92.

Corburn, J., Curl, S., and Arredondo, G. (2014). A health-in-all-policies approach addresses many of Richmond, California's place-based hazards, stressors. *Health Affairs*, 33(11), 1905–13.

Corburn, J. (2017). Urban place and health equity: critical issues and practices. *International Journal of Environmental Research and Public Health*, 14(2), 117.

Corburn, J. (2005). *Street Science: Community Knowledge and Environmental Health Justice*. Cambridge: MIT Press.

Corburn, J. (2009). *Toward the Healthy City: People, Places and the Politics of Planning*. Cambridge: MIT Press.

Corburn, J. (2013). *Healthy City Planning: From Neighbourhood to National Health Equity*. London: Routledge.

Curtain, P.D. (1985). Medical knowledge and urban planning in tropical Africa. *The American Historical Review*, 90(3), 594–613.

Du Bois, W.E.B. (1906). *The Health and Physique of the Negro American: Report of a Social Study Made Under the Direction of Atlanta University*. [Online] Available at: http://www.archive.org/details/healthphysiqueof00dubo

Duffy, J. (1990). *The Sanitarians: A History of American Public Health*. Chicago, IL: University of Illinois Press.

Duhl, L.J. and Sanchez, A.K. (1999). *Healthy Cities and the City Planning Process*. [Online] Available at: https://apps.who.int/iris/handle/10665/108252

Dye, C. (2008). Health and urban living. *Science*, 319(5864), 766–9.

Ellaway, A., Benzeval, M., Green, M., Leyland, A., and Macintyre, S. (2012). Getting sicker quicker: does living in a more deprived neighbourhood mean your health deteriorates faster? *Health and Place*, 8(2), 132–7.
Freiler, A., et al. (2013). Glossary for the implementation of Health in All Policies (HiAP). *Journal of Epidemiology and Community Health*, 67, 1068–72.
Friel, S., Hancock, T., Kjellstrom, T., McGranahan, G., Monge, P., and Roy, J. (2011). Urban health inequities and the added pressure of climate change: an action-oriented research agenda. *Journal of Urban Health*, 88(5), 886–95.
Frumkin, H. (2003). Healthy Places: exploring the evidence. *American Journal of Public Health*, 93, 1451–56.
Fullilove, M.T. (2004). *Root Shock: How Tearing Up City Neighborhoods Hurts America and What We Can Do About It*. New York: Ballantine.
Galea, S., Freudenberg, N., and Vlahov, D. (2005). Cities and population health. *Social Science and Medicine*, 60, 1017–33.
Garside, P.L. (1988). Unhealthy areas: town planning, eugenics and slums, 1890–1945. *Planning Perspectives*, 3, 24–46.
Gilbert, A. (2007). The return of the slum: does language matter? *International Journal of Urban and Regional Research*, 31(4), 697–713.
Grant, M., et al. (2017). Cities and health: an evolving global conversation. *Cities & Health*, 25, 1–9.
Hachmann, S., Arsanjani, J.J., and Vaz, E. (2018). Spatial data for slum upgrading: volunteered geographic information and the role of citizen science. *Habitat International*, 72, 18–26.
Hajat, A., et al. (2015). Examining the cross-sectional and longitudinal association between diurnal cortisol and neighborhood characteristics: evidence from the multi-ethnic study of atherosclerosis. *Health Place*, 34, 199–206.
Home, R.K. (2013). *Of Planting and Planning: The Making of British Colonial Cities*. London: Routledge.
Jasanoff, S. (2004). *States of Knowledge: The Co-Production of Science and Social Order*. London: Routledge.
Kjellstrom, T., Mercado, S., Sattherthwaite, D., McGranahan, G., Friel, S., and Havemann, K. (2007). *Our Cities, Our Health, Our Future: Acting on Social Determinants for Health Equity in Urban Settings*. Kobe: World Health Organization, Centre for Health Development.
Krieger, N. (2011). *Epidemiology and the People's Health: Theory and Context*. Cambridge: Oxford University Press.
Leavitt, J.W. (1982). *The Healthiest City: Milwaukee and the Politics of Health Reform*. Madison, WI: University of Wisconsin Press.
Lilford, R.J., et al. (2017). Improving the health and welfare of people who live in slums. *The Lancet*, 389(10068), 559–70.
McEwen, C.A. and McEwen, B.S. (2017). Social structure, adversity, toxic stress, and intergenerational poverty: an early childhood model. *Annual Review of Sociology*, 43, 445–72.
Peterson, J. (1979). The impact of sanitary reform upon american urban planning, 1840–1890. *Journal of Social History*, 13, 83–103.
Porter D. (1999). *Health, Civilization and the State: A History of Public Health from Ancient to Modern Times*. London: Routledge.
Price, L.H., Kao, H.T., Burgers, D.E., Carpenter, L.L., and Tyrka, A.R. (2013). Telomeres and early-life stress: an overview. *Biologic Psychiatry*, 73(1), 15–23.
Rao, V. (2006). Slum as theory: the South/Asian city and globalization. IJURR, 30(1), 225–32.
Rosen, G. (1993). *A History of Public Health*, rev. ed. Baltimore, MD: Johns Hopkins University Press.
Shankardass K, et al. (2018). The implementation of Health in All Policies initiatives: a systems framework for government action. Health Research Policy and Systems, 16(1), 26.

Storm, I., Harting, J., Stronks, K., and Schuit, A.J. (2014). Measuring stages of health in all policies on a local level: the applicability of a maturity model. *Health Policy*, 114(2–3), 183–91.

Susser, M. and Stein, Z. (2009). *Eras in Epidemiology: The Evolution of Ideas*. Oxford: Oxford University Press.

Szreter, S. (2005). *Health and Wealth. Studies in History and Policy*. New York: Rochester University Press.

United Nations (2018). *Revision of World Urbanization Prospects. Population Division of the United Nations Department of Economic and Social Affairs (UN DESA)*. [Online] Available at: https://esa.un.org/unpd/wup/

Velasquez-Manoff, M. (2012). *An Epidemic of Absence: A New Way of Understanding Allergies and Autoimmune Diseases*. New York: Scribner.

World Health Organization (WHO)/United Nations Human Settlements Programme (WHO/UN HABITAT) (2016). *Global Report on Urban Health: Equitable, Healthier Cities for Sustainable Development*. Geneva: WHO.

World Health Organization (WHO)/United Nations Human Settlements Programme (UN HABITAT) (2010). *Hidden Cities: Unmasking and Overcoming Health Inequities in Urban Settings*. Kobe: World Health Organization Centre for Health Development.

SECTION 8

Major health problems

8.1 Epidemiology and prevention of cardiovascular disease
Nathan D. Wong and Wenjun Fan

8.2 Cancer epidemiology and public health
8.3 Paolo Boffetta, Zuo-Feng Zhang, and Carlo La Vecchia Chronic obstructive pulmonary disease and asthma
Craig M. Riley, Jessica Bon, and Alison Morris

8.4 Obesity
Anna Peeters and Tim Lobstein

8.5 Physical activity and public health
8.6 Nyssa T. Hadgraft, Neville Owen, and Paddy C. Dempsey Diabetes mellitus
Farah Naz Khan, Nida Izhar Shaikh, K.M. Venkat Narayan, and Mohammed K. Ali
8.7 Public mental health and suicide
Danuta Wasserman and Kristian Wahlbeck

8.8 Dental public health
Amira S. Mohamed and Peter G. Robinson

8.9 Musculoskeletal disorders
Lope H. Barrero and Alberto J. Caban-Martinez

8.10 Neurological diseases, epidemiology, and public health
Walter A. Kukull, Kumeren Govender, and James Bowen

8.11 Infectious diseases and prions
8.12 Davidson H. Hamer, Amira Khan, and Zulfiqar A. Bhutta Sexually transmitted infections
Noah Kojima and J.D. Klausner

8.13 Acquired immunodeficiency syndrome (AIDS)
Quarraisha Abdool Karim, Urisha Singh, Cheryl Baxter, and Salim S. Abdool Karim
8.14 Tuberculosis
Roxana Rustomjee

8.15 Malaria
Frank Baiden, Keziah L. Malm, and Fred Binka

8.16 Viral alcoholic and fatty liver diseases
Ehud Zigmond and Daniel Shouval

8.17 Emerging and re-emerging infections
David L. Heymann and Vernon J.M. Lee

8.18 Bioterrorism

Peter Katona

8.19 Genetic epidemiology

Elizabeth H. Young and Manjinder S. Sandhu

8.1

Epidemiology and prevention of cardiovascular disease

Nathan D. Wong and Wenjun Fan

Introduction

Cardiovascular diseases (CVDs) comprise the leading cause of morbidity and mortality, with recent 2015 estimates noting 422.7 million cases and 17.9 million deaths (Roth et al. 2017) and expected to grow to 23.6 million by the year 2030; 80% of these deaths occur in lower and middle income countries (WHO 2011). It is the largest contributor to non-communicable diseases (NCDs) that are now responsible for the largest share of morbidity and mortality worldwide. The incidence of CVD, including coronary heart disease, peripheral arterial disease, heart failure, and stroke, as well as the prevalence of key risk factors, varies greatly according to geographical region, gender, and ethnic background. Multiple longitudinal epidemiological studies have provided valuable insights into the natural history and risk factors associated with the development and prognosis of CVD. Randomized clinical trials have demonstrated the value of management of several key risk factors for both the primary and secondary prevention of CVD. This chapter discusses the epidemiology of CVD, its associated risk factors and evidence for their control, assessment of CVD risk, and the evidence behind the control of CVD risk factors for the prevention of CVD.

Definitions, incidence, and distribution

Cardiovascular disease comprises many conditions, including coronary heart disease, heart failure, rheumatic fever/rheumatic heart disease, stroke, and congenital heart disease. Ischaemic heart disease is typically the

predominant manifestation of CVD and responsible for 46% of deaths due to CVD in men and 38% in women, followed closely by cerebrovascular disease at 34% and 37%, respectively (WHO 2011). Certain regions of the world, however, show ischaemic heart disease to be the cause of cardiovascular deaths in more than 50% of persons, with significant regional variations (Roth et al. 2017) (Fig. 8.1.1). While the burden of CVD was highest in Western countries during much of the 20th century, the highest mortality rates of CVD now occur among certain central Asian and Middle Eastern regions (Roth et al. 2017) (Fig. 8.1.2). By country, total death rates (per 100,000) from CVD, CHD, and stroke for the years 2012–2015 are highest in the Russian Federation for both men (1,041, 709, and 204, respectively) and women (433, 286, and 106, respectively) and among the lowest in Israel for men (96, 45, and 20, respectively) and in France for women (40, 9, and 12, respectively), with intermediate rates in the US (234, 124, and 27, respectively, in men and 117, 48, and 20, respectively, in women) (Benjamin et al. 2018).

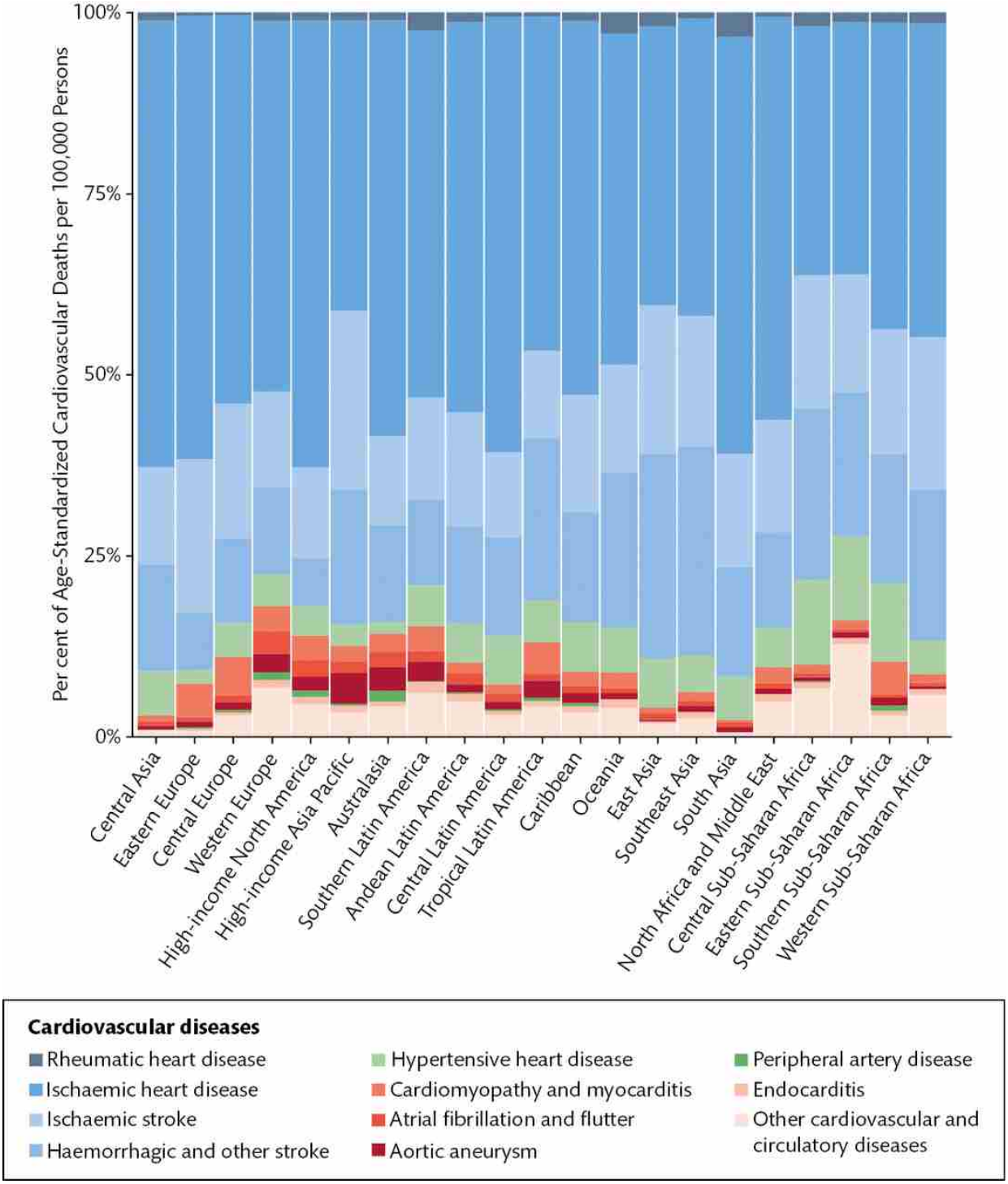


Fig. 8.1.1 Relative distribution of age-standardized cardiovascular deaths by cause for 21 global burden of disease regions.

From Roth GA, Johnson C, Abajobir A, et al. Global, Regional, and National Burden of Cardiovascular Diseases for 10 Causes, 1990 to 2015. *J Am Coll Cardiol*. 2017 Jul 4;70(1):1–25 under a CC BY 4.0 license.

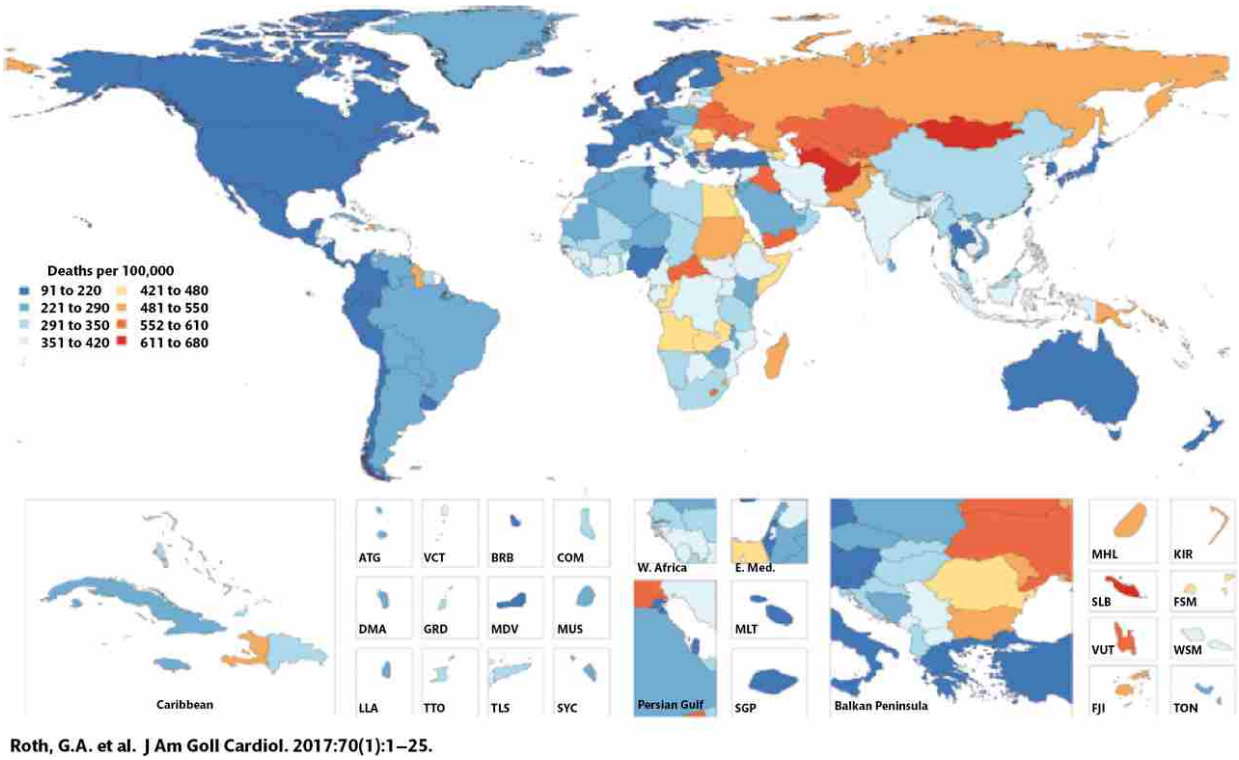


Fig. 8.1.2 Age-standardized death rate from cardiovascular diseases by country, 2015.
From Roth GA, Johnson C, Abajobir A, et al. Global, Regional, and National Burden of Cardiovascular Diseases for 10 Causes, 1990 to 2015. *J Am Coll Cardiol*. 2017 Jul 4;70(1):1–25 under a CC BY 4.0 license.

In the United States, the overall prevalence of CVD in the years 2011–2014 (including hypertension) increases dramatically from 13.5% in men between the ages 20 and 39 years to 84.4% for men above 80+ years old; the corresponding figures for women were 11.5% and 86.5% (Benjamin et al. 2018).

Of note, however, is the substantial reduction in CVD mortality rates in the US over the past three decades, both in men and in women, although absolute mortality from CVD in women, since 1985 has exceed that observed in men, until recent years when the number of deaths has been fewer in women than in men. These declines have been attributed both to improved treatments for CVD and its associated risk factors and to improvements in lifestyle factors such as substantial declines in cigarette smoking, although the increasing obesity epidemic is expected to negate some of these reductions in CVD.

Myocardial infarction, angina pectoris, and sudden coronary death are the major clinical manifestations of CHD. CHD initially presents as sudden coronary death in approximately one-third of cases. Other forms of documented CHD include procedures done as a result of documented significant atherosclerosis, such as coronary artery bypass grafting (CABG) or percutaneous coronary interventions (PCIs), including angioplasty and stenting. Persons with documented significant disease from a coronary angiogram, echocardiogram, nuclear myocardial perfusion, magnetic resonance imaging, or computed tomography angiographic or coronary calcium scan can also designate the presence of CHD; however, because the definitions used to define significant CHD vary and these findings often do not result in hospitalization or hard CHD events, such persons are not normally counted as incident or prevalent CHD, particularly for the purposes of endpoints in epidemiologic studies or clinical trials. Non-fatal or fatal myocardial infarction or sudden coronary death are most typically included as ‘hard’ CHD endpoints, while ‘total’ CHD may additionally include angina requiring hospitalization a well as PCI or CABG.

CHD prevalence rates vary dramatically by age and gender in the United States. The most recent statistics from 2011 to 2014 showed that 16.5 million US adults aged 20 years and older have CHD; this varies widely by race and gender within the United States. For example, for males, while the overall prevalence was 7.4%, the prevalence in non-Hispanic white males was 7.7%, in black males 7.1%, in Mexican-American males 5.9%, and in Asian males 5.0%. The corresponding figures for females were 5.3% overall, 5.3% in non-Hispanic white females, 5.7% in black females, 6.1% in Mexican-American females and 2.6% in Asian females. An American experiences a heart attack every 40 seconds. CHD makes up more than half of all CVD deaths in men and women <75 years of age, with a lifetime risk (≥2 major risk factors) of developing CHD after age 55 years, of 37.5% for men and 18.3% for women (Benjamin et al. 2018).

Other important manifestations of CVD include cerebrovascular disease (including stroke and transient ischaemic attacks), heart failure, atrial fibrillation, and peripheral arterial disease, in particular. Somewhat less common but still of significance are valvular heart disease, rheumatic heart disease, and bacterial and valvular endocarditis.

Cerebrovascular disease includes stroke and transient ischaemic attacks (TIA). Stroke prevalence in the United States ranges from <1% in those <40

years of age to 14% in both genders in those aged 80 years and over. An estimated 7.2 million Americans aged 20 years and over have had a stroke; each year an estimated 795,000 persons experience a new or recurrent stroke, with approximately 610,000 of these cases being first attacks. Of all strokes, about 87% are ischaemic and 10% are intracerebral haemorrhagic strokes, and 3% subarachnoid haemorrhage strokes (Benjamin et al. 2018). The self-reported prevalence of physician-diagnosed TIAs is about 2.3%, translating to approximately 5 million persons, but the true prevalence is likely to be greater as many with neurological symptoms consistent of TIA fail to report them to their healthcare provider. About 12% of all strokes are preceded by a TIA.

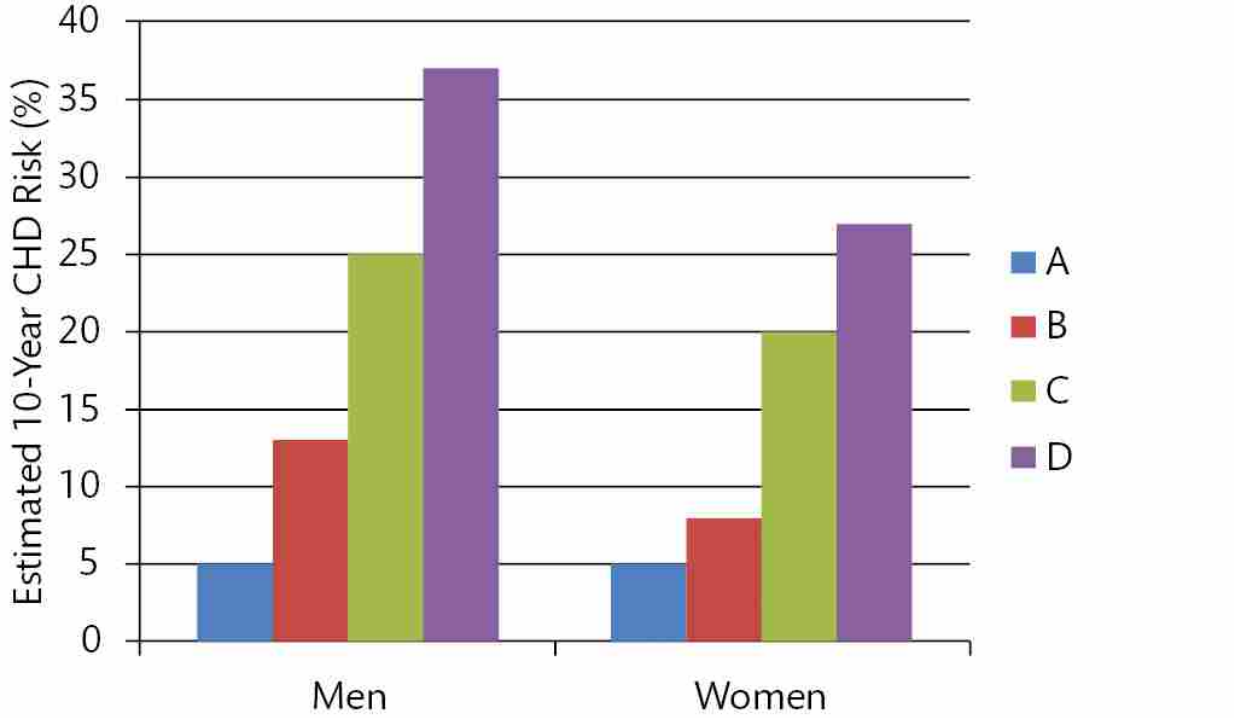
While mortality rates due to total CVD, and especially CHD and stroke, have declined substantially over the past 30 years, hospitalizations due to total CVD have increased, much of this fuelled by the more than doubling in the number of hospitalizations due to heart failure between 1980 and 2010 (1,023,000 in 2010 in the United States). It is projected that this will increase by another 25% by the year 2030. The prevalence of heart failure ranges from <1% in those under age 40 to 14.1% of men and 13.4% of women aged 80 years and over.

Peripheral arterial disease is highly prevalent affecting approximately 8.5 million Americans aged 40 years and over and is associated with significant morbidity and mortality; it is most common in older persons over age 55 and is asymptomatic in most cases, being diagnosed by a Doppler tool measuring the ankle brachial index, where a value of <0.9 is diagnostic of peripheral arterial disease. The estimated prevalence in the US population is 7.2%; only about 10% of those with peripheral arterial disease actually have the classic symptoms of intermittent claudication (or leg pain) (Benjamin et al. 2018).

Estimates of the prevalence of valvular heart disease, which includes aortic, mitral, and tricuspid valve disorders range from 0.3% in those under the age of 45 years to 11.7% in those aged 75 years and over. Rheumatic heart disease while uncommon in developed countries, affects more than 15 million individuals in Africa, Asia, and the Pacific, causing over 200,000 deaths annually. Infective endocarditis is relatively rare and results from formation of non-bacterial thrombotic endocarditis on the surface of a cardiac valve, bacteraemia, and adherence of the bacteria in the bloodstream; the estimated risk in the general population is as low as 1 case per 14 million dental procedures (Benjamin et al. 2018).

Risk factors for cardiovascular disease

The Framingham Heart Study, the seminal epidemiologic study of CVD, is a large longitudinal investigation begun in 1948 in the small town of Framingham, Massachusetts, USA. The original cohort of 5,209 participants aged 30–62 years of age received biennial physical examinations, risk factor assessments, and surveillance for CVD events (Wong and Levy 2013). The Framingham study was instrumental in the original identification of many of the factors that are associated with an increased risk of CVD, and in 1961, coined for the first time, the term ‘risk factors’ that is widely utilized in many different fields of medicine today (Kannel et al. 1961). The increased awareness of major risk factors for CVD initially identified by the Framingham Heart Study and by other researchers provided the impetus for important public health initiatives against smoking in the 1960s, hypertension in the 1970s, and hypercholesterolemia in the 1980s. More recently, obesity and physical inactivity have also been recognized as key risk factors for CVD. Diabetes is also now widely regarded as a CHD risk equivalent and the importance of a clustering of major cardiometabolic risk factors, commonly referred to as the metabolic syndrome, has received significant attention from the research and clinical community (Grundy et al. 2005). Risk factors often cluster together, and the number of risk factors present and their co-occurrence and is directly related to the incidence of CHD (Fig. 8.1.3). Importantly, it was the Framingham Heart Study that first introduced the concept of multivariable or global risk assessment for coronary heart disease (see also section to follow on global risk assessment) (Wilson et al. 1998).



	A	B	C	D
Blood Pressure (mmHg)	120/80	140/90	140/90	140/90
Total Cholesterol (mg/dl)	200	240	240	240
HDL Cholesterol (mg/dl)	50	50	40	40
Diabetes	No	No	Yes	Yes
Cigarettes	No	No	No	Yes

Fig. 8.1.3 Estimated 10-year CHD risk in 55-year-old adults according to levels of various risk factors: Framingham Heart Study

Source: data from Wilson PWF et al. Prediction of Coronary Heart Disease Using Risk Factor Categories. *Circulation* 1998 97: 1837–47.

Family history and genomics. A family history of premature CHD is a well-established, but unmodifiable risk factor for future CHD, and can sometimes be the crucial and single most important risk factor in predisposing an individual to early CHD (Hopkins et al. 2000; Williams et al. 2001). A large proportion of heart attacks or strokes occurring at a young age are felt to be attributable to inherited or familial predisposition. Hence, knowledge of an individual's family history can help guide preventive efforts. A family history of premature CHD is generally defined as having a

male first-degree relative experiencing a first manifestation of CHD under the age of 45, or a female first-degree relative experiencing CHD under the age of 55. The number of affected relatives with premature CHD is also felt to be an important factor, since those with 1 affected relative can be shown to have a fourfold greater risk of CHD, while those with 2 or more affected relatives may have more than a 12-fold greater risk of CHD compared to those without any affected relatives (Hopkins et al. 2000). Moreover, it has been shown that 35% of all early CHD occurs in just 3.2% of families, all of whom have a strong positive family history of CHD. Familial hypercholesterolemia (FH) is an autosomal dominant condition associated with very high levels of total and LDL-cholesterol (total cholesterol levels exceeding 500 are commonly reported) where individuals inherit from one or both parents (homozygous) a defective LDL-cholesterol receptor gene. Affected individuals have been known to have heart attacks and die by age 20 and FH is among the most widely studied genetic conditions responsible for CHD. Other genetic defects responsible for hypertension, obesity, diabetes, and other major cardiovascular risk factors have also been identified and are the subject of major investigations (Williams et al. 2001).

In recent years, with the ability to sequence the entire human genome, key single nucleotide polymorphisms (SNPs) have been identified and linked to the likelihood of CHD, although with a much higher threshold required to confirm a statistically significant relationship ($p = 10^{-8}$) (Schunkert et al. 2011). One that has received significant attention is 9p21 because of its strong association with CHD (Palomaki et al. 2010). However, the utility of such SNPs in improving the ability to predict CHD over traditional risk factors is only modest and current guidelines have not recommended genomic screening due to the thus far limited demonstrated clinical utility.

Of great recent interest is the emerging role of Mendelian randomization studies to estimate the effect of lifetime exposure to a given risk factor (or absence of risk factor) on risk of CHD. For instance, in one meta-analysis of studies involving 312,321 participants, examining the effect of long-term exposure to lower LDL-cholesterol on the risk of CHD mediated by 9 polymorphisms in 6 different genes found such naturally random allocation to lower LDL-C was associated with a 55% lower risk of CHD for each mmol/l lower LDL-C, which is approximately three times greater than that observed from statin treatment started later in life (Ference et al. 2012).

Diabetes and metabolic syndrome. Diabetes mellitus is a major risk factor for CVD, and is associated with a greater risk for CHD, stroke, chronic kidney, and peripheral vascular disease (Kaseta and Sowers 2005). There are wide variations in prevalence according to gender and ethnicity. In 2011–2014, among US adults, the prevalence of physician-diagnosed diabetes mellitus was lowest in whites (7.4% in females and 8.0% in males) and highest in non-Hispanic Blacks (13.6% in females and 14.1% in males) (Benjamin et al. 2018).

More than three-fourths of those with diabetes die of cardiovascular complications, most notably myocardial infarction and stroke, although peripheral arterial disease and heart failure are among the most common initial first CVD manifestations in persons with diabetes. Given this point, the National Cholesterol Education Program designated diabetes as a CHD risk equivalent (Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults 2001) because the risk of CHD in those with diabetes without known heart disease had been shown to be similar to recurrent CHD events in those with CHD (but without known diabetes), although more recent studies are showing that this is not always the case, with recent meta-analyses showing those with diabetes (without CHD) have about half the risk of future CHD events as those with CHD (but without diabetes). Population-based studies such as Framingham show about a 2–3-fold greater risk of CVD events in men with diabetes as compared to men without diabetes, but that in women, the corresponding figure is a 3–7-fold greater risk.

Diabetes is typically diagnosed by a glycated haemoglobin (HbA1c) level of 6.5% or greater, fasting glucose $\geq$126 mg/dl (7 mmol/l) or glucose level of $\geq$200 mg/dl (11.1 mmol/l) from the non-fasting state or after a 2-hour post-load glucose challenge test (American Diabetes Association 2018a). Pre-diabetes is diagnosed by a fasting glucose of 100–125 mg/dl or casual glucose of 140–199 mg/dl.

Clinical trial evidence to show whether intensive glycaemic control in persons with diabetes lowers CVD event rates has shown mixed results. The United Kingdom Prospective Diabetes Study (UKPDS) originally showed a borderline, non-significant ($p = 0.052$) 16% reduction in risk of myocardial infarction among those with newly diagnosed type 2 diabetes who were treated with intensive glucose-lowering therapy, but the 10-year post-interventional follow-up of this study, which was recently published, showed

that these benefits continued with a 15% significant ($p = 0.01$) reduction in risk of myocardial infarction (Holman et al. 2008). A similar, extended post-interventional follow-up of the Diabetes Control and Complications Trial of Type 1 diabetes, which did not initially show a significant reduction in CVD events at the end of the randomized trial, showed years later a continued effect of the original intensive glucose-lowering treatment, with a 42% reduction in the risk of any cardiovascular event (The Diabetes Control and Complications Trial 2005). Such a continued post-trial effect of the randomized glycaemic therapy has been termed a 'glycaemic legacy' effect. More recently, however, three important randomized clinical trials (the Action to Control Cardiovascular Risk in Diabetes (ACCORD), ADVANCE and the Veterans Affairs Diabetes Trial (VADT)) failed to show that intensive glucose-lowering to a HbA1c range of 6–6.5%, significantly reduced CVD event rates, as compared to more standard maintenance of the HbA1c in the 7–9% range. In fact, those randomized to the intensive therapy group in ACCORD actually had a higher rate of CVD mortality (The Action to Control Cardiovascular Risk in Diabetes Study Group 2008), although this increased risk was restricted to those with prior macrovascular disease and where there was unsuccessful lowering of the glucose as a result of the intensive therapy. Those with less complicated diabetes (e.g. patients who were more recently diagnosed and without prior macrovascular disease) actually showed a benefit in the primary endpoint. Moreover, an important lipid substudy of ACCORD tested the efficacy of adding the fibric acid derivative fenofibrate or placebo to ongoing statin therapy and found no benefit in terms of reducing CVD event rates (although there was a benefit seen in the subgroup with high triglycerides and low HDL-C) (ACCORD Study Group 2000a). In addition, in the blood pressure substudy of ACCORD, intensive blood pressure therapy to achieve a systolic blood pressure <120 mmHg, as compared with <140 mmHg, was also found to provide no significant benefit in terms of CVD risk reduction, although there was a significant benefit in stroke reduction (ACCORD Study Group 2010b).

Most recently, however, newer diabetes agents, in particular recent trials of SGLT-2 inhibitors and GLP-1 receptor agonists have finally demonstrated benefit in cardiovascular outcomes trials. SGLT-2 inhibitors lower blood glucose by urinary excretion and have other positive benefits, such as on blood pressure, arterial stiffness, and weight/adiposity, and oxidative stress.

The first of these, empagliflozin, was tested in the EMPA-REG outcomes trial among persons with diabetes and established CVD, demonstrating a significant 14% reduction in the primary composite cardiovascular outcome, but a greater 38% reduction in cardiovascular death and 35% reduction in heart failure hospitalizations (Zinman et al. 2015). Later, the CANVAS trial involving canagliflozin showed a similar benefit in the primary composite cardiovascular outcome and in heart failure hospitalizations, but no significant benefit in cardiovascular mortality (Neal et al. 2017). GLP-1 receptor agonists are incretin-based therapies that stimulate insulin secretion and suppress glucagon secretion and may have other effects including reducing inflammation, blood pressure, body weight, and improving beta-cell function. The LEADER trial involving liraglutide is the first cardiovascular outcomes-driven trial of a GLP-1 reception agonist to show a benefit with a significant 13% reduction in the primary composite cardiovascular outcome (Marso et al. 2016).

While the concept of the metabolic syndrome has waned over recent years, this condition referring to a constellation of cardiometabolic risk factors that are associated with a greater risk of developing future diabetes and CVD has been useful for the purposes of defining those with multiple risk factors who may benefit from a more intensified, unified approach at risk factor modification. Studies have documented that 30–40% of adults have metabolic syndrome in the United States, with even higher prevalence rates in some regions such as in the Middle East. Certain definitions of the metabolic syndrome place insulin resistance or abdominal obesity as the necessary condition, with additional conditions including elevated blood pressure, low HDL-cholesterol, elevated triglycerides, and impaired fasting glucose making up the definition. The American Heart Association/National Heart Lung and Blood Institute definition (Grundy et al. 2005) requires the presence of at least three of the following five criteria: abdominal obesity defined by a waist circumference >35 inches in women or >45 inches in men; HDL-cholesterol <40 mg/dl in men or <50 mg/dl in women, fasting triglycerides $\geq$150 mg/dl; elevated blood pressure of $\geq$130 mmHg systolic or $\geq$85 mmHg diastolic or on hypertensive therapy; or impaired fasting glucose defined as $\geq$100 mg/dl (5.6 mmol/L) or on hypoglycaemic therapy. It is important to recognize that other definitions exist, such as that from the International Diabetes Federation which requires the presence of abdominal obesity as measured by lower waist circumference cut-off points in

European Caucasians (>94 cm in men and >80 cm in women), and among most Asian groups and those of Central and South American ancestry (where >80 cm in women and >90 cm in men define increased waist circumference) (Alberti et al. 2005).

Numerous studies have shown an increased risk of future CVD events in persons with the metabolic syndrome. Among United States adults (Malik et al. 2004), a stepwise increase in risk for CHD, CVD, and total mortality has been shown relating to the presence of metabolic syndrome alone, metabolic syndrome with diabetes or pre-existing CVD, and metabolic syndrome with both CVD and diabetes (Fig. 8.1.4).

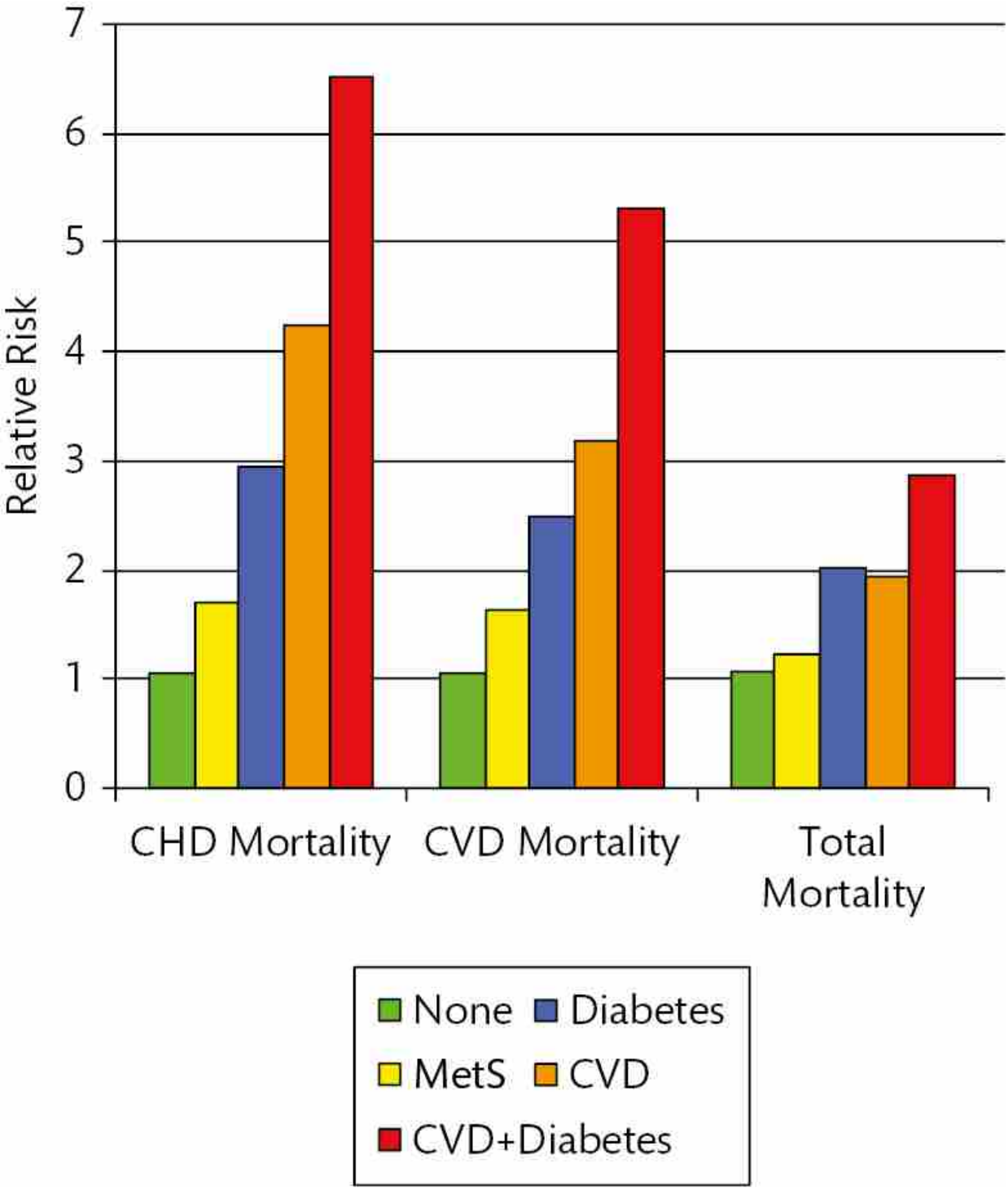


Fig. 8.1.4 Relative risks for mortality from coronary heart disease, cardiovascular disease, and all causes in United States Adults associated with the presence of metabolic syndrome with and without diabetes: NHANES II: 1976–80 follow-up study (mean 13 years follow-up).

From Malik S, Wong ND, Franklin SS, et al. Impact of the metabolic syndrome on mortality from coronary heart disease, cardiovascular disease, and all causes in United States Adults. *Circulation*; 2004: 110: 1239–44, with permission from Wolters Kluwer.

Hypertension. Elevated blood pressure, and particularly systolic blood pressure, is strongly and positively related to the risk of future CHD and stroke. Hypertension is currently defined as a systolic blood pressure of 130 mmHg or higher, diastolic blood pressure of 80 mmHg or higher (or on pharmacologic treatment to lower blood pressure). Levels of 130–139 mmHg systolic or 80–89 mmHg diastolic are defined as Stage 1 hypertension, and levels of 140 mmHg or higher systolic or 90 mmHg or higher diastolic are defined as Stage 2 hypertension based on the new 2017 ACC/AHA guidelines (Whelton et al. 2018). Based on US NHANES data from 2011 to 2014, this results in the identification of 45.6% of US adults (103.3 million) with hypertension, of which 36.2% (81.9 million) would be recommended for treatment, and among those treated 53.4% (29.2 million) would still be above target blood pressure goal (Muntner et al. 2018). There is a strong age-related increase in prevalence of hypertension, and based on the older cut-point of 140/90 mmHg, prevalence ranges from under 10% in those <35 years of age and progressively increases with age to more than 70% in those aged 75 years and over with higher prevalences in men before age 55, but higher prevalences in women after age 65 (reaching 81.2% among those aged 75 years and over). Of note, isolated diastolic hypertension (systolic blood pressure of <140 mmHg but diastolic blood pressure of 90 mmHg or higher) is the most common form of hypertension under age 50, while isolated systolic hypertension (systolic blood pressure of 140 mmHg or higher, but diastolic blood pressure <90 mmHg) predominates over the age of 60 years (Franklin et al. 2001).

For every blood pressure increase of 20/10 mmHg (systolic/diastolic) above 115/75 mmHg, the risk of death from CHD, stroke, or CVD doubles. Numerous clinical trials have shown that lowering blood pressure substantially decreases the risk of future cardiovascular events, stroke, and end-stage renal disease. A large meta-analysis of active treatment (with initial low dose diuretic, beta blocker, ACE inhibitor or calcium antagonist) compared to placebo showed significant reductions in the risk for heart failure (36%), stroke (33%), CHD (16%), CVD death (15%), overall CVD (25%) and total mortality (12%) (Prospective Studies Collaborative 2002; Elliott 2005). Of note, the recent landmark Systolic Blood Pressure Intervention Trial (SPRINT) showed those non-diabetic but higher cardiovascular risk participants targeted to a systolic blood pressure <120 mmHg compared to <140 mmHg (with actual achieved blood pressures of

121 mmHg and 136 mmHg, respectively) was stopped early due to definitive benefit with the intervention group showing a 25% reduction in the primary composite CVD endpoint with a substantial 38% reduction in heart failure hospitalization and 43% reduction in cardiovascular mortality. All-cause mortality was also significantly reduced by 27% (SPRINT Research Group et al. 2015).

Dyslipidaemia. Increased levels of total and low density lipoprotein (LDL)-cholesterol have long been recognized as major risk factors for CHD. A direct curvilinear relation exists between total and low density lipoprotein (LDL)-cholesterol and the risk of CHD. For example, compared to those with total cholesterol levels of 200 mg/dl, those with levels of 240 mg/dl are at approximately twofold greater risk, while those with levels of 300 mg/dl have a fourfold increased risk. However, there is significant overlap in total or LDL-cholesterol levels between those who experience CHD events versus those who do not and approximately one-third of heart attacks occur in persons with “normal” levels of total cholesterol below 200 mg/dl. Importantly, a low level of high density lipoprotein-cholesterol (typically defined as <40 mg/dl in men and <50 mg/dl in women), regardless of level of total cholesterol, is strongly associated with an increased risk of CHD (Castelli et al. 1986). For example, across levels of total cholesterol from <200 mg/dl to ≥260 mg/dl, the 14-year incidence of CHD varies only from 11.2 to 12.5%, whereas when HDL-C >60 mg/dl, these rates are 3.7% to 3.8%, respectively.

Table 8.1.1 Prevalence of elevated total and LDL-cholesterol and low HDL-cholesterol by gender and ethnicity, US adults aged 20 years and over, National Health and Nutrition Examination Survey 2011–2014

Population	Total cholesterol ≥200 mg/dl	Total cholesterol ≥240 mg/dl	LDL-cholesterol ≥130 mg/dl	HDL-cholesterol <40 mg/dl
All	39.7%	11.9%	30.3%	18.7%
Males	37.0%	10.6%	30.0%	27.9%
Females	42.0%	13.0%	30.4%	10.0%
White males	37.0%	10.8%	29.3%	28.4%
White females	43.4%	13.8%	32.1%	10.3%
Black males	32.6%	7.3%	29.9%	20.7%
Black females	36.1%	9.6%	27.9%`	8.0%
Mexican-American males	43.1%	13.6%	36.6%	30.7%
Mexican-American females	41.2%	12.5%	28.7%	11.8%
Asian males	39.9%	10.8%	29.2%	25.0%
Asian females	40.5%	11.2%	25.0%	6.7%

Source: data from Benjamin EJ, Virani SS, Callaway CW, et al. American Heart Association Council on Epidemiology and Prevention Statistics Committee and Stroke Statistics Subcommittee. Heart Disease and Stroke Statistics-2018 Update: A Report from the American Heart Association. Circulation. 2018; 137(12):e67–e492.

Other lipid abnormalities include elevated serum triglyceride levels (the normal level is considered to be <150 mg/dl), small particle size LDL-cholesterol (known as small dense LDL-cholesterol) and elevated lipoprotein (a) levels. There has also been recent interest in non-HDL-cholesterol since it captures all the atherogenic apolipoprotein B-containing lipoproteins and does not require, unlike LDL-cholesterol, measurements to be made after fasting. Several epidemiological studies have demonstrated a stronger relationship of non-HDL-C compared to LDL-cholesterol with incidence of CHD.

Table 8.1.1 shows the prevalence of elevated total and LDL-cholesterol and low HDL-cholesterol levels among US adults aged 20 years and over in 2011–2014 (Benjamin et al. 2018).

Numerous primary and secondary prevention clinical trials have documented the efficacy of lowering LDL-cholesterol, particularly with HMG-CoA reductase inhibitor drugs ('statins'), which can lead to approximately 25–35% reductions in CHD and CVD incidence, where the lower the LDL-C achieved, the lower were the event rates (Rosensen 2004). These trials also showed no threshold level of LDL-C reductions where the CVD event rates begin to flatten. A large meta-analysis of 14 statin trials comprising 90,056 subjects showed that over a 5-year treatment period, every mmol/L reduction (approximately 40 mg/dl) in LDL-cholesterol achieved, was associated with reductions in all-cause mortality of 12%, coronary mortality of 18%, myocardial infarction and CHD death of 23%, stroke incidence of 17%, all major vascular events of 21%, and with no difference (relative risk 1.00) in cancer incidence (Baigent et al. 2005). Importantly, several trials have shown the additional benefit of reducing the LDL-C, with high dose statin therapy, to levels below 100 mg/dl. The largest of these, the Treat to New Targets (TNT) study showed that the achieved LDL-C of 77 mg/dl with high dosage atorvastatin therapy, when compared to achieved LDL-C of 101 mg/dl with low dose atorvastatin therapy, resulted in an additional 22% reduction in major cardiovascular events (LaRosa et al. 2005).

Given that most randomized clinical trials of LDL-C lowering have demonstrated no more than 40% reductions in CVD incidence, addressing the remaining 'residual' risk has been of great interest. One factor that has been specifically investigated has been whether the raising of HDL-C levels will provide further benefit in reducing CVD events beyond that afforded by LDL-C lowering, given the strong protective effect of increased HDL-C levels suggested by observational studies. The AIM-High Trial (AIM-HIGH Investigators et al. 2011) was designed to be a rigorous test of the HDL hypothesis, randomizing 3,414 men and women with a prior CVD event and low HDL-C, and who had well-controlled LDL-C levels averaging 74 mg/dl on a statin, to extended-release niacin or placebo. LDL-C levels were kept similar between the groups, and the on-treatment HDL-C level averaged 5 mg/dl higher in those on niacin. However, the trial was discontinued early due to futility, with no significant reduction in CVD events observed (HR =

1.02, p = 0.79). The failure of this trial, as well as that of a second much larger niacin trial, HPS2-Thrive as well as the failure of three of four major trials of cholesterol ester transferase protein (CETP) inhibitors have raised significant questions as to the benefit of raising HDL-C levels, particularly in patients who are already well-controlled on statins.

In recent years, evidence has accumulated regarding the cardiovascular benefits of newer non-statin therapies. In the IMPROVE-IT trial of persons with a recent (within 10 days) acute coronary syndrome, the addition of the cholesterol absorption inhibitor ezetimibe to simvastatin compared to simvastatin alone resulted in a 6% relative risk reduction for the primary composite cardiovascular endpoint, although given the high-risk study population, the absolute risk reduction was 2%, resulting in a number needed to treat of 50 for the prevention of a CVD event (Cannon et al. 2015). In addition, with the availability of proprotein convertase subtilisin kexin type 9 (PCSK9) monoclonal antibody therapies evolocumab and alirocumab, which provide average 50–60% reductions in LDL-C above maximal tolerated statin therapy, recently reported cardiovascular outcomes trials (Sabatine et al. 2017; American College of Cardiology 2018) have demonstrated further 15% relative risk reductions for the composite CVD endpoint, with greater benefit being observed with time and in those with higher risk features (e.g. more recent myocardial infarction or number of prior cardiovascular events).

Cigarette smoking. Tobacco smoking is among the leading preventable causes of death globally. Numerous studies have linked tobacco use to the incidence of, and mortality from, CVD, with approximately a half-million deaths annually in the United States being attributed to tobacco use. Further, environmental tobacco smoke (‘second-hand smoke’) is responsible for approximately 40,000 deaths from heart disease annually in the United States. Cigarette smokers are 2 to 4 times more likely to develop CHD than non-smokers. Also, their risk of stroke is doubled and risk for peripheral vascular disease is more than 10 times higher than that of non-smokers (Luepker and Lando 2005). The most recent data among US adults in 2015 show an overall prevalence of smoking of 15.1%, with a wide variation by ethnic groups, ranging from 7% in Asian to 21.9% in American Indian/Alaska Natives (Benjamin et al. 2018).

Obesity and physical inactivity. Recent estimates showed that two-thirds of US adults are overweight or obese (body mass index of 25 kg/m^2 or

higher), with 30% being obese (body mass index of 30 kg/m^2 or higher). Since the 1960s, there has been a dramatic nearly threefold increase in obesity prevalence among men from 10.7% to 34.4% and more than a twofold increase in women from 15.7% to 36.1%. Moreover, 17.4% of children aged 6–11 and 20.6% aged 12–17 are defined to be obese based on latest survey data from NHANES 2013–2014. Further recent estimates from NHIS 2015 showed that half (49.7%) of the US adult population (52.9% of men and 46.7% of women) met nationally recommended amounts of physical activity (150 minutes of moderate, or 75 minutes of vigorous activity or an equivalent combination per week) (Benjamin et al. 2018).

Obesity has been shown by numerous studies to be associated with an approximate 1.5- to 2-fold increase in risk of death from CHD, with the increase in risk beginning below the 25 kg/m^2 cut-off point for overweight. Numerous studies also show approximately 20–40% lower risks of mortality and cardiovascular events associated with increased levels of physical activity or measured fitness. Several cardiovascular risk factors are linked to obesity, including hypertension, dyslipidaemia (including low HDL-cholesterol levels), type 2 diabetes, obstructive sleep apnoea, and hyperinsulinemia. Increases in physical fitness have also been shown to be linked to increases in HDL-cholesterol levels and reductions in systolic and diastolic blood pressure, insulin resistance, and glucose intolerance. Abdominal obesity, most commonly indicated by a waist circumference of >40 inches in men or >35 inches in women, is a major component of the metabolic syndrome (McCowen and Blackburn 2005). *Measures of cardiovascular health and CVD risk*. In recent years, the American Heart Association (AHA) (Benjamin et al. 2018) has promoted a goal to improve the cardiovascular health of all Americans by 20% while reducing mortality from heart disease and stroke by 20% by the year 2020. These goals introduce the concept of promoting cardiovascular health which is more positive and motivating than the concept of preventing *cardiovascular disease*, and have focused on examining the proportion of individuals (and the relation to CVD risk) who achieve one or more “ideal levels” based on the following seven metrics (also known as AHA’s Life’s Simple Seven™): (1) cigarette smoking (non-smoking is ideal); (2) physical activity (150 minutes or more moderate intensity or equivalent exercise per week is ideal); (3) body mass index (<25 kg/m^2 is ideal); (4) healthy diet (achieving at least 4 of 5 key dietary components focusing on fruit/vegetable, fish, fibre, and

sodium intake and sweetened beverage intake); (5) cholesterol (<200 mg/dl ideal in adults, <170 mg/dl in children), 6) blood pressure (<120/80 mmHg is ideal); and (7) fasting plasma glucose (<100 mg/dl is ideal). Fewer than 20% of adults and only about 40% of children meet five or more of these criteria for ideal cardiovascular health and fewer than 1% of adults are at ideal levels of all seven metrics of cardiovascular health. Moreover, CVD incidence varies more than 10-fold according to the number of cardiovascular health behaviours and factors that are present (Folsom et al. 2011) (Fig. 8.1.5).

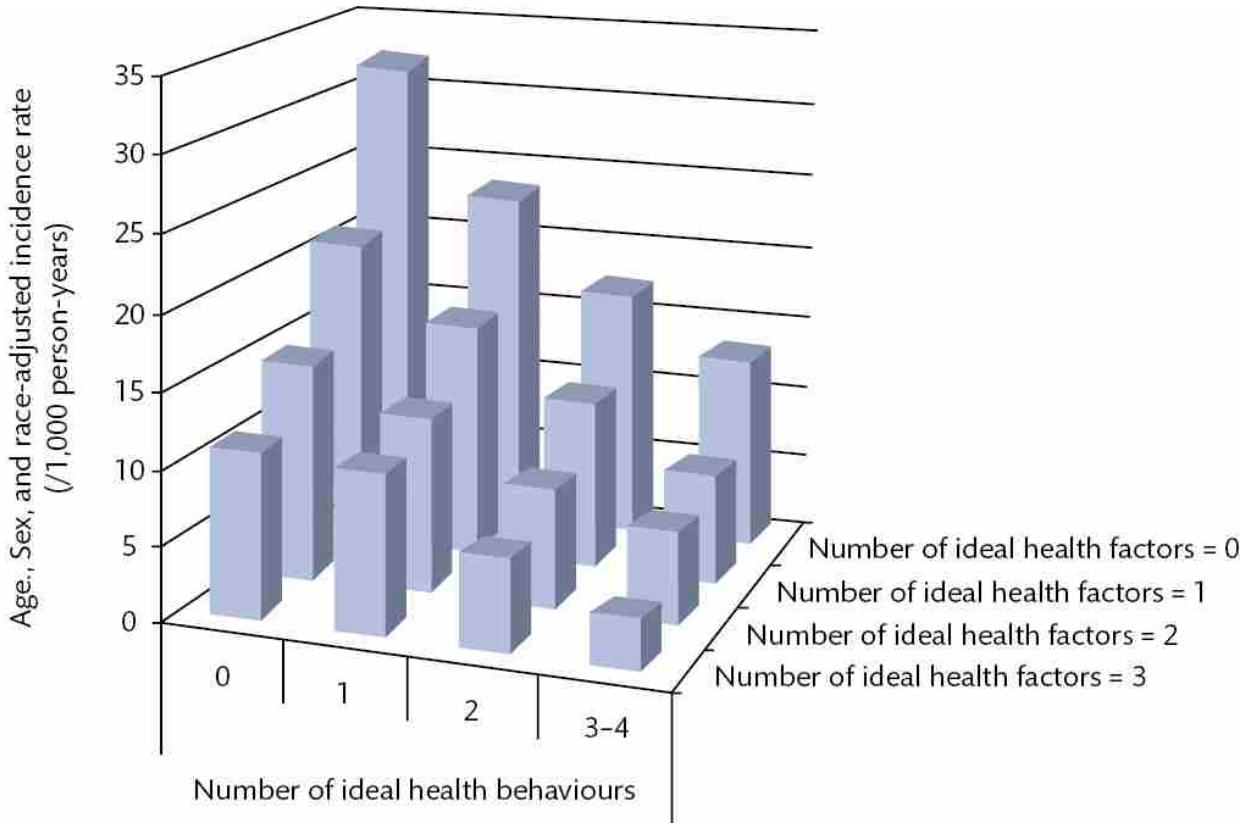


Fig. 8.1.5 Incidence of CVD according to the number of ideal health behaviours and health factors. Reprinted from Folsom AR, Yatsuya H, Nettleton JA, et al. Community prevalence of ideal cardiovascular health by the American Heart Association definition and relation with cardiovascular disease incidence. *J Am Coll Cardiol* 2011; 57: 1690–6 with permission from Elsevier.

Risk factors for stroke

Important risk factors for stroke include elevated blood pressure, diabetes mellitus, disorders of heart rhythm, low HDL-cholesterol, cigarette smoking, family history, and chronic kidney disease (WHO 2011). Elevated blood pressure is the most important risk factor for both ischaemic stroke and intracranial haemorrhage, with about 77% of those with a first stroke having a blood pressure of 140/90 mmHg or higher. The Framingham Heart Study demonstrated the importance of mid-life antecedent hypertension as well as systolic hypertension as important predictors of stroke (Romero and Wolf 2013). Even prehypertension is associated with incident stroke. Randomized clinical trials consistently show the efficacy of blood pressure control for prevention of stroke. Diabetes is also a strong predictor of stroke, although mainly before age 55 years of age in blacks and before age 65 years of age in whiles. Atrial fibrillation is another powerful risk factor for stroke, independently increasing risk approximately fivefold at all ages, with nearly one-quarter of strokes attributable to atrial fibrillation in the ninth decade of life. While elevated total cholesterol levels are not strongly predictive of stroke, some studies have found low HDL-cholesterol levels to predict risk, but clinical trials of statin therapy consistently show benefits of such therapy for preventing stroke. Prospective studies have also shown moderate to vigorous physical activity to be associated with an overall 35% reduction in risk of ischaemic stroke compared with no physical activity. A family history of parental ischaemic stroke before the age of 65 has also been shown to be associated with a threefold increase in ischaemic stroke in offspring, even after adjustment for other risk factors. Finally, increased levels of creatinine and reduced glomerular filtration rate (eFGR) have been shown to predict future stroke and symptoms.

Risk factors for heart failure

Hypertension is among the most prominent of risk factors for heart failure, followed by a preceding myocardial infarction, elevated B-type natriuretic peptide levels, and urinary albumin-to-creatinine ratio. In the Physicians Health Study, the lifestyle risk of heart failure was greater in those with hypertension and less in those with healthy lifestyle factors, including normal weight, not smoking, regular exercise, moderate alcohol intake, consumption of breakfast cereals, fruits, and vegetables (Djousse et al. 2009). The Framingham Heart Study was key in demonstrating the poor

prognosis associated with heart failure in the community, with 1 of 2 individuals dying within 5 years of diagnosis and the strong association of hypertension conferring nearly an eightfold risk in men and fourfold risk in women for developing heart failure. They also documented other predictors of new onset heart failure, including elevated plasma natriuretic peptide levels, asymptomatic left ventricular systolic dysfunction, and increased left ventricular diastolic dimension (Mahmood and Wang 2013).

Risk factors for peripheral arterial disease

Peripheral arterial disease shares many of the same risk factors that coronary heart disease does, although diabetes and cigarette smoking are stronger risk factors for peripheral arterial disease than for coronary heart disease with odds ratios in the range of 3 to 4. Another study showed hypertension, diabetes, chronic kidney disease, and smoking were associated with incident peripheral arterial disease. Meta-analysis of 16 cohort studies indicated people at with low ankle brachial index (ABI) (<0.9) had around threefold risk of all-cause death as well as cardiovascular death in both sexes.

Cardiovascular disease risk assessment

Global risk scores for cardiovascular disease risk assessment. In order to understand an individual's risk of CVD or CHD, it is important to understand an individual's "global risk" of developing the condition. Most risk scores developed over the past few decades, including those from the Framingham Heart Study in the United States as well as those from other countries have involved the determination of the future risk of developing CHD in the next 10 years. These have been used to stratify a person's risk for consideration or intensification of risk factor management, especially for lipid-lowering therapy. More recently, algorithms for total CVD risk from Framingham have been developed, which are more global as they incorporate the combined risk for CHD, stroke, heart failure, and peripheral arterial disease (D'Agostino et al. 2008). In late 2013, the ACC/AHA released the pooled cohort risk calculator, which allows for the prediction of 10-year risk (in those aged 40–79 years of age) and lifetime risk (in those aged 20–59 years of age) of atherosclerotic cardiovascular disease (ASCVD)

comprising of non-fatal or fatal myocardial infarction or stroke (Goff et al. 2014), and which is currently the standard for risk assessment recommended today for clinical practice.

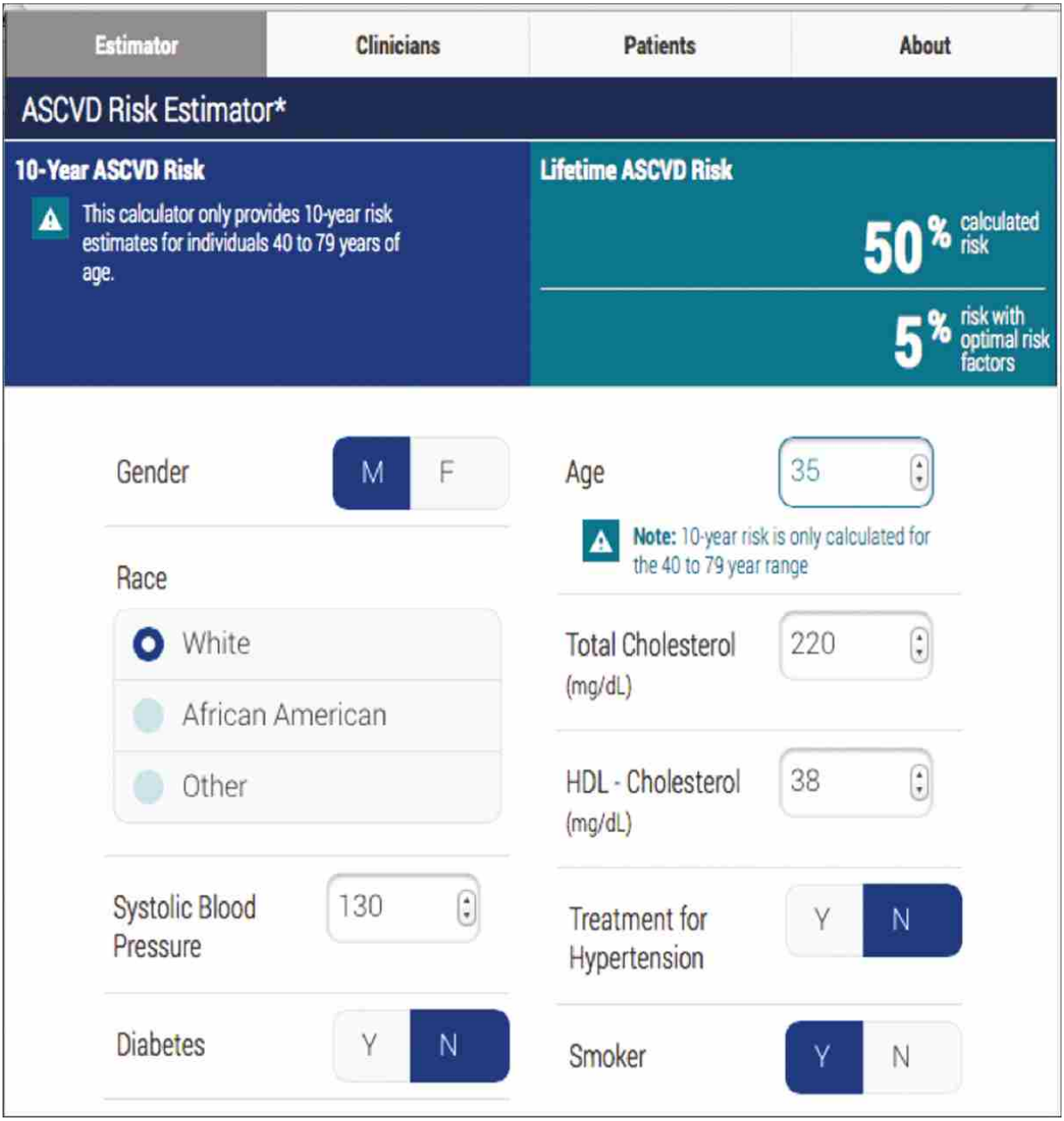


Fig. 8.1.6 Pooled cohort risk calculator.
From https://clincalc.com/cardiology/ascvd/pooledcohort.aspx.

In those where a global risk estimate is obtained, depending on the risk status, the appropriate intensity of treatment for given risk factors is considered (see next under cholesterol management). While estimation of global risk as described earlier is recommended, it may be best considered a starting point in risk assessment and a strong emphasis is to use the risk calculator as part of the patient-clinician 'risk discussion' on the best ways to lower cardiovascular risk through both lifestyle and use of appropriate preventive therapies. In addition, the presence of other risk factors not included in the risk algorithms, such as abdominal obesity, elevated triglycerides (if severe enough), impaired fasting glucose, or a strong positive premature family history, if present, however, could be used by the clinician to stratify an individual's risk or required intensity of treatment upward. The ACC/AHA 2014 Guideline on Cardiovascular Risk Assessment (Goff et al. 2014) specifically notes that if the treatment decision is uncertain on the basis of the risk calculator that the following additional measures can be used for consideration to further inform the treatment decision: (1) family history of premature CVD (which should always be done as part of patient evaluation); (2) high-sensitivity C-reactive protein (with a level of >2 mg/L indicating higher risk); (3) coronary calcium scoring (with a score of >300 mg/dl or >75th %tile for age and sex warranting a high risk level); or (4) ABI (with a level of <0.9 indicating higher risk and diagnostic of peripheral arterial disease) (see next).

Inflammatory measures for cardiovascular disease risk. High levels of inflammatory markers, including high-sensitivity C-reactive protein (hs-CRP), a marker of systemic inflammation, as well as lipoprotein phospholipase associated A2 (LpPla2), a measure of vascular inflammation have been shown in numerous studies to be independently related to the risk of future cardiovascular events. IL-6, other interleukins, myeloperoxidase, and tumour necrosis factor-alpha (TNF-alpha) also have inflammatory activity and have been shown to be associated with cardiovascular events in some studies, as have other biomarkers such as brain naturietic peptide (BNP) and troponin levels. Levels of C-reactive protein exceeding 2 mg/L have been recommended as a higher risk level that could inform the treatment decision beyond global risk scoring (Goff et al. 2014). There are no current recommendations for the measurement of other inflammatory markers or biomarkers for the assessment of CVD risk in asymptomatic adults, however. The recent CANTOS trial recently provided proof of

concept for the role of reducing/modifying inflammation in high risk individuals for the reduction of CVD events (Ridker et al. 2017). In this trial, high risk persons with known CVD were randomly assigned on top of optimal treatment to the monoclonal antibody to Il-1β, canakinumab for a median follow-up of 3.7 years and showed a significant 15% relative risk reduction in the primary CVD endpoint.

Subclinical cardiovascular disease assessment. The ability to image or directly measure the atherosclerotic burden has been of great interest but whether these measures provide added value over less expensive and easier traditional office-based risk factor measures (including global risk assessment scores) has been an important question in cardiovascular epidemiology and prevention over the past 15 years (Greenland et al. 2010). Such measures must have a high level of sensitivity and specificity for detection of the disease, must be reproducible and cost-effective, and should provide added clinical utility over current office-based risk assessment.

Increased carotid intima medial thickness (CIMT) as measured from carotid B-mode ultrasound, has been shown to be related to the risk of future CVD events from numerous prospective epidemiologic studies over the past 25 years, including the Atherosclerosis Risk in Communities (ARIC), Cardiovascular Health Study (CHS), and the Multiethnic Study of Atherosclerosis (MESA). CIMT has also been used as a surrogate endpoint in numerous clinical trials involving lipid-lowering and other interventions. Moreover, the presence of carotid plaques provides additional information and together with CIMT measurements have modest clinical utility for reclassification of risk in asymptomatic intermediate risk adults. While the most recent ACC/AHA Guidelines on Cardiovascular Risk Assessment do not recommend the measurement of CIMT alone for the assessment of CVD risk (Goff et al. 2014), due to its modest improvement in risk prediction from recent meta-analyses, more recent European guidelines have emphasized the assessment of carotid plaque together with CIMT.

ABI measured from ankle and arm blood pressures utilizing a Doppler, provides for an assessment of peripheral arterial disease, which is diagnosed to be present when the ABI is less than 0.9. A low ABI has also been shown to be related to increased total and CVD mortality in numerous observational studies. However, as ABI tends not to be low until after the age of 60 years, it is therefore most useful as a screening tool for peripheral arterial disease in older persons. It is also a recommendation for further informing the

treatment decision if uncertain from global risk assessment alone (Goff et al. 2014).

Measures of coronary artery calcium as assessed by computed tomography (CT), have been shown in numerous prospective studies, to be associated with the future risk of CHD and CVD events, independent of standard risk factors (Greenland et al. 2010). Coronary calcium measures also provide incremental predictive value and added reclassification of risk over global risk assessment for the prediction of cardiovascular events. In the MESA study, it has been shown to provide greater incremental risk prediction among intermediate risk individuals (from C-statistic improvement) than any other major biomarker or subclinical disease screening test (Detrano et al. 2008). CAC testing is also felt to be the most useful of the current approaches for improving risk assessment in intermediate risk subjects (Goff et al. 2014).

There are also numerous other tests which have varying levels of evidence relating them to CVD event risk. These include assessments of endothelial function, pulse wave velocity, brachial artery reactivity, and imaging of soft plaque such as by CT coronary angiography. Most of these other tests, however, do not have the quantity, strength, or consistency of evidence that CIMT, ABI, or CAC assessment do, and hence there is not the same level of consensus for recommending their use for the screening of asymptomatic adults.

Preventive strategies for cardiovascular disease risk reduction

Cardiovascular epidemiology and disease prevention encompass an extensive field investigating the distribution and variation of CVD conditions, most notably coronary heart disease and stroke, their risk factor determinants, and strategies at the population and individual level aimed at preventing the development or recurrence of CVD. Epidemiologic approaches to studying CVD provide us with the tools for preventive efforts at the individual and population level. *Primordial prevention* is aimed at prevention of the risk factors for CVD, such as efforts aimed to prevent hypertension, obesity, or dyslipidaemia. *Primary prevention*, which focuses on the modification of these and other known risk factors, is aimed at preventing the clinical manifestations of CVD, such as myocardial infarction and stroke. *Secondary prevention* focuses on those who already have

manifestations of disease, but where aggressive control of risk factors can have a major impact in preventing recurrences of disease. Concerted efforts between governmental agencies, the community, and the private sector are required to best address the continuing epidemic of CVD.

Preventive strategies can be considered using the 'ABCDE' approach addressing aspirin/antiplatelet therapy (A), blood pressure and hypertension (B), cholesterol and dyslipidaemia (C), diet and nutrition (D), and exercise and physical activity (E). For secondary prevention among persons with pre-existing CVD, guidelines have been published by the American Heart Association/American College of Cardiology (Smith et al. 2011).

Aspirin and Antiplatelet Therapy. Aspirin prophylaxis (81–162 mg/day) has been recommended for persons at intermediate risk, for the prevention of CHD in men and prevention of stroke in women, including those with diabetes or metabolic syndrome. Additionally, among those with pre-existing CVD, aspirin is recommended in combination with clopidogrel due to the added benefit in prevention of secondary events. The use of alternative antiplatelet therapies is also discussed (Smith et al. 2011).

Blood pressure control. The most recent AHA/ACC guidelines (Whelton et al. 2018) have recommended the standard blood pressure goal of treatment of those with hypertension to be <130/80, with consideration of a diuretic, calcium channel blocker, or ACE inhibitor/ARB as the initial therapy in most patients, with the recommendation to add a second agent, including as a fixed-dose combination as needed to reach goal, especially when the blood pressure exceeds 20 mmHg systolic or 10 mmHg diastolic from target. Such pharmacologic treatment is recommended in most persons with hypertension at or above 140/90 mmHg, or 130/80 mmHg when the 10-year ASCVD risk from the pooled cohort risk calculator is at 10% or greater, thus persons between these numbers need such an assessment of their ASCVD risk. These recent guidelines also place a greater emphasis on home or ambulatory blood pressure monitoring for the evaluation of possible white coat or masked hypertension.

Cholesterol and lipid control. The 2014 ACC/AHA Guideline for the Management of Blood Cholesterol (Stone et al. 2014) has emphasizes defining four 'statin eligible' groups of patients: those with (1) clinical ASCVD; (2) LDL-C of 190 mg/dl or greater; (3) diabetes; and (4) 10-year ASCVD risk of 7.5% or greater based on the pooled cohort risk calculator. It has specifically removed LDL-C targets that were present in prior guidelines

because of the lack of clinical trial evidence for the use of such targets. Instead, the focus is on adequate intensity of statin medication, where for those with clinical ASCVD <75 years of age, LDL-C of 190 mg/dl, or diabetes and a 10-year ASCVD risk of 7.5% or greater, a high intensity statin is recommended with the expectation of a 50% or greater LDL-C reduction. In older persons with ASCVD, diabetes with <7.5% 10-year risk of ASCVD, or primary prevention subjects of 10-year ASCVD risk of 7.5% or greater, a moderate intensity statin is suitable. A recent non-statin consensus pathway update to these guidelines (Lloyd-Jones et al. 2017) has provided further guidance as to when these 'targets' of percentage (e.g. ≥50%) LDL-C reduction or specific values (e.g. <70 mg/dl in higher risk persons) have not been achieved that non-statin therapies such as ezetimibe or PCSK9 monoclonal antibody therapy can be recommended.

Cigarette smoking prevention and intervention. Smoking cessation has long been documented to reduce the risk of CVD. The risk of CHD is reduced by 50% within one year of smoking cessation and to that of a never-smoker within 15 years. It is recommended that at each consultation with a healthcare provider the visit should be used as an opportunity to address the 5A's of smoking cessation counselling which include 'Ask' (assess tobacco use at every visit); 'Advise' (strongly urge quitting); 'Attempt' (try to identify smokers ready to quit); 'Assist' (aid the patient in quitting); and 'Arrange' (for follow-up contacts). Interventions for smoking are varied. For youth, school, and community-based prevention programmes, state, and federal initiatives, as well as cessation assistance are available. For adults, behavioural treatment, self-help approaches, and pharmacologic therapy are used, with varying levels of success. For example, quit rates vary from 6% with physician advice only to 40% in those who participate in group programmes (40).

Diabetes control. Recommendations for the treatment of diabetes to optimize cardiovascular risk reduction is available from a variety of sources. For example, the latest recommendations from the American Diabetes Association (2018b) addresses antiplatelet therapy, glycaemic control, blood pressure control, lipid control (discussed earlier), weight control, and lifestyle management. Moreover, similar guidelines have been published by the European Association for the Study of Diabetes (EASD). Multifactorial intervention focusing on lipid, blood pressure, and glycaemic control has been shown to decrease the risk of future CVD events by more than 50%

(Gaede et al. 2008). This is a very important message given that in recent reports, only one tenth of those with diabetes have achieved target treatment levels for HbA1c, blood pressure, and LDL-C (Wong et al. 2012). While the goal of glycaemic control is a HbA1c level of less than 7% in most uncomplicated diabetics, and even 6–6.5% if achievable without side effects, given the results of recent trials, HbA1c levels of <8% are accepted as an appropriate goal for those with pre-existing micro or macrovascular disease or difficult-to-control diabetes (American Diabetes Association 2018b). Appropriate intensity statin therapy similar to the ACC/AHA guidelines discussed earlier is also considered appropriate. In addition, in persons requiring additional HbA1c reduction beyond initial therapy (especially when 9% or greater), for those with known ASCVD, these most recent ADA guidelines recommend the use of a drug shown to have cardiovascular outcomes benefit such as empagliflozin, canagliflozin, or liraglutide (although possible benefit is considered for metformin and pioglitazone).

Obesity management. The ACC/AHA/Obesity Society guideline for the management of overweight and obesity (Jensen et al. 2014) provides guidelines for maintaining ideal body weight. In particular, it notes even modest weight reduction can have beneficial effects on many of the cardiometabolic risk factors. It also recommends up to 14 sessions with a lifestyle interventionalist to provide maximum benefit in terms of weight reduction.

Exercise and physical activity. The recent ACC/AHA Guideline on Lifestyle Management (Eckel et al. 2014) recommends 150 minutes per week of moderate intensity physical activity in at least 3–4 sessions per week each lasting approximately 40 minutes, for LDL-C and blood pressure reduction. Regular physical activity can help maintain healthy blood pressure, weight, lipid levels, low levels of inflammation (C-reactive protein) and insulin sensitivity. A pedometer can be an effective tool for intervention with one guideline, for example, recommending walking at least 10,000 steps a day. For secondary prevention, a prescribed exercise programme is recommended, and participation initially in a supervised cardiac rehabilitation programme is advised for those recovering from a recent CHD event.

Food and nutrition. Lifestyle management remains the cornerstone for reducing CVD risk including achieving and maintaining optimal risk factor levels. The ACC/AHA Guideline on Lifestyle Management (Eckel et al.

2014) recommends the consumption of a dietary pattern that emphasizes intake of vegetables, fruits, and whole grains; includes low-fat dairy products, poultry, fish, legumes, non-tropical vegetable oils and nuts; and limits intake of sweets, sugar-sweetened beverages, and red meats. It notes this dietary pattern should be adapted to appropriate calorie requirements, personal and cultural food preferences, and nutrition therapy for other medical conditions (including diabetes) and to consider following plans such as the Dietary Approaches to Stop Hypertension (DASH) dietary pattern, the US Department of Agriculture (USDA) Food Pattern, or the AHA Diet. It is also recommended to consume no more than 6% calories in saturated fat, and to avoid when possible the intake of trans fats. Moreover, a sodium intake of no more than 2,400 mg per day is recommended, with an alternative target being to reduce sodium take by 1,000 mg per day.

Specific additional recommendations for secondary prevention of CVD. Besides lipid control, blood pressure, antiplatelet, diabetes control, and lifestyle recommendations discussed earlier, guidelines for secondary prevention also include therapy with a beta blocker, renin angiotensin, and aldosterone system blockers, as well as annual influenza vaccination (Smith et al. 2011). Specific systems implemented in the hospital setting, such as pre-printed discharge instructions, reminders on charts (which can be automated in the case of electronic medical records systems) are key to ensure maximal adherence to recommended therapies for secondary prevention of CHD.

Conclusions

CVDs account for the greatest burden of morbidity and mortality worldwide, and both in developed and in developing countries. Key cardiovascular risk factors, including hypertension, cigarette smoking, high blood glucose, physical inactivity, obesity, and elevated cholesterol are (in that order) the top leading causes of death worldwide. It has been estimated that elimination of obesity, unhealthy diets, and physical inactivity could reduce up to 80% of heart disease, stroke, and diabetes (Roth et al. 2017).

In a recent Presidential Advisory (Smith et al. 2012) issued by the presidents of the World Heart Federation, European Society of Cardiology, AHA, and other leading cardiac societies, the call to reduce non-communicable deaths by 25% by the year 2025 was noted (of which nearly

half are due to CVD) and key global targets recommended for adoption were: (1) a 10% reduction in the prevalence of insufficient physical activity; (2) a 25% reduction in the prevalence of raised blood pressure; (3) a 30% reduction in salt/sodium intake with recommended goal of achieving <5 grams salt intake (2,000 mg sodium) per day; and (4) a 30% relative reduction in the prevalence of current tobacco smoking. Also receiving support was a reduction of saturated fat intake by 15%, halving the prevalence of obesity, reducing alcohol intake by 10%, reducing elevated cholesterol by 20%, increasing by 50% the number of eligible persons to receive drug therapy to prevent heart attacks and strokes, and increasing availability of basic technologies and generic essential medicines for prevention of CVD.

Concerted global efforts aimed at coordinating care for those with pre-existing CVD (secondary prevention), primary prevention, identifying persons with cardiovascular risk factors and getting them the necessary treatments, as well as primordial prevention aimed at prevention of obesity and other major risk factors in the first place, are critical for reducing the morbidity and mortality associated with CVDs.

REFERENCES

ACCORD Study Group (2010a). Effects of combination lipid therapy in type 2 diabetes mellitus. *New England Journal of Medicine*, **362**, 1563–74.

ACCORD Study Group (2010b). Effects of intensive blood-pressure control in type 2 diabetes mellitus. *New England Journal of Medicine*, **362**, 1575–85.

AIM-HIGH Investigators, et al. (2011). Niacin in patients with low HDL cholesterol levels receiving intensive statin therapy. *New England Journal of Medicine*, **365**, 2255–67.

Alberti, K.G., Zimmet, P., and Shaw, J. (2005). IDF Epidemiology Task Force Consensus Group, the metabolic syndrome—a new worldwide definition. *The Lancet*, **366**, 1059.

American College of Cardiology (2018). *Evaluation of Cardiovascular Outcomes After an Acute Coronary Syndrome During Treatment with Alirocumab—ODYSSEY OUTCOMES.* [Online] Available at: https://www.acc.org/latest-in-cardiology/clinical-trials/2018/03/09/08/02/odyssey-outcomes

American Diabetes Association (2018a). Classification and diagnosis of diabetes: standards of medical care in diabetes—2018. *Diabetes Care*, 41(Suppl 1), S13–S27.

American Diabetes Association (2018b). Standards of medical care in diabetes—2018. *Diabetes Care*, 41(Suppl 1), S1–159.

Baigent, C., et al. (2005). Efficacy and safety of cholesterol-lowering treatment: prospective meta-analysis of data from 90,056 participants in 14 randomised trials of statins. *The Lancet*, **366**, 1267–78.

Benjamin, E.J., et al. (2018). American Heart Association Council on Epidemiology and Prevention Statistics Committee and Stroke Statistics Subcommittee. Heart disease and stroke

statistics—2018 update: a report from the American Heart Association. *Circulation*, 137(12), e67–492.
Cannon, C.P., et al. (2015). Ezetimibe added to statin therapy after acute coronary syndromes. *New England Journal of Medicine*, 372(25), 2387–97.
Castelli, W.P., Garrison, R.J., Wilson, P.W., Abbott, R.D., Kalousdian, S., and Kannel, W.B. (1986). Incidence of coronary heart disease and lipoprotein cholesterol levels. The Framingham Study. *Journal of the American Medical Association*, 256(20), 2835–8.
D'Agostino, R.B., et al. (2008). General cardiovascular risk profile for use in primary care: the Framingham Heart Study. *Circulation*, **117**, 743–53.
Detrano, R., et al. (2008). Coronary calcium as a predictor of coronary events in four racial or ethnic groups. *New England Journal of Medicine*, **358**, 1336–45.
Djousse, L., Driver, J.A., and Gaxiano, J.M. (2009). Relation between modifiable lifestyle factors and lifetime risk of heart failure. *Journal of the American Medical Association*, **302**, 394–400.
Eckel, R.H., et al. (2014). American College of Cardiology/American Heart Association Task Force on Practice Guidelines. 2013 AHA/ACC guideline on lifestyle management to reduce cardiovascular risk: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *Journal of the American College of Cardiology*, 63(25 Pt B), 2960–84.
Elliott, W.J. (2005). Cardiovascular events in clinical trials of antihypertensive drugs vs. placebo/no treatment: a meta-analysis. *Journal of Hypertension*, 23(Suppl 2), S273.
Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (2001). Executive Summary of the Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, evaluation, and treatment of high blood cholesterol in adults (adult treatment panel III). *Journal of the American Medical Association*, **285**, 2486–97.
Ference, B.A., et al. (2012). Effect of long-term exposure to lower low-density lipoprotein cholesterol beginning early in life on the risk of coronary heart disease: a Mendelian randomization analysis. *Journal of the American College of Cardiology*, **60**, 2631–9.
Folsom, A.R., et al. (2011). Community prevalence of ideal cardiovascular health by the American Heart Association definition and relation with cardiovascular disease incidence. *Journal of the American College of Cardiology*, **57**, 1690–6.
Franklin, S.S., Jacobs, M.J., Wong, N.D., L'Italien, G.L., and Lapuerta, P. (2001). Predominance of isolated systolic hypertension among middle-aged and elderly US Hypertensives—analysis based on NHANES III. *Hypertension*, **37**, 869–74.
Gaede, P., Lund-Andersen, H., Parving, H.H., and Pedersen, O. (2008). Effect of multifactorial intervention on mortality in type 2 diabetes. *New England Journal of Medicine*, **358**, 580–91.
Goff, D.C. Jr, et al. (2014). American College of Cardiology/American Heart Association Task Force on Practice Guidelines. 2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *Journal of the American College of Cardiology*, 63(25 Pt B), 2935–59.
Greenland, P., et al. (2010). ACCF/AHA Guideline for Assessment of Cardiovascular Risk in asymptomatic adults: a report of the American College of Cardiology Foundation/American Heart Association Task Force on practice guidelines. *Journal of the American College of Cardiology*, **56**, 50–103.
Grundy, S.M., et al. (2005). Diagnosis and management of the metabolic syndrome: an American Heart Association/National Heart, Lung, and Blood Institute Scientific statement. *Circulation*, **112**, 2735–2752.
Holman, R.R., et al. (2008). 10-year follow-up of intensive glucose control in type 2 diabetes. *New England Journal of Medicine*, **359**, 1577–89.
Hopkins, P.N., Hut, S.C., and Wu, L.L. (2000). Chapter 6. Family history and genetic factors. In: Wong, N.D., Black, H.R., and Gardin, J.M. (eds.). *Preventive Cardiology: A Practical Approach*. New York, NY: McGraw Hill.

Jensen, M.D., et al. (2014). 2013 AHA/ACC/TOS guideline for the management of overweight and obesity in adults: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines and The Obesity Society. *Journal of the American College of Cardiology*, 63(25 Pt B), 2985–3023.

Kannel, W.B., et al. (1961). Factors of risk in development of coronary heart disease—six year follow-up experience: the Framingham Study. *Annals of Internal Medicine*, **55**, 33–50.

Kaseta, J. and Sowers, J.R. (2005). Chapter 9. Diabetes and the metabolic syndrome. In: Wong, N.D., Black, H.R., and Gardin, J.M. (eds.) *Preventive Cardiology: A Practical Approach*. New York, NY: McGraw Hill.

LaRosa, J.C., et al. (2005). Treating to New Targets (TNT) Investigators. Intensive lipid lowering with atorvastatin in patients with stable coronary disease. *New England Journal of Medicine*, **352**, 1425–35.

Lloyd-Jones, D.M., et al. (2017). Focused update of the 2016 ACC Expert Consensus decision pathway on the role of non-statin therapies for LDL-cholesterol lowering in the management of atherosclerotic cardiovascular disease risk: a report of the American College of Cardiology Task Force on expert consensus decision pathways. *Journal of the American College of Cardiology*, 70(14), 1785–822.

Luepker, R.V. and Lando, H.A. (2005). Tobacco use, passive smoking, and smoking cessation interventions. In: Wong, N.D., Black, H.R., and Gardin, J.M. (eds.) *Preventive Cardiology: A Practical Approach*. New York: McGraw Hill.

Mahmood, S.S. and Wang, T.J. (2013). The epidemiology of congestive heart failure: contributions from the Framingham Heart Study. *Global Heart*, **8**, 77–82.

Malik, S., et al. (2004). Impact of the metabolic syndrome on mortality from coronary heart disease, cardiovascular disease, and all causes in United States Adults. *Circulation*, **110**, 1239–44.

Marso, S.P., et al. (2016). LEADER Steering Committee, LEADER Trial Investigators. Liraglutide and cardiovascular outcomes in type 2 diabetes. *New England Journal of Medicine*, 375(4), 311–22.

McCowen, K.C. and Blackburn, G.L. (2005). Chapter 10. Obesity and weight control. In: Wong, N.D., Black, H.R., Gardin, J.M. (eds.) *Preventive Cardiology: A Practical Approach*. New York: McGraw Hill.

Muntner, P., et al. (2018). Potential U.S. population impact of the 2017 ACC/AHA high blood pressure guideline. *Journal of the American College of Cardiology*, 71(2), 109–18.

Neal B, et al. (2017). CANVAS Program Collaborative Group. Canagliflozin and cardiovascular and renal events in type 2 diabetes. *New England Journal of Medicine*, 377(7), 644–57.

Palomaki, G.E., Melillo, S., and Bradley, L.A. (2010). Association between 9p21 genomic markers and heart disease. A meta-analysis. *Journal of the American Medical Association*, **303**, 648–56.

Prospective Studies Collaborative (2002). Age-specific relevant of usual blood pressure to vascular mortality: a meta-analysis of individual data for one million adults in 61 prospective studies. *The Lancet*, **360**, 1903.

Ridker, P.M., et al. (2017). CANTOS Trial Group. Antiinflammatory therapy with canakinumab for atherosclerotic disease. *New England Journal of Medicine*, 377(12), 1119–31.

Romero, J.R. and Wolf, P.A. (2013). Epidemiology of stroke. Legacy of the Framingham Heart Study. *Global Heart*, **8**, 67–75.

Rosensen, R.S. (2004). Statins: can the new generation make an impression? *Expert Opinion on Emerging Drugs*, 9(2), 269–79.

Roth, G.A., et al. (2017). Global, regional, and national burden of cardiovascular diseases for 10 causes, 1990 to 2015. *Journal of the American College of Cardiology*, 70(1), 1–25.

Sabatine, M.S., et al. FOURIER Steering Committee and Investigators (2017). Evolocumab and clinical outcomes in patients with cardiovascular disease. *New England Journal of Medicine*, 376(18), 1713–22.

Schunkert, H., Konig, I.R., and Kathiresan, S. (2011). CARDIoGRAM Consortium. Large scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nature Genetics*, **43**, 333–8.

Smith, S.C. Jr, et al. (2011). AHA/ACCF secondary prevention and risk reduction therapy for patients with coronary and other atherosclerotic vascular disease: 2011 update: a guideline from the American Heart Association and American College of Cardiology Foundation endorsed by the World Heart Federation and the Preventive Cardiovascular Nurses Association. *Journal of the American College of Cardiology*, **58**, 2432–46.

Smith, S.C. Jr, et al. (2012). World Heart Federation, American Heart Association, American College of Cardiology Foundation, European Heart Network, European Society of Cardiology. Our time: a call to save preventable death from cardiovascular disease (heart disease and stroke). *Journal of the American College of Cardiology*, **60**, 2343–8.

SPRINT Research Group, et al. (2015). A randomized trial of intensive versus standard blood-pressure control. *New England Journal of Medicine*, 373(22), 2103–16.

Stone, N.J., et al. (2014). American College of Cardiology/American Heart Association Task Force on Practice Guidelines. 2013 ACC/AHA guideline on the treatment of blood cholesterol to reduce atherosclerotic cardiovascular risk in adults: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *Journal of the American College of Cardiology*, 63(25 Pt B), 2889–934.

The Action to Control Cardiovascular Risk in Diabetes Study Group (2008). Effects of intensive glucose lowering in type 2 diabetes. *New England Journal of Medicine*, **358**, 2545.

The Diabetes Control and Complications Trial/Epidemiology of Diabetes Interventions and Complications Study Research Group (2005). Intensive diabetes treatment and cardiovascular disease in patients with type 1 diabetes. *New England Journal of Medicine*, **353**, 2643–53.

Whelton, P.K., et al. (2018). 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA Guideline for the prevention, detection, evaluation, and management of high blood pressure in adults: a report of the American College of Cardiology/American Heart Association Task Force on clinical practice guidelines. *Journal of the American College of Cardiology*, 71(19), e127–248.

Williams, R.R., et al. (2001). Usefulness of cardiovascular family history data for population-based preventive medicine and medical research (The Health Family Tree Study and the NHLBI Family Heart Study). *American Journal of Cardiology*, **87**, 129–35.

Wilson, P.W.F., et al. (1998). Prediction of coronary heart disease using risk factor categories. *Circulation*, **97**, 1837–47.

Wong, K., et al. (2012). Comparison of demographic factors and cardiovascular risk factor control among US adults with type 2 diabetes by insulin treatment classification. *Journal of Diabetes Complications*, **26**, 169–74.

Wong, N.D. and Levy, D. (2013). Legacy of the Framingham Heart Study: rationale, design, initial findings and implications. *Global Heart*, 8(1), 3–9.

World Health Organization (WHO) (2011). *Global Atlas on Cardiovascular Disease Prevention and Control*. Mendis, S., Puska, P., and Norrving, B. (eds.) Geneva: World Health Organization.

Zinman B, et al. (2015). EMPA-REG OUTCOME Investigators. Empagliflozin, cardiovascular outcomes, and mortality in type 2 diabetes. *New England Journal of Medicine*, 373(22), 2117–28.

8.2

Cancer epidemiology and public health

Paolo Boffetta, Zuo-Feng Zhang, and Carlo La Vecchia

Introduction to cancer epidemiology and public health

Classification of neoplasms is based on the International Classification of Diseases-Oncology, now in its third edition (Fritz et al. 2000), into topographical categories (according to the organ where the neoplasm arises) and morphological categories (according to the characteristics of the cells). In addition to localization and morphology, neoplasms, which are characterized by abnormal growth, can be classified according to genetic and molecular characteristics, clinical behaviour, and response to therapy.

Benign and malignant neoplasms

Benign neoplasms represent localized growths of tissue with predominantly normal characteristics: in many cases they cause minor symptoms and are amenable to surgical therapy. They can be fatal because of their clinical importance when they occur in organs in which compression is possible and surgery cannot be easily performed (mainly the brain), and when they produce hormones or other substances with a systemic effect (e.g. adrenaline (epinephrine) produced by phaeochromocytoma). Because relatively little is known about the distributions and causes of most benign neoplasms, they will not be further discussed in this chapter.

Malignant neoplasms are characterized by progressive growth of abnormal cells with structural and functional alterations with respect to the normal tissue. In some cases, the alterations can be so serious that it becomes difficult to identify the tissue of origin. A peculiarity of most malignant neoplasms is their ability to invade the nearby tissue, and to migrate and colonize other organs via blood and lymph vessel penetration

and transport. The presence, extension, and control of metastases are often the critical factors to determine the success of therapy and the survival of cancer patients.

Carcinogenesis

Increasingly, neoplasms are characterized at the molecular level according to phenotypic aspects (e.g. presence of receptors, expression of genes) and genetic alterations (e.g. inherited or acquired mutations in one or multiple genes).

The understanding of the molecular and cellular mechanisms of carcinogenesis has greatly advanced in recent years. According to a widely accepted model, normal cells have to acquire six characteristics to become fully malignant (Hanahan and Weinberg 2011). These include the ability to produce growth signals (several known oncogenes mimic growth signalling), the lack of sensitivity to antigrowth signals (the retinoblastoma [Rb] protein and its homologues play a key role in the ability of the cell to decide whether to proliferate, to be quiescent, or to enter into a post-mitotic state, based on external signalling), resistance towards programmed cell death or apoptosis (in many cases via inactivation of the p53 protein), immortalization (normal cells have a limited replication potential that is related to the length of the telomeres: in malignant cells, overexpression of telomerase circumvents it), stimulation of blood vessel production (by changing the balance of angiogenesis inducers and countervailing inhibitors), and ability to invade and metastasize.

The acquisition of these neoplastic characteristics typically occurs by alterations of relevant genes, but an inability to maintain genomic integrity (so-called genomic instability, which includes reduced ability to repair DNA damage) is an additional feature of malignant cells, as accumulation of random mutations in genes involved in all the functions just mentioned would be a too rare event for the development of cancer during the normal lifespan. Inflammation also fosters multiple hallmark functions, and consequently promotes carcinogenesis. A final point to consider is the heterogeneity of neoplastic genetic alterations: the acquisition of the different biological capabilities of the neoplastic cell can appear at different times, and the particular sequence in which capabilities are acquired can vary widely, even among the same type of malignant tumours. In addition,

neoplasms contain a repertoire of normal cells in their stroma, including immune cells, which contribute to create the tumour microenvironment.

Several inherited conditions carry a very high risk of one or several cancers. High-penetrance genes (or high-risk genes) have been identified through family-based and other linkage studies (Table 8.2.1). These conditions, however, are rare and explain only a small proportion of human cancers. Genetic factors, including a large number of low-penetrance variants, however, are likely to play an important role in interacting with non-genetic factors to determine individual cancer risk.

The long process of carcinogenesis justifies the efforts to develop and apply screening approaches for early detection and treatment of selected subclinical neoplasms in healthy individuals.

Table 8.2.1 Examples of high penetrance cancer predisposition genes

Gene	Syndrome	Main targets
RET	Multiple endocrine neoplasia 2	Medullary thyroid carcinoma, pheochromocytoma, parathyroid adenoma
MET	Familial papillary renal cancer syndrome	Papillary renal cancer
APC	Familial adenomatous polyposis	Colorectal cancer
VHL	von Hippel-Lindau syndrome	Clear-cell renal carcinoma, hemangioblastoma, retinal angioma, pheochromocytoma
WT1	Wilms' tumour syndrome	Bilateral Wilms' tumour
RB1	Hereditary retinoblastoma	Retinoblastoma
NF1	Neurofibromatosis 1	Neurofibroma, neurofibrosarcoma, optic glioma
NF2	Neurofibromatosis 2	Vestibular schwannoma, meningioma
p53	Li–Fraumeni syndrome	Sarcoma, leukaemia, breast, brain, lung, pancreas and skin cancers, others
p16/DCK4	Hereditary melanoma syndrome	Melanoma
PTCH	Nevoid basal cell carcinoma syndrome	Basal cell carcinoma
MEN1	Multiple endocrine syndrome 1	Tumours of parathyroids, gastrointestinal endocrine tissues and anterior pituitary
BRCA1	Hereditary breast-ovarian cancer syndrome	Breast, ovary, prostate, and colon cancer
BRCA2	Hereditary breast-ovarian cancer syndrome	Breast (also male) and ovary cancer
PTEN	Cowden syndrome	Hamartoma, breast, and thyroid cancer
hMSH2, hMLH1, hPMS1, hPMS2	Hereditary non-polyposis colon	Colon, endometrium, ovary, stomach cancers, others

Gene	Syndrome	Main targets
ATM	Ataxia-telangiectasia	Leukaemia, lymphoma, breast cancer
XP(A-G)	Xeroderma pigmentosum	Skin cancer
BLM	Bloom syndrome	Leukaemia, lymphoma, most cancers
FAC, FAA	Fanconi anaemia	Acute myeloid leukaemia, others
WRN	Werner syndrome	Sarcoma, melanoma, thyroid carcinoma

Cancer epidemiology and distribution

The development of cancer epidemiology has formed the basis for knowledge about the causes and the possible preventive strategies which has greatly advanced during the last several decades. The identification of the determinants of cancer relies on two complementary approaches, the epidemiological and the experimental.

Table 8.2.2 reports the lowest and highest country-specific incidence rates of selected cancers, as well as the 20th and the 80th percentiles of the respective rankings, as estimated in the Globocan 2018 project (Bray et al. 2018; Ferlay et al. 2019). For all cancers combined, the ratios of 80th percentile to the 20th percentile are 2.8 for men and 2.0 for women. In both sexes, the highest 80th/20th percentile ratio among common cancers was for lung cancer (8.8 in men, 6.2 in women). This comparison is based on stable figures, and indicates larger variations among very-high-risk and very-low-risk areas. The highest ratios in Table 8.2.2 represent extreme geographic differences that deserve epidemiological investigation into environmental and genetic factors in the development of cancer.

Table 8.2.2 Ratio of the 20th and 80th percentile in the ranking of country-specific estimated age-standardized incidence rates of selected cancers—Globocan 2018 (Bray et al. 2018)

Cancer	Men	Women
All cancers	2.8	2.0
Oral cavity, pharynx	4.5	2.7
Nasopharynx	32	13
Oesophagus	4.6	5.3
Stomach	3.7	2.7
Colon-rectum	5.0	3.7
Liver	3.0	2.5
Pancreas	4.8	5.0
Lung	8.8	6.2
Melanoma	8.1	15
Breast	–	2.6
Cervix	–	3.6
Endometrium	–	5.3
Ovary	–	2.1
Prostate	4.9	–
Bladder	6.4	4.0
Kidney	10	6.0
Nervous system	6.3	6.4
Non-Hodgkin’s lymphoma	2.7	2.9
Leukaemia	2.7	2.6

Analytical studies (case–control and cohort) have shown the causal role of specific exposures in the aetiology of several malignant neoplasms. One limitation of the epidemiological approach, which may prove of critical

importance in trying to detect comparatively small increases in risk, as in the case of environmental pollutants, is that even in the best conditions it is impossible to confidently identify by epidemiological means an increase in risk smaller than about 20% and serious problems arise in the interpretation of increases below 50%, as the biases inherent in any observational study are of at least this order of magnitude (Boffetta et al. 2015; Lagiou et al. 2018).

The identification of carcinogens via experimental studies relies on three types of tests: (i) long-term (often lifetime) carcinogenicity tests in experimental animals, most commonly rodents (mice, rats, and hamsters); (ii) short-term tests assessing the effects of chemical agents on a variety of end points belonging to three general classes: DNA damage, mutagenicity, and chromosome damage; and (iii) mechanistic tests, aimed at identifying the intermediate steps in the compound-specific carcinogenic process.

These tests are valuable to the extent that such effects may reflect underlying events in the carcinogenic process. Indeed, consistent positivity in tests measuring DNA damage, mutagenicity and chromosomal damage is usually regarded as indicating potential carcinogenicity of the tested agent. In addition, a number of factors, including inflammation (Mantovani et al. 2019), hormones and overweight/obesity, show carcinogenic effects through non-genotoxic pathways, for which mechanisms of carcinogenesis are less well defined and their biomarkers are sparse (Boffetta and Islami 2013).

Results of laboratory tests constitute useful supporting evidence when adequate epidemiological data for the carcinogenicity of an environmental agent exists, but they become all the more essential when the epidemiological evidence is non-existent or inadequate in quality or in quantity. In the latter case, although no universally accepted criteria exist to automatically translate data from long-term animal tests or short-term tests to cancer hazard in humans, an evaluation of the risk can be made on a judgemental basis using all available scientific evidence. This policy has been applied by the International Agency for Research on Cancer (IARC) in a systematic programme of evaluation of the carcinogenic risk of chemicals to humans. Agents are commonly classified in group 1 when the evidence of their carcinogenicity in humans, derived from epidemiological studies, is considered sufficient. They are classified in group 2A when the evidence in humans is limited and the agent is an experimental carcinogen and in group 2B when evidence of carcinogenicity in humans is limited and less than sufficient evidence of carcinogenicity in experimental animals (IARC 2018).

Since 1972, IARC has published 124 volumes presenting evaluations and re-evaluations for 1,014 chemical, physical, and biological agents and groups of agents, as well as mixtures and exposure circumstances, such as occupations. A total of 120 agents have been classified in group 1, 82 in group 2A, and 311 in group 2B, and 500 in group 3 (not classifiable as to its carcinogenicity to humans). The complete list of agents, with their evaluations can be found on the Monographs website (https://monographs.iarc.fr).

Temporal changes in incidence rates, particularly when they take place over a few decades, are incompatible with a genetic explanation. However, the possibility of environmental exposure changes as well as their possible interaction with genetic factors might contribute partially to temporal changes in incidence rates. Recorded incidence rates are also affected by changes in disease classification, diagnostic precision, and registration quality. Mortality rates are, in addition, affected by changes in treatment effectiveness. However, marked trends like the one for lung cancer mortality (Malvezzi et al. 2017) and other fatal neoplasms with poor prognosis are most likely to reflect real time trends in cancer rates, pointing to the importance of environmental factors.

Etiological risk factors of cancer

A number of non-genetic agents and exposures are associated with an increased risk of malignant neoplasms, but their importance varies in different population settings and for neoplasms from different organs.

Tobacco and related agents

Tobacco smoking is the single major cause of human cancer worldwide (IARC 2004; GBD 2017). Tobacco-related cancers include cancers of the lung, nasal cavity, larynx, oral cavity, pharynx, oesophagus, stomach, liver, pancreas, colorectum, ovary (mucinous histotype), uterine cervix, kidney, and bladder, as well as myeloid leukaemia. In high-income countries, tobacco smoking has been estimated to cause approximately 30% of all human cancers (Doll and Peto 2005), though—given smoking cessation in men—this proportion may tend to somewhat decline now. In many medium- and low-income countries, the burden of tobacco-related cancer is still

lower, given the relatively later start of the tobacco epidemics, which will result in a greater number of cancers in the future. All smoking tobacco products are carcinogenic, including cigarettes, cigars, cigarillos, pipe tobacco, waterpipe (hookah), bidi, and a large number of local products. Tobacco smoke contains a large number of carcinogens: the main groups are polycyclic aromatic hydrocarbons (PAHs), a group of compounds formed during the combustion of organic substances, and tobacco-specific nitrosamines, that are produced from nicotine during the curing of the tobacco leaf. Exposure to second-hand (passive) tobacco smoking is a risk factor for lung cancer (IARC 2004). An effect of second-hand smoking on other organs, such as nasopharynx, oral cavity, oesophagus, and larynx, is also suspected and plausible.

Use of smokeless tobacco products causes cancer of the oral cavity and other head and neck organs; an effect on other digestive organs is possible. The carcinogenicity of smokeless tobacco products appears to be due to their content in nitrosamines (Boffetta et al. 2008). Smokeless tobacco products represent a particularly important cause of cancer in South Asia, but they are widely used in other regions as well.

Areca nut is the main constituent of chewing products that contain a few slices of the nut wrapped in a betel leaf along with lime (calcium hydroxide), spices, and possibly tobacco. Chewing of areca nut, with or without tobacco, is a risk factor for cancer of the head and neck and the oesophagus (Mehrtash et al. 2017).

Opium, the dried latex obtained from the opium poppy contains two main groups of alkaloids: phenanthrenes such as morphine (10–16% of total weight) and codeine, which are the main psychoactive constituents, and isoquinolines, which have no central nervous system effects. The smoking of opium, which does not involve the burning of the agent but its heating, resulting in the vaporization of the active alkaloids, has been associated with increased risk of oesophageal and other digestive and respiratory cancers (Kamangar et al. 2014). Despite the fact that opium is illegal in almost all countries, smoking for recreational use remains prevalent in certain populations in West and Central Asia.

Diet and nutrition

Despite considerable research efforts in recent years, the exact role of dietary factors in causing or protecting against human cancer remains largely unquantified. There is sufficient evidence of an increased risk of liver cancer with exposure to aflatoxin, a contaminant produced by some fungi in some fungi in Southeast China, Taiwan, Southeast Asia, and sub-Saharan Africa. According to a recent comprehensive review on nutrition and cancer (Table 8.2.3) (WCRF/AIRC 2018), high intake of processed meat was considered to have convincing evidence for an increased risk of colorectal cancer, and high intake of red meat to probably increase the risk of the same cancer. High intake of food containing dietary fibres and whole grains probably decreases the risk of colorectal cancer, and that of non-starchy vegetables and fruits probably decreases the risk of cancer of upper aerodigestive tract.

Aflatoxin, a product of fungi of the *Aspergillus* genus, causes liver cancer in humans (Henry et al. 2002): contamination occurs when cereals and legumes are stored under hot and humid conditions. Exposure is widespread, but is particularly important in sub-Saharan Africa, Southeast Asia, and East Asia.

Table 8.2.3 Assessment of associations between dietary factors and human cancer (WCRF/AIRC, 2018)

Factor (high intake)	Oral cavity, pharynx, larynx	Oesophagus	Stomach	Colon, rectum	Lung	Breast	Prostate
Non-starchy vegetables or fruits	(–)	(–)					
Whole grains				(–)			
Foods containing dietary fibres				(–)			
Red meat				(+)			
Processed meat				+			
Dairy products				(–)			
Foods preserved by salt			(+)				

Direction of the effect: + increased risk; – decreased risk.

Strength of the evidence: no brackets: convincing evidence of an association.

Parentheses: probable association.

Intake of *Aristolochia sp.* results in exposure to aristolochic acid, a carcinogen of the upper urinary tract, including the renal pelvis (Matin et al. 2014). Exposure occurs via the use of these herbs in traditional herbal medicine, in particular in Asia, as well as contamination of other vegetables, such as wheat, as it was shown in the regions of Central Europe in which the so-called Balkan Endemic Nephropathy, another condition caused by exposure to aristolochic acid, was prevalent (Stiborová et al. 2016).

Alcohol drinking

Alcohol drinking is an established carcinogen, with no apparent safe dose. All types of alcoholic beverages cause cancer. The evidence is strongest for cancers of the oral cavity, pharynx, larynx, oesophagus, liver, pancreas,

colorectum, and female breast (WCRF/AIRC 2018). Globally, alcohol drinking was estimated to cause 7.2% of all cancers in men and 3.5% in women, for a total of 770,000 cases per year (Praud et al. 2016).

Coffee, tea, and maté drinking

Coffee drinking has been associated with a decrease in cancer risk: the evidence is strongest for liver and endometrial cancer, and it is suggestive for other cancers including colorectal and breast cancer, as well as melanoma (Alicandro et al. 2017). The mechanisms underlying such protective effect are not known, though possibly linked to favourable metabolic effects on glycemia, and non-causal explanations (e.g. better diet) cannot be excluded. The evidence linking tea drinking and cancer risk is inadequate: early reports of a possible protective effect of green tea have not been confirmed by subsequent research (Yi et al. 2019).

Drinking of maté, which is prevalent in some South American countries, has been associated with an increased risk of oesophageal cancer (Loria et al. 2009). Since maté is often drank at high temperature, this is probably due in part to a thermal effect, since also hot tea drinking at has been associated with oesophageal cancer risk (Chen et al. 2015).

Anthropometric characteristics

There is convincing evidence that adult body fatness is associated with increased risk of cancers of the oesophagus (adenocarcinoma), liver, pancreas, colorectum, breast (postmenopausal), endometrium, and kidney, and a causal association is probable for cancers of the head and neck, stomach, gallbladder, ovary and prostate (WCRF/AIRC 2018). Attained height is causally associated with risk of colorectal, breast, and ovarian cancer.

Physical activity

Regular vigorous workplace or recreational physical activity (e.g. at least 30 minutes per day) decreases the risk of colon cancer; a protective effect is also likely for breast and endometrial cancer (WCRF/AIRC 2018). The magnitude of risk reduction for colon and breast cancer is in the order of

40%, and a dose–response relationship has been shown for both neoplasms. Although regular physical activity contributes to weight control, the epidemiological evidence suggests that two factors also act independently. The mechanisms through which physical activity contributes to cancer prevention are not fully understood, but they may include enhancement of immune function, interference with sex steroids, and insulin and IGF-I pathways.

Infectious agents

As a whole, infectious agents represent the second most important risk factor for cancer worldwide. For 11 infectious agents (7 viruses, 1 bacterium, and 3 parasites) there is sufficient evidence that they cause cancer in humans (Table 8.2.4). Several others are suspected to cause cancer (Table 8.2.4). The proportion of total cancers attributable to infectious agents has been estimated to be 15.4% (Plummer et al. 2016); liver, stomach and cervical cancers are responsible for 86% of infection-related cancer. The burden of carcinogenic infections affects disproportionately low- and middle-income countries (23.4% of all cancers vs. 9.2% in high-income countries).

Table 8.2.4 Associations between infections and human cancer (Plummer et al. 2016)

Agent	Target organs*
Viruses	
Hepatitis B virus	Liver, bile duct (leukaemia/lymphoma)
Hepatitis C virus	Liver, bile duct, leukaemia/lymphoma
Hepatitis D virus	Liver
Human papilloma virus type 16	Cervix, vulva, vagina, penis, anus, oral cavity, tonsil, pharynx, (larynx)
Human papilloma virus type 18	Cervix, (oral cavity), (penis), (anus), (vulva)
Human papilloma virus types 31, 35, 39, 45, 51, 52, 56, 58, 59	Cervix
Human papilloma virus type 33	Cervix, (anus), (vulva)
Human papilloma virus types 6, 11	(Larynx)
Human papilloma virus types 26, 53, 66, 67, 68, 70, 73, 82	(Cervix)
Human papilloma virus types 30, 34, 69, 85, 97	
Human papilloma virus, genus-beta types (except 5 and 8) and gamma types	(Skin)
Human papillomavirus types 5, 8 (in patients with epidermodysplasia verruciformis)	(Skin)
Human immunodeficiency virus 1	Anus, Kaposi sarcoma, cervix, eye, leukaemia/lymphoma, (liver/bile duct), (skin), (vulva), (vagina), (penis)
Human immunodeficiency virus 2	Kaposi sarcoma, non-Hodgkin’s lymphoma
Human T-cell lymphotrophic virus I	Adult T-cell leukaemia/lymphoma
Human T-cell lymphotrophic virus II	
Epstein–Barr virus	Nasopharynx, leukaemia/lymphoma, (stomach)
Human herpes virus 8	(Kaposi’s sarcoma), leukaemia/lymphoma

Agent	Target organs*
Merkel cell polyomavirus	(Skin)
Bacteria	
Helicobacter pylori	Stomach, leukaemia/lymphoma
Parasites	
Schistosoma haematobium	Bladder
Schistosoma japonicum	(Colorectum, liver/bile duct)
Schistosoma mansoni	
Opistorchis viverrini	Liver/bile duct
Opistorchis felineus	
Clonorchis sinensis	Liver/bile duct
Malaria (*Plasmodium falciparum*)	(Leukaemia/lymphoma)

* suspected target organs in parentheses.

Occupation and pollution

While some occupational carcinogens (e.g. *bis*-chloro methylethers) represent today a historic curiosity, exposure is still widespread for carcinogens such as asbestos, crystalline silica and mixtures of PAHs (Table 8.2.5). Estimates of the burden of cancer attributable to established occupational carcinogens in high-income countries result in figures in the order of 3–5% in men and 0.5–2.5% in women (Boffetta et al. 2009; Brown et al. 2018; Islami et al. 2018). Estimates for low- and middle-income countries are uncertain because of lack of valid data on prevalence of exposure to occupational carcinogens.

Table 8.2.5 Occupational carcinogens (Boffetta et al. 2010)

Agents, mixture, circumstance	Main industry, use
Individual agents and groups of agents	
4-Aminobiphenyl	Pigment
Arsenic and arsenic compounds	Glass, metal, pesticide
Asbestos	Insulation, filter, textile
Benzene	Chemical, solvent
Benzidine, dyes metabolized to benzidine	Pigment
Benzo[a]pyrene	Combustion fumes
Beryllium and beryllium compounds	Aerospace
Bis(chloromethyl)ether and chloromethyl methyl ether*	Chemical intermediate
1,3 Butadiene	Plastics, rubber
Cadmium and cadmium compounds	Dye/pigment
Chromium[VI] compounds	Metal plating, dye/pigment
Dioxin	Chemical
Ethylene oxide	Sterilant
Formaldehyde	Plastics, textiles, laboratory agent
Mustard gas*	War gas
2-Naphthylamine	Pigment
Nickel compounds	Metallurgy, alloy, catalyst
PCB-126	Chemical
Plutonium-239 and its decay products	Nuclear industry
Radium-226 and its decay products*	Luminizing industry
Radium-228 and its decay products*	Luminizing industry
Radon-222 and its decay products	Mining
Silica, crystalline	Stone cutting, mining, glass

Agents, mixture, circumstance	Main industry, use
Solar radiation	Agriculture
Talc containing asbestiform fibres	Paper, paints
Trichloroethylene	Solvent, dry cleaning, metal
Vinyl chloride	Plastics
X- and γ-radiation	Medical
Mixtures	
Coal-tar pitches	Construction, electrode
Coal-tars	Fuel, construction, chemical
Diesel engine exhaust	Transport
Leather dust	Leather
Mineral oils, untreated	Metal
Shale-oils	Fuel
Soot	Pigment
Wood dust	Wood
Exposure circumstances	
Aluminium production	
Auramine, manufacture of	Pigment
Boot and shoe manufacture and repair	
Coal gasification	
Coal-tar distillation	
Coke production	
Haematite mining (underground) with exposure to radon	
Iron and steel founding	
Magenta production*	Pigment
Painter (occupational exposure as a)	
Rubber industry	

Agents, mixture, circumstance	Main industry, use
Strong inorganic-acid mists containing sulphuric acid	Metallurgy

* Agent mainly of historical interest.

Air and water pollution

High level of indoor and outdoor air pollution can cause lung cancer: the evidence is strong for non-smoking women experiencing high level of fumes from cooking and heating in poorly ventilated households: studies have mainly been conducted in Asia (Bruce et al. 2015). Air pollution may contain lung carcinogens such as PAHs, and has been linked to an increased risk of lung cancer, although the quantification of the risk is problematic. Estimates of the burden of cancer due to outdoor air pollution are uncertain, but typically below 1% (Boffetta et al. 2015; Brown et al. 2018; Islami et al. 2018).

Contamination of drinking water by high level of arsenic, primarily from natural sources, has been causally associated with an increased risk of cancers of the lung and bladder, and possibly other organs such as skin and kidney in studies conducted in highly contaminated areas of South America, South Asia, and East Asia (Karagas et al. 2015). No apparent effect was shown at low concentrations (below 100 µg/L), which are prevalent is most regions of the world. Chlorination of drinking water results in low-level formation of chloroform and related compounds, called trihalomethanes, which are carcinogenic in experimental systems (Richardson et al. 2007). An increased risk of bladder cancer from consumption of chlorinated water has been suggested, but the evidence is not conclusive (Hrudey et al. 2015). Contamination of drinking water by microcystins has been associated to excess risk of liver and colorectal cancer, but the evidence is inconclusive (Martínez Hernández et al. 2009).

Ionizing and non-ionizing radiation

The estimates of the contribution of ionizing radiation to human cancer in high-income countries are on the order of 2–4% (Brown et al. 2018; Doll and Peto 2005). Several neoplasms, in particular acute lymphocytic leukaemia, acute and chronic myeloid leukaemia, and cancers of the breast,

lung, bone, brain, and thyroid are produced as a result of high-level exposure to ionizing radiation (IARC 2000). Ultraviolet (UV) non-ionizing radiation concerns exposure to solar radiation and artificial sources of UV radiation for cosmetic and medical purposes. Skin cancer, both melanoma and non-melanoma, results from exposure to UV radiation (Narayanan et al. 2010). Between 1 and 4% of all cancers are attributed to UV exposure in high-income countries (Boffetta et al. 2009; Brown et al. 2018), depending on the estimates of prevalence of exposure, besides the skin characteristics of different racial and ethnic groups.

Reproductive factors

The epidemiological evidence of a carcinogenic effect of reproductive factors is strongest for breast cancer: early age at menarche, low parity, late age at first pregnancy, and late age at menopause are all associated with an increased risk, while spontaneous and induced abortions are not (Tamimi et al. 2018). In addition, breastfeeding protects from breast cancer, with a stronger effect for cancers that are negative for hormone receptors (Islami et al. 2015). Apart from age at first pregnancy, the same reproductive factors seem to exert an effect on endometrial and ovarian cancer risk similar to that played on breast cancer, while the evidence of an effect on other cancers is inadequate. Estimates of the contribution of reproductive factors to the global burden of cancer differ greatly (from 1% to 15%), depending on the underlying assumptions, especially on the choice of the reference population (Boffetta et al. 2009; Doll and Peto 2005; Brown et al. 2018; Islami et al. 2018).

Hormones

Increased levels of endogenous oestrogens, prevalent in overweight postmenopausal women, are associated with an increased risk of breast and endometrial cancers, and a similar effect is likely to be played by endogenous androgens (Crous-Bou et al. 2018; Tamini et al. 2018). The role of other hormones, such as progesterone and prolactin, in these cancers is not clearly known, nor is the role of endogenous androgens in prostate cancer.

There is a large body of evidence that growth hormones, in particular, insulin and IGF-I, have a strong effect on the risk of breast, colon, pancreas, endometrium, prostate, and possibly other cancers (Faulds et al. 2012). An effect of other endogenous hormones such as growth hormone on cancer risk is plausible but not fully elucidated in epidemiological studies (Clayton et al. 2011).

There is also strong evidence supporting an increase in cancer risk following exposure to exogenous hormones. Current and recent (up to 10 years) use of oral contraceptives entails a small increase in breast cancer risk, but no excess risk is apparent 10 or more years after cessation of use (CGHFBC 1996). Long-term use of oral contraceptives is associated with an increased risk of liver cancer, while the risk of endometrial and ovarian cancer is decreased following oral contraceptive use (Cibula et al. 2010). Postmenopausal hormonal therapy increases the risk of breast and endometrial cancer (CGHFBC 1997). In the case of breast cancer, the effect is stronger for combined oestrogen–progestogen combinations than for other types of hormonal therapy (Beral et al. 2003). The evidence for other organs is inconclusive. Tamoxifen is widely used for treatment of breast cancer: beyond its therapeutic effects, it decreases the risk of contralateral breast cancer, but it increases the risk of endometrial cancer (Crous-Bou et al. 2018; Tamini et al. 2018). Studies of transgender men and women might also be indicative of potential impact of sex hormones in the development of cancer. Because of relatively large dosage of sex hormone therapy for transgender individuals, the risk of cancer would increase and the latency might be much shorter than other hormonal factors (de Blok et al. 2019).

Perinatal factors

High birth weight is possibly linked with an increased risk of premenopausal breast cancer (WCRF/AIRC 2018). In general, perinatal factors have been proposed to cause a sizeable proportion of human cancers, but the estimate is subject to uncertainty. The implications of these findings for preventive strategies will be clarified by a more detailed understanding of the underlying carcinogenic mechanisms (Xue and Michels 2007).

Medical procedures and drugs

In addition to postmenopausal hormonal therapy, oral contraceptives, and tamoxifen, other drugs may cause cancer. Many cancer chemotherapy drugs are active on the DNA, in order to block the replication of cancer cells. This, however, might result in damage to normal cells, including cancer transformation. The main neoplasm associated with chemotherapy treatment is leukaemia, although the risk of solid tumours is also increased. A second group of carcinogenic drugs includes immunosuppressive agents, which have been studied in particular in transplanted patients (Gutierrez-Dalmau and Campistol 2007). Non-Hodgkin lymphoma and leukaemia are the main neoplasms caused by these drugs. Phenacetin-containing analgesics increase the risk of cancer of the renal pelvis (Piper et al. 1985).

There is strong evidence from observational studies that aspirin reduces the risk of colorectal cancer (Cuzick et al. 2015), an effect probably shared by other non-steroidal anti-inflammatory drugs. No precise estimates are available for the global contribution of drug use to human cancer. It is unlikely, however, that drugs represent more than 1% in high-income countries (Doll and Peto 2005). Furthermore, the benefits of such therapies are usually much greater than the potential cancer risk.

Use of ionizing radiation for diagnostic purposes is likely to carry a small risk of cancer, which has been demonstrated only for childhood leukaemia following intrauterine exposure (IARC 2000). Computed tomography (CT) scan in children may also increase subsequent cancer risk (Pearce et al. 2012). Radiotherapy increases the risk of cancer in the irradiated organs. There is no clear evidence of an increased cancer risk following other medical procedures, including surgical implants.

Medical conditions

Changes in immunological function are likely to play an important role in human cancer, but epidemiological studies have been largely unable to identify specific factors determining an increased or a decreased risk. Both severe immunosuppression and immunostimulation are associated with an elevated risk of cancer (Engels and Hildesheim 2018). Individuals infected with the human immunodeficiency virus (HIV) and patients undergoing immunosuppressive treatments, such as transplant recipients, are at increased risk of non-Hodgkin lymphoma, Kaposi sarcoma, skin cancer, and infectious-related cancers such as cervical cancer (Zucchetto et al. 2016;

Engels and Hildesheim 2018). In addition, patients suffering from systemic autoimmune diseases are also at increased risk of lymphoma and possibly other neoplasms (Engels and Hildesheim 2018). The significance of less severe disturbances of the immunological competence is poorly known.

Several chronic inflammatory conditions represent a risk factor for cancer: the epidemiological evidence is particularly strong in the case of colorectal cancer following inflammatory bowel disease and of lymphoma following chronic infectious diseases such as tuberculosis, malaria, and herpes zoster. In addition, gastroesophageal reflux is an important cause of adenocarcinoma of the lower oesophagus, a neoplasm whose incidence is increasing in many regions (Abnet et al. 2018), and tuberculosis is a risk factor for lung cancer (Malhotra et al. 2016).

The global burden of neoplasms

The number of new cases of cancer which occurred worldwide in 2018 has been estimated at 18.1 million, of which 9.5 million among men and 8.6 million among women (Bray et al. 2018). The corresponding figure for cancer deaths was 9.6 million (5.4 million in men, 4.2 million in women). Lung, colorectal, stomach, liver, and breast cancer are the most common causes of cancer death and together they comprise half of the total burden; among men, lung, liver, stomach, colorectal, and prostate cancer are the most common causes of cancer deaths, while breast, lung, colorectal, cervix uteri, and stomach cancer are the most common causes of cancer deaths among women.

One complementary approach in assessing the global burden of neoplasms is to weigh the years of life with disability and add them to the years lost because of premature death (disability-adjusted life-years, DALYs). According to the Global Burden of Disease Study, cancer was responsible for 213.2 million DALY in 2016, of which 98% was due to premature deaths and 2% from living with disabilities (GBD 2018).

Distribution, aetiology, and prevention of selected cancers

The following paragraphs discuss the epidemiology of the 10 most common cancers; they are ordered based on numbers of estimated deaths based on

Globocan 2018 (Bray et al. 2018).

Lung cancer

In 2018 lung cancer was the most frequent cancer in the world as there were 2,093,876 new lung cancer cases, accounting for approximately 11.6% of the global cancer burden (Bray et al. 2018; Malhotra et al. 2018). Lung cancer was also the most common cause of cancer death, causing 1,761,007 deaths (18.4% of the total) in 2018.

Incidence and mortality rates for lung cancer are approximately twice as high among men as among women. Lung cancer is the most common cancer in men worldwide (14.5% of cases, 22% of deaths), with the highest age-standardized incidence rates observed in Micronesia/Polynesia (52.2/100,000) and Eastern Europe (49.3/100,000)) (Bray et al. 2018). The geographic pattern of incidence rates among women is slightly different, and generally lower, reflecting historic tobacco consumption patterns. The highest age-standardized incidence rates for women are observed in North America (30.7/100,000) and Northern Europe (26.9/100,000). The geographical patterns in mortality closely follow those in incidence due to the high fatality and relative lack of variability in survival in different regions of the world.

The main histological types of lung cancer include squamous cell carcinoma, small cell carcinoma, adenocarcinoma, and large cell carcinoma; the last decades have witnessed a decrease in the predominant squamous cell carcinoma type and an increase in adenocarcinomas.

The geographical and temporal patterns of lung cancer incidence are to a large extent determined by consumption of tobacco one or two decades earlier. A carcinogenic effect of tobacco smoke on the lung was demonstrated in the 1950s and has been recognized by public health and regulatory authorities since the mid-1960s (IARC 2004). The risk of lung cancer among smokers relative to the risk among never-smokers is on the order of over 20- to 40-fold. As compared to continuous smokers, the excess risk levels off in ex-smokers after quitting, but a residual excess risk is likely to persist in long-term quitters throughout life. An association has been shown in many studies between exposure to involuntary smoking and lung cancer risk in non-smokers. The magnitude of the excess risk among non-

smokers exposed to involuntary smoking is on the order of 20% (IARC 2004).

The UK Million Women Study demonstrated that the relative risks (RRs) of lung cancer were 10.5 for current smokers of ten cigarettes per day, 22.0 for 15, and 36.0 for 20 or more cigarettes per day (Pirie et al. 2013). Thus, even moderate smokers have substantial excess lung cancer mortality. This overall risk reflects the contribution of the different aspects of tobacco smoking: average consumption, duration of smoking, time since quitting, age at start, type of tobacco product, and inhalation pattern, with duration being the dominant factor.

The lifetime cumulative risk of lung cancer of a continuous smoker is 16% and it is reduced to 10%, 6%, 3%, and 2% among those who stopped at ages 60, 50, 40, and 30, respectively (Peto et al. 2000). In the United States, stopping smoking before the age of 40 years would reduce the risk of death associated with continued smoking by about 90% (Jha et al. 2013).

An increase in tobacco consumption is paralleled some decades later by an increase in the incidence of lung cancer; similarly, a decrease in consumption is followed by a decrease in incidence, as observed in the United States and other countries in which the tobacco epidemics has reached its maturity. A notable exception is China and other East Asian countries, with relatively high rates being recorded among women (21.9/100,000), probably due to very high prevalence of passive smoking and high levels of indoor air pollution.

The risk of lung cancer is increased among workers employed in several industries and occupations. For several of these high-risk workplaces, the agent (or agents) responsible for the increased risk have been identified. Of these, asbestos and combustion fumes comprising mixtures of PAHs are the most important. Occupational agents were responsible in the past for an estimated 5–10% of lung cancers in industrialized countries, though this proportion is likely to be smaller in recent years.

Pulmonary tuberculosis has demonstrated an increased risk to lung cancer, although it is not clear whether it is due to the chronic inflammatory status of the lung parenchyma or the specific action of the *Mycobacterium*. High levels of fibres and dusts might result in lung fibrosis (e.g. silicosis and asbestosis), along with chronic bronchitis and emphysema, conditions that also entail an increase in the risk of lung cancer.

Other concerns relate to the presence of urban air posing as a susceptible risk factor; however, the excess is unlikely to be larger than 20% in most polluted urban areas (Hamra et al. 2014).

There is conclusive evidence that exposure to ionizing radiation increases the risk of lung cancer (IARC 2000). Atomic bomb survivors and patients treated with radiotherapy for ankylosing spondylitis or breast cancer are at moderately increased risk of lung cancer, while studies of nuclear industry workers exposed to relatively low levels provided no evidence of an increased risk of lung cancer. Underground miners exposed to radioactive radon and its decay products, which emit α-particles, have been consistently found to be at increased risk of lung cancer (IARC 2001). The risk increased with estimated cumulative exposure and decreased with attained age and time since cessation of exposure (Lubin et al. 1994). Exposure to radon decay products mainly occurs in environmental setting (e.g. household exposure from soil).

Despite the large number of studies, no dietary factors have been conclusively linked to lung cancer risk. However, greater intake of green vegetable, fruits, food containing carotenoids, beta-carotene, retinol, isoflavones, and fibre are associated with suggested evidence of protective effects on lung cancer (WCRF/AICR 2018).

No high-risk genes for lung cancer have been identified, that can be of epidemiological and clinical importance. Genome-wide association studies (GWAS) with large sample size have identified polymorphisms of genes located within 15q25 (*CHRNA3* and *CHRNA5*), 5p15.33 (*TERT-CLPTM1L)*, and 6p21.33 (*BAT3*) to be associated with the risk of lung cancer. However, future studies are still needed to identify heritable risk factors and molecular pathways and signature of lung carcinogenesis (Marshall and Christiani 2013).

Prevention of lung cancer

Control of tobacco smoking (including involuntary smoking) remains the key strategy for the prevention of lung cancer. Reduction in exposure to occupational and environmental carcinogens (in particular indoor air pollution and radon), are additional preventive opportunities. There is strong evidence that low-dose CT chest scan may detect early lesions in high-risk

individuals (Moyer 2014); the benefit on mortality is estimated to be around 20% reduction (Rota et al. 2019).

Colorectal cancer

Cancers of the colon and rectum accounted in 2018 for an estimated 1,849,518 new cases and 880,792 deaths worldwide (Bray et al. 2018; Keum et al. 2018). They represent the third most frequent malignant disease in terms of incidence (10% of the total) and the second for mortality (9%). Colorectal cancer is the third most common cancer in men (10.9% of the total) and the second most common cancer in women (9.5% of the total).

The majority of cases of colorectal cancer occur in high-income regions of the world, but the incidence is rapidly increasing in countries in social and economic transition, in particular in Asia; geographical patterns are similar in men and women. Generally, rates are slightly higher among men than women. The highest age-standardized rates are observed in Northern and Central Europe (up to 70/100,000 for men and 40/100,000 for women), followed by Australia, Argentina, and North America. There is considerable geographical variation in incidence of colorectal cancer, and the lowest rates in the world are observed in South Asia and Western Africa (less than 5/100,000 for men and 3/100,000 for women) (Bray et al. 2018). A small increase in the incidence of colon cancer has been observed during the last few decades in most populations, but Western Europe over the last two decades (Siegel et al. 2014). Studies of migrant populations have shown that the risk of colon cancer approaches that of the country of adoption within one generation; the incidence is higher in urban than in rural populations.

The predominant histological type of malignant neoplasms of the colorectum is adenocarcinoma. This neoplasm is usually preceded by an adenomatous polyp, or adenoma, less frequently by a small area of flat mucosa exhibiting various grades of dysplasia. The malignant potential of an adenoma is increased by a surface diameter greater than 1 cm, by villous (rather than tubular) organization and by severe cellular dysplasia. Carriers of one adenoma larger than 1 cm have a two to four times increased risk of developing colon cancer; this risk is further doubled in carriers of multiple adenomas. On a topographical basis, the prevalence of adenomas detected during colonoscopy closely parallels the incidence of colon cancer.

Several studies have associated tobacco smoking with an increased risk of colorectal adenoma. For colon cancer, a modest increased risk following prolonged heavy smoking has been shown in some of the largest prospective studies (IARC 2004). High alcohol consumption (50 g alcohol/day) has been associated with moderately increased risk of colorectal cancer (Bagnardi et al. 2015).

Patients with ulcerative colitis and Crohn's disease are at increased risk of colorectal cancer. The overall RR has been estimated in the range of 5–20, and it is higher for young age at diagnosis, severity of the disease, and presence of dysplasia. The contribution of shared genetic and environmental factors in the genesis of the two inflammatory conditions and of colorectal cancer is not known. Diabetes and cholecystectomy have been associated with a moderate (1.5–2-fold) increased risk of (right-sided) colon cancer, possibly due to continuous secretion of bile. Patients with one cancer of the colon have a twofold risk to develop a second primary tumour in the colon or rectum, and the relative (though not the absolute) risk is greater for early age at first diagnosis (Levi et al. 2008). In women, an association has been shown also with cancers of the endometrium, ovary, and breast, possibly due to shared hormonal or dietary factors.

Several pathways which incorporate genetic instability (chromosomal instability (CIN) and microsatellite instability (MSI)), and inherited mutations of the adenomatous polyposis coli (*APC*) gene and mismatch repair genes pose as risk factors for colon cancer (Arends 2013). More than 100 susceptibility variants have been identified in multiple genetic regions through GWAs and sequencing studies (Huyghe et al. 2019). The majority of these studies were conducted in European descendants; however, studies in East Asian populations largely confirmed these genetic factors (Lu et al. 2019).

Several rare hereditary conditions are characterized by a very high incidence of colon cancer. Familial adenomatous polyposis, due to inherited or *de novo* mutation in the *APC* gene on chromosome 5, is characterized by a very high number of colonic adenomas and a cumulative incidence of colon or rectal cancer close to 100% by age 55. Gardner's syndrome, Turcot syndrome, and juvenile polyposis feature among the rarer diseases having this association. However, all these hereditary conditions, although very serious for the affected patients, account for no more than 1 per cent of colon cancers in the general population.

Lynch syndrome I, also known as hereditary non-polyposis colorectal cancer (HNPCC), is characterized by an increased risk of cancer of the proximal (right) colon, and is due to inherited mutation in one of two genes involved in DNA repair. Contrastingly, Lynch syndrome II has also an increased risk of extracolonic neoplasms, mainly of the endometrium and the ovary. Overall, hereditary non-polyposis colon cancer may account for a sizeable proportion of cases of colon cancer in Western populations. In addition to these hereditary conditions, first-degree relatives of colon cancer patients have a two- to threefold increased risk of developing a cancer of the colon or the rectum.

Overweight, obesity, and lack of physical activity are important risk factors of colorectal cancer (WCRF/AIRC 2018). The underlying mechanisms are not fully elucidated, but these factors may explain up to 20% of cases of the disease in high-income populations (Islami et al. 2018).

A large number of studies have been conducted to identify dietary factors other than alcohol associated with colorectal cancer. The evidence is strongest for the association between an increased risk and intake of processed meat. Less convincing associations have been reported for a protective effect of whole grains, dietary fibres, vegetables and fruit, dairy products and calcium, and for a carcinogenic effect of red meat intake.

Vitamin D, and in particular its most active form, 25(OH) D, has been inversely related to colorectal cancer risk (Giovannucci 2005). The preventive effect is only possible when the vitamin D receptor (VDR) is present so that the mechanisms to increase VDR expression would enhance the preventive ability of vitamin D (Stubbins et al. 2012). Reduction of the incidence of colorectal cancer is also associated with the increased use of aspirin and other non-steroidal anti-inflammatory drugs (Bosetti et al. 2012; Ferrández et al. 2012). Additionally, hormone therapy in menopause and other female hormones, including oral contraceptives (OC), have been inversely related to colon cancer risks, and hence may also play some protective role (ACS 2011a).

Prevention of colorectal cancer

Increased physical activity, avoidance of overweight and obesity, reduced alcohol drinking, and quitting smoking are the main tools for the primary prevention of colorectal cancer. While aspirin use could prevent colorectal

cancer as well, its use is not routinely recommended because of its spectrum of side effects.

Surveillance via flexible colonoscopy, involving removal of adenomas, is a secondary preventive measure. An additional approach consists in the detection of faecal occult blood. The method suffers from low specificity and, to a lesser extent, low sensitivity, in particular in the ability to detect adenomas. However, randomized trials have shown a reduced mortality from colorectal cancer after an annual test, although this is achieved at a high cost due to an elevated number of false positive cases, who must then undergo colonoscopy. Randomized trials have also demonstrated the efficacy of sigmoidoscopy in reducing colorectal cancer mortality, primarily for the distal colon.

Current guidelines recommend screening for average-risk individuals between ages 50 and 75 using colonoscopy (every 10 years), flexible sigmoidoscopy (every 5 years) or faecal occult blood test (FOBT, annually or biennially) (Benard et al. 2018; Wolf et al. 2018). However, there are disparities between recommendations issues by different national and international agencies.

Cancer of the stomach

Accounting for an estimated 1,033,701 new cases or 5.7% of the global cancer burden, stomach cancer is the sixth most frequent cancer in the world and the third most common cause of cancer death worldwide (782,685 deaths in 2012, 8.2%) in 2018 (Bray et al. 2018; Ye et al. 2018). This marks a substantive decrease in incidence since 1975 when stomach cancer was the most common neoplasm worldwide. Incidence rates are about twice as high in men as in women, with age-standardized rates highest (above 15/100,000 in men and 7/100,000 in women) in Eastern Asia, and Eastern Europe. Low-incidence areas include Africa, North America, and Northern Europe. The majority of cases occur in developing countries, with 44% of all new cases and almost 50% of deaths coming from China alone.

An increased risk of gastric cancer is associated with *Helicobacter pylori* infection. The biological plausibility of a causal association is also supported by a strong association between *H. pylori* and precancerous lesions, including chronic and atrophic gastritis and dysplasia. Given that the prevalence of infection is very high, especially in low- and middle-income

countries and among older cohorts, *H. pylori* can explain a large proportion of all cases of gastric cancer, or over 4% of all cancer cases globally (Plummer et al. 2016). However, only a small fraction of infected individuals are diagnosed as having stomach cancer. Persistent *H. pylori* infection due to a combination of virulence factors and immune subversion and manipulation mechanisms is associated with gastric carcinogenesis (Salama et al. 2013).

Another important cause of stomach cancer is tobacco smoking. Smokers have a 50–60% increased risk of stomach cancer, as compared to non-smokers. This relationship would indicate that smoking is responsible for approximately 10% of all cases (IARC 2004).

Throughout the world there is a consistent correlation between consumption of salt and salted foods and stomach cancer incidence. A large number of studies that have examined this relationship have generally found an increased risk of approximately twofold for frequent consumption of salt and salted foods. The relationship is biologically plausible, given that salt may lead to damage to the protective mucosal layer of the stomach.

A notably striking feature of gastric cancer is the dramatic decline in its incidence and mortality which has been observed in most high-income countries over the past century. The decline is apparent for both sexes. This continuous dramatic decline, as well as the results from migrant studies, suggests a strong environmental influence on the disease.

Reasons for the generalized decline in gastric cancer rates are complex and not completely understood. Almost certainly, these include a more varied and affluent diet and better food conservation, including refrigeration, as well as the effective treatment of *H. pylori* infection. Whether improved diagnosis and treatment have also played some role on the favourable trends in gastric cancer, particularly over most recent calendar periods, however, remains open to question.

A total of 11 variants were significantly associated with gastric cancer risk in a systematic review of GWAS: MUC1 rs2070803 at 1q22 (diffuse carcinoma subgroup), MTX1 rs2075570 at 1q22 (diffuse), prostate stem cell antigen (PSCA) rs2294008 at 8q24.2 (non-cardia), PRKAA1 rs13361707 5p13 (non-cardia), PLCE1 rs2274223 10q23 (cardia), TGFBR2 rs3087465 3p22 (Asian), PKLR rs3762272 1q22 (diffuse), PSCA rs2976392 (intestinal), GSTP1 rs1695 11q13 (Asian), CASP8 rs3834129 2q33 (mixed), and TNF rs1799724 6p21.3 (mixed) (Mocellin et al. 2015). The strongest

evidence is for variants of the *PSCA* gene, which have been identified in multiple populations, and may interact with H. pylori infection (Cai et al. 2017). These genes might regulate cell growth, differentiation, apoptosis, and angiogenesis and be associated with the development of stomach cancer.

Other intervention trials conducted in a Chinese population known to be micronutrient deficient observed that a combination supplement of beta-carotene, vitamin E and selenium did result in a small reduction in the risk of stomach cancer (Qiao et al. 2009), but recent findings on the issue on other, better nourished populations, are largely negative (Plummer et al. 2007).

Primary prevention of stomach cancer by dietary means is feasible by encouraging high-risk populations to decrease consumption of cured meats and salt preserved foods. Prevention may also be feasible through eradication of *H. pylori* infection, particularly in childhood and adolescence, and by avoiding mother to child transmission. Screening and early detection of stomach cancer have been developed in Japan and South Korea with the use of X-ray photofluorography to identify early lesions, followed by gastroscopy.

Cancer of the liver

The estimated worldwide number of new cases of liver cancer in 2018 is 841,080 (4.7% of the total), that of deaths is 781,631 (8.2%). More than 80% of the liver cancer burden is in middle and low-income countries (over 50% from China alone) (Bamia et al. 2018; Bray et al. 2018). Most low-income regions of the world, with the exception of South-Central and Western Asia, experience a high incidence of liver cancer. The highest rates (above 20/100,000 in men and above 6.5/100,000 in women) are recorded in Eastern and South-Eastern Asia, and Northern Africa. The lowest rates are in South-Central Asia, where age-standardized rates are below 5/100,000 in men and 2/100,000 in women. Intermediate rates (5–10/100,000 in men) are observed in areas of Southern Europe and North America. Rates are two- to threefold higher in men than women, and the difference is stronger in high-incidence than in low-incidence areas. The majority of liver cancers are hepatocellular carcinomas (HCC), which originates from hepatocytes; a second important type is cholangiocarcinoma, which originates from the biliary ducts.

Chronic infections with hepatitis B virus (HBV) and hepatitis C virus (HCV) are the main causes of HCC. The risk increases with early age at infection (in high-risk countries, most HBV infections occur perinatally or in early childhood), and the presence of cirrhosis is a pathogenic step. The estimated relative risk of developing HCC among infected subjects, compared to uninfected, ranged between 10 and 50 in different studies. On a global scale, over 75% of liver cancer cases are related to HBV and HCV infections, though their prevalence varies in different high-risk areas HCV infections, though their prevalence varies in different high-risk areas.

Ecological studies have shown that the incidence of HCC correlates not only with HBV and HCV infection, but also with contamination of foodstuffs with aflatoxins, a group of mycotoxins produced by the fungi *Aspergillus flavus* and *Aspergillus parasiticus*, which cause liver cancer in many species of experimental animals (Turner et al. 2012). Contamination originates mainly from improper storage of cereals, peanuts, dried corns, and other vegetables and is prevalent in particular in Africa, South-East Asia, China, and Taiwan.

(Alcoholic) cirrhosis is probably the most important risk factor for HCC in populations with low prevalence of HBV and HCV infection and low exposure to aflatoxins, such as North America and northern Europe (La Vecchia 2007). The association between tobacco smoking and HCC is now established, with a RR of the order of 1.5 to 2 for tobacco smoking on liver carcinogenesis (IARC 2004).

Use of OC increases the risk of liver adenomas, and is associated with the risk of HCC, although the absolute risk is likely to be small. Case reports have associated use of anabolic steroids with development of liver cancer, but the evidence is not conclusive (Cibula et al. 2010).

Diabetes is also related to an excess risk of HCC, and the increased prevalence of overweight and obesity, and consequently of diabetes and non-alcoholic fatty liver disease (NAFLD), in several populations may have had some role in recent unfavourable trends of HCC in North America and other areas of the world. Combined exposure to overweight and diabetes led to a RR of HCC of 4.75, after allowance for hepatitis, alcohol, and other recognized possible confounding factors (Bosetti et al. 2014).

The excess risk of liver cancer associated with overweight/obesity and diabetes has been related to the development of NAFLD (Sanyal et al. 2010). NAFLD is characterized by excess fat accumulation in the liver, and ranges

from isolated hepatic steatosis to non-alcoholic steatohepatitis (NASH), the more aggressive form of fatty liver disease, which can progress to cirrhosis and HCC. However, NAFLD/NASH increases HCC risk even in the absence of cirrhosis (Bosetti et al. 2014).

Coffee drinking has been associated with a reduced risk of liver cancer, with a RR in the order of 0.66 (Bravi et al. 2017).

The incidence of cholangiocarcinoma is increasing in many regions of the world (Khan et al. 2019). In several areas of Southeast and East Asia cholangiocarcinoma is the main form of liver cancer and is linked to infection with the liver flukes, *Opistorchis viverrini* and *Chlonorchia sinensis*. Other risk factors of this type of cancer include primary sclerosing cholangitis, a rare inflammatory condition, as well as infection with HBV and HCV. A role of tobacco smoking, diabetes and overweight/obesity is suspected (Khan et al. 2019).

Prevention of liver cancer

In high-prevalence areas, HBV vaccination has to be introduced in the perinatal period. In the last few decades, many countries from Asia, Southern Europe, and, to a lesser extent, Africa have expanded their national childhood vaccination programme to include HBV. Reduction of the incidence of liver cancer in young adults was observed in Qidong, a high-risk area of liver cancer in China, because of both HBV vaccination and aflatoxin B_1 reduction in foods (Sun et al. 2013). A similar primary preventive approach is not available for HCV; however, new antiviral therapies have become available in recent years, which clear the infection and are expected to reduce the subsequent risk of liver cancer (Waziry et al. 2017).

Control of aflatoxin contamination of foodstuffs represents another important preventive measure. While this is easily achieved in high-income countries, its implementation is limited by economic and logistic factors in many high-prevalence regions. Control of alcohol drinking and tobacco smoking represent additional primary preventive measures.

Since about half of HCC, but not normal adult liver, secrete the fetal antigen α-fetoprotein, the detection of this marker has been proposed as a screening method. However, no population-based studies are currently available showing a decreased mortality from liver cancer in screened

populations. Approaches combining ultrasound and MRI are explored for early detection of liver cancer in high-risk individuals with cirrhosis.

Cancer of the breast

Breast cancer is the most important cause of cancer diagnosis and death among women, and accounted for an estimated 2,088,848 new cases and 626,679 deaths worldwide in 2018 (Bray et al. 2018; Tamini et al. 2018). The incidence of breast cancer is relatively low (less than 25/100,000) in most countries of Middle Africa, and some countries of East and South-Central Asia. The highest rates (70–100/100,000) are recorded in North America, Australia and New Zealand, Northern and Western Europe. The incidence of breast cancer has grown rapidly during the last few decades in many low-resource countries and slowly in high-income countries. In high-income countries, age-specific incidence rates increase sharply after age 45 and the rate in women aged 65+ is over 250/100,000. However, incidence rates in medium- and low-income countries are between 50 to 100/100,000 after age 45 (Fig. 8.2.1). Mortality rates remained fairly stable between 1960 and 1990 in most of Europe and the Americas; appreciable declines have occurred since the early 1990s. Incidence appears to increase linearly with age up to menopause, after which a further increase is less distinct (in high-income countries) or almost absent (in low-income countries). Women from high social class have consistently higher rates (about 30–50%) than women from low social class.

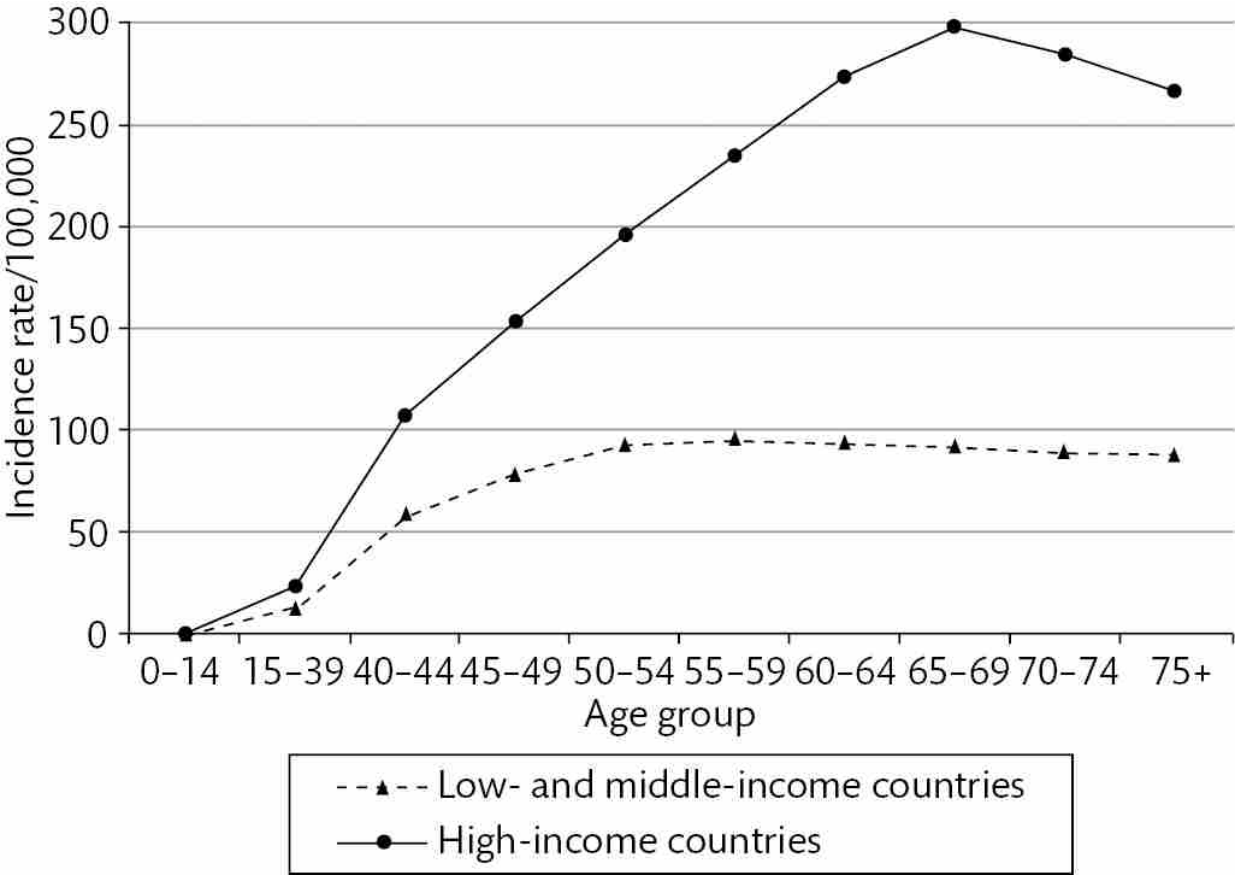


Fig. 8.2.1 Age-specific incidence of breast cancer by region of the world, 2012.
Source: data from Ferlay, J., et al. GLOBOCAN 2012 v1.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11, International Agency for Research on Cancer, Lyon, France, Copyright © IARC 2013, available from http://globocan.iarc.fr.

The combined evidence from reproductive factors points towards an important role for endogenous hormones in breast carcinogenesis. A direct assessment of the role of oestrogens and testosterone is also available from recent prospective studies collecting data from biological samples. Oestradiol concentrations in the blood have been directly associated with breast cancer risk in postmenopausal women, whereas data are fewer and results are less consistent in premenopausal women. The association might be stronger with oestrogen and progesterone receptor positive tumours. Comparable findings have been reported for measures of testosterone and other androgens, but the data are inconsistent for all endogenous hormones across major cohort studies.

Less than 1 per cent of all cases of breast cancer occur in men (Ruddy & Winer 2013). The incidence provides limited evidence of geographical and

interracial variations, with no clear correlation with incidence in women. Conditions involving high oestrogen levels, such as gonadal dysfunction, alcohol abuse, and obesity, are likely risk factors for breast cancer in men. *BRCA2* mutations are more frequent than *BRCA1* in male familial breast cancers.

Women suffering from the two most common benign breast diseases, fibrocystic disease and fibroadenoma, carry a two- to threefold increased risk of breast cancer. While these lesions are not likely to represent pre-neoplastic conditions, they share with breast cancer epithelial proliferation, linked to hormonal alterations.

A history of breast cancer in first-degree relatives is associated with a two- to threefold increased risk of the disease. Most of the role of familial history is likely to result from low-penetrance genes associated with hormonal metabolism and regulation, and DNA damage and repair. In the family linkage studies, breast cancer risk is greatly increased in carriers of mutations of several high-penetrance and low-frequency genes, including *BRCA1*, *BRCA2*, *PTEN*, and *TP53*. The mutations of these genes are associated with high risk of breast cancer, defined as inherited breast cancer syndrome. Although the cumulative lifetime risk in carriers of these genes is over 50%, they are rare in most populations and explain only a small fraction (2–5%) of total cases. There are exceptions, however, such as Ashkenazi Jews, among whom high-risk *BRCA1* or *BRCA2* mutations are responsible for an estimated 12% of breast cancers. In addition, a group of low penetrance and low-frequency DNA repair genes are related to moderate risk of breast cancer, including *CHEK2*, *ATM*, *BRIP*, and *PALB2*. Both high- and moderate-risk genes account for approximately 25% of cases with family history of breast cancer. A large number of GWAS have identified more than 100 susceptibility variants in genes involved in metabolism, apoptosis, cell cycle regulation, and mitochondrial regulation and other cancer-related functions (Low et al. 2018).

There are similarities and differences in risk profiles between triple negative (TNBC) and ER+ breast cancer based on results of the Women's Health Initiative study. Similar risk factors included body mass index (BMI), lack of physical activity, and breast density (Phipps et al. 2011a). Nulliparity was associated with decreased risk of TNBC but increased risk of ER+ breast cancer. Among parous women, the number of births was positively associated with TNBC and inversely associated with ER+ disease. Ages at

menarche and menopause were modestly associated with risk of ER+ but not TNBC (Phipps et al. 2011b). Tobacco smoking was not associated TNBC, but a risk factor for ER+ disease. Alcohol drinking is a risk factor specifically for ER+ breast cancer (Kabat et al. 2011).

Many lifestyle factors have been investigated as possible causes of breast cancer. Despite very extensive research, no dietary factors have been conclusively linked to risk of breast cancer (WCRF/AICR 2018). Alcohol drinking is an established etiological factor for breast cancer, in particular for the postmenopausal form of the disease (WCRF 2018). Consumption of three or more alcoholic drinks per day carries an increased risk on the order of 30–50%, with each daily drink accounting for about a 10% higher risk. Overweight, obesity, and adult weight gain are established risk factors of postmenopausal breast cancer, while the effect on the premenopausal form of the disease might go on the opposite direction (WCRF/AIRC 2018). It is likely that both overweight and heavy alcohol drinking act on breast cancer through mechanisms involving hormonal level or metabolism. Tobacco smoking has been inconsistently related to breast cancer, and risk quantification is not easy. A high level of physical activity, on the other hand, is likely to moderately decrease the risk (WCRF/AIRC 2018). Studies of occupational factors and of exposure to organochlorine pesticides have failed to provide evidence of an aetiological role (Salehi et al. 2008).

Survival from breast cancer has slowly increased in high-income countries, such as the United States, Canada, Sweden, Cuba, and Japan, where improvements in screening practices and treatments have achieved 5-year survival rates of 80–90%. In contrast, 5-year survival rates are between 60% and 79% in middle-income countries and below 60% in low-income countries (Coleman et al. 2008). In the United States, breast cancer survival in the black population is consistently poorer than the white population (Allemani et al. 2013).

Prevention of breast cancer

Primary prevention of breast cancer has been attempted via nutritional intervention, involving reduction of energy intake, reduction of the proportion of calories from fat, reduction of alcohol consumption, and increase in fruit and vegetable consumption. No evidence of efficacy has been produced so far. However, control of weight gain, particularly of

postmenopausal women, would have favourable implications in breast cancer risk.

The most suitable approach for breast cancer control is secondary prevention through mammography. The effectiveness of screening by mammography in women older than 50 years has been demonstrated, and programmes have been established in various countries, which however are not consistent among them (Jørgensen et al. 2017). The effectiveness in women younger than 50 is not yet demonstrated. Other screening techniques, including ultrasound or breast self-examination, have not been proven to reduce breast cancer mortality (Siu et al. 2016).

Cancer of the oesophagus

Approximately 572,000 new cases and 509,000 deaths of oesophageal cancer occurred worldwide in 2018 (Abnet et al. 2018; Bray et al. 2018). The geographical distribution of oesophageal cancer is characterized by very wide variations within relatively small areas. Very high rates (over 50/100,000) are recorded for both men and women in northern Iran, various provinces of eastern China, and certain areas of Central Asia (Torre et al. 2016). Intermediate rates in men (10–50/100,000) occur in South Africa, Eastern Africa, Eastern Asia, some provinces in China, and regions of Central Asia.

In all these high-risk areas, squamous cell carcinoma (SqCC) is the predominant histological type. Ethnic factors are suggested by the fact that populations at higher risk in Central Asia are of Turkish or Mongolian origin. Generally, rates in men are two- to fourfold higher than in women, although men and women have comparable rates in the very high-risk populations. In many high-risk areas, a decrease in the incidence of SqCC of the oesophagus has occurred during recent decades. In northern Europeans and white people in the United States, an increase in incidence was largely observed for adenocarcinoma of the lower oesophagus.

Tobacco smoking and alcohol drinking account for 90% or more of the cases of oesophageal SqCC in Western Europe and North America. However, this proportion is lower in low- and middle-income countries, particularly in selected high-risk areas of Asia and South America. In a large case–control study in China, both tobacco smoking and alcohol drinking accounted for only 40% of SqCC cases (Wu et al. 2011a).

The risk in heavy smokers relative to non-smokers is on the order of 5–10 in the Western populations. However, the RR estimate is much lower in the Asian populations. A recent large case–control study in Jiangsu, China reported that for individuals smoking more than 30 pack-years, the adjusted odds ratio (OR) was 1.85 (1.49–2.29) (Wu et al. 2011b). A strong relationship has been shown for the duration of smoking and average consumption. Quitting smoking substantially reduces the risk: the RR declines within 5 years after quitting, remains substantially elevated at least 10 years after cessation, and declines by 40% thereafter (Bosetti et al. 2006; Wu et al. 2011b). Thus, cessation of smoking could have an appreciable impact in reducing SqCC oesophageal cancer, and represents an obvious priority for prevention and public health purposes. Smoking black tobacco, high-tar, and hand-rolled cigarettes, as well as pipes, might exert a stronger effect than smoking other products. Chewing tobacco represents an important risk factor in India and southern Africa, but its role has not been confirmed in Central Asia. In the latter region, smoking and eating opium is a reason for the high incidence rates (Shakeri et al. 2012). Snuff use has also been related to an excess risk of oesophageal cancer, with RR of 3.5 for SqCC in a Swedish cohort study based on 10 non-smoking cases (Zendehdel et al. 2008; Lee 2011).

Areca nut chewing has been reported as an independent risk factor for SqCC in Asia and it may have additive interactions with tobacco smoking on the risk of the disease. Carcinogenic nitrosamines derived from areca nut are formed in the saliva of chewers and have been suggested as a risk factor for oral cancer and oesophageal cancer (Akhtar 2012).

In a meta-analysis of 40 case–control and 13 cohort studies, the pooled RRs for oesophageal SqCC were 1.38 for light drinking (<12.5 g/day), 2.62 for 12.5–50 g/day, and 5.54 for greater than 50 g/day. Among never-smokers, there was no association for light drinkers, but the RR rose to 3.1 for heavy drinkers (Bagnardi et al. 2015). The association with alcohol might be modified by genetic factors such as aldehyde dehydrogenase 2 (*ALDH2*) polymorphisms in Asian populations (Islami et al. 2011; Wu et al. 2013). It is unclear whether there are differences in the carcinogenic potency of different alcoholic beverages, but the most common beverage in each population is the most strongly related one. A reduction in the excess risk of oesophageal SqCC is suggested 15–20 years after quitting. Furthermore, the effect of alcohol is independent from the effect of tobacco, and the

interaction between the two exposures fits a multiplicative or additive model (Wu et al. 2011b).

Maté drinking has been associated with an increased risk of oesophageal cancer (Loria et al. 2009); hot tea drinking has also been associated with oesophageal cancer risk (Chen et al. 2015).

Familial aggregation of oesophageal cancer has been occasionally shown, with joint segregation of a gene responsible for *keratosis palmaris et plantaris* (tylosis). Studies of families without the tylosis gene have not provided evidence of an important role of other high-penetrance genetic susceptibility factors in oesophageal cancer. However, recent GWAS studies in Chinese and Japanese populations implicate certain low-penetrance genes in oesophageal cancer development. Reported genes encoded enzymes involved in the metabolism of alcohol or tobacco, DNA damage repair, and other carcinogenesis pathways (Cui et al. 2009; Wang et al. 2010; Wu et al. 2011).

Other active carcinogens might be mycotoxins or N-nitroso compounds. Mycotoxins, including fumonisin B1, have also been detected in mouldy corn from high-risk areas in China and southern Africa. In addition, in Japan, eating bracken fern has been associated with an elevated oesophageal cancer risk. The elucidation of dietary factors implicated in oesophageal carcinogenesis, in particular the possible role of mycotoxins and N-nitroso compounds (including endogenously formed nitrosamines), would represent an important step in the understanding and prevention of this disease. Oesophageal cancer is related to ionizing radiation, in particular in women irradiated for breast cancer.

Reduced intake of fresh fruits and vegetables also appears to represent a risk factor for oesophageal cancer. A similar effect has been suggested for low intake of fish, and high intake of red and processed meat. In general, a dietary pattern rich in foods from animal origin and poor in vitamins and fibre appears to increase oesophageal cancer risk (Bravi et al. 2012). The available data do not, however, allow us to establish the potentially preventive role of specific micronutrients from fruits and vegetables, and the results of most chemo-preventive trials with retinol, riboflavin, vitamin E, zinc, and selenium have failed to show a benefit.

Patients suffering from Plummer–Vinson syndrome, a sideropenic dysphagia due to deficit of iron, riboflavin, and other vitamins, had an increased incidence of hypopharyngeal and oesophageal cancers.

Oesophageal cancer risk is also increased among coeliac disease patients, possibly because of nutritional deficiencies. Subjects with a family history of oesophageal cancer have about threefold excess risk after adjustment for tobacco and alcohol, and the RR rose to over 100-fold in heavy drinkers and smokers with positive family history in first-degree relatives (Garavello et al. 2005).

Adenocarcinoma occurs in the lower third of the oesophagus. Its incidence has sharply increased in the last four decades in most Western countries and has largely affected white people. Barrett's oesophagus, a columnar metaplasia of the epithelium, is strongly associated with the subsequent development of adenocarcinoma. The main risk factor for Barrett's oesophagus and oesophageal adenocarcinoma is persistent reflux esophagitis. It is predominantly seen in countries such as the United States or Scotland.

Overweight, obesity, lack of physical activity, and gastroesophageal reflux are featured risk factors in adenocarcinoma. In a meta-analysis of 22 studies (Turati et al. 2013) compared with normal weight individuals, the pooled RR was 1.7 for overweight (BMI 25–30) and 2.3 for obese ones (BMI >30). The increased prevalence of overweight and inactivity in North America and northern Europe may partly or largely, explain the increased frequency of adenocarcinoma (La Vecchia et al. 2002; Buas and Vaughan 2013).

Tobacco smoking has also shown a consistent association with oesophageal adenocarcinoma. The pooled RR from a meta-analysis of 33 studies was 1.8 for ever smokers and 2.36 for current smokers (Tramacere et al. 2011; Buas and Vaughan 2013). In contrast, there is no material association between alcohol drinking and oesophageal adenocarcinoma (Tramacere et al. 2012). Epidemiological evidence suggests a protective role for high intake of fruits and vegetables and an unfavourable role for intake of salty food.

Prevention of oesophageal cancer

Avoidance of tobacco smoking and elevated alcohol drinking remains the main preventive approach in reducing the burden of oesophageal SqCC in Western populations. Improved diet, in particular increased consumption of fresh fruits and vegetables, might also contribute to prevention. Avoidance of tobacco, control of obesity, increased physical activity, and treatment of

reflux are considered the main issues for the prevention of oesophageal adenocarcinoma. Incomplete understanding of the role of other factors complicates the elaboration of preventive strategies in many high-risk regions, although decreasing intake of hot beverages and avoidance of tobacco and opium might be important (Wu et al. 2011a).

Cancer of the pancreas

The great majority of malignant neoplasms of the pancreas are adenocarcinomas which originate from the exocrine portion of the organ secreting digestive enzymes (Benetou et al. 2018). Rare pancreatic neoplasms include tumours (of uncertain clinical behaviour) of the endocrine portion, which secrete insulin and glucagon, as well as lymphomas and sarcomas.

Geographical and temporal variations exist in the sensitivity and specificity of clinical diagnosis and in the proportion of histological verification of pancreatic cancer cases. Even when comparing populations living in the same place at the same time (e.g. different social classes or age groups), differential access to healthcare might affect incidence and mortality data.

The disease accounts for an estimated 459,000 new cases in 2018, 48% of which (219,000) occurred in more developed regions (Bray et al. 2018) where the highest rates (close to 20/100,000 in men and 15/100,000 in women) are recorded among black people in the United States and European countries. The lowest rates, which may suffer from underdiagnosis, are recorded in South-Central Asia and Middle and Eastern Africa (below 2/100,000 in men and 1/100,000 in women). In the United States, the incidence ratio of black versus white people were 1.26 and 1.36 for men and women, respectively (ACS 2013).

The median of 5-year survival was 5.2 per cent with a range of 2.2–9.6% in countries with different economic status (ACS 2011b). From 2014 to 2018, the 5-year survival of pancreatic cancer has increased from 6 per cent to 9%, suggesting improvements in medical treatment (Rawla et al. 2019). However it remains on the most fatal cancers. Given its very poor survival, mortality rates closely parallel incidence rates. Rates are about 50% higher in men than in women. An increase in incidence and mortality has taken place since the 1970s, particularly in Europe, that can be attributed in part to

diagnostic improvements. However, incidence and mortality have levelled off and declined over recent years in men in the United States, Canada, and United Kingdom (Lucas et al. 2016).

Urban populations have higher rates than rural ones, but this may again reflect differences in quality of diagnosis. Migrant population studies suggest that first-generation migrants from low-to high-risk areas experience, after 15 or 20 years, rates that are even higher than those of the country of migration, suggesting an important role of environmental exposures occurring late in life (Benetou et al. 2018).

The best known risk factor for pancreatic cancer is tobacco smoking. The risk in smokers is two- to threefold higher than that in non-smokers, and a dose–response relationship and a favourable effect of quitting smoking have been shown in many populations. It has been estimated that 20–30% and 10% of cases in men and women are attributable to tobacco smoking (IARC 2004).

Interethnic differences in pancreatic cancer risk were evaluated in the Multiethnic Cohort. The RRs for pancreatic cancer were 1.60 (95% CI 1.30–1.98) for Native Hawaiians; 1.33 (95% CI 1.15–1.54) for Japanese Americans; 1.20 (95% CI 1.01–1.42) for African Americans when compared to European Americans (Huang et al. 2019). Some of these differences can be explained by differences in smoking habits.

Several medical conditions have been studied with respect to their associations with subsequent risk of pancreatic cancer. A history of pancreatitis increases the risk more than tenfold, with little difference between the alcoholic and non-alcoholic forms of the disease. An increased risk has also been shown in several studies of diabetic patients; the RR is likely to fall in the range of 1.3–2 and is higher in the short term after diagnosis of diabetes. Gastrectomy patients are at a two- to threefold increased risk of pancreatic cancer; the association does not appear to be confounded by tobacco smoking.

Studies have reported on the significant association between alcohol and pancreatic cancer; however, not all have addressed it. The current evidence indicates a weak effect of heavy alcohol drinking, but the evidence is not sufficient (Lucenteforte et al. 2012).

Nutritional and dietary factors have been suggested to be related to pancreatic cancer, including obesity and low physical activity, low intake of foods containing folate, red meat, and a low intake of vegetables and fruit

(WCRF/AICR 2018). The issue of nutrition, diet, and pancreatic cancer remains, however, largely undefined. Early reports of an association between coffee consumption and pancreatic cancer risk have not been confirmed by larger, more recent investigations, probably because of selection bias in the early studies (WCRF/AICR 2018).

A familial history of cancer of the pancreas is present in 8–10% of patients, suggesting a possible role for genetic factors. Specific hereditary conditions carrying an increased risk of pancreatic cancer include the Li–Fraumeni syndrome, hereditary non-polyposis colon cancer, and a group of rare hereditary pancreatitis that involve deficiency in enzyme metabolism: these conditions, however, explain only a small proportion of cases in the general population.

GWAS conducted on pancreatic cancer have identified more than 15 variants that are associated with susceptibility to the disease (Amundadottir 2016). These variants explain part of the familial aggregation of the disease (Childs et al. 2016).

There is no effective cure for pancreatic cancer, with the exception of surgery for a small number of patients. Screening methods are not available. Primary prevention is the only available tool for this disease: avoidance of smoking is the major practicable way for reducing the number of cases. Control of obesity and diabetes may be additional preventive measures.

Head and neck cancer (cancers of oral cavity, pharynx, and larynx)

Tumours of the oral cavity, oropharynx, hypopharynx, and larynx are grouped under the rubric of head and neck cancer because of shared histology (the majority are SqCC) and predominant aetiological factors (tobacco smoking and alcohol drinking). However, tumours of the lip, nasopharynx, and salivary glands have distinct aetiological factors. We will focus our discussion on cancers of the oral cavity, oropharynx, hypopharynx, and larynx in this section.

There were an estimated 706,000 new cases of cancers of the head and neck worldwide in 2018, of whom 355,000 are from the oral cavity, 177,000 from the larynx, and 174,000 from the oro- and hypopharynx. The estimated number of deaths was 358,000 (Bray et al. 2018).

Cancer of the oral cavity, oropharynx, and hypopharynx

The incidence of cancers of the oral cavity varies over 20-fold between high-risk areas (e.g. Papua New Guinea, Maldives, Sri Lanka, Pakistan, Bangladesh, and India) and low-risk areas (e.g. China, Vietnam, Korea, and Haiti) (Bray et al. 2018; Rider et al. 2018). In all populations, rates in the oral cavity and pharynx in men exceed those in women by a factor of 2–8. Incidence rates of both sites increased in Europe and the Americas until the late 1980s and have levelled off or declined in most countries over the last decade. When looking at subsites within the oral cavity and the pharynx, cancer of the oropharynx and hypopharynx account for as many or more cases than cancer of the oral cavity in high-risk European populations. Cancers of the tongue, floor of the mouth, and other parts of the oral cavity represent the majority of cases in India and the United States.

Tobacco and alcohol are the most important risk factors for cancer of the oral cavity and pharynx. In Western populations, smoking represents the main use of tobacco, and the RRs of oral cancer among smokers compared to non-smokers are on the order of 3–10. The risk is higher for heavy smokers, long-term smokers, and smokers of black tobacco or high-tar cigarettes. Cigar and pipe smoking also poses a risk, while stopping smoking is followed by a decrease in risk. In India, chewing tobacco is the main risk factor for oral cancer, although *bidi* and cigarette smoking also contribute to the risk. In the United States and Europe, use of smokeless tobacco has also been implicated in the development of oral cancer (Boffetta et al. 2008).

Consumption of alcoholic beverages increases the risk of oral and pharyngeal cancer (Islami et al. 2010; Tramacere et al. 2010; Turati et al. 2010). Relative to abstainers and light drinkers, the RR in heavy drinkers is on the order of 10. The effects of tobacco smoking and alcohol drinking are multiplicative or larger; that is, the effect of exposure to both is close to or greater than the product of their individual effects. The combined effect of tobacco smoking and alcohol drinking accounts for 80% of cancers of the oral cavity and pharynx in high-income countries. Similarly, tobacco chewing and smoking and their combination are responsible for a large proportion of these cancers in India.

With reference to alcohol drinking, additional issues are worth discussing. First, oral cancer risks show a clear decline after stopping smoking. The pattern of risk after stopping drinking remains unclear, though it appears that an appreciable excess risk persists for several years. Second, although

ethanol is the main carcinogenic ingredient in alcoholic beverages, it remains unclear whether different types of alcohol beverages have different influences on oral carcinogenesis. For example, some fruit-based spirits may be associated with higher risks than beer or wine (Boffetta and Hashibe 2006). This could explain some of the exceedingly high rates in countries like Hungary or Slovakia, where fruit-derived hard spirits are commonly consumed.

Human papilloma virus (HPV) DNA, detected in pre-neoplastic and neoplastic lesions of the oropharynx, is associated with over a 100-fold excess risk of oropharyngeal cancer (Pytynia et al. 2014). The association is less strong for cancer of the oral cavity and is suspected as an aetiological factor for cancer of the hypopharynx. The role of HPV in head and neck cancer has become increasingly important since the 1990s in Europe and North America, mainly because of changes in sexual practices, while it remains less important in other regions of the world. The role of other oral cancer risk factors, as well as diagnosis and disease management, is smaller than that of tobacco or alcohol, and remains largely undefined. For example, dietary and nutritional factors, such as fruit and vegetable consumption, have shown a protective effect on oral cancer risk. Moreover, earlier diagnosis and improved treatment on oral cancer survival may have favourably influenced oral cancer death rates over time.

Poor oral hygiene and ill-fitting dentures are likely additional risk factors for oral cancer. Several occupations have been sporadically reported to confer an increased risk of oral and pharyngeal cancer. The evidence is consistent only for employment as a waiter and/or bartender, probably reflecting an increased consumption of alcohol and exposure to environmental tobacco smoke (ETS).

The role of genetic susceptibility in oral carcinogenesis is probably modest. High-risk families have been reported only occasionally. However, a role is likely for low-penetrance factors, such as increased sensitivity to mutagens and genetic polymorphisms of enzymes implicated in the metabolism of alcohol (alcohol dehydrogenase and aldehyde dehydrogenase) (Lewis and Smith 2005). GWAS in Caucasian populations have identified five variants, including three variants located within the alcohol dehydrogenase genes (*ADH7*, *ADH1B*, and *ADH1C*), one 4q21 variant located near DNA repair genes *HEL308* and *FAM175A*, and one 12q24 variant located near the *ALDH2* gene to be associated with head and neck

cancer (Mckay et al. 2011). One GWAS with limited sample size conducted in an Indian population on chewing tobacco-related oral cancers did not identify any variant which reached genome-wide significance level (Bhatnagar et al. 2012). GWAS on head and neck cancer are relatively limited and replication studies in populations other than Caucasians are needed.

Prevention of oral and pharyngeal cancers

Avoidance of tobacco (smoking, chewing, and snuffing) and of excessive alcohol drinking represent the main preventive measures for cancers of the oral cavity and pharynx. Primary prevention through prophylactic HPV vaccination is promising and would involve both girls and boys (Kreimer 2014). It is unclear whether additional benefits might be obtained from an increase in fruit and vegetable intake and improvement of oral hygiene. Avoidance of excessive exposure to solar radiation would represent the main preventive approach for lip cancer. In populations at high risk of nasopharyngeal cancer from China and possibly other countries, avoidance of salted fish and other preserved food, in particular as weaning food, should be recommended.

Oral inspection aimed to identify pre-neoplastic lesions might be an effective approach for secondary prevention of oral cancer. The inspection can be performed by medically certified professionals, but also, in particular in high-risk areas from developing countries, such as India, by specifically trained health workers. A large-scale preventive trial demonstrated a reduction in oral cancer mortality during its 15-year follow-up, with larger reductions in smokers and drinkers and in those adhering to repeated rounds of screening (Sankaranarayanan et al. 2013).

Cancer of the larynx

More than 90% of cancers of the larynx are SqCCs; the majority originate from the supraglottic and glottic regions of the organs (Malhotra et al. 2018). The incidence in men is high (10/100,000 or more) in Cuba, Central Europe, and Central Asia, while the lowest rates (<1/100,000) are recorded in Africa. The incidence in women is ≤1/100,000 in most populations. In most high-income countries, rates have declined in men over the last two decades.

Up to 80% of cases of laryngeal cancer in high-income countries are attributable to tobacco smoking, alcohol drinking, and the interaction between the two factors (Malhotra et al. 2018). The effect of tobacco, with risks in smokers on the order of 10 relative to non-smokers, seems to be stronger for glottic than supraglottic neoplasms. Studies in several populations have shown a dose-response relationship and a beneficial effect of quitting smoking (Pelucchi et al. 2008). Smoking black-tobacco cigarettes confers a stronger risk than smoking blond-tobacco cigarettes. Studies from India have also reported an effect of chewing tobacco-containing products. The effect of alcohol is stronger for supraglottic tumours than for tumours at other sites. It is not clear, however, whether different alcoholic beverages exert a different carcinogenic effect.

There are suggestions of a protective effect exerted by a high intake of fruits and vegetables, although the evidence is not conclusive and the data regarding specific micronutrients, such as carotenoids and vitamin C, are inadequate. Data concerning a possible effect of other foods are not consistent.

Occupational exposure to asbestos and to mists of strong inorganic acids, in particular of sulphuric acid, are suspected causes of laryngeal cancer. A possible effect has been suggested for other occupational exposures, including nickel and ionizing radiation, but the evidence is not conclusive.

An etiological role of HPV infection in laryngeal cancer has been suggested. Laryngeal papillomatosis, a condition characterized by multiple benign papillomas caused by infection with HPV types 6 and 11, confers an increased risk of laryngeal cancer. A meta-analysis of studies assessing the presence of HPV-DNA reported strong associations between HPV infection and laryngeal SqCC, with a summary OR of 5.39 (95% confidence interval (CI) 3.25–8.94). A similar association was observed for HPV-16 (OR 6.07; 95% CI 3.44–10.70) (Li et al. 2013).

There is no evidence of strong genetic factors in laryngeal carcinogenesis. However, as for other head and neck cancers, polymorphisms for enzymes implicated in the metabolism of alcohol and tobacco and DNA repair might represent susceptibility factors (McKay et al. 2011).

Five-year survival rates are on the order of 65% in high-income countries and 40% in developing countries (Ellis et al. 2012). These patients are at very high risk of developing a second primary tumour in the oral cavity, pharynx, and lung. While shared risk factors are likely to play an important

role, it is plausible that host factors are also partially responsible. For example, an increased sensitivity to mutagens has been shown in lymphocytes of laryngeal cancer patients, in particular those with multiple tumours, as compared to controls. A report indicated that *p53*-related genes are related to susceptibility of second primary cancer in patients with primary head and neck cancer (Jin et al. 2013)

Control of tobacco smoking and excessive alcohol drinking, possibly together with an increased intake of fruits and vegetables, would prevent the majority of cases of laryngeal cancer in most populations. Control of exposure to known and suspected occupational carcinogens is an additional measure for exposed workers. Prophylactic HPV vaccination is promising (Kreimer 2014). No screening methods are currently available for laryngeal cancer.

Prostate cancer

Prostate cancer is the second most common cancer and the fifth leading cause of death from cancer in men worldwide with an estimated 1,276,000 newly diagnosed cases and 359,000 deaths in 2018 (Bray et al. 2018; Wilson and Mucci 2018). Almost 60% of the cases (754,000) occurred in high-income regions. It is the most common male malignant neoplasm in Oceania (79/100,000), North America (74/100,000), and Europe (62/100,000). The incidence is lower in low- and middle-income regions, such as Southern Asia (5.0/100,000). However, rates in Africa are relatively high (e.g. 32/100,000 in Western Africa), despite the low-income level of the region. Mortality rates of prostate cancer show less variability among regions, suggesting that the number of non-fatal cases diagnosed in different countries varies depending on screening and other diagnostic procedures.

The incidence of prostate cancer is highly dependent on the adoption of prostate-specific antigen (PSA) testing. Prostate cancer incidence has shown substantial changes following the introduction of PSA testing, with major increases due to the detection of large number of prevalent cases, followed by substantial declines. The changes in trends have been much smaller for mortality, but both in the United States and in Western Europe, peak rates were observed in the early 1990s, with a levelling off and a decline thereafter, approaching 20% in the EU (Bosetti et al. 2011).

The recent trends in prostate cancer mortality in high-income are consistent with a favourable impact of improved diagnosis, as well as of advances in therapy, on prostate cancer mortality in these countries (Hashim et al. 2016).

Prospective studies failed to provide convincing evidence of an increased risk linked to increased level of testosterone or other sexual hormones. An increased risk of prostate cancer was reported in retrospective studies following a history of benign prostatic hypertrophy, but no excess risk was found in prospective studies. If an association exists between prostatic hypertrophy and cancer, it can be due to shared aetiological factors or to a common pathological process.

It has been suggested that the risk of the disease increases with number of sexual partners and of encounters with prostitutes, and with a previous history of syphilis and gonorrhoea. Serological studies of HPV 16 and HPV 18 have shown an increased risk among positive subjects, but the findings are inconsistent. It is not clear at present, however, whether syphilis and HPV are causal factors or markers of infection with sexually transmitted agents, and hence general markers of inflammation.

A possible protective role of high intake of vegetables has been suggested in a few studies; high intake of meat, dairy products, total fat, and saturated fat might represent a risk factor. The evidence concerning other dietary factors, including fruit intake and intake of specific micronutrients, is, however, largely inconclusive at present. These include lycopene, a retinoid present in particular in tomatoes which has been found to be associated with a reduced risk in a few (but not other) studies, and calcium which has been associated with an elevated risk, possibly on account of its influence on vitamin D balance. There is, in fact, biological evidence that the most active form of vitamin D, 25(OH) D, has a favourable role on prostate cancer, but epidemiological data are inconsistent (Gilbert et al. 2011). Vitamin D may prevent prostate cancer progression, and hence the most aggressive forms of the disease (Li et al. 2007), and there is also an observation by a meta-analysis that an increased risk exists among individuals with elevated serum level of insulin-like growth factor 1 (Chen et al. 2009). An increased risk of the advanced prostate cancer has been reported among subjects with a high BMI (Discacciati et al. 2012). Overall, data on nutritional factors and prostate cancer are largely inconsistent (WCRF/AIRC 2018).

Carriers of *BRCA1* and *BRCA2* mutations have a four- to fivefold increased risk of prostate cancer with especially a two- to threefold increased risk of developing the same neoplasm in first-degree relatives. It is also similarly identified in small magnitudes in breast and colon cancers. More than 60 low-penetrance genetic variants entailing an increased risk of prostate cancer have been identified through GWAS (Rebbeck 2017), in particular within the 8q24 and possibly the 17q12 regions. The 8q24 region appears to be implicated in prostate cancer risk in all race/ethnic groups, while heterogeneity in other variants appear to explain some of the differences risk, in particular among people of African descent (Rebbeck 2017).

The wide geographical variability of prostate cancer suggests that environmental factors likely related to diet and other lifestyle factors, such as physical activity, are important determinants of the disease. Primary prevention, however, is hampered by the fragmentary knowledge of its precise causes. Secondary prevention has been proposed, based on measurement of PSA and digital rectal examination. There is little evidence from controlled trials that either procedure decreases the mortality from prostate cancer (Boyle and Brawley 2009; Welch and Albertsen 2009; Andriole et al. 2012; Schroder et al. 2012). Despite this lack of evidence, these procedures, in particular the PSA testing, have gained popularity in many countries, and are the cause of the steep increase in number of diagnosed cases since the mid-1980s in North America and other high-income countries. It is unclear how much of the decrease in mortality reported since the mid-1990s in the United States and in Western Europe can be attributed to a beneficial effect of (unplanned) use of PSA testing, but it is likely due mainly to improved management and treatment of the disease, including better surgery, radiotherapy, and medical therapy. In 2012, the United States Preventive Services Task Force (USPSTF) made a final recommendation against PSA based screening for healthy men because there is ‘moderate or high certainty that the service has no benefit or that the harms outweigh the benefits’ (Moyer 2012). The USPSTF issued a Grade D to discourage the use of the test by men of all ages for screening purpose. This recommendation is likely to have had an impact on recent trends in prostate cancer incidence (Fleshner et al. 2017).

Cancer of the uterine cervix

The estimated worldwide burden of cervical cancer in 2018 was approximately 570,000 new cases and 311,000 deaths (Bray et al. 2018; Sundström et al. 2018). On a global scale, cervical cancer is the fourth cause of cancer diagnosis and death in women; in particular, it is a major public health problem in middle- and low-income regions, where about 87% of cases (497,000, second most common cancer) and 91% of deaths (282,000, third most common cause of cancer death) occurred. Incidence rates are high (over 25/100,000) in sub-Saharan Africa and Melanesia. Rates are lowest (less than 7/100,000) in Western Asia and Northern Africa. Incidence and mortality rates have decreased steadily in high-income countries, but an upturn in incidence had been observed among young women in a few of these. Few data on temporal trends are available from middle- and low-income countries, but incidence has likely decreased during recent decades in those areas of the world, too. In high-income countries, rates increase up to age 60, while in middle- and low-income countries there is little increase above age 50. Cervical cancer hits preferentially women of lower education and social class.

Most invasive cancers are SqCCs or mixed adeno-squamous tumours. Chronic high-risk HPV infections regresses spontaneously in 60% of cases. Only about 15% of these patients develop, typically within 3–4 years, a cervical squamous intraepithelial neoplasia Grade 1 (CIN-1): this process is facilitated by cofactors such as cigarette smoking; immunosuppression; other cervical infections, such as chlamydia and herpes simplex virus, and oral contraceptive use. If left untreated, about 10% of CIN-1 patients progress to squamous intraepithelial neoplasia grades 2 and 3 (CIN2 and CIN 3). Overall, development of invasive cancer occurs in about 1% of women with high-risk HPV infection, and the required time-period is approximately 25–40 years (Basu et al. 2018).

Chronic infection with HPV is a necessary cause of cervical cancer. Using sensitive molecular techniques, virtually all tumours are positive for the HPV (Clifford et al. 2005). Different types of HPV exist, and those associated with cervical cancer are mainly types 16, 18, 31, 45, and 58 where HPV 16 is the main cervical carcinogen in most populations, while the distribution of other types varies by geographical region. The host response to HPV infection is important in determining its carcinogenic effect; immunosuppression, as present in transplanted patients and HIV

infected women, increases the risk of dysplasia, carcinoma *in situ*, and invasive neoplasms.

Sexual history of women (early age at first intercourse and increased number of sexual partners) and of their male partners (high number of sexual partners, presence of genital diseases, and contact with prostitutes) are risk factors for cervical cancer in many populations. They reflect an increased likelihood of HPV infection, as well as the duration of HPV infection. As for most carcinogens, in fact, duration of exposure is the major determinant of subsequent cancer risk (Plummer et al. 2016).

Tobacco smoking has also an independent effect on cervical carcinogenesis with a RR of 1.5–1.6 for current smokers, also once HPV infection is taken into account (IARC 2004; Moralejo 2009). A possible protective effect of a diet rich in fruits and vegetables has been suggested in a few studies, but the role of diet on cervical cancer risk is probably modest and largely undefined. Studies of infection with other agents, in particular *Chlamydia* and herpes simplex 2, have failed to provide consistent evidence of an effect independent from HPV. An increased risk, on the order of twofold, has been detected among long-term current or recent users of OC, which is not completely explained by sexual behaviour or HPV infection. However, there is no residual association 5–10 years after stopping OC use (La Vecchia and Bosetti 2003). Consequently, the public health implications of OC use on cervical cancer risk are limited in time (Cibula et al. 2010). Condom and (perhaps) diaphragm use, on the other hand, exert a protective effect, possibly via prevention of HPV infection.

Cytological examination of exfoliated cervical cells (the Papanicolaou smear, Pap test) is effective in identifying precursor lesions that can be removed, resulting in a substantial decrease in incidence of and mortality from invasive cancer. Cytological smears are not largely applicable, however, in countries with limited availability of cytologists and pathologists, including many countries with a high prevalence of HPV infection and high incidence of invasive cancer. Alternative approaches for secondary prevention have therefore been proposed, including visual inspection of the cervix with possible enhancement of precursor lesions by visual inspection with acetic acid (VIA). VIA performs as well as the Pap test in terms of sensitivity and specificity and can be used in low-income areas (Bradford and Goodman 2013). Use of HPV-DNA testing appears now, however, to be more specific and sensitive than the Pap test, and is,

therefore, likely to replace the Pap smear as the first screening method in the near future. The primary method for prevention of cervical cancer for future generations, however, is HPV vaccination. Prophylactic HPV vaccines have been progressively introduced worldwide since 2006, and large clinical trials and post-marketing studies have shown their effectiveness and safety against HPV-related lesions and CIN-1 (Lehtinen and Dillner 2013). Three vaccines are currently available (bivalent, quadrivalent, and non-avalent); they all contain antigens of HPV 16 and 18. More than 100 countries have introduced some vaccination programme; and an impact on HPV-related diseases is already visible in high-income countries that started vaccination earlier. Introduction of vaccines in low- and middle-income countries poses economic challenges, but would have a major impact on the cancer burden (de Sanjosé et al. 2019). In addition, smoking cessation may play a role since tobacco smoking is an independent aetiological factor of cervical cancer.

GWAS identified multiple loci within the major histocompatibility complex (MHC) region at 6p21.3 to be associated with cervical cancer, indicating the importance of impaired immune response in cervical cancer development (Chen and Gyllensten 2015).

Principles of primary and secondary cancer prevention

Primary prevention

Many determinants of malignant neoplasms, including UV radiation, ionizing radiation, tobacco smoking, alcohol drinking, overweight, a number of viruses and parasites, and a number of chemicals, industrial processes, and occupational exposures, are sufficiently well established to constitute logical priorities for preventive action. Two more reasons add weight to this priority: some of the agents are responsible for sizeable proportions of the cancers occurring today, and for many agents it is, in principle, feasible to reduce or even to completely eliminate exposure. If this is taken as the objective of preventive action, some practical points are helpful in guiding such action.

First, although epidemiological data in most cases do not allow a direct estimate of the risk of cancer at low doses, it is reasonable (at least from a preventive point of view) to assume that the dose (exposure)–risk relationships for agents acting through damage to DNA are linear with no

threshold (Peto et al. 1991). Second, the carcinogenic effect is not equally dependent on the dose rate (dose per unit of time) and on duration of exposure. For example, in regular smokers, the incidence rate of lung cancer depends more strongly on duration of exposure, increasing with the fourth power of it, than on dose rate, increasing only with the first or second power of it.

Furthermore, as illustrated earlier, the carcinogenic process may be represented as a succession of stages, taking place in the time span from first exposure to a carcinogenic agent to the appearance of clinical cancer. In its simplest form, as first brought out in mouse skin carcinogenesis experiments, the multistage process reduces to two stages: an irreversible 'initiation' stage inducing malignant cells, and a 'promotion' stage which propagates these cells into a malignant growth. A third stage of 'progression', characterized by an increased rate of growth and metastases, as well as an increase in chromosomal changes in the cell, has also been observed. Formal statistical multistage models of carcinogenesis have provided a useful framework to interpret on a common basis of (postulated) mechanisms both experimental and epidemiological observations. As the stages are assumed to occur in a specific sequence, some may be described as 'early' and some as 'late'. Epidemiological observations indicate that, for example, smoking has both an early stage effect, as indicated by the existence of a minimum interval of several years before an increase in risk of lung cancer becomes manifest, and a late stage effect, as indicated by the decrease in risk (with respect to continuous smokers) soon after stopping smoking.

Cancer prevention strategies have evolved from a predominant environmental and lifestyle approach to a model that matches individual-oriented actions with public health interventions. Advances in identifying, developing, and testing agents with the potential either to prevent cancer initiation, or to inhibit or reverse the progression of initiated lesions support this approach. Encouraging laboratory and epidemiological studies, along with studies of secondary end points using intermediate surrogate markers in prevention trials, have provided a scientific rationale for the hypothesis. Promising results have been reported for various types of cancer, in particular among high-risk individuals (Greenwald 2005; Dunn and Greenwald 2010).

Secondary and chemoprevention

Given the limitations still constraining the primary prevention of many cancers, early detection needs to be considered as a secondary and alternative option, based on the reasonable expectation that the earlier the diagnosis and the stage at which a malignancy is discovered, the better the prognosis if treated. This implies that an effective treatment for the disease exists and that the less advanced the cancer at the pre-clinical stage, the better the scope for treatment, and the better the prognosis. This latter aspect cannot be taken for granted.

Before a screening programme can be adopted on a large scale, a number of other requirements need to be fulfilled. First of all, a screening test (that is, a relatively simple and rapid test aimed at the presumptive identification of preclinical disease) must be available that is capable of correctly identifying cases and non-cases. In other words, both sensitivity and specificity should be high, approaching 100%. While high sensitivity is obviously important, given that the very purpose of screening is to pick up, if possible, all cases of a cancer in its detectable preclinical phase, it is specificity that plays a dominant role in the practical utilization of the test within a defined population. As the prevalence of a preclinical cancer to be screened in well-defined populations is often in the range of 1 to 10 per 1,000, if a test is used with a specificity of 95%, then 5% of results will be false positives. In other words, for every case which will turn out at the diagnostic work-up to be a true cancer (assuming 100% sensitivity), there will be 5–50 cases falsely identified as such and ultimately found not to be cancers. This situation is likely to prove unacceptable due to too high psychological and economic costs. One solution is an increase in specificity, for example, by developing better tests or combinations of tests, or by changing the criterion of positivity of a given test to make it more stringent (this necessarily decreases sensitivity). In addition, one might select populations with relatively high prevalence of the cancer (‘high-risk’ groups), so as to increase the number of the true positives. Whatever the group on which the programme operates, additional requirements are that the test is safe, less invasive, easily and rapidly applicable, and acceptable in a broad sense to the population to be examined. It has also to be cheap, but what is or is not cheap is better evaluated within a cost-effectiveness analysis of different ways of preventing a cancer case or death, an issue not further discussed here.

If these requirements are met, still nothing is known about the possible net benefit in outcome deriving from the screening programme (in fact, screening test plus diagnostic work-up plus treatment, as applied in a given population). To evaluate benefit, several measures of outcome can be assessed. An early one, useful but not sufficient, is the distribution by stage of the detected cancer cases which, if the programme is ultimately to be beneficial, should be shifted to earlier, less invasive stages of the disease in comparison with the distribution of the cases discovered through ordinary medical care. A second measure of outcome is the survival of cases detected at screening compared with the survival of cases detected through ordinary medical care. This is a superficially attractive but usually equivocal criterion, to the extent that screening may only advance the time of diagnosis (and therefore the apparent survival time), without postponing the time of death (‘lead-time bias’). A final outcome (and the main test of the programme) is the site-specific cancer mortality in the screened population compared with the mortality in the unscreened population.

Correct, unbiased comparison of this outcome, and thus unbiased measure of the effect of the screening programme, should in principle be made within the framework of a randomized controlled trial, in which two groups of subjects are randomly allocated to the screening programme and to no screening (that is, receiving only the existing medical care system) or to two alternative screening programmes, for instance, entailing different tests or different intervals between periodical examinations. However, largely due to pressures to adopt on a large-scale screening programmes hoped to be effective, a situation has often arisen where withholding screening to a group has been regarded as unethical or socially unacceptable, thus preventing the conduct of a proper experiment. Very few randomized trials evaluating the effectiveness of screening programmes are available. Comparisons are made through non-randomized experiments or through observational studies.

In addition to lead-time bias, three types of bias are peculiar to the assessment of screening programmes. Because of self-selection, persons who elect to receive early detection may be different from those who do not: for instance, they may belong to better educated classes, be generally healthier and health conscious, and this could produce a longer survival independent of any effect of early detection. In addition, cancers with longer preclinical phases, which may mean less biological aggressiveness and better prognosis, are, in any case, more likely to be intercepted by a programme of periodic

screening than cancers with a short preclinical phase, and a rapid, aggressive clinical course (length bias). Finally, because of criteria of positivity adopted to maximize yield of early cases, a number of lesions which in fact would never become malignant growths are included as ‘cases’, thus falsely improving the survival statistics (overdiagnosis bias) (Croswell et al. 2010).

Chemoprevention can also be considered for primary and secondary prevention of cancer, but data are negative or inconsistent for most micronutrients or other substances considered.

Conclusions

Neoplasms are a group of diverse diseases with complex distributions in human populations and with different etiological factors. Current knowledge of the causes of human neoplasms and the development of control strategies have led to the elaboration of lists of recommendations for their prevention (Table 8.2.6). A comprehensive strategy for cancer control might lead to the avoidance of a sizeable proportion of human cancers, and the greatest benefit can be achieved via tobacco control. However, such a strategy would imply major cultural, societal, and economic changes. More modest objectives for cancer prevention should focus on the neoplasms and the exposures that are prevalent in any given population. For example, vaccination of children against HBV is likely to be the most cost-effective cancer prevention action in many countries of Africa and Asia.

Table 8.2.6 European Code Against Cancer, fourth edition (Schüz et al. 2015)

1.	Do not smoke. Do not use any form of tobacco
2.	Make your home smoke-free. Support smoke-free policies in your workplace
3.	Take action to be a healthy body weight
4.	Be physically active in everyday life. Limit the time you spend sitting
5.	Have a healthy diet: • Eat plenty of whole grains, pulses, vegetables, and fruits • Limit high-calorie foods (foods high in sugar or fat) and avoid sugary drinks • Avoid processed meat; limit red meat and foods high in salt
6.	If you drink alcohol of any type, limit your intake. Not drinking alcohol is better for cancer prevention
7.	Avoid too much sun, especially for children. Use sun protection. Do not use sunbeds
8.	In the workplace, protect yourself against cancer-causing substances by following health and safety instructions
9.	Find out if you are exposed to radiation from naturally high radon levels in your home; take action to reduce high radon levels
10.	For women: • Breastfeeding reduces the mother's cancer risk. If you can, breastfeed your baby • Hormone replacement therapy (HRT) increases the risk of certain cancers. Limit use of HRT
11.	Ensure your children take part in vaccination programmes for: • Hepatitis B (for newborns) • Human papillomavirus (HPV) (for girls)
12.	Take part in organized cancer screening programmes for: • Bowel cancer (men and women) • Breast cancer (women) • Cervical cancer (women)

Neoplasms will continue to be a major source of human disease and death. Considerable efforts are made in the public and private domains to develop

effective therapeutic approaches. Even if major discoveries in the clinical management of cancer patients will be accomplished in the near future, the changes will mainly affect the affluent part of the world population. Prevention of the known causes of cancer remains the most promising approach to reducing the consequences of cancer, in particular in countries with limited resources. Control of tobacco smoking and of smokeless tobacco products, moderation in alcohol consumption, processed and red meat intake, increased ingestion of vegetable and fruits, reduced overweight and obesity, increased physical activity, avoidance of exposure to solar radiation, control of known occupational carcinogens, and vaccination, eradication, and treatment against known carcinogenic infectious agents are the main approaches.

REFERENCES

Abnet, C.C., Nyrén, O., and Adami, H.O. (2018). Esophageal cancer. In: Adami, H.O., Hunter, D.J., Lagiou, P., and Mucci, L. (eds). *Textbook of Cancer Epidemiology* (3rd ed.). New York, NY: Oxford University Press, pp. 110–212.

Akhtar, S. (2012). Areca nut chewing and esophageal squamous-cell carcinoma risk in Asians: a meta-analysis of case–control studies. *Cancer Causes Control*, **24**, 257–65.

Alicandro, G., Tavani, A., and La Vecchia, C. (2017). Coffee and cancer risk: a summary overview. *European Journal of Cancer Prevention*, **26**, 424–32.

Allemani, C., et al. (2013). Breast cancer survival in the US and Europe: a CONCORD high-resolution study. *International Journal of Cancer*, **132**, 1170–81.

Amundadottir, L.T. (2016). Pancreatic cancer genetics. *International Journal of Biological Sciences*, **12**, 314–25.

Andriole, G.L., et al. (2012). Prostate cancer screening in the randomized Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial: mortality results after 13 years of follow-up. *Journal of the National Cancer Institute*, **104**, 125–32.

Arends, M.J. (2013). Pathways of colorectal carcinogenesis. *Applied Immunohistochemistry & Molecular Morphology*, **21**, 97–102.

Bagnardi, V., et al. (2015). Alcohol consumption and site-specific cancer risk: a comprehensive dose-response meta analysis. *British Journal of Cancer*, **112**, 580–93.

Bamia, C., Stuver, S., and Mucci, L. (2018). Cancer of the liver and biliary tract. In: Adami, H.O., Hunter, D.J., Lagiou, P., and Mucci, L. (eds). *Textbook of Cancer Epidemiology* (3rd ed.). New York, NY: Oxford University Press, pp. 277–308.

Basu, P., Taghavi, K., Hu, S.Y., Mogri, S., and Joshi, S. (2018). Management of cervical premalignant lesions. *Current Problems in Cancer*, **42**, 129–36.

Benard, F., Barkun, A.N., Martel, M., and von Renteln, D. (2018). Systematic review of colorectal cancer screening guidelines for average-risk adults: summarizing the current global recommendations. *World Journal of Gastroenterology*, **24**, 124–38.

Benetou, V., Ekbom, A., and Mucci, L. (2018). Pancreatic cancer. In: Adami, H.O., Hunter, D.J., Lagiou, P., and Mucci, L. (eds). *Textbook of Cancer Epidemiology* (3rd ed.). New York, NY: Oxford University Press, pp. 309–25.

Beral, V., Million Women Study Collaborators (2003). Breast cancer and hormone-replacement therapy in the Million Women Study. *The Lancet*, **362**, 419–27.
Bhatnagar, R., Dabholkar, J., and Saranath, D. (2012). Genome-wide disease association study in chewing tobacco associated oral cancers. *Oral Oncology*, **48**, 831–5.
Boffetta, P. and Hashibe, M. (2006). Alcohol and cancer. *Lancet Oncology*, **7**, 149–56.
Boffetta, P., Hecht, S., Gray, N., Gupta, P., and Straif, K. (2008). Smokeless tobacco and cancer. *Lancet Oncology*, **9**, 667–75.
Boffetta P, et al. (2009). The causes of cancer in France. *Annals of Oncology*, **20**, 550–5.
Boffetta, P., et al. (2010). An estimate of cancers attributable to occupational exposures in France. *Journal of Occupational and Environmental Medicine*, **52**, 399–406.
Boffetta, P. and Islami, F. (2013). The contribution of molecular epidemiology to the identification of human carcinogens: current status and future perspectives. *Annals of Oncology*, **24**, 901–8.
Boffetta, P., La Vecchia, C., and Moolgavkar, S. (2015). Chronic effects of air pollution are probably overestimated. *Risk Analysis*, **35**, 766–9.
Bosetti, C., Bertuccio, P., Chatenoud, L., Negri, E., La Vecchia, C., and Levi, F. (2011). Trends in mortality from urologic cancers in Europe, 1970– 2008. *European Urology*, **60**, 1–15.
Bosetti, C., Gallus, S., Garavello, W., and La Vecchia, C. (2006). Smoking cessation and the risk of oesophageal cancer: an overview of published studies. *Oral Oncology*, **42**, 957–64.
Bosetti, C., Rosato, V., Gallus, S., Cuzick, J., and La Vecchia, C. (2012). Aspirin and cancer risk: a quantitative review to 2011. Ann Oncol 23:1403–15.
Bosetti, C., Turati, F., and La Vecchia, C. (2014). Hepatocellular carcinoma epidemiology. *Best Practice Research and Clinical Gastroenterology*, **28**, 753–70.
Boyle, P. and Brawley, O.W. (2009). Prostate cancer: current evidence weighs against population screening. *CA: A Cancer Journal for Clinicians*, **59**, 220–4.
Bravi, F., et al. (2012). Dietary patterns and the risk of esophageal cancer. *Annal of Oncology*, **23**, 765–70.
Bravi, F., Tavani, A., Bosetti, C., Boffetta, P., and La Vecchia, C. (2017). Coffee and the risk of hepatocellular carcinoma and chronic liver disease: a systematic review and meta-analysis of prospective studies. *European Journal of Cancer Prevention*, **26**, 368–77.
Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A., and Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, **68**, 394–424.
Brown, K.F., et al. (2018). The fraction of cancer attributable to modifiable risk factors in England, Wales, Scotland, Northern Ireland, and the United Kingdom in 2015. *British Journal of Cancer*, **118**, 1130–41.
Bruce, N., et al. (2015). Does household use of biomass fuel cause lung cancer? A systematic review and evaluation of the evidence for the GBD 2010 study. *Thorax*, **70**, 433–41.
Buas, M.F. and Vaughan, T.L. (2013). Epidemiology and risk factors for gastroesophageal junction tumors: understanding the rising incidence of this disease. *Seminars in Radiation Oncology*, **23**, 3–9.
Cai, M., et al. (2017). Environmental factors, seven GWAS-identified susceptibility loci, and risk of gastric cancer and its precursors in a Chinese population. *Cancer Medicine*, **6**, 708–20.
Chen, D. and Gyllensten U. (2015). Lessons and implications from association studies and post-GWAS analyses of cervical cancer. *Trends in Genetics*, **31**, 41–54.
Chen, W., et al. (2009). Phenotypes and genotypes of insulin-like growth factor 1, IGF-binding protein-3 and cancer risk: evidence from 96 studies. *European Journal of Human Genetics*, **17**, 1668–75.
Chen, Y., et al. (2015). Consumption of hot beverages and foods and the risk of esophageal cancer: a meta-analysis of observational studies. *BMC Cancer*, **15**, 449
Childs, E.J., et al. (2016). Association of common susceptibility variants of pancreatic cancer in higher-risk patients: a PACGENE study. *Cancer Epidemiology, Biomarkers & Prevention*, **25**,

1185–91.
Cibula, D., et al. (2010). Hormonal contraception and risk of cancer. *Human Reproduction Update*, **16**, 631–50.
Clayton, P.E., Banerjee, I., Murray, P.G., and Renehan, A.G. (2011). Growth hormone, the insulin-like growth factor axis, insulin and cancer risk. *Nature Reviews & Endocrinology*, **7**, 11–24.
Clifford, G.M., et al. (2005). Worldwide distribution of human papillomavirus types in cytologically normal women in the International Agency for Research on Cancer HPV prevalence surveys: a pooled analysis. *The Lancet*, **366**, 991–8.
Coleman, M.P., et al. (2008). Cancer survival in five continents: a worldwide population-based study (CONCORD). *Lancet Oncology*, **9**, 730–56.
Collaborative Group on Hormonal Factors in Breast Cancer (CGHFBC) (1996). Breast cancer and hormonal contraceptives: further results. *Contraception*, 54(3 Suppl), 1S–106S.
Collaborative Group on Hormonal Factors in Breast Cancer (CGHFBC) (1997). Breast cancer and hormone replacement therapy: collaborative reanalysis of data from 51 epidemiological studies of 52,705 women with breast cancer and 108,411 women without breast cancer. *The Lancet*, **350**, 1047–59.
Croswell, J.M., Ransohoff, D.F., and Kramer, B.S. (2010). Principles of cancer screening: lessons from history and study design issues. *Seminars in Oncology*, **37**, 202–15.
Crous-Bou, M., De Vivo, I., and Lagiou P. (2018). Endometrial cancer. In: Adami, H.O., Hunter, D.J., Lagiou, .P., and Mucci, L. (eds). *Textbook of Cancer Epidemiology* (3rd ed.). New York, NY: Oxford University Press, pp. 441–57.
Cui, R., et al. (2009). Functional variants in ADH1B and ALDH2 coupled with alcohol and smoking synergistically enhance esophageal cancer risk. *Gastroenterology*, **137**, 1768–75.
Cuzick, J., et al. (2015). Estimates of benefits and harms of prophylactic use of aspirin in the general population. *Annals of Oncology*, **26**, 47–57.
de Blok, C.J.M., et al. (2019). Breast cancer risk in transgender people receiving hormone treatment: nationwide cohort study in the Netherlands. *British Medical Journal*, 365, l1652.
De Sanjosé, S., Brotons, M., LaMontagne, D.S., and Bruni, L. (2019). Human papillomavirus vaccine disease impact beyond expectations. *Current Opinions and Virology*, **39**, 16–22.
Discacciati, A., Orsini, N., and Wolk, A. (2012). Body mass index and incidence of localized and advanced prostate cancer—a dose-response meta-analysis of prospective studies. *Annals of Oncology*, **23**, 1665–71.
Doll, R. and Peto, R. (2005). Epidemiology of cancer. In: Warell, D.A., Cox, T.M., and Firth, J.D. (eds). *Oxford Textbook of Medicine* (4th ed.). Oxford, UK: Oxford University Press, pp. 193–218.
Dunn, B.K. and Greenwald, P. (2010). Cancer prevention I: introduction. *Seminars in Oncology*, **37**, 190–201.
Ellis, L., Rachet, B., Birchall, M., and Coleman, M.P. (2012). Trends and inequalities in laryngeal cancer survival in men and women: England and Wales 1991–2006. *Oral Oncology*, **48**, 284–9.
Engels, E.E. and Hildesheim, A. (2018). Immunological factors. In: Thun, M., Linet, M.S., Cerhan, J.R., Haiman, C.A., and Schottenfeld, D. (eds.). *Cancer Epidemiology and Prevention* (4th ed.) New York, NY: Oxford University Press, pp. 461–87.
Faulds, M.H., et al. (2012). Metabolic diseases and cancer risk. *Current Opinion in Oncology*, 24:58–61.
Ferlay, J., et al. (2019). Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods. *International Journal of Cancer*, **144**, 1941–53.
Ferrández, A., Piazuelo, E., and Castells, A. (2012). Aspirin and the prevention of colorectal cancer. *Best Practice in Research and Clinical Gastroenterology*, **26**, 185–95.
Fleshner, K., Carlsson, S.V., and Roobol, M.J. (2017). The effect of the USPSTF PSA screening recommendation on prostate cancer incidence patterns in the USA. *Nature Reviews in Urology*, **14**, 26–37.

Fritz, A., et al. (eds) (2000). *International Classification of Disease for Oncology (ICD-O)* (3rd ed.). Geneva, Switzerland: WHO.

Garavello, W., et al. (2005). Family history of cancer, its combination with smoking and drinking, and risk of squamous cell carcinoma of the esophagus. *Cancer Epidemiology, Biomarkers & Prevention*, **14**, 1390–3.

Gilbert, R., et al. (2011). Associations of circulating and dietary vitamin D with prostate cancer risk: a systematic review and dose-response meta-analysis. *Cancer Causes Control*, **22**, 319–40.

Giovannucci, E. (2005). The epidemiology of vitamin D and cancer incidence and mortality: a review (United States). *Cancer Causes Control*, **16**, 83–95.

Global Burden of Disease 2015 Tobacco Collaborators (GBD) (2017). Smoking prevalence and attributable disease burden in 195 countries and territories, 1990–2015: a systematic analysis from the Global Burden of Disease Study 2015. *The Lancet*, **389**, 1885–906.

Global Burden of Disease Cancer Collaboration, et al. (GBD) (2018). Global, regional, and national cancer incidence, mortality, years of life lost, years lived with disability, and disability-adjusted life-years for 29 cancer groups, 1990 to 2016: a systematic analysis for the Global Burden of Disease Study. *JAMA Oncology*, **4**, 1553–68.

Greenwald, P. (2005). The future of cancer prevention. *Seminars Oncology Nursing*, **21**, 296–98.

Gutierrez-Dalmau, A. and Campistol, J.M. (2007). Immunosuppressive therapy and malignancy in organ transplant recipients: a systematic review. *Drugs*, **67**, 1167–98.

Hamra, G.B., et al. (2014). Outdoor particulate matter exposure and lung cancer: a systematic review and meta-analysis. *Environmental Health Perspectives*, **122**, 906–11.

Hanahan, D. and Weinberg, R.A. (2011). Hallmarks of cancer: the next generation. *Cell*, **144**, 646–74.

Hashim, D., et al. (2016). The global decrease in cancer mortality: trends and disparities. *Annals of Oncology*, **27**, 926–33.

Henry, S.H., Bosch, F.X., and Bowers, J.C. (2002). Aflatoxin, hepatitis and worldwide liver cancer risks. *Advances in Experimental Medicine and Biology*, 504:229–33.

Hrudey, S.E., et al. (2015). Evaluating evidence for association of human bladder cancer with drinking-water chlorination disinfection by-products. *Journal of Toxicology and Environmental Health B Critical Reviews*, **18**, 213–41.

Huang, B.Z., et al. (2019). Interethnic differences in pancreatic cancer incidence and risk factors: the multiethnic cohort. *Cancer Medicine*, **8**, 3592–603.

Huyghe, J.R., et al. (2019). Discovery of common and rare genetic risk variants for colorectal cancer. *Nature Genetics* 51, 76–87.

International Agency for Research on Cancer (IARC) (2000). *Ionizing Radiation, Part 1: X-radiation and Gamma-radiation, and Neutrons*. *IARC Monographs on the Evaluation of Carcinogenic Risks to Humans*, 75. Lyon, France: International Agency for Research on Cancer.

International Agency for Research on Cancer (IARC) (2001). *Ionizing Radiation, Part 2: Some Internally Deposited Radionuclides*. *IARC Monographs on the Evaluation of Carcinogenic Risks to Humans*, 78. Lyon, France: International Agency for Research on Cancer.

International Agency for Research on Cancer (IARC) (2004). *Tobacco Smoke and Involuntary Smoking*. *IARC Monographs on the Evaluation of Carcinogenic Risks to Humans*, Vol 83. Lyon: International Agency for Research on Cancer.

International Agency for Research on Cancer (IARC) (2018). Preamble to the IARC Monographs [Online]. *IARC Monographs on the Evaluation of Carcinogenic Risks to Humans*. Lyon, France: International Agency for Research on Cancer. [https://www.iarc.fr/wp-content/uploads/2018/07/Monographs-QA.pdf]

Islami, F., et al. (2018). Proportion and number of cancer cases and deaths attributable to potentially modifiable risk factors in the United States. *CA: A Cancer Journal for Clinicians*, **68**, 31–54.

Islami, F., et al. (2011). Alcohol drinking and esophageal squamous cell carcinoma with focus on light-drinkers and never-smokers: a systematic review and meta-analysis. *International Journal of Cancer*, **129**, 2473–84.
Islami, F., et al. (2010). Alcohol drinking and laryngeal cancer: overall and dose risk relation-a systematic review and meta-analysis. *Oral Oncology*, **46**, 802–10.
Islami, F., et al. (2015). Breastfeeding and breast cancer risk by receptor status--a systematic review and meta-analysis. *Annals of Oncology*, **26**, 2398–407.
Jha, P., et al. (2013). 21st-century hazards of smoking and benefits of cessation in the United States. *New England Journal of Medicine*, **368**, 341–50.
Jin, L., et al. (2013). Genetic variants in p53-related genes confer susceptibility to second primary malignancy in patients with index squamous cell carcinoma of head and neck. *Carcinogenesis*, **34**, 1551–7.
Jørgensen, K.J., et al. (2017). Overview of guidelines on breast screening: why recommendations differ and what to do about it. *Breast*, **31**, 261–9.
Kabat, G.C., et al. (2011). Smoking and alcohol consumption in relation to risk of triple-negative breast cancer in a cohort of postmenopausal women. *Cancer Causes Control*, **22**, 775–83.
Kamangar, F., Shakeri, R., Malekzadeh, R., and Islami, F. (2014). Opium use: an emerging risk factor for cancer? *Lancet Oncology*, 15, e69–77.
Karagas, M.R., Gossai, A., Pierce, B., and Ahsan, H. (2015). Drinking water arsenic contamination, skin lesions, and malignancies: a systematic review of the global evidence. *Current Environment Health Reports*, **2**, 52–68.
Keum, N., Wu, K., Giovannucci, E., and Hunter, D.J. (2018). Colorectal cancer. In: Adami, H.O., Hunter, D.J., Lagiou, P., and Mucci, L. (eds). *Textbook of Cancer Epidemiology* (3rd ed.). New York, NY: Oxford University Press, pp. 243–76.
Khan, S.A., Tavolari, S., and Brandi, G. (2019). Cholangiocarcinoma: epidemiology and risk factors. *Liver International*, 39 Suppl 1, 19–31.
Kreimer, A.R. (2014). Prospects for prevention of HPV-driven oropharynx cancer. *Oral Oncology*, **50**, 555–9.
Lagiou, P., Trichopoulos, D., and Adami, H.O. (2018). Concepts in cancer epidemiology and etiology. In: Adami, H.O., Hunter, D.J., Lagiou, P., and Mucci, L. (eds). *Textbook of Cancer Epidemiology* (3rd ed.). New York, NY: Oxford University Press, pp. 111–33.
La Vecchia, C. (2007). Alcohol and liver cancer. *European Journal of Cancer*, **16**, 495–97.
La Vecchia, C. and Bosetti, C. (2003). Oral contraceptives and cervical cancer: public health implications. *European Journal of Cancer Prevention*, **12**, 1–2.
La Vecchia, C., Negri, E., Lagiou, P., and Trichopoulos, D. (2002). Oesophageal adenocarcinoma: a paradigm of mechanical carcinogenesis? *International Journal of Cancer*, **102**, 269–70.
Lee, P.N. (2011). Summary of the epidemiological evidence relating snus to health. *Regulatory Toxicology and Pharmacology*, **59**, 197–214.
Lehtinen, M. and Dillner, J. (2013). Clinical trials of human papillomavirus vaccines and beyond. *Nature Reviews and Clinical Oncology*, **10**, 400–10.
Levi, F., Boffetta, P., and La Vecchia, C. (2008). High constant incidence rates of second primary neoplasms. *European Journal of Cancer Prevention*, **17**, 385–88.
Lewis, S.J. and Smith G.D. (2005). Alcohol, ALDH2, and esophageal cancer: a meta-analysis which illustrates the potentials and limitations of a Mendelian randomization approach. *Cancer Epidemiology, Biomarkers & Prevention*, **14**, 1967–71.
Li, H., et al. (2007). A prospective study of plasma vitamin D metabolites, vitamin D receptor polymorphisms, and prostate cancer. *PLoS Medicine*, 4, e103.
Li, X., et al. (2013). Human papillomavirus infection and laryngeal cancer risk: a systematic review and meta-analysis. *Journal of Infectious Disease*, **207**, 479–88.
Loria, D., Barrios, E., and Zanetti, R. (2009). Cancer and yerba mate consumption: a review of possible associations. *Revista Panamericana de Salud Pública*, **25**, 530–9.

Low, S.K., Zembutsu, H., and Nakamura, Y. (2018). Breast cancer: the translation of big genomic data to cancer precision medicine. *Cancer Science*, **109**, 497–506.

Lu, Y., et al. (2019). Large –scale genome-wide association study of East Asians identifies loci associated with risk for colorectal cancer. *Gastroenterology*, **156**, 1455–66.

Lubin, J.H., et al. (1994). Radon exposure in residences and lung cancer among women: combined analysis of three studies. *Cancer Causes Control*, **5**, 114–28.

Lucas, A.L., et al. (2016). Global trends in pancreatic cancer mortality from 1980 through 2013 and predictions for 2017. *Clinical Gastroenterology and Hepatology*, **14**, 1452–62.

Lucenteforte, E., et al. (2012). Alcohol consumption and pancreatic cancer: a pooled analysis in the International Pancreatic Cancer Case–Control Consortium (PanC4). *Annals of Oncology*, **23**, 374–82.

Malhotra, J., Malvezzi, M., Negri, E., La Vecchia, C., and Boffetta, P. (2016). Risk factors for lung cancer worldwide. *European Respiratory Journal*, **48**, 889–902.

Malhotra, J., Boffetta, P., and Mucci, L. (2018). Cancer of the lung, larynx, and pleura. In: Adami, H.O., Hunter, D.J., Lagiou, P., and Mucci, L. (eds). *Textbook of Cancer Epidemiology* (3rd ed.). New York, NY: Oxford University Press, pp. 327–53.

Malvezzi, M., et al.(2017). European cancer mortality predictions for the year 2017, with focus on lung cancer. *Annals of Oncology*, **28**, 117–23.

Mantovani, A., Ponzetta, A., Inforzato, A., and Jaillon, S. (2019). Innate immunity, inflammation and tumour progression: double-edged swords. *Journal of Internal Medicine*, **285**, 524–32.

Marshall, A.L. and Christiani, D.C. (2013). Genetic susceptibility to lung cancer—light at the end of the tunnel? *Carcinogenesis*, **34**, 487–502.

Martínez Hernández, J., López-Rodas, V., and Costas, E. (2009). Microcystins from tap water could be a risk factor for liver and colorectal cancer: a risk intensified by global change. *Medical Hypotheses*, **72**, 539–40.

Matin, S.F., et al. (2014). Highlights from the first symposium on upper tract urothelial carcinoma. *Urology Oncology*, **32**, 309–16.

McKay, J.D., et al. (2011). A genome-wide association study of upper aerodigestive tract cancers conducted within the INHANCE consortium. *PLoS Genetics*, 7, e1001333.

Mehrtash, H., et al. (2017). Defining a global research and policy agenda for betel quid and areca nut. *Lancet Oncology*, 18, e767–75.

Mocellin, S., Verdi, D., Pooley, K.A., and Nitti, D. (2015). Genetic variation and gastric cancer risk: a field synopsis and meta-analysis. *Gut*, **64**, 1209–19.

Moralejo, D. (2009). Smoking increased risk of cervical cancer, independent of infection with high-risk HPV types. *Evidence Based Nursing*, **12**, 122.

Moyer, V.A. and US Preventive Services Task Force (2012). Screening for prostate cancer: US Preventive Services Task Force recommendation statement. *Annals of Internal Medicine*, **157**, 120–34.

Moyer, V.A. and US Preventive Services Task Force (2014). Screening for lung cancer: US Preventive Services Task Force recommendation statement. *Annals of Internal Medicine*, **160**, 330–8.

Narayanan, D.L., Saladi, R.N., and Fox, J.L. (2010). Ultraviolet radiation and skin cancer. *International Journal of Dermatology*, **49**, 978–86.

Pearce, M.S., et al. (2012). Radiation exposure from CT scans in childhood and subsequent risk of leukaemia and brain tumours: a retrospective cohort study. *The Lancet*, **380**, 499–505.

Pelucchi, C., Gallus, S., Garavello, W., Bosetti, C., and La Vecchia, C. (2008). Alcohol and tobacco use, and cancer risk for upper aerodigestive tract and liver. *European Journal of Cancer Prevention*, **17**, 340–4.

Peto, R., Darby, S., Deo, H., Silcocks, P., Whitley, E., and Doll, R. (2000). Smoking, smoking cessation, and lung cancer in the UK since 1950: combination of national statistics with two case–control studies. *British Medical Journal*, **321**, 323–9.

Peto, R., Gray, R., Brantom, P., and Grasso, P. (1991). Effects on 4080 rats of chronic ingestion of N-nitrosodiethylamine or N-nitrosodimethylamine: a detailed dose–response study. *Cancer Research*, **51**, 6415–51.

Phipps, A.I., et al. (2011a). Body size, physical activity, and risk of triple-negative and estrogen receptor-positive breast cancer. *Cancer Epidemiology, Biomarkers & Prevention*, **20**, 454–63.

Phipps, A.I., et al. (2011b). Reproductive history and oral contraceptive use in relation to risk of triple-negative breast cancer. *Journal of National Cancer Institute*, **103**, 470–7.

Piper, J.M., Tonascia, J., and Matanoski, G.M. (1985). Heavy phenacetin use and bladder cancer in women aged 20 to 49 years. *New England Journal of Medicine*, **313**, 292–5.

Pirie, K., Peto, R., Reeves, G.K., Green, J., Beral, V., and Million Women Study Collaborators (2013). The 21st century hazards of smoking and benefits of stopping: a prospective study of one million women in the UK. *The Lancet*, **381**, 133–41.

Plummer, M., et al. (2007). Chemoprevention of precancerous gastric lesions with antioxidant vitamin supplementation: a randomized trial in a high-risk population. *Journal of the National Cancer Institute*, **99**, 137–46.

Plummer, M., de Martel, C., Vignat, J., Ferlay, J., Bray, F., and Franceschi, S. (2016). Global burden of cancers attributable to infections in 2012: a synthetic analysis. *Lancet Global Health*, 4, e609–16.

Praud, D., et al. (2016). Cancer incidence and mortality attributable to alcohol consumption. *International Journal of Cancer*, **138**, 1380–7.

Pytynia, K.B., Dahlstrom, K.R., and Sturgis, E.M. (2014). Epidemiology of HPV-associated oropharyngeal cancer. *Oral Oncology*, **50**, 380–6.

Qiao, Y.L., et al. (2009). Total and cancer mortality after supplementation with vitamins and minerals: follow up of the Linxian General Population Nutrition Intervention Trial. *Journal of the National Cancer Institute*, **101**, 507–18.

Rawla, P., Sunkara, T., and Gaduputi, V. (2019). Epidemiology of pancreatic cancer: global trends, etiology and risk factors. *World Journal of Oncology*, **10**, 10–27.

Rebbeck, T.R. (2017). Prostate cancer genetics: variation by race, ethnicity, and geography. *Seminars of Radiation Oncology*, **27**, 3–10.

Richardson, S.D., Plewa, M.J., Wagner, E.D., Schoeny, R., and Demarini, D.M. (2007). Occurrence, genotoxicity, and carcinogenicity of regulated and emerging disinfection by-products in drinking water: a review and roadmap for research. *Mutation Research*, **636**, 178–242.

Rider, J., Brennan, P., and Lagiou, P. (2018). Oral and pharyngeal cancer. In: Adami, H.O., Hunter, D.J., Lagiou, P., and Mucci, L. (eds). *Textbook of Cancer Epidemiology* (3rd ed.). New York, NY: Oxford University Press, pp. 137–57.

Rota, M., Pizzato, M., La Vecchia, C., and Boffetta, P. (2019). Efficacy of lung cancer screening appears to increase with prolonged intervention: results from the MILD trial and a meta-analysis. *Annals of Oncology*, 30(7), 1040–3.

Ruddy, K.J. and Winer, E.P. (2013). Male breast cancer: risk factors, biology, diagnosis, treatment, and survivorship. *Annals of Oncology*, **24**, 1434–43.

Salama, N.R., Hartung, M.L., and Müller, A. (2013). Life in the human stomach: persistence strategies of the bacterial pathogen Helicobacter pylori. *Nature Reviews and Microbiology*, **11**, 385–99.

Salehi, F., Turner, M.C., Phillips, K.P., Wigle, D.T., Krewski, D., and Aronson, K.J. (2008). Review of the etiology of breast cancer with special attention to organochlorines as potential endocrine disruptors. *Journal of Toxicology and Environmental Health, Part B: Critic Reviews*, **11**, 276–300.

Sankaranarayanan, R., et al. (2013). Long term effect of visual screening on oral cancer incidence and mortality in a randomized trial in Kerala, India. *Oral Oncology*, **49**, 314–21.

Sanyal, A., Poklepovic, A., Moyneur, E., and Barghout, V. (2010). Population-based risk factors and resource utilization for HCC: US perspective. *Current Medical Research and Opinion*, **26**, 2183–91.

Schroder, F.H., et al. (2012). Prostate-cancer mortality at 11 years of follow-up. *New England Journal of Medicine*, **366**, 981–90.

Schüz, J., et al. (2015). European Code against Cancer 4th Edition: 12 ways to reduce your cancer risk. *Cancer Epidemiology*, 39 Suppl 1, S1–10.

Shakeri, R., et al. (2012). Is opium a real risk factor for esophageal cancer or just a methodological artifact? Hospital and neighborhood controls in case–control studies. *PLoS One*, 7, e32711.

Siegel, R., Ma, J., Zou, Z., and Jemal, A. (2014). Cancer statistics, 2014. *CA: A Cancer Journal for Clinicians*, **64**, 9–29.

Siu, A.L., US Preventive Services Task Force. (2016). Screening for breast cancer: US Preventive Services Task Force recommendation statement. *Annals of Internal Medicine*, **164**, 279–96.

Stiborová, M., Arlt, V.M., and Schmeiser, H.H. (2016). Balkan endemic nephropathy: an update on its aetiology. *Archives of Toxicology*, **90**, 2595–615.

Stubbins, R.E., Hakeem, A., and Núñez, N.P. (2012). Using components of the vitamin D pathway to prevent and treat colon cancer. *Nutrition Reviews*, **70**, 721–9.

Sun, Z., et al. (2013). Dramatic reduction of liver cancer incidence in young adults: 28 year follow-up of etiological interventions in an endemic area of China. *Carcinogenesis*, **34**, 1800–5.

Sundström, K., Dillner, J., and Adami, H.O. (2018). Cervical cancer. In: Adami, H.O., Hunter, D.J., Lagiou, P., and Mucci, L. (eds). *Textbook of Cancer Epidemiology* (3rd ed.). New York, NY: Oxford University Press, pp. 421–39.

Tamimi, R., Hankinson, S., and Lagiou, P. (2018). Breast cancer. In: Adami, H.O., Hunter, D.J., Lagiou, P., and Mucci, L. (eds). *Textbook of Cancer Epidemiology* (3rd ed.). New York, NY: Oxford University Press, pp. 381–419.

Torre, L.A., Siegel, R.L., Ward, E.M., and Jemal, A. (2016). Global cancer incidence and mortality rates and trends-an update. *Cancer Epidemiology, Biomarkers & Prevention*, **25**, 16–27.

Tramacere, I., et al. (2010). A meta-analysis of alcohol drinking and oral and pharyngeal cancers. Part 1: overall results and dose-risk relation. *Oral Oncology*, **46**, 497–503.

Tramacere, I., La Vecchia, C., and Negri, E. (2011). Tobacco smoking and esophageal and gastric cardia adenocarcinoma: a meta-analysis. *Epidemiology*, **22**, 344–9.

Tramacere, I., et al. (2012). A meta-analysis on alcohol drinking and esophageal and gastric cardia adenocarcinoma risk. *Annals of Oncology*, **23**, 287–97.

Turati, F., et al. (2010). A meta-analysis of alcohol drinking and oral and pharyngeal cancers. Part 2: results by subsites. *Oral Oncology*, **46**, 720–6.

Turati, F., Tramacere, I., La Vecchia, C., and Negri, E. (2013). A meta-analysis of body mass index and esophageal and gastric cardia adenocarcinoma. *Annals of Oncology*, **24**, 609–17.

Turner, P.C., Flannery, B., Isitt, C., Ali, M., and Pestka, J. (2012). The role of biomarkers in evaluating human health concerns from fungal contaminants in food. *Nutrition Research Reviews*, **25**, 162–79.

Wang, L.D., et al. (2010). Genome-wide association study of esophageal squamous cell carcinoma in Chinese subjects identifies susceptibility loci at PLCE1 and C20orf54. Nat Genet 42:759–63.

Waziry, R., et al. (2017). Hepatocellular carcinoma risk following direct-acting antiviral HCV therapy: a systematic review, meta-analyses, and meta-regression. J Hepatol 67:1204–12.

Welch, H.G. and Albertsen, P.C. (2009). Prostate cancer diagnosis and treatment after the introduction of prostate-specific antigen screening: 1986–2005. *Journal of the National Cancer Institute*, **101**, 1325–9.

Wilson, K.M. and Mucci, L. (2018). Prostate cancer. In: Adami, H.O., Hunter, D.J., Lagiou, P., and Mucci, L. (eds). *Textbook of Cancer Epidemiology* (3rd ed.). New York, NY: Oxford University Press, pp. 481–523.

Wolf, A.M.D., et al. (2018). Colorectal cancer screening for average-risk adults: 2018 guideline update from the American Cancer Society. *CA: A Cancer Journal for Clinicians*, **68**, 250–81.

World Cancer Research Fund, American Institute for Cancer Research (WCRF/AIRC) (2018). *Diet, Nutrition, Physical Activity and Cancer: A Global Perspective*. London: WCRF.

Wu, C., et al. (2011). Genome-wide association study identifies three new susceptibility loci for esophageal squamous-cell carcinoma in Chinese populations. *Nature Genetics*, **43**, 679–84.

Wu, M., et al. (2013). Single nucleotide polymorphisms of ADH1B, ADH1C and ALDH2 genes and esophageal cancer: a population-based case–control study in China. *International Journal of Cancer*, **132**, 1868–77.

Wu, M., et al. (2011a). A large proportion of esophageal cancer cases and the incidence difference between regions are attributable to lifestyle risk factors in China. *Cancer Letters*, **308**, 189–96.

Wu, M., et al. (2011b). Smoking and alcohol drinking increased the risk of esophageal cancer among Chinese men but not women in a high-risk population. *Cancer Causes Control*, **22**, 649–57.

Xue, F. and Michels, K.B. (2007). Intrauterine factors and risk of breast cancer: a systematic review and meta-analysis of current evidence. *Lancet Oncol*, **8**, 1088–100.

Ye, W., Nyrén, O., and Adami, H.O. (2018). Stomach cancer. In: Adami, H.O., Hunter, D.J., Lagiou, P., and Mucci, L. (eds). *Textbook of Cancer Epidemiology* (3rd ed.). New York, NY: Oxford University Press, pp. 213–42.

Yi, M., et al. (2019). Tea consumption and health outcomes: umbrella review of meta-analyses of observational studies in humans. *Molecular Nutrition & Food Research*, 63, e1900389.

Zendehdel, K., et al. (2008). Risk of gastroesophageal cancer among smokers and users of Scandinavian moist snuff. *International Journal of Cancer*, **122**, 1095–9.

Zucchetto, A., et al. (2016). Non-AIDS-defining cancer mortality: emerging patterns in the late HAART era. *Journal of Acquired Immune Deficiency Syndromes*, **73**, 190–6.

8.3

Chronic obstructive pulmonary disease and asthma

Craig M. Riley, Jessica Bon, and Alison Morris

Introduction

The most common non-malignant respiratory conditions characterized by airway dysfunction are referred to collectively as obstructive airway diseases and include chronic obstructive pulmonary disease (COPD) and asthma. Both conditions are highly prevalent and have increased dramatically worldwide in the past few decades. COPD is now the fourth leading cause of death worldwide (WHO 2007), and both COPD and asthma have a profound impact on quality of life for patients and their families. In this chapter, we review definitions, epidemiology, and risk factors for COPD and asthma.

Definitions

COPD

Traditionally, COPD has been defined as irreversible airflow obstruction due to chronic bronchitis and emphysema that progresses over time (Petty 2006). The Ciba Guest Symposium (Ciba Foundation Guest Symposium 1959) and the American Thoracic Society symposium (Committee on Diagnostic Standards for Nontuberculous Respiratory Diseases 1962) proposed the first formal definitions of chronic bronchitis and emphysema respectively. Current definitions are still largely based on these initial definitions; that is, chronic bronchitis is defined clinically as the presence of chronic productive cough for at least 3 consecutive months in 2 consecutive years, and emphysema is a pathological diagnosis of an increase in the size of the distal airspaces and destruction of their walls without obvious fibrosis (American Thoracic Society Committee on Diagnostic Standards 1962). The most

recent definitions of COPD as reported in the American Thoracic Society (ATS), European Respiratory Society (ERS), and the Global Initiative for Chronic Obstructive Lung Disease (GOLD) guidelines (Agusti et al. 2017) emphasize the inflammatory response to noxious particles and gases as the predominant pathological feature of the disease and have parted from the definition of COPD as being a syndrome of chronic bronchitis and emphysema. The widely accepted GOLD definition (Agusti et al. 2017) states that COPD is:

> a common, preventable and treatable disease that is characterized by persistent respiratory symptoms and airflow limitation that is due to airway and/or alveolar abnormalities usually caused by significant exposure to noxious particles or gases. (Agusti et al. 2017)

Although no longer specifically included in the definition, chronic bronchitis and emphysema remain important causes of the chronic airflow limitation characteristic of COPD. The three main components (inflammation, airflow limitation that is not fully reversible, and a gradual loss of lung function over time) represent the major pathophysiological events leading to the symptoms typically expressed by those with COPD: chronic and progressive cough, sputum production, and dyspnoea.

Clinical assessment of COPD

Spirometry is an essential tool in the clinical diagnosis of COPD. COPD is still traditionally diagnosed based on the post-bronchodilator ratio of the forced expiratory volume in 1 second (FEV1, the volume of air exhaled in the first second of expiration) to the forced vital capacity (FVC, the maximum volume of air that can be forcibly expired). A cut-point of 70% continues to be widely used (Agusti et al. 2017), but increasing attention has been paid to utilizing the lower limit of normal as a cut-off for the FEV_1/FVC ratio to define obstruction (Vaz Fragoso et al. 2016). Use of this cut-off has the potential advantage of limiting underdiagnosis in younger adults and overdiagnosis in the elderly, but to date, a clear clinical advantage to utilizing the lower limit of normal over a fixed cut-off has not been demonstrated (Calverley et al. 2018). Bronchodilator treatment prior to spirometry is important as it establishes whether obstruction is irreversible and thus helps to distinguish COPD from asthma. In COPD with mild or moderate obstruction, the FVC is generally unaffected, or only marginally

affected; however, FEV_1 is reduced. In severe obstruction, air-trapping may lead to a reduction in FVC as well as FEV_1, which can be differentiated from restriction by analysis of flow volume loops or lung volume measurements by nitrogen washout or plethysmography. Patients with COPD also frequently have a low diffusing capacity for carbon monoxide (DLco) because of the destruction of lung tissue resulting in decreased gas transfer.

The degree of severity of obstruction in those with FEV_1/FVC less than 70% is usually based on the patient's FEV_1. The 2018 GOLD criteria classify obstruction severity in COPD into four stages (Agusti et al. 2017) (Box 8.3.1). While the degree of obstruction is useful for monitoring the progression of disease and correlates strongly with mortality (Mannino et al. 2003), FEV_1 is less predictive of other patient-centred outcomes including symptom burden and acute exacerbations (Han et al. 2013). As these patient-centred outcomes are the major targets of common treatment modalities, recent GOLD guideline updates have developed risk-stratification tools and treatment algorithms that incorporate symptom scores and exacerbation risk independently of spirometry discussed next (Agusti et al. 2017).

Box 8.3.1 GOLD criteria of COPD severity

- Stage I, mild: $FEV_1 \geq 80\%$ predicted.
- Stage II, moderate: $50\% \leq FEV_1 \geq 80\%$ predicted.
- Stage III, severe: $30\% \leq FEV_1 \geq 50\%$ predicted.
- Stage IV, very severe: $FEV_1 < 30\%$ predicted.

Global Initiative for Chronic Obstructive Lung Disease (GOLD). Global Strategy for the Diagnosis, Management and Prevention of Chronic Obstructive Pulmonary Disease: 2018 Report.

Epidemiologic assessment of COPD

The methodology used to assess COPD prevalence in epidemiologic cohorts and population-based studies often leads to underestimation of the true prevalence. In population-based surveys, COPD is often defined on the basis of: (1) self-report of a doctor diagnosis of COPD, bronchitis, or emphysema; (2) self-report of respiratory symptoms; and (3) spirometry with or without

prior bronchodilator treatment. Self-report of a clinical diagnosis significantly underestimate the true disease prevalence (Halbert et al. 2006). More accurate approaches incorporate both spirometry, exposure history, and respiratory symptom assessment (Buist et al. 2007).

The variable use of bronchodilators significantly complicates large population-based spirometry surveys, and many studies unfortunately do not collect post-bronchodilator measurements. Failing to check for reversibility of airflow obstruction (using pre- and post-bronchodilator spirometry) may result in an overestimation of the prevalence of COPD. In population studies comparing pre- and post-bronchodilator spirometry to examine COPD prevalence, using post-bronchodilator values lowered the estimated rate of airflow obstruction by 30–50% (Tilert et al. 2013). Therefore, the use of post-bronchodilator spirometry to determine the diagnosis of COPD in population-based studies is strongly recommended.

Asthma

The word ‘asthma’ comes from a Greek word meaning ‘panting’ (Keeney 1964), but reference to asthma-like symptoms can also be found in ancient Egyptian, Hebrew, and Indian medical writings (Ellul-Micallef 1976). There were clear observations of patients experiencing attacks of asthma in the second century, and evidence of disordered lung anatomy as far back as the seventeenth century (Willis 1678). The definition of asthma initially proposed at the Ciba Foundation conference in 1959 (Ciba Foundation Guest Symposium 1959) and endorsed by the ATS in 1962 (American Thoracic Society Committee on Diagnostic Standards 1962) is that ‘asthma is a disease characterized by wide variation over short periods of time in resistance to flow in the airways of the lung’. Although these features receive lesser prominence in some current definitions given the now-recognized importance of airways inflammation, they are still integral to the current Global Initiative for Asthma (GINA) description of asthma as:

> a heterogeneous disease, usually characterized by chronic airway inflammation. It is defined by the history of respiratory symptoms such as wheeze, shortness of breath, chest tightness and cough that vary over time and in intensity, together with variable expiratory airflow limitation. (GINA 2018)

This definition of asthma encompasses both the pathophysiological mechanisms (in chronic airway inflammation and variable airflow

obstruction due to bronchial hyperreactivity) and their downstream effects on patients by which providers clinically identify and diagnose this disease.

Clinical assessment of asthma

There is no single test or pathognomonic feature that defines the presence or absence of asthma. Furthermore, the variability of the condition means that evidence of it may not be present at the initial assessment. Thus, a diagnosis of asthma is made on the basis of the clinical history in combination with physical examination and respiratory function tests over a period of time. As a result, asthma is subject simultaneously both to significant under- and overdiagnosis. Studies have shown both a significantly lower prevalence of physician-diagnosed asthma in the community as compared to the prevalence of asthma-like symptoms, suggesting that 20–70% of asthmatics are undiagnosed (Aaron et al. 2018). Furthermore, there are several disorders that have signs and symptoms (such as wheeze) that can be confused with or mimic aspects of asthma and lead to overdiagnosis in approximately 1/3 of cases (Aaron et al. 2017); common conditions mistaken for asthma include gastroesophageal reflux disease, chronic bronchitis, lower respiratory tract infections, bronchiectasis, vocal cord dysfunction, obesity, and COPD.

There are a number of tests that facilitate the diagnosis and monitoring of asthma. Measurements of lung function are the most frequently used and provide important information on airflow variability, reversibility, and severity. Airflow limitation is generally measured using spirometry or a peak expiratory flow (PEF) meter. PEF meters are inexpensive and easy to use, but they are less precise than spirometry and may underestimate the degree of airflow limitation. Therefore, spirometry is preferred for diagnosis when available. Because of the importance of bronchial hyperreactivity and bronchodilator reversibility, guidelines emphasize the importance of conducting spirometry or PEF measurements both before and after bronchodilator treatment when assessing lung function in asthma (GINA 2018). Pre- and post-bronchodilator testing is important since it establishes if obstruction is irreversible and allows distinction from COPD. Reversibility of FEV_1 of $\geq$ 12% and $\geq$ 200 mL from the pre-bronchodilator value is generally accepted as a valid indication of asthma (GINA 2018). However, due to the highly variable nature of the condition, repeated lung function tests are required, as some patients may not display reversibility when their

asthma is well-controlled, and reversibility may be present in non-asthmatics.

In individuals with asthma symptoms and normal lung function, bronchial hyperresponsiveness (BHR) testing may be used as a diagnostic aid. BHR constitutes airway narrowing to non-specific stimuli, such as exercise, cold air, and chemical irritants. Although BHR is related to asthma, it may occur independently of asthma, and vice versa, which makes BHR testing insufficient to diagnose asthma in isolation.

An increasing number of tests are available to measure non-invasive markers of airway inflammation, including sputum induction tests, exhaled nitric oxide, and measurements of inflammatory markers in exhaled breath condensate. While these tests may be useful in establishing asthma phenotypes and identifying patients who are more likely to respond to optimal treatment, they have as yet to be convincingly demonstrated to aid in asthma diagnosis (Colak et al. 2018).

There have been several approaches used to categorize asthma (Wenzel 2012). Until recently, asthma has most commonly been clinically classified on the basis of severity using GINA criteria (GINA 2006), which subdivided asthma into four categories (intermittent, mild persistent, moderate persistent, and severe persistent). More recent guidelines (GINA 2018) have suggested that clinical categorization should be based on the current level of asthma control (controlled, partly controlled, and uncontrolled), which may more adequately address the individual clinical needs of the patient. Other categorizations of asthma exist based on type of inflammation seen such as atopic, eosinophilic, or TH-2 high (Tran et al. 2016). Although these asthma endotypes may overlap, they may represent different mechanistic pathways and respond differently to treatment.

Defining asthma in epidemiological surveys

Defining and diagnosing asthma in population-based epidemiological surveys poses even greater difficulties than defining asthma in individuals. As a result, asthma prevalence surveys usually focus on self- or parental-reported ‘asthma symptoms’ rather than diagnosed asthma. This approach allows a large number of participants to be rapidly surveyed without great cost. Of the symptoms clinically associated with asthma, epidemiological studies have shown that wheezing is the most important symptom for the

identification of asthma, and thus questionnaires used to assess asthma prevalence are often based on wheezing. An alternative approach to symptom questionnaires has been to use more ‘objective’ measures such as BHR testing, either alone or in combination with questionnaires. BHR may not be more reliable than questionnaire alone, in part due to the variability of BHR at the time of assessment.

Overlap syndrome

Although reversibility of airway obstruction (or the lack thereof) is the primary clinical characteristic used to define either asthma or COPD, it is increasingly recognized that patients may have features of both diseases (Woodruff et al. 2017). It is often difficult to clearly distinguish and diagnose the two conditions, particularly in older patients. These observations have led to the description of asthma-COPD overlap (ACO), in which obstructive respiratory disease shows features of both asthma and COPD (Woodruff et al. 2017). The epidemiology of ACO is not well-understood, and attempts to characterize its prevalence and associations have been limited by different definitions across various studies. Approximately 17% of patients diagnosed with either asthma or COPD would meet criteria for ACO, and these patients are at higher risk for exacerbation, hospitalization, and disability compared to either population alone (Vaz Fragoso et al. 2017; Woodruff et al. 2017). Further work needs to be done to clearly define this syndrome to better identify long-term outcomes and efficacious treatment strategies as ACO patients have generally been excluded from treatment trials of asthma and COPD.

Epidemiology, risk factors, and prevention of COPD

Epidemiology of COPD

COPD is a major global cause of mortality and morbidity, with exacerbations leading to frequent hospitalization. Despite the large personal, societal, and economic burden, it has been difficult to accurately estimate the true prevalence and it is often underdiagnosed. According to the World Health Organization, the global burden of COPD is estimated at 210 million

people with 65 million suffering from moderate to severe disease (WHO 2007). It is also a major cause of healthcare costs and lost productivity.

European epidemiological studies of COPD show prevalence estimates for COPD of 2 to greater than 10% (WHO 2007) (Fig. 8.3.1). These differences in prevalence estimates are presumed to be attributable to differences in risk exposures or population characteristics, but methods and definitions used to measure COPD may also play a role. Definitions of COPD used in these studies vary from a doctor's diagnosis of COPD, to more rigid definitions based on pathology and/or pulmonary lung function testing.

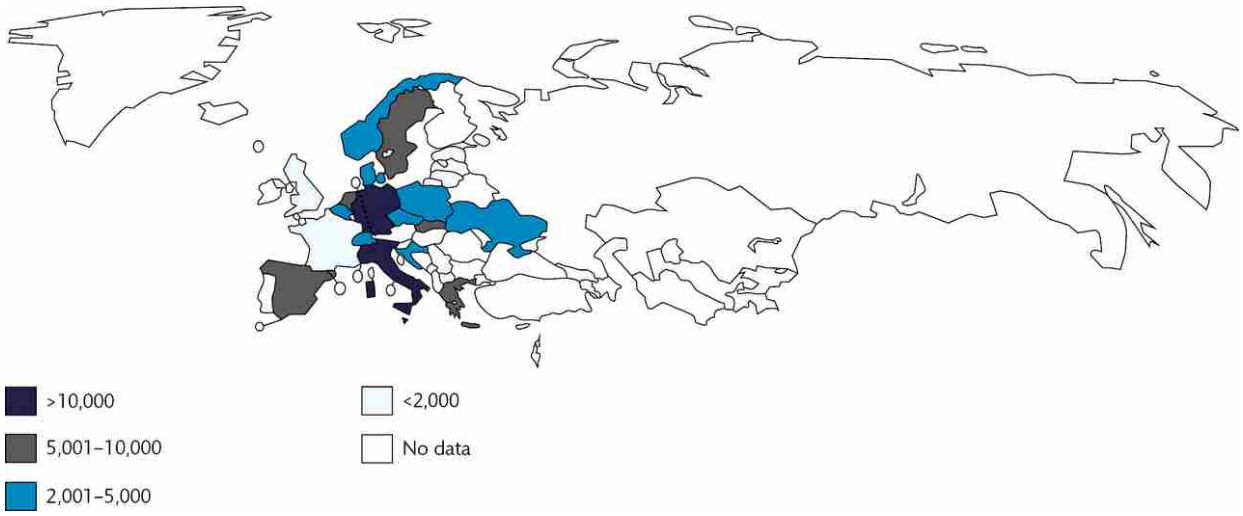


Fig. 8.3.1 Prevalence rate (1/100,000) of COPD in Europe.
Reproduced with permission from World Health Organization, Global surveillance, prevention and control of chronic respiratory diseases,
http://apps.who.int/iris/bitstream/handle/10665/43776/9789241563468_eng.pdf

Halbert and co-workers (Halbert et al. 2003) assessed the reasons for conflicting prevalence estimates in the literature. They selected studies that had estimated population-based COPD prevalences and clearly described methods used to obtain these estimates. In total, 32 studies, representing 17 countries and eight WHO regions, presenting prevalence data for COPD were identified and reviewed. Prevalence estimates were based on spirometry (11 studies), respiratory symptoms (14 studies), patient-reported disease (10 studies), or expert opinion. The reported prevalence of COPD ranged from 0.23 to 18.3%, with the lowest prevalences (0.2 to 2.5%) being those based on expert opinion. Sixteen studies had measured rates that could

confidently be extrapolated to an entire region or country. All 16 studies were conducted in Europe or North America, and in most, the prevalence was between 4% and 10%. However, in some studies, COPD prevalence may be higher. For example, preliminary data from 13,301 participants in the LifeLines cohort study performed in the Netherlands showed that 1920 out of 13,301 subjects had GOLD stage I or higher (14%) (i.e. 1,402 had GOLD stage I, 494 GOLD stage II, and 24 GOLD stage III/IV). In a later study, Halbert et al. (Halbert et al. 2006) identified 101 COPD prevalence estimates from 28 countries during the 1990–2004 time period. The pooled prevalence of COPD was 7.6%; the prevalence of chronic bronchitis alone was 6.4%, and the prevalence of emphysema alone was 1.8%. The pooled prevalence of COPD based on spirometry was 8.9%.

Although the most accurate prevalence data are from high-income countries (HCs), the WHO estimated that almost 90% of the 3.22 million deaths estimated to be attributable to COPD in 2015 occurred in low- and middle-income countries (LMICs) (WHO 2013). Few epidemiologic studies have provided truly comparable international prevalence estimates for COPD. The establishment of the Burden of Obstructive Lung Disease (BOLD) Initiative to assess the global burden of COPD allows more valid comparisons of the prevalence of COPD over time and across ethnopolitical regions (Buist et al. 2007). This study included 9,425 participants from 12 centres in 12 countries including China, Turkey, Austria, South Africa, Iceland, Germany, Poland, Norway, Canada, the United States, Philippines, and Australia. The study showed considerable variation in COPD prevalences, with GOLD stage II or higher COPD in women ranging from 5.1% in Guangzhou, China, to 16.7% in Cape Town, South Africa. In men, it ranged from 8.5% in Reykjavik, Iceland to 22.2% in Cape Town, South Africa. Using a similar study design, the Latin American Project for the Investigation of Obstructive Lung Diseases (PLATINO) studied the prevalence of COPD in 5,315 study participants in five Latin American centres (Mexico, Venezuela, Brazil, Chile, and Uruguay) (Menezes et al. 2005). Crude prevalence rates of COPD ranged from 7.8% in Mexico City, Mexico, to 19.7% in Montevideo, Uruguay. Age and smoking did not fully explain the international variation in disease prevalence, suggesting a role for additional risk factors.

Future prevalence and burden of COPD

Assessing the future prevalence and burden of COPD requires the consideration of changing population distributions and smoking habits. A recent study (Burgel et al. 2018*)* modelled COPD prevalence by 2025 based on data from France and predicted a fourfold increase in prevalence of more severe COPD with disproportionate increases in women and individuals over the age of 75. Globally, COPD is projected to be the third-leading cause of death in both men and women with an estimated 4.57 million deaths annually, a 42% increase (WHO 2007). While this increase is driven primarily by LMICs where the mortality share of COPD is expected to increase primarily due to increases in smoking, the population mortality rate of COPD is projected to increase in every economic group studied (WHO 2013). Mortality rates are at risk to substantially increase with economic progress and the adoption of Western practices, particularly in Asia, Africa, and the Middle East (Bilano et al. 2015). In addition to mortality, COPD morbidity is on the rise. COPD is projected to be the seventh cause of morbidity worldwide by 2030 (WHO 2007*)*. Analysis from the Global Burden of Diseases, Injuries and Risk Factors 2015 study suggests that sub-Saharan Africa and South and Southeast Asia suffer from the highest disability-adjusted life year burdens of COPD (GBD 2015 Chronic Respiratory Disease Collaborators 2017).

What is most striking about the epidemiology of COPD has been the change in prevalence in women (Aryal et al. 2013). Due to increased smoking rates in women and women taking up smoking at younger ages in recent decades, women now are about equal to men in COPD prevalence with similar mortality (WHO 2007). This trend is expected to be continue in LMICs in response to both increases in smoking, continued biomass exposure, and increasing air pollution. Nonetheless, women with COPD remain less likely to be diagnosed and treated (Miravitlles et al. 2005).

Risk factors for COPD

Cigarette smoking

COPD usually occurs with exposure to an inhalational irritant. In developed countries, this exposure is predominantly cigarette smoke. Cigarette smoke contains thousands of toxic chemicals including nicotine, acetone, and tar. Smoke inhalation stimulates an inflammatory reaction with recruitment of inflammatory cells to the lung that results in loss of lung parenchyma and

airway obstruction. Smoking also leads to impaired mucociliary clearance, increased oxidative stress, and protease/antiprotease imbalance. Second-hand smoke (SHS) exposure increases risk of COPD and negatively impacts lung function by similar mechanisms.

There is a well-established association between current and cumulative smoking and COPD, and adult smokers experience a faster FEV_1 decline than non-smokers. This excess decline in lung function may return to normal levels of ageing-related decline after smoking cessation, emphasizing the importance of early tobacco cessation interventions. Few data are available on the incidence of COPD according to detailed smoking habits. Analysis of the Copenhagen City Heart Study confirmed a higher incidence of COPD in smokers than non-smokers, with 27% of participants who continued to smoke over a 25-year period developing COPD as defined by GOLD stage II–IV COPD, compared with 5.7% of non-smokers (Lokke et al. 2006). The longitudinal Vlagtwedde/Vlaardingen Cohort Study (with a follow-up of 25 years) also showed that persistent cigarette smokers, recidivist smokers, variable pipe/cigar smokers, and subjects who stopped smoking had an increased risk of developing COPD compared to never smokers. The risks of developing COPD in sustained ex-smokers, starters, brief smokers, and persistent pipe/cigar smokers were not significantly different.

Less is known about the relative risk of non-cigarette smoking habits, including cigars, traditional pipes, water pipes, marijuana, and electronic cigarettes and vaporizers. Observational studies of these modalities will be challenging given the heterogeneity of exposure burden across brands and devices; however, it is likely that relative risk is significant given similar immune and inflammatory effects to cigarette smoking. At this time, evidence about safety and risk is insufficient to recommend alternate inhaled substances as a means of tobacco cessation.

Indoor biomass exposures

Although cigarette smoking is the most common inhalational exposure related to COPD, biomass exposure from fuels burned for heating and cooking (WHO 2007) can lead to COPD as well. Such indoor air exposures are important risk factors in low-income countries where biomass fuels (wood and crop residues) and coal fuels are often used for heating and cooking without appropriate ventilation. WHO estimates that approximately 3 billion people, predominantly in LMICs, cook with solid fuels and

kerosene resulting in household air pollution and increased COPD risk (WHO 2007). It is estimated that this exposure results in 25% of premature deaths from COPD in LMICs (WHO 2007). Recently, it has been recognized that poor inner-city and rural areas of wealthier countries have similarly high rates of indoor biomass exposure (Levy et al. 2018). While COPD is also thought to be related to outdoor pollution levels, current evidence is not conclusive regarding the magnitude of this risk (Schikowski et al. 2014).

Occupational exposures

Occupational exposure to dusts, gases, and fumes are also associated with the rate of decline in FEV_1 and the development of COPD. Some examples of occupational risk factors for COPD are coal dust, silica dust, oil mists, welding fumes, and organic dusts including cotton, grain, and wood. It is estimated that 15 to more than 20% of COPD is caused by occupational factors (Lytras et al. 2018). For example, the Third National Health and Nutrition Examination Survey in the United States found that approximately 20% of COPD cases in the United States were attributable to work-related exposures (Hnizdo et al. 2002). The proportion was even higher (approximately 30%) in never smokers. Individuals who have occupational exposures and also smoke have an even greater COPD risk.

HIV

Pulmonary disease has been a cause of significant morbidity and mortality in HIV-infected individuals since the beginning of the AIDS epidemic. COPD and emphysema were reported to be more common in HIV-infected individuals prior to the advent of effective antiretroviral therapy (ART)(Diaz et al. 2000) and prevalence of lung function imparirment has not decreased significantly in recent years. Population studies in the Veterans Administration found that a diagnosis of COPD was more common in HIV-infected individuals than in similar HIV-uninfected individuals in the current era (Crothers et al. 2011), and studies of pulmonary function have demonstrated that HIV is an independent risk factor for impairment of diffusing capacity for carbon monoxide, a lung function abnormality that reflects emphysema/COPD as well as pulmonary vascular disease (Crothers 2013; Fitzpatrick et al. 2013). In addition, over half of outpatient HIV-

infected cohorts report respiratory symptoms (Gingo et al. 2014). Although COPD occurs in HIV-infected individuals who have never smoked, it is still most common in smokers, making smoking cessation critical in this population.

Genetic risk factors

The contribution of genetics to COPD is exemplified by the heritable disorder alpha-1 antitrypsin deficiency (AATD). Individuals with AATD have a mutation in the SERPINA1 gene that results in abnormal structure of or decreases in endogenous levels of alpha-1 antitrypsin, a critical antiprotease whose absence can lead to breakdown of lung tissue and COPD in early life (Piitulainen and Eriksson 1999). The interplay between inhalational exposure and genetics is well-illustrated by the observation not all patients with AATD will develop COPD, and those who smoke experience lung function decline 2–4 times faster than never smokers with AATD (Piitulainen and Eriksson 1999). Similarly, not all individuals with a heavy cigarette smoking burden develop COPD.

Other than the single gene AATD, genetic factors in COPD are likely multifactorial and have been difficult to study. Since COPD is predominantly expressed in later years, the parents of individuals with COPD have often already died, and the children of individuals with COPD are likely to be too young to have significant fixed airway obstruction. Initial genetic studies on COPD included small numbers of subjects and applied different definitions of disease, hampering a valid comparison between studies and increasing the likelihood of spurious results. Many positive results from genetic studies of COPD have therefore not been replicated. Furthermore, lack of replication might also be due to the apparent difficulties of extrapolating results from one population to another, such as Asians and Caucasians.

Most studies on genetics of COPD have focused on genes that are involved in the processing of various tobacco smoke products, oxidative stress, or in the protease/antiprotease balance. In particular, variations in the genes encoding enzymes that detoxify cigarette smoke products (and are thus protective against oxidative stress) may explain why only a proportion of smokers develop COPD. Candidate genes identified in this way include *EPHX1*, *ADAM33*, *MMP1*, *GSTP1*, *GSTT1*, *GSTM1*, *HMOX1*, and *SERPINA1* (Postma et al. 2011).

Genome-wide association (GWA) studies are applied to identify genetic loci that are associated with disease using markers throughout the genome, rather than selected on the basis of their assumed biological function. The first COPD GWA study identified variants in the nicotinic acetylcholine receptor (nAChR) subunit genes, *CHRNA3/5* (Pillai et al. 2009) which were confirmed in later studies (Kim and Lee 2015). SNPs in this nicotinic cluster have previously been related to smoking habits and addiction, raising the question of whether these genes are directly involved in lung function loss leading to COPD *per se*, or act indirectly through effects on nicotine addiction. Other GWA studies have identified *IREB2*, which regulates cellular iron that is associated with pulmonary artery enlargement in COPD, *AGER*, the soluble protein product of which (sRAGE) is a potential biomarker for emphysema presence and progression, and hedgehog interacting protein (*HHIP*) which plays a role in lung development and injury response (Kim and Lee 2015). A recent whole genome sequencing study of patients with severe COPD confirmed findings of variants of HHIP and CHRNA3 genes as influential (Prokopenko et al. 2018). GWA studies have also been complemented and validated by quantitative trait loci (QTL) analyses, which allow incorporation of gene expression, methylation, and many other data types (Morrow et al. 2018). To date, discovery of these genetic variations has not led to novel therapies.

COPD prevention

The mainstay of COPD prevention is reducing exposure to cigarette smoke and to indoor air pollutants. Barriers to prevention efforts include the underdiagnosis of COPD, lack of public awareness of the disease, and inadequate policies for preventing exposures. In LMICs, respiratory disease prevention efforts may be more focused on communicable pulmonary diseases such as tuberculosis and pneumonia. The Centers for Disease Control identified four main goals for COPD prevention (CDC 2011) (Box 8.3.2). The WHO recommends five policies for decreasing tobacco use that would be expected to impact COPD incidence (Box 8.3.2).

Box 8.3.2 Centers for Disease Control goals for COPD prevention

- Surveillance and evaluation: Improving public health data
- Public health research and prevention strategies: Improving understanding of COPD
- Programmes and policies: Improving collaboration among stakeholders
- Communication: Improving awareness of COPD

World Health Organization policies for controlling tobacco use

- Promotion of smoke-free environments
- Support for tobacco cessation programmes
- Health warnings on cigarettes
- Bans on tobacco advertisers
- Higher taxes on tobacco products

From Centers for Disease Control and Prevention. Public Health Strategic Framework for COPD Prevention. Atlanta, GA: Centers for Disease Control and Prevention; 2011. Available at www.cdc.gov/copd; World Health Organization. World Health Statistics 2008.

Public health efforts aimed at reducing smoking rates and encouraging smoking cessation are of great importance. Improved ventilation measures to reduce indoor pollutants in houses decrease both the prevalence and morbidity of COPD (Zhou et al. 2014). For example, a systematic review and meta-analysis of improved biomass cookstoves found decreased COPD and respiratory symptoms in women (Thakur et al. 2018). The greatest benefits in COPD prevention may be in targeting those patients at risk for impaired lung development or rapid lung function decline in the setting of harmful exposures—assuming that they can be identified *a priori*, which remains the topic of ongoing research efforts.

Epidemiology, risk factors, and prevention of asthma

Epidemiology of asthma

Like COPD, asthma is prevalent worldwide with an estimated 300 million people affected (WHO 2007). Its prevalence varies by country (Fig. 8.3.2), and it is increasing worldwide, with exact rates depending on environmental exposures, lifestyle factors, and urbanization (WHO 2007). Given the strong relationship of asthma with allergy, its increasing prevalence mirrors that of allergy. Despite availability of multiple therapeutic agents for asthma, mortality is about 250,000 deaths worldwide each year (WHO 2007). Asthma is also highly morbid, leading to both missed workdays and missed school given its frequency in children. Both morbidity and mortality from

asthma can be linked to lack of access to medical care and adequate therapies.

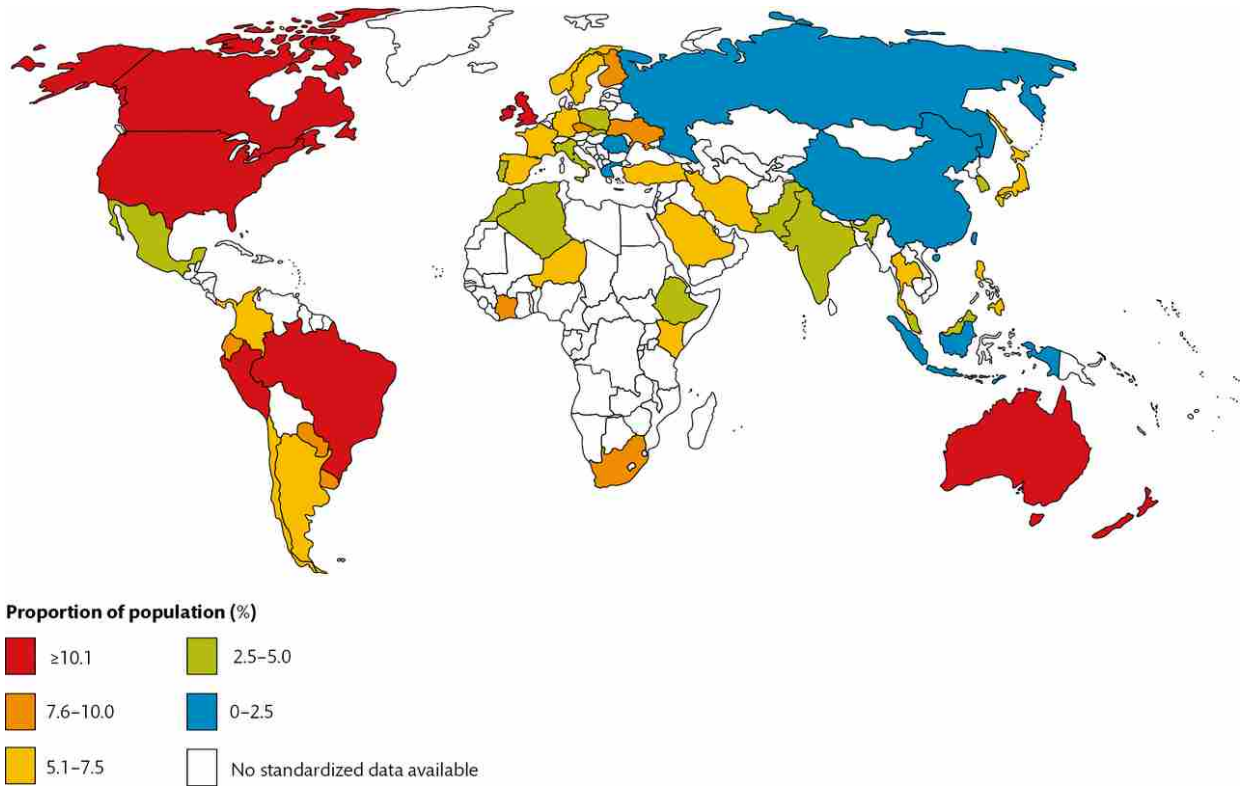


Fig. 8.3.2 World map of the prevalence of clinical asthma.
Reproduced with permission from World Health Organization, Global surveillance, prevention and control of chronic respiratory diseases.
http://apps.who.int/iris/bitstream/handle/10665/43776/9789241563468_eng.pdf

Two multidecade survey studies, the European Community Respiratory Health Survey (ECRHS) and the International Study of Asthma and Allergies in Childhood (ISAAC), started in the early 1990s and provided insight into changing asthma prevalence trends. Both studies showed striking international differences in asthma symptom prevalence with a particularly high prevalence of reported asthma symptoms in English-speaking countries including the British Isles, New Zealand, Australia, the United States, and Canada (Asher et al. 1995; Burney et al. 1996; Asher et al. 1998; Beasley et al. 1998). Among the non-English-speaking European countries, both studies showed high asthma prevalence in Western Europe, with lower prevalences in Eastern and Southern Europe. These studies also provided insights into

asthma in Africa and Asia. These areas generally showed relatively low asthma prevalence (Fig. 8.3.2). In particular, prevalence was low in developing countries such as China and Indonesia whereas more affluent Asian countries such as Singapore and Japan showed relatively high asthma prevalence rates. The ISAAC survey also showed that centres in Latin America had particularly high symptom prevalence. Interestingly, Latin America showed higher prevalences than Spain itself, in contrast to the general tendency for more affluent countries to have higher prevalence rates.

Future prevalence and burden of asthma

Global asthma trends have exhibited either slight increases or plateaus in prevalence. While many theories have been proposed for divergent studies, there is no clear explanation. Countries with similar socioeconomic and genetic compositions have experienced divergent prevalence rate progression (de Korte-de Boer et al. 2015; Backman et al. 2017). The asthma prevalence in many affluent countries has peaked or even begun to decline, whereas asthma symptom prevalence continues to rise in less affluent countries. Changes in prevalence of asthma may result from changes in methods of diagnosis, but changes in environmental risk factors may also play a role.

Risk factors for asthma

Atopy

Atopy, or the presence of IgE antibodies to specific allergens, is strongly associated with asthma within populations. About half of asthma cases are linked to atopy (Arbes et al. 2007), although there is a lower population attributable risk (~20%) in less affluent populations and risk is variable (Weinmayr et al. 2007). IgE levels in serum are associated with a diagnosis of asthma and with presence of airway hyperresponsiveness. Although atopy is a risk factor for asthma, it represents a biological response to various exposures (e.g. allergen exposure) which may be modified by susceptibility factors (genetic and/or environmental).

Allergen exposure

In sensitized asthmatics, allergen exposure can trigger asthma attacks and prolonged exposure can lead to the prolongation and exacerbation of symptoms; however, a causative role of allergen exposure in development of asthma is less clear. The key study linking allergen exposure in infancy to the subsequent development of asthma is that of Sporik et al. (1990) who followed 67 children with a family history of atopy. Subsequent longitudinal birth cohort studies found little or no association between early dust mite allergen exposure and asthma later in childhood (Tepas et al. 2006). A recent systematic review reported that visible mould or mould odour was associated with respiratory symptoms, wheezing, and development of asthma in children (Caillaud et al. 2018).

Gender and ethnicity

There are a variety of demographic factors that are associated with asthma including age, gender, and ethnicity. Asthma incidence and prevalence are consistently lower in females than in males before age 12 years, whereas during adolescence and adulthood there is evidence of higher incidence and prevalence in females (Kimbell-Dunn et al. 1999). The reasons for this disparity are unclear, but could be related to later age of asthma onset in childhood and adolescence in females and differences between boys and girls in exposures, airway size, atopy, and frequency and severity of lower respiratory tract illnesses. These differences could also be due to hormonal influences. Hormones likely play a role as demonstrated by premenstrual asthma which manifests as worsening of asthma symptoms around the time of menstruation and the observation that pregnancy can often increase or decrease asthma control. Ethnicity also appears to influence asthma risk. For example, Puerto Ricans in the US have a high rate of asthma while Mexicans in the US have very low rates, a finding that has been termed the ‘Hispanic paradox’ (Rosser et al. 2014).

Socioeconomic status

The impact of socioeconomic status on asthma is conflicting. Studies in the 1960s and 1970s suggested that asthma is more common in children in the higher socioeconomic status (SES); however, severe asthma appears to be more common in children in lower income groups (Stewart et al. 2001) and in some disadvantaged ethnic groups (Rosser et al. 2014). These findings

could represent either a greater prevalence of asthma in disadvantaged groups, or increased severity due to environmental factors (e.g. SHS, nutrition, occupational exposures) or inadequate disease management and poor access to healthcare.

Obesity

Studies have shown an association between body weight and asthma in both adults and children with prospective studies suggesting that obesity precedes asthma (Ronmark et al. 2005; Noal et al. 2011). However, findings supporting this association have not been entirely consistent (Noal et al. 2011). The evidence linking obesity and atopy is also conflicting with some studies suggesting that obesity is associated with asthma independent of atopy (Visness et al. 2010; Holguin et al. 2013) and with the identification of asthma clusters containing predominantly obese, non-atopic patients (Moore et al. 2010).

Diet

Diet throughout the lifespan may impact asthma risk and severity. Many studies including several infant cohort studies have investigated the effects of breastfeeding on allergies and asthma with some studies showing protective effects, some showing no effect, and others suggesting that breastfeeding is a risk factor (Friedman and Zeiger 2005; Kramer et al. 2007; Matheson et al. 2007). Other nutritional factors may also play a role in the aetiology of asthma. In particular, it has been speculated that the increase in asthma prevalence may be due to a change in dietary patterns in the past few decades, that is, as cultures have become more ‘Westernized’ they have shifted from growing and consuming locally grown foods to consuming more processed foods and less fresh produce. This hypothesis is supported by results from the ISAAC Phase II and other studies (Chatzi and Kogevinas 2009; Nagel et al. 2010) that found that fruit, vegetable, and fish consumption was associated with a lower prevalence of current wheeze and asthma. Development of some atopic disorders has been associated with exposures to n-6 versus n-3 fatty acids, but this finding was not consistent for all atopy/allergies (Almqvist et al. 2007; Notenboom et al. 2011). Maternal folate supplementation during pregnancy has also been suggested as a risk factor for asthma in childhood.

Vitamin D

There has been conflict in studies suggesting an association between vitamin D deficiency and asthma risk. Several studies based on dietary questionnaire findings have associated higher levels of maternal vitamin D with lower rates of asthma in children (Muehleisen and Gallo 2013) and low serum vitamin D levels in asthmatic children and adults with asthma severity and poor asthma control (Korn et al. 2013). A meta-analysis of double-blind, randomized, controlled trials of vitamin D found that vitamin D resulted in decreased rate of asthma exacerbations that required corticosteroid treatment (Joliffe et al. 2017*)*, but not all trials examining vitamin D supplementation have not been successful in decreasing asthma rates or improving outcomes.

Outdoor air pollution

The role of outdoor air pollutants (particulate matter (PM), ozone, nitrogen dioxide (NO_2), and sulphur dioxide) in asthma has been extensively studied (Guarnieri and Balmes 2014). An association between measures of distance to major roads or traffic density and asthma symptoms has been demonstrated in a number of European countries (WHO 2005). Also, a large number of studies have reported associations between direct measurements of air pollution levels and exacerbation of pre-existing asthma, both in children and adults (WHO 2005). Some studies, including a birth cohort study (Brauer et al. 2007), have also suggested that air pollution may cause *new onset* of asthma and allergic disease. In particular, several large prospective studies have suggested a role for ozone (McDonnell et al. 1999; McConnell et al. 2002), although significant associations with some asthma outcomes were also shown for PM2.5, soot, and NO_2 (Brauer et al. 2007).

Indoor air pollution

Little is currently known about the contribution of indoor air pollutants (other than SHS) to the incidence and prevalence of asthma. Nitrogen dioxide from burning fossil fuels has received the most attention while sulphur dioxide from burning sulphur-containing coal or gas, mosquito coil smoke, and formaldehyde from wood preparation have also been considered. Particulates from open or closed wood and coal burning fires have received less attention in HCs (Oluwole et al. 2017), but have been studied in LMICs where very high indoor levels have been encountered. Damp indoor

environments and indoor fungal exposure may also play a role as demonstrated in a large number of studies conducted across many geographical regions (Douwes and Pearce 2003; Karvonen et al. 2018). Further study is needed to determine whether programmes to decrease indoor PM concentrations or humidity levels can prove effective in decreasing asthma prevalence in addition to decreasing respiratory morbidity (Beasley et al. 2015).

Tobacco

Although smoking is a primary risk factor for COPD, it may also influence asthma, both if asthmatic patients smoke or if they are exposed to SHS either before or after birth. The evidence for a role of tobacco smoke in asthma is strongest for increases in severity in children who already have asthma; however, an increasing number of studies also show positive associations between SHS and incidence of asthma (Accordini et al. 2012). Several meta-analyses concluded that household SHS exposure was associated with the incidence of new-onset asthma and incidence of wheezing in younger and older children (Vonk et al. 2003; Burke et al. 2012).

Occupational exposures

Occupational asthma is the most common occupational respiratory disease in HCs. Estimates of the total proportion of adult occupational asthma range from 2 to 15% in the United States, 15% in Japan (Chan-Yeung and Malo 1994), 5% in Spain (Kogevinas et al. 1996), 2–3% in New Zealand (Fishwick et al. 1997), and 2–6% in the United Kingdom (Meredith and Nordman 1996). A study in the United States estimated that new-onset asthma may be caused by occupational exposures in as many as one of six adult patients (Mazurek et al. 2013). More than 250 agents have been identified as causes of occupational asthma. Some of the most common occupational asthmagens include flour/grain dusts, wood dusts, latex allergens, and isocyanates (Schenker et al. 1998).

Infections

The increase in allergy and asthma contrasts strongly with a reduced prevalence of infectious diseases and corresponding decreased exposure to non-infectious micro-organisms. This reduction in exposure to micro-

organisms has been hypothesized as a possible explanation for the allergy and asthma epidemic (Lambrecht and Hammad 2017) and is commonly referred to as the ‘hygiene hypothesis’. The hygiene hypothesis has been prompted by evidence that overcrowding and unhygienic conditions were associated with a lower prevalence of atopy, eczema, hay fever, and asthma. The hygiene hypothesis may not fully explain the current time and geographical trends in asthma prevalence, but the changes associated with Westernization may be contributing to the global increases in asthma susceptibility and prevalence. It remains unclear as to what is the most important factor in protection from allergies and asthma: dose, route (oral–faecal versus airborne), type of infection (viral versus bacterial versus parasitic) or location (upper versus lower airways or digestive tract).

Viruses

Viral infections are common causes of exacerbations of asthma and associated with hospital admissions in both children and adults. In fact, respiratory viral infections are detected in the majority of asthma exacerbations; of these, about 60% are rhinoviruses (Johnston 2007). Viral infections may also be involved in the development of asthma, but the evidence is less clear. Longitudinal studies have shown that both respiratory syncytial virus and rhinovirus infection increases the risk of subsequent recurrent wheezing and asthma in early childhood (Lemanske et al. 2005; Feldman et al. 2015). The mechanisms of viral-induced asthma are poorly understood, but some evidence suggests that neutrophilic inflammation, impaired innate immune responses and T_H17 signalling dysregulation play crucial roles (Johnston 2007; Das et al. 2017).

Bacteria

Exposures to different types of non-pathogenic bacteria may have differential effects on atopy and asthma beginning at birth. Perinatal microbial exposure with vaginal vs. caesarean delivery is associated with decreased likelihood of developing childhood asthma (Kim et al. 2019), but the specific bacteria involved and whether the relationship is causal is not yet known. Additionally, studies in farming and non-farming environments have suggested that diversity of microbial species, rather than exposure to one particular microbial component, may be particularly protective

(Heederik and von Mutius 2012). Consistent with these findings are observations that daycare centres—known to be associated with a lower prevalence of allergies and asthma in attending children—are associated with considerable bacterial diversity (Lee et al. 2007) as have households with daycare attendees, or with cats and dogs (Maier et al. 2010). Live bacterial exposure may not be required as levels of pro-inflammatory microbial agents (such as bacterial endotoxin) have a significant inverse association between indoor endotoxin levels and atopic sensitization, hay fever, and atopic asthma (Stein et al. 2016).

The microbiome

With the advent of next generation sequencing, we are able to identify many more microbes in environments in the human body than previously detected by traditional culture-based methods. The microbiome, i.e. the total microbes and their genetic content in the human body, may impact asthma development and progression (Lynch and Boushey 2016). It is possible that the microbial composition and function of the lung microbiome alter local immune response and contributes to asthma. Enrichment of certain bacteria in the lungs is associated with inflammation, and studies have found differences in lung bacterial communities in individuals with asthma compared to health controls with enrichment of bacteria such as *Prevotella* and *Haemophilus* (Depner et al. 2017; Durack et al. 2017, 2018). The gut microbiome may also impact asthma as the microbial communities in the gut have profound effects on systemic immunity and can lead to tolerance or inflammation. Dysbiosis during infancy and childhood is associated with wheezing, atopy, and development of asthma (Fujimara et al. 2016). While efforts to change the gut microbiome with probiotics have shown some success in eczema and potentially atopy (Wickens et al. 2013; Cabana et al. 2017), similar benefits in asthma have not been demonstrated and no attempts to directly alter the lung microbiome have been performed. Whether therapeutic alterations in the microbiome or direct targeting of regulatory T cell processes may benefit asthma control or prevention requires further study.

Genetic factors

Asthma is multifactorial in origin and influenced by multiple genes and environmental factors. The heritable component of asthma has been variably estimated to be between 35 and 95% (Ober and Yao 2011). It is not inherited in a simple Mendelian fashion, and particular genetic factors may affect one or more aspects of the complex etiological processes potentially involved in asthma including atopic sensitization, BHR, airway inflammation, and innate immunity. Investigating possible genes for these individual etiological factors is also fraught with difficulties, as control of these factors is also multifactorial. Multiple genes may lead to the same phenotype in different individuals or multiple genes contribute to asthma in a single person. Asthma development, exacerbations, and progression may all involve different environmental triggers and genetic factors, further complicating genetic studies.

Despite the considerable heterogeneity observed, a number of genome linkage and GWA studies have identified several chromosomal regions significantly associated with asthma susceptibility. Several genes on the 17q21 locus including the *ORMDL3/GSDMB* genes have been identified in multiple populations as associated with asthma (Moffatt et al. 2007; Moffatt et al. 2010). In many cases, the involvement of the products of these genes in asthma pathophysiology is feasible, as they are associated with a variety of inflammatory or epithelial/cell-contact pathways and include genes for cytokines such as IL-13 and IL-33. As with COPD, the biological plausibility of these associations has also more recently been strengthened with epigenomic QTL studies (Nicodemus-Johnson et al. 2016; Dijk et al. 2018). However, even when incorporating epigenetic factors, these genes make only a small contribution to overall disease risk on an individual basis, and there may be issues due to the use of different asthma classifications in different studies. While recent meta-analyses of GWA studies have leveraged large cohort sizes into discoveries of further novel disease-associated loci (Demenais et al. 2018), some gene variants may only be associated with certain asthma subtypes and will be difficult to identify with present methodology.

Asthma prevention

Risk factors for asthma may be more difficult to modify than those for COPD, and the heterogeneous nature of the disease complicates public health initiatives. As most asthma begins in childhood, the majority of

prevention efforts have been focused on younger children, although asthma can also present in adulthood. Given potential roles for exposures to indoor allergens and chemicals, strategies targeting the home environment may be effective although these programmes have not been proven as primary prevention (Tanno et al. 2017). Secondary prevention efforts are targeted at improving asthma control and include efforts aimed at decreasing environmental triggers and in providing appropriate asthma controller medications. Encouraging parental smoking cessation is clearly important in both primary and secondary prevention as well as in prevention of COPD in the parent. Programmes targeting improvement in training health professionals, establishing treatment guidelines, and improving public awareness may decrease healthcare costs for asthma. Given the links of asthma with the microbiome, use of probiotics may be helpful as a primary or secondary prevention measure, but their efficacy has not been demonstrated in large-scale studies. As with COPD, prevention efforts face different obstacles in LMICs and HCs.

Treatment of COPD and asthma

The initiation and evaluation of treatment for patients with COPD or asthma is complex. Specific recommendations regarding therapy can be found in disease-specific international (GOLD, GINA, etc) and regional guidelines (Agusti et al. 2017, GINA 2018). All patients with chronic respiratory disease benefit from avoidance of inhalational irritants (particularly cigarette smoke), immunization against common respiratory pathogens (i.e. pneumococcus, influenza), and careful evaluation for comorbid conditions. Patients with mild, intermittent symptoms from either asthma or COPD can be managed with short-acting, as-needed bronchodilators that can be taken during or before activities or exposures that provoke respiratory symptoms. Persistent asthma is managed with inhaled corticosteroids as first-line therapy, with long-acting bronchodilators and other medications including low-dose theophylline and leukotriene receptor antagonists often added in stepwise fashion until disease control is achieved. In individuals with poorly controlled asthma despite these therapies, tiotropium, a long-acting muscarinic antagonist, has been shown to modestly improve lung function and to increase the time to severe exacerbation requiring oral steroid use. Anti-immunoglobin E treatment and anti-interleukin 5 treatment can

likewise be considered in individuals with poorly controlled allergic or severe eosinophilic asthma, respectively (GINA 2018). In contrast, daily COPD pharmacotherapy starts with long-acting bronchodilators; inhaled corticosteroids are typically added for recurrent exacerbations. ACO treatment, due to increased disease severity and combined features of persistent airflow obstruction and airway inflammation (Vaz Fragoso et al. 2017), usually includes both long-acting bronchodilators and inhaled steroids. Patients with functional status limitations can benefit significantly from specialized pulmonary rehabilitation programmes, though the benefits are better established in COPD (Mendes et al. 2010; McCarthy et al. 2015) than asthma.

Conclusions

What do these epidemiological findings tell us about the future of asthma and COPD? The prevalence of COPD is still increasing, and future projections indicate that the global burden of COPD will increase even further. Cigarette smoking remains the greatest cause of COPD, and increasing use in the developing world will ensure that tobacco remains one of the most significant threats to global public health throughout the twenty-first century. Other major causes of COPD including occupational and indoor exposures, particularly among the poor in all countries, remain underappreciated and inadequately addressed. The downstream impact of evolving social trends like electronic cigarettes remains unclear, but demands further study. Different genetic susceptibility to lung function decline from inhalational exposures leads differential likelihood of developing COPD. International differences in COPD prevalence are likely to persist, and research into these disparities may help in targeting smoking cessation and other public health measures towards the highest-risk subgroups.

Global asthma rates continue to increase with wide variation between countries and regions, and the main drivers of asthma development remain incompletely understood. Current public health efforts are thus limited to improving asthma control rather than targeted towards disease prevention. Shifting global patterns in exposure, diet, and occupation are all likely involved in altering the interplay between immunity and environment—through the “hygiene hypothesis” or otherwise—and further study is

required to parse these complex interactions to understand how public health interventions can more effectively decrease the global burden of asthma.

REFERENCES

Aaron, S.D., et al. (2018). Under-diagnosis and over-diagnosis of asthma. *American Journal of Respiratory and Critical Care Medicine*, 197(8), 1065–7.

Aaron, S.D., et al. (2017). Re-evaluation of diagnosis in adults with physician-diagnosed asthma. *Journal of the American Medical Association*, **317**, 269–79.

Accordini, S., et al. (2012). The role of smoking in allergy and asthma: lessons from the ECRHS. *Current Allergy and Asthma Reports*, **12**, 185–91.

Agusti, A., et al. (2017). *Global Strategy for Prevention, Diagnosis and Management of COPD: 2018 Report*. Global Initiative for Chronic Obstructive Lung Disease. [Online] Available at: https://goldcopd.org/gold-reports/

Almqvist, C., et al. (2007). Omega-3 and omega-6 fatty acid exposure from early life does not affect atopy and asthma at age 5 years. *Journal of Allergy and Clinical Immunology*, **119**, 1438–44.

American Thoracic Society Committee on Diagnostic Standards (1962). Definitions and classification of chronic bronchitis, asthma and pulmonary emphysema. *American Review of Respiratory Disease*, **85**, 762–8.

Arbes, S.J., Gergen, P.J., Vaughn, B., and Zeldin D.C. (2007). Asthma cases attributable to atopy: results from the Third National Health and Nutrition Examination Survey. *Journal of Allergy and Clinical Immunology*, **120**, 1139–45.

Aryal, S., Diaz-Guzman, E. and Mannino, D.M. (2013). COPD and gender differences: an update. *Translational Research*, **162**, 208–18.

Asher, M.I., et al. (1998). Worldwide variations in the prevalence of asthma symptoms: International Study of Asthma and Allergies in Childhood (ISAAC). *European Respiratory Journal*, **12**, 315–35.

Asher, M.I., et al. (1995). International Study of Asthma and Allergies in Childhood (ISAAC): rationale and methods. *European Respiratory Journal*, **8**, 483–91.

Backman, H., et al. (2017). Increased prevalence of allergic asthma from 1996 to 2006 and further to 2016—results from three population surveys. *Clinical and Experimental Allergy*, **47**, 1426–35.

Beasley, R., et al. (1998). Worldwide variation in prevalence of symptoms of asthma, allergic rhinoconjunctivitis and atopic eczema: ISAAC. *The Lancet*, **351**, 1225–32.

Beasley, R., Semprini, A. and Mitchell, E.A. (2015). Risk factors for asthma: is prevention possible? *The Lancet*, **386**, 1075–85.

Bilano, V., et al. (2015). Global trends and projections for tobacco use, 1990–2025: an analysis of smoking indicators from the WHO Comprehensive Information Systems for Tobacco Control. *The Lancet*, **385**, 966–76.

Brauer, M., et al. (2007). Air pollution and development of asthma, allergy and infections in a birth cohort. *European Respiratory Journal*, **29**, 879–88.

Buist, A.S., et al. (2007). International variation in the prevalence of COPD (the BOLD Study): a population-based prevalence study. *The Lancet*, **370**, 741–50.

Burgel, P.R., et al. (2018) An attempt at modeling COPD epidemiological trends in France. *Respiratory Research*, **19**, 130.

Burke, H., et al. (2012). Prenatal and passive smoke exposure and incidence of asthma and wheeze: systematic review and meta-analysis. *Pediatrics*, **129**, 735–44.

Burney, P., et al. (1996). Variations in the prevalence of respiratory symptoms, self-reported asthma attacks, and use of asthma medication in the European community respiratory health survey (ECRHS). *European Respiratory Journal*, **9**, 687–95.
Cabana, M.D., et al. (2017). Early probiotic supplementation for eczema and asthma prevention: a randomized controlled trial. *Pediatrics*, 140, e20163000.
Caillaud, D., et al. (2018). Indoor mould exposure, asthma, and rhinitis: findings from systemic reviews and recent longitudinal studies. *European Respiratory Reviews*, **27**, 170137.
Calverley, P.M.A., et al. (2018). The effect of defining chronic obstructive pulmonary disease by the lower limit of normal of FEV1/FVC ratio in tiotropium safety and performance in Respimat participants. *Annals of the American Thoracic Society*, **15**, 200–8.
Centers for Disease Control and Prevention (CDC) (2011). Public Health Strategic Framework for COPD Prevention. Atlanta, GA: Centers for Disease Control and Prevention. [Online] Available at: https://www.cdc.gov/copd
Chan-Yeung, M. and Malo, J.-L. (1994). Epidemiology of occupational asthma. In: Busse, W. and Holgate, S.T. (eds.) *Asthma and Rhinitis*. Oxford: Blackwell Scientific, pp. 44–57.
Chatzi, L. and Kogevinas, M. (2009). Prenatal and childhood Mediterranean diet and the development of asthma and allergies in children. *Public Health Nutrition*, **12**, 1629–34.
Ciba Foundation Guest Symposium (1959). Ciba Foundation Guest Symposium (1959). Terminology definitions, classification of chronic pulmonary emphysema and related conditions. *Thorax*, **14**, 286–99.
Colak, Y., et al. (2018). Combined value of exhaled nitric oxide and blood eosinophils in chronic airway disease: the Copenhagen General Population Study. *European Respiratory Journal*, **52**, 1800616.
Crothers, K.A., et al. (2011). HIV infection and risk for incident pulmonary diseases in the combination antiretroviral therapy era. *American Journal of Respiratory and Critical Care Medicine*, **183**, 388–95.
Crothers, K.A. (2013). HIV infection is associated with reduced pulmonary diffusing capacity. *Journal of Acquired Immune Deficiency Syndrome*, **64**, 271–8.
Committee on Diagnostic Standards for Nontuberculous Respiratory Diseases, American Thoracic Society (1962). Definitions and classification of chronic bronchitis, asthma and pulmonary emphysema. *American Review of Respiratory Disease*, **85**, 762–9.
Das, S., et al. (2017). Respiratory syncytial virus infection of newborn CX3CR1-deficient mice induces a pathogenic pulmonary innate immune response. *JCI Insight*, 2, e94605.
Diaz, P.T., et al. (2000). Increased susceptibility to pulmonary emphysema among HIV-seropositive smokers. *Annals of Internal Medicine*, **132**, 369–72.
de Korte-de Boer, D., et al. (2015). Stabilising prevalence trends of eczema, asthma and rhinoconjunctivitis in Dutch schoolchildren (2001–2010). *Allergy*, **70**, 1669–73.
Demenais, F., et al. (2018). Multiancestry association study identifies new asthma risk loci that colocalize with immune-cell enhancer marks. *Nature Genetics*, **50**, 42–53.
Depner, M., et al. (2017). Bacterial microbiota of the upper respiratory tract and childhood asthma. *Journal of Allergy and Clinical Immunology*, **139**, 826–34.
Dijk, F.N., et al. (2018). Genetic regulation of IL1RL1 methylation and IL1RL1-a protein levels in asthma. *European Respiratory Journal*, **51**, 1701377.
Douwes, J. and Pearce, N. (2003). Is indoor mold exposure a risk factor for asthma? *American Journal of Epidemiology*, **158**, 203–6.
Durack, J., et al. (2017). Features of the bronchial microbiome associated with atopy, asthma, and responsiveness to inhaled corticosteroid treatment. *Journal of Allergy and Clinical Immunology*, **140**, 63–75.
Durack, J., et al. (2018). Bacterial biogeography of adult airways in atopic asthma. *Microbiome*, **6**, 104.

Ellul-Micallef, R. (1976). Asthma: a look at the past. *British Journal of Diseases of the Chest*, **70**, 112–16.
Feldman, A.S., et al. (2015). Toward primary prevention of asthma. Reviewing the evidence for early-life respiratory viral infections as modifiable risk factors to prevent childhood asthma. *American Journal of Respiratory and Critical Care Medicine*, **191**, 34–44.
Fishwick, D., et al. (1997). Occupational asthma in New Zealanders: a population based study. *Occupational & Environmental Medicine*, **54**, 301–6.
Fitzpatrick, M., et al. (2013). HIV infection is associated with diffusing capacity impairment in women. *Journal of the Acquired Immune Deficiency Syndrome*, **64**, 284–8.
Friedman, N.J. and Zeiger, R.S. (2005). The role of breast-feeding in the development of allergies and asthma. *Journal of Allergy and Clinical Immunology*, **115**, 1238–48.
Fujimara, K.E., et al. (2016). Neonatal gut microbiota associates with childhood multisensitized atopy and T cell differentiation. *Nature Medicine*, **22**, 1187–91.
Gingo, M.R., et al. (2014). Pulmonary symptoms and diagnoses are associated with HIV in the MACS and WIHS cohorts. *BMC Pulmonary Medicine*, **14**, 75.
Global Burden of Disease 2015 Chronic Respiratory Disease Collaborators (2017). Global, regional, and national deaths, prevalence, disability-adjusted life years, and years lived with disability for chronic obstructive pulmonary disease and asthma, 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015. *Lancet Respiratory Medicine*, **5**, 691–706.
Global Initiative for Asthma (GINA) (2006). *Global Strategy for Asthma Management and Prevention* [Online]. Available at: http://www.ginasthma.org/.
Global Initiative for Asthma (GINA) (2018). *Global Strategy for Asthma Management and Prevention* [Online]. Available at: http://www.ginasthma.org/.
Guarnieri, M. and Balmes, J.R. (2014). Outdoor air pollution an asthma. *The Lancet*, **383**, 1581–92.
Halbert, R.J., et al. (2003). Interpreting COPD prevalence estimates: what is the true burden of disease? *Chest*, **123**, 1684–92.
Halbert, R.J., et al. (2006). Global burden of COPD: systematic review and meta-analysis. *European Respiratory Journal*, **28**, 523–32.
Han, M.K., et al. (2013). GOLD 2011 disease severity classification in COPDGene: a prospective cohort study. *Lancet Respiratory Medicine*, **1**, 43–50.
Heederik, D. and von Mutius, E. (2012). Does diversity of environmental microbial exposure matter for the occurrence of allergy and asthma? *Journal of Allergy and Clinical Immunology*, **130**, 44–50.
Hnizdo, E., et al. (2002). Association between chronic obstructive pulmonary disease and employment by industry and occupation in the US population: a study of data from the Third National Health and Nutrition Examination Survey. *American Journal of Epidemiology*, **156**, 738–46.
Holguin, F., et al. (2013). An association between L-arginine/asymmetric dimethyl arginine balance, obesity, and the age of asthma onset phenotype. *American Journal of Respiratory and Critical Care Medicine*, **187**, 153–9.
Johnston, S.L. (2007). Innate immunity in the pathogenesis of virus-induced asthma exacerbations. *Proceedings of the American Thoracic Society*, **4**, 267–70.
Joliffe, D.A., et al. (2017). Vitamin D supplementation to prevent asthma exacerbations: a systematic review and meta-analysis of individual participant data. *Lancet Respiratory Medicine*, **5**, 881–90.
Karvonen, A.M., et al. (2018). Early age exposure to moisture damage and systemic inflammation at the age of 6 years. *Indoor Air*, **28**, 450–8.
Keeney, E.L. (1964). The history of asthma from Hippocrates to Meltzer. *Journal of Allergy & Clinical Immunology*, **35**, 215–26.

Kim, H., Sitarik, A. R., Woodcroft, K., Johnson, C. C., Zoratti, E. (2019). Birth mode, breastfeeding, pet exposure, and antibiotic use: associations with the gut microbiome and sensitization in children. *Current Allergy and Asthma Reports*, **19**(4):22. doi: 10.1007/s11882-019-0851-9.

Kim, W.J. and Lee, S.D. (2015). Candidate genes for COPD: current evidence and research. *International Journal of Chronic Obstructive Pulmonary Disease*, **10**:2249–55.

Kimbell-Dunn, M., Pearce, N.E., and Beasley, R. (1999). Asthma. In: Hatch, M. (ed.) *Women and Health*. San Diego, CA: Academic Press, pp. 724–39.

Kogevinas, M., et al. (1996). The risk of asthma attributable to occupational exposures—a population-based study in Spain. *American Journal of Respiratory and Critical Care Medicine*, **154**, 137–43.

Korn, S., et al. (2013). Severe and uncontrolled adult asthma is associated with vitamin D insufficiency and deficiency. *Respiratory Research*, **14**, 25.

Kramer, M.S., et al. (2007). Effect of prolonged and exclusive breast feeding on risk of allergy and asthma: cluster randomised trial. *British Medical Journal*, **335**, 815.

Lambrecht, B.N. and Hammad, H. (2017). The immunology of the allergy epidemic and the hygiene hypothesis. *Nature Immunology*, **18**, 1076–83.

Lee, L., Tin, S., and Kelley, S.T. (2007). Culture-independent analysis of bacterial diversity in a child-care facility. *BMC Microbiology*, **7**, 27.

Lemanske, R.F., Jr., et al. (2005). Rhinovirus illnesses during infancy predict subsequent childhood wheezing. *Journal of Allergy and Clinical Immunology*, **116**, 571–7.

Levy, J.I., et al. (2018). Established and emerging environmental contributors to disparities in asthma and chronic obstructive pulmonary disease. *Current Epidemiology Reports*, **5**, 114–24.

Lokke, A., et al. (2006). Developing COPD: a 25 year follow up study of the general population. *Thorax*, **61**, 935–9.

Lynch, S.V. and Boushey, H.A. (2016). The microbiome and development of allergic disease. *Current Opinions in Allergy and Clinical Immunology*, **16**, 165–71.

Lytras, T., et al. (2018). Occupational exposures and 20-year incidence of COPD: the European Community Respiratory Health Survey. *Thorax*, 73(11), 1008–15.

Maier, R.M., et al. (2010). Environmental determinants of and impact on childhood asthma by the bacterial community in household dust. *Applied and Environmental Microbiology*, **76**, 2663–7.

Mannino, D.M., et al. (2003). Lung function and mortality in the United States: data from the First National Health and Nutrition Examination Survey follow up study. *Thorax*, **58**, 388–93.

Matheson, M.C., et al. (2007). Breast-feeding and atopic disease: a cohort study from childhood to middle age. *Journal of Allergy and Clinical Immunology*, **120**, 1051–7.

Mazurek, J.M., et al. (2013). Occupational asthma incidence: findings from the behavioral risk factor surveillance system asthma call-back survey—United States, 2006–2009. *Journal of Asthma*, **50**, 390–4.

McCarthy B., et al. (2015). Pulmonary rehabilitation for chronic obstructive pulmonary disease. *Cochrane Database of Systematic Reviews*, 2, CD003793.

McConnell, R., et al. (2002). Asthma in exercising children exposed to ozone: a cohort study. *The Lancet*, **359**, 386–91.

McDonnell, W.F., et al. (1999). Long-term ambient ozone concentration and the incidence of asthma in nonsmoking adults: the AHSMOG Study. *Environmental Research*, **80**, 110–21.

Mendes, F.A., et al. (2010). Effects of aerobic training on psychosocial morbidity and symptoms in patients with asthma: a randomized clinical trial. *Chest*, **138**, 331–7.

Menezes, A.M., et al. (2005). Chronic obstructive pulmonary disease in five Latin American cities (the PLATINO study): a prevalence study. *The Lancet*, **366**, 1875–81.

Meredith, S. and Nordman, H. (1996). Occupational asthma: measures of frequency from four countries. *Thorax*, **51**, 435–40.

Miller, M.R., et al. (2005). Standardisation of spirometry. *European Respiratory Journal*, **26**, 319–38.
Miravitlles, M., et al. (2005). Characteristics of a population of COPD patients identified from a population-based study. Focus on previous diagnosis and never smokers. *Respiratory Medicine*, **99**, 985–95.
Moffatt, M.F., et al. (2010). A large-scale, consortium-based genomewide association study of asthma. *The New England Journal of Medicine*, **363**, 1211–21.
Moffatt, M.F., et al. (2007). Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. *Nature*, **448**, 470–3.
Moore, W.C., et al. (2010). Identification of asthma phenotypes using cluster analysis in the Severe Asthma Research Program. *American Journal of Respiratory and Critical Care Medicine*, **181**, 315–23.
Morrow, J.D., et al. (2018). Human lung DNA methylation quantitative trait loci colocalize with chronic obstructive pulmonary disease genome-wide association loci. *American Journal of Respiratory and Critical Care Medicine*, **197**, 1275–84.
Muehleisen, B. and Gallo, R.L. (2013). Vitamin D in allergic disease: shedding light on a complex problem. *Journal of Allergy and Clinical Immunology*, **131**, 324–9.
Nagel, G., et al. (2010). Effect of diet on asthma and allergic sensitisation in the International Study on Allergies and Asthma in Childhood (ISAAC) Phase Two. *Thorax*, **65**, 516–22.
Nicodemus-Johnson, J., et al. (2016). DNA methylation in lung cells is associated with asthma endotypes and genetic risk. *JCI Insight*, 1, e90151.
Noal, R.B., et al. (2011). Childhood body mass index and risk of asthma in adolescence: a systematic review. *Obesity Reviews*, **12**, 93–104.
Notenboom, M.L., et al. (2011). Maternal fatty acid status in pregnancy and childhood atopic manifestations: KOALA Birth Cohort Study. *Clinical & Experimental Allergy*, **41**, 407–16.
Ober, C. and Yao, T.C. (2011). The genetics of asthma and allergic disease: a 21st century perspective. *Immunological Reviews*, **242**, 10–30.
Oluwole, O., et al. (2017). Household biomass fuel use, asthma symptoms severity, and asthma underdiagnosis in rural schoolchildren in Nigeria: a cross-sectional observational study. *BMC Pulmonary Medicine*, **17**, 3.
Petty, T.L. (2006). The history of COPD. *International Journal of COPD*, **1**, 3–14.
Piitulainen, E. and Eriksson, S. (1999). Decline in FEV1 related to smoking status in individuals with severe alpha 1-antitrypsin deficiency (PiZZ). *European Respiratory Journal*, **13**, 247–51.
Pillai, S.G., et al. (2009). A genome-wide association study in chronic obstructive pulmonary disease (COPD): identification of two major susceptibility loci. *PLoS Genetics*, 5, e1000421.
Postma, D.S., et al. (2011). Asthma and chronic obstructive pulmonary disease: common genes, common environments? *American Journal of Respiratory and Critical Care Medicine*, **183**, 1588–94.
Prokopenko D., et al. (2018) Whole genome sequencing in severe chronic obstructive pulmonary disease. *American Journal of Respiratory Cell and Molecular Biology*, 59(5), 614–22.
Ronmark, E., et al. (2005). Obesity increases the risk of incident asthma among adults. *European Respiratory Journal*, **25**, 282–8.
Rosser, F.J., Forno, E., Cooper, P.J., and Celedón, J.C. (2014). Asthma in Hispanics. An 8-year update. *American Journal of Respiratory and Critical Care Medicine.* 189, 1316–27.
Schenker, M.B., et al. (1998). Respiratory health hazards in agriculture. *American Journal of Respiratory and Critical Care Medicine*, 158, S1–S76.
Schikowski, T., Mills, I.C., and Anderson, H.R. (2014). Ambient air pollution—a cause for COPD? *European Respiratory Journal*, **43**, 250–63.
Sporik, R., et al. (1990). Exposure to house-dust mite allergen (Der-P-I) and the development of asthma in childhood—a prospective-study. *The New England Journal of Medicine*, **323**, 502–7.

Stein M.M., et al. (2016). Innate immunity and asthma risk in Amish and Hutterite farm children. *New England Journal of Medicine*, **375**, 411–21.
Stewart, A.W., et al. (2001). The relationship of per capita gross national product to the prevalence of symptoms of asthma and other atopic diseases in children (ISAAC). *International Journal of Epidemiology*, **30**, 173–9.
Tanno, L.K., et al. (2017). Implementation gaps for asthma prevention and control. *Respiratory Medicine*, **130**, 13–19.
Tepas, E.C., et al. (2006). Sensitization to aeroallergens and airway hyperresponsiveness at 7 years of age. *Chest*, **129**, 1500–8.
Tilert, T., et al. (2013). Estimating the U.S. prevalence of chronic obstructive pulmonary disease using pre- and post-bronchodilator spirometry: the National Health and Nutrition Examination Survey (NHANES) 2007–2010. *Respiratory Research*, **14**, 103.
Thakur M., et al. (2018). Impact of improved cookstoves on women's and child health in low and middle income countries: a systematic review and meta-analysis. *Thorax*, 73(11), 1026–40.
Tran, T.N., et al. (2016). Overlap of atopic, eosinophilic, and TH2-high asthma phenotypes in a general population with current asthma. *Annals of Allergy Asthma and Immunology*, **116**, 37–42.
Vaz Fragoso, C.A., et al. (2016). Phenotype of spirometric impairment in an ageing population. *American Journal of Respiratory and Critical Care Medicine*, **193**, 727–35.
Vaz Fragoso, C.A., et al. (2017). Asthma-COPD overlap syndrome in the U.S.: a prospective population-based analysis of patient-reported outcomes and health care utilization. *International Journal of Chronic Obstructive Pulmonary Disease*, **12**, 517–27.
Visness, C.M., et al. (2010). Association of childhood obesity with atopic and nonatopic asthma: results from the National Health and Nutrition Examination Survey 1999–2006. *Journal of Asthma*, **47**, 822–9.
Vonk, J.M., et al. (2003). Risk factors associated with the presence of irreversible airflow limitation and reduced transfer coefficient in patients with asthma after 26 years of follow up. *Thorax*, **58**, 322–7.
Weinmayr, G., et al. (2007). Atopic sensitization and the international variation of asthma symptom prevalence in children. *American Journal of Respiratory and Critical Care Medicine*, **176**, 565–74.
Wenzel, S.E. (2012). Asthma phenotypes: the evolution from clinical to molecular approaches. *Nature Medicine*, **18**, 716–25.
Wickens, K., et al. (2013). Early supplementation with Lactobacillus rhamnosus HN001 reduces eczema prevalence to 6 years: does it also reduce atopic sensitization? *Clinical and Experimental Allergy*, **43**, 1048–57.
Willis, T. (1678). *Practice of Physick, Pharmaceutice Rationalis or the Operations of Medicine in Humane Bodies*. London: Dring, Harper and Leigh.
Woodruff, P.G., et al. (2017). American Thoracic Society/National Heart, Lung, and Blood Institute asthma-chronic obstructive pulmonary disease overlap workshop report. *American Journal of Respiratory and Critical Care Medicine*, **196**, 375–81.
World Health Organization (WHO) (2005). *Air Quality Guidelines, Global Update 2005. Particulate Matter, Ozone, Nitrogen Dioxide and Sulphur Dioxide*. Copenhagen: WHO Regional Office for Europe.
World Health Organization (WHO) (2007). Global surveillance, prevention and control of chronic respiratory diseases. [Online] Available at http://apps.who.int/iris/bitstream/handle/10665/43776/9789241563468_eng.pdf
World Health Organization (WHO) (2013). Global health estimates summary tables. [Online] Available at: http://www.who.int/healthinfo/global_burden_disease/projections/en/
Zhou, Y., et al. (2014). Lung function and incidence of chronic obstructive pulmonary disease after improved cooking fuels and kitchen ventilation: a 9-year prospective cohort study. *PLoS Medicine*, 11, e1001621.

8.4

Obesity

Anna Peeters and Tim Lobstein

Introduction to obesity

Globally, the numbers of adults experiencing obesity has increased over sixfold since 1975, from some 100 million to around 670 million in 2016 (Anonymous 2017). In the same period the numbers of children and adolescents living with obesity have risen more than tenfold (from 11m to 125m). While viewed as a condition affecting developed economies, over 15%, some 105m of adults with obesity, live in China and India. Over 20% of children with obesity live in China alone.

Many middle-income countries are now experiencing a nutritional "double burden" of obesity and undernourishment (stunting, underweight, and/or micronutrient deficiencies) simultaneously. These trends are of major concern for population health and healthcare costs and capacity around the world, with obesity increasing the risk of a range of chronic diseases, disability, and early death. There is increasing international agreement on the need to take action to improve population diets and body weight, and a number of countries are now implementing obesity prevention policies such as warnings on food labelling and taxes on sugary drinks. However, further action at global political as well as local community levels are needed to avoid the projected future burden on health, well-being, and productivity.

Definitions of obesity

Commonly used measures for population health

The World Health Organization (WHO) defines obesity as 'abnormal or excessive fat accumulation that may impair health' (WHO 2000). While a number of methods measure an individual's body fat percentage with a high

degree of accuracy, such as doubly labelled water, dual-energy X-ray absorptiometry, and magnetic resonance imaging, the cost of these methods usually prohibits their use in population surveys (Peeters et al. 2016). Instead, such surveys usually rely on anthropometric markers which are simple and reproducible to measure and have a high correlation with percentage body fat. The primary anthropometric measures are weight, height, and waist and hip circumferences. Most commonly, weight and height are recorded and combined to create an index of body mass (the BMI), calculated as weight (in kilograms) divided by the square of height (in metres).

While body mass index (BMI) is valuable for estimating risk of obesity in population groups and for showing trends across time or disparities between groups, BMI alone should not be relied on for diagnostic purposes for individuals: high levels of muscle mass may raise BMI without indicating a risk of excess adiposity, while low levels of muscle mass may put someone at risk of excess adiposity within the 'normal' BMI range.

Adults

Obesity was originally defined by analysis undertaken by the US Metropolitan Life Insurance Company relating BMI values to mortality (James 1976). A BMI of approximately 25 kg/m^2 was identified as the cut-off above which mortality rates increased, with a BMI of 30 kg/m^2 at 20% above this upper 'normal' BMI limit of 25 kg/m^2. In 1995, WHO accepted BMI as the method for assessing degrees of underweight and overweight and took a BMI lower limit of 18.5 kg/m^2 for distinguishing normal from underweight (Table 8.4.1) (WHO 1995).

However, it is recognized that all common epidemiological measures of obesity at a population level have limitations. From a public health perspective, the precise choice of cut-off point is arbitrary: the health hazards are progressive, age dependent, and not substantially changed at any particular cut-off point. The classic comorbidities intrinsically linked to an excess BMI, for example, diabetes, hypertension, gall stones, and coronary heart disease are linearly related to BMI from a BMI nadir of about 19 or 20 kg/m^2.

While anthropometric measures such as weight and height provide simpler, less expensive, assessment of body composition than methods, such as imaging and bioelectrical impedance analysis, they do not provide a direct measure of body fatness, and as such which method is best has been much debated (Tanamas et al. 2016). Furthermore, for the same BMI, Europids generally have

less body fat than Asians, but more than Africans or Pacific Islanders, which has led to suggestions that a lower BMI threshold should be used to define overweight and obesity in Asians, or that further subcategories of BMI should trigger public health action in different populations (WHO Expert Consultation 2004).

Furthermore, as people age, fat mass tends to increase while muscle mass decreases. As a consequence, BMI may remain stable as people age despite an increase in body fat, particularly among the elderly. Sometimes referred to as sarcopenic obesity (Stenholm et al. 2008) the implication is that lower thresholds of BMI should trigger interventions for obesity prevention and treatment in older populations.

Various alternative or additional measures have been advocated to improve estimates of adiposity and/or to provide better prediction of ill-health and risk of chronic diseases. These include waist circumference (WC), waist–hip ratio (WHR), waist-height ratio, neck circumference, conicity index, and body adiposity index. Most of these involve computation of convenience-ratios, which do not correspond to biological interpretation. They are also hard to visualize, a particular obstacle to public understanding in applications for health promotion. Where the range of body fat is very large (as in most population surveys), WC correlates surprisingly strongly with total body fat, explaining 70.4–77.8% of variance, so it is a little better than BMI, at least in men.(Lean et al. 1996) It is also a slightly better predictor of metabolic risk and chronic diseases than BMI, partly by better identifying individuals at increased health risk through greater total body fat, and partly by also including those with central adiposity who are not identified by BMI (Lean et al. 1995). The simplicity of BMI remains its major strength, resulting in recommendation by the WHO that ‘BMI provides the most useful population-level measure of overweight and obesity’ (WHO 2011). However, recent data has indicated that the overlap between those populations categorized as high risk adiposity by WC and BMI are increasingly differing over time, suggesting further consideration may be needed of new, combined measures (Gearon et al. 2018).

Table 8.4.1 provides the international cut-points for BMI and WC in adults. While there are known ethnic differences in level of risk at different cut-points, the WHO recommends all countries report against these cut-points to allow for international comparisons (WHO Expert Consultation 2004).

Table 8.4.1 Body mass index and waist cut-points in adults

	International cut-points*		Additional BMI cut-points for reporting and action
	Waist circumference	BMI	BMI
Underweight		<18.5 kg m^{-2}	<16 kg m^{-2}
			≥16 to <17 kg m^{-2}
			≥17 to <18.5 kg m^{-2}
Normal weight	<80 cm (women)	18.5 to <25 kg m^{-2}	≥18.5 to <23 kg m^{-2}
	<94 cm (men)		≥23 to <25 kg m^{-2}
Overweight	80 to <88 cm (women)	25 to <30 kg m^{-2}	≥25 to <27.5 kg m^{-2}
	94 to <102 cm (men)		≥27.5 to <30 kg m^{-2}
Obese	≥88cm (women)	≥30 kg m^{-2}	
	≥102 cm (men)		
Obese Class I			≥30 to <32.5 kg m^{-2}
			≥32.5 to <35 kg m^{-2}
Obese Class II			≥35 to <37.5 kg m^{-2}
			≥37.5 to <40 kg m^{-2}
Obese Class III			≥40 kg m^{-2}

BMI, body mass index (an indicator of weight for height (kg m^{-2}).

* International cut-points were derived from Europid populations. The WHO recommends use of international cut-points by all countries, and additional use of ethnic-specific cut-points where available.

Source: data from WHO Expert Consultation 2004.

Children

Classifying overweight and obesity in children is more difficult that with adults due to their rapid growth and body changes as they age. Classification systems

based on national population surveys have been developed and are in use in several countries. The International Obesity Task Force (IOTF) published a multinationally based classification system for children in 2000, and extended in 2012, which has been in widespread use since (Cole et al. 2000; Cole and Lobstein 2012). More recently the WHO developed a set of classification standards for children under 5 years, and a further set of classification reference values for older children (WHO 2018c, WHO 2018d). The two international classification systems show broad agreement but the prevalence of overweight and obesity in a given population can differ somewhat according to age and sex across these classification systems. It is recommended that estimates of excess adiposity in child populations should be reported using both IOTF and WHO classification systems to allow backward comparability and estimates of trends.

Obesity as a disease

Over the past five years a number of national medical associations have identified a need to classify obesity as a disease. In the WHO's expert report *Obesity: Preventing and Managing the Global Epidemic* (WHO 2000) obesity is described as a chronic disease, but the description has been challenged, not least because it would classify millions of adults and children as ill. In 2013, the American Medical Association published its support for describing obesity as a disease, arguing that while such a declaration would have no legal implications, it would pave the way for those with obesity to be better treated within the American medical system.

This position statement reflected the growing opinion globally that people with obesity are not receiving the timely, affordable, evidence-based treatment that would improve their health and quality of life. While treatment options and pathways of care exist, it has been shown that many people with obesity do not receive adequate treatment and management due to reasons ranging from discrimination by health professionals to inadequate health system funding (All-Party Parliamentary Group on Obesity 2018).

The rationale for identifying obesity as a disease includes the potential to drive broader healthcare coverage for both the prevention and treatment of obesity, in part through increasing recognition of the biologically irreversible nature of weight gain.

In 2017, the World Obesity Federation released a position statement supporting the identification of obesity as a chronic relapsing progressive disease process (Bray et al. 2017). This position statement argued that 'obesity fits the epidemiological model of a disease process except that the toxic or

pathological agent is food rather than a microbe'. As such they argued for improved treatment and management globally due to the continuous effort needed to control the relapsing disease process of obesity.

At the same time the World Obesity Federation published an accompanying article acknowledging that declaring obesity to be a disease could have a variety of benefits (Lobstein et al. 2017). It argued for the beneficial effects for people living with obesity who require access to medical advice and support, and that declaring obesity a disease would strengthen the argument for dealing with the social determinants, obesogenic environments, and systemic causes of individual weight gain. Furthermore, it may help to counter the commonly held assumption that obesity was a 'lifestyle choice' and entirely an individual's responsibility.

As more health organizations around the world declare obesity as a disease, it will be important to monitor whether this action supports the achievement of better prevention, management, and treatment of overweight and obesity.

Global trends in overweight and obesity

Trends in the prevalence in obesity

As at 2016, some 390 million women and 281 million men were living with obesity globally, and a further 1.3 billion adults were overweight, non-obese (BMI 25–30 kg/m^2) (NCD Risk Factor Collaboration 2017). Among children and youth (ages 5–19y) some 50 million girls and 74 million boys were living with obesity, and a further 213 million overweight, non-obese youth (Box 8.4.1).

Box 8.4.1 Estimates of global obesity prevalence

The figures cited in this chapter are based on the World Health Organization's Global Health Observatory database (see http://www.who.int/gho/ncd/risk_factors/overweight/en/), provided by the NCD Risk Factor Collaboration (see http://ncdrisc.org/index.html). The methods used to make national and regional estimates include interpolations between studies undertaken at different times and cross-matching countries for their demographic similarities. Adjustments are made for survey data which rely on self-reported weights and heights. Classification of BMI in childhood is based on the WHO-recommended cut-offs. Similar methods have been used for the Global Burden of Disease analyses (see http://www.healthdata.org/gbd) although in this case the classification of BMI in childhood is based on the IOTF cut-offs.

When did obesity become a major public health concern? If we look back through the literature we find Kannel and Gordon in the USA (Kannel and Gordon 1974) suggesting in 1974, ‘There is a massive preoccupation with obesity in affluent countries and both the afflicted and their physicians have been engaged in frustrating efforts to correct it’. In the same year, Garrow in the UK (Garrow 1974) stated ‘Obesity is by far the most common and serious nutritional disease in Britain and other affluent societies’.

In that year obesity (as currently defined) was affecting around 13% of the adult population in the USA, and under 7% of adults in the UK. In the next four decades obesity rates in those countries more than trebled, and by 2016 the disease was affecting over 36% of adults in the USA and 28% in the UK, along with 29% in Australia, 27% in South Africa, 29% in Mexico and 35% in Egypt.

These latter two figures are perhaps the most alarming, as the rate of increase in several middle-income economies has been far steeper than that in high-income countries. As Figure 8.4.1 shows (NCD Risk Factor Collaboration 2017) traditionally affluent countries have shown a gradual rise in obesity prevalence from the early 1970s, whereas in several middle-income countries the rise in adult obesity prevalence is more recent, and more dramatic, and is taking place in a context where weight management services may be less accessible and healthcare pathways less developed.

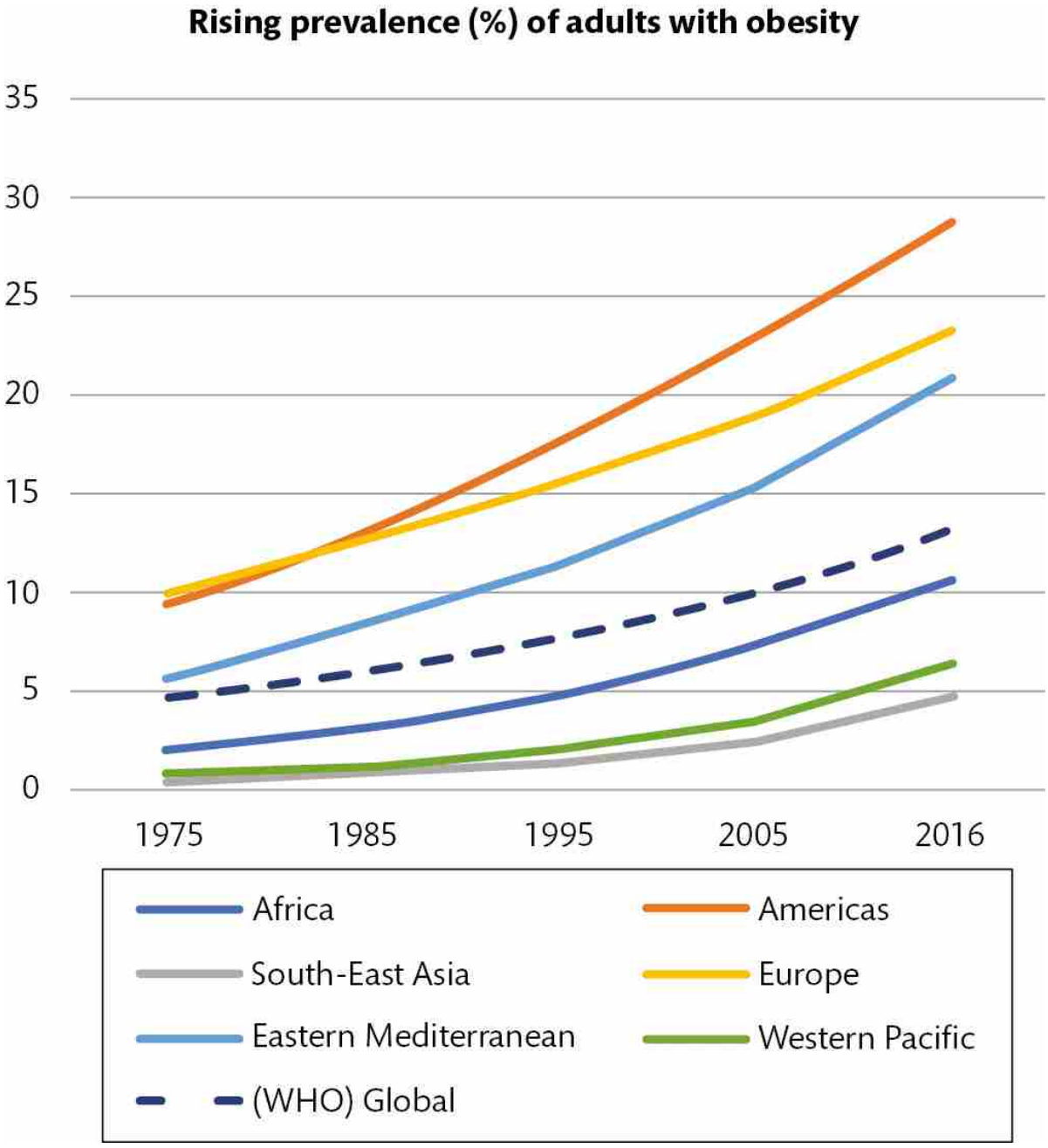


Fig. 8.4.1 Trends in adult obesity prevalence 1975–2016.
Source: data from N.C.D. Risk Factor Collaboration. Worldwide trends in body-mass index, underweight, overweight, and obesity from 1975 to 2016: a pooled analysis of 2416 population-based measurement studies in 128.9 million children, adolescents, and adults. *Lancet* 2017; 390: 2627–42.

As the principle drivers of obesity—including sedentary behaviour, urbanization, and unhealthy and ultraprocessed food consumption—become more globalized, so the rise in obesity prevalence can be seen in populations exposed to these drivers. Estimates undertaken for the Global Burden of

Disease study indicate that in the last decade the countries experiencing the most rapid increase in obesity prevalence are the emerging markets of SE Asia, Latin America, and parts of Africa (Table 8.4.2).

Table 8.4.2 Top 20 countries with the highest rate of increase in adult obesity prevalence 2010–2014

	% increase in 4 years
Myanmar	29.3
Uganda	26.7
Cambodia	26.1
Lao People's Democratic Republic	26.0
Burundi	24.4
Viet Nam	23.5
Sri Lanka	22.9
Malawi	22.6
Rwanda	22.6
United Republic of Tanzania	21.4
Republic of Korea	21.4
Indonesia	20.9
Timor-Leste	20.6
Zambia	20.4
Ethiopia	20.0
Kenya	19.5
Mozambique	18.6
Liberia	18.3
Madagascar	18.0
Sierra Leone	17.7

Source: data from WHO Global Health Observatory.

Childhood obesity

Although obesity is still uncommon among children, its prevalence has shown a similar pattern of increase across countries to that of adults (NCD Risk Factor Collaboration 2017) We have noted that there are several methods for defining obesity in children, and while direct comparisons between surveys may not always be robust, trends using the same definitions are strongly indicative of the rising scale of the problem.

As with adults, long-term increases in childhood obesity can be seen for most high-income countries, and more rapid increases found in low- and middle-income economies. Several regions are of particular concern: the countries of Latin America, from Mexico southwards, have shown a doubling of childhood obesity levels in just 20 years, while adolescents in the Middle East from Egypt through to the Gulf are showing obesity prevalence levels around 15–20%, and some Pacific islands are experiencing adolescent obesity prevalence above 30% (Table 8.4.3).

Table 8.4.3 Countries with highest adolescent obesity prevalence 2016: adolescents aged 10–19 years

	2016
Nauru	31.7
Cook Islands	30.3
Palau	29.4
Niue	27.6
Tuvalu	25.3
Tonga	24.9
Marshall Islands	24.4
Kuwait	22.8
Kiribati	20.9
United States of America	20.7
Samoa	19.9
Micronesia (Federated States of)	18.7
Qatar	17.6
Saudi Arabia	16.7

Source: data from WHO Global Health Observatory.

Age and social disparities

While variations in prevalence levels between countries have been high, and can partly be explained by a country's stage of economic development, disparities between population subgroups within countries reveal additional factors of public health concern.

The clearest variation across population group is by age (Fig. 8.4.2). Obesity prevalence increases by age in virtually all countries studied, with a decline in prevalence shown only in the age groups above 65 years in most high-income countries. This accumulation through the life course reflects the difficulty in losing weight once gained, along with the rising prevalence in all age groups from one decade to the next. It means that national averages can mask a more serious crisis: in the UK the average obesity prevalence level is 26% but among

adults aged 55–64y it is 34% (NHS Digital 2017a) and therefore, all else being equal, over a third of adults can expect to become obese during their lives. In the USA the figure for older adults rises above 45% for the general population, and over 57% for non-Hispanic black women.

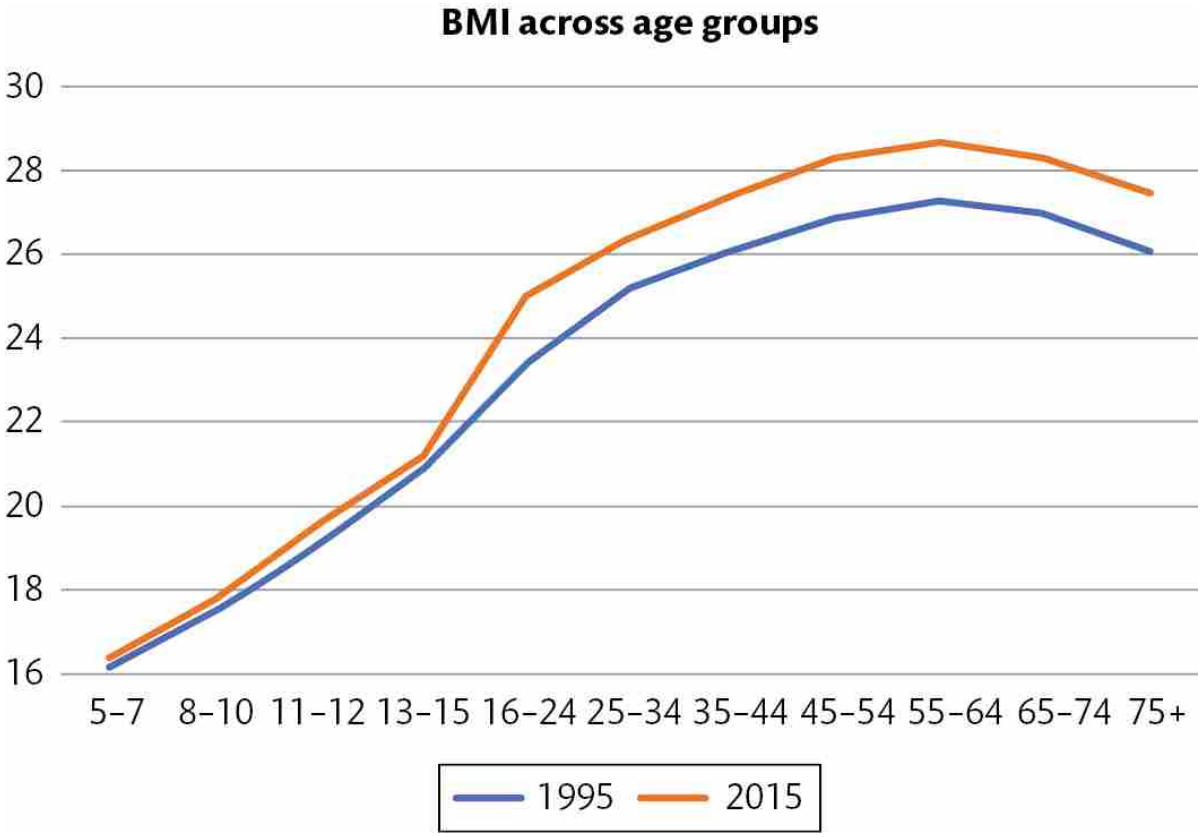
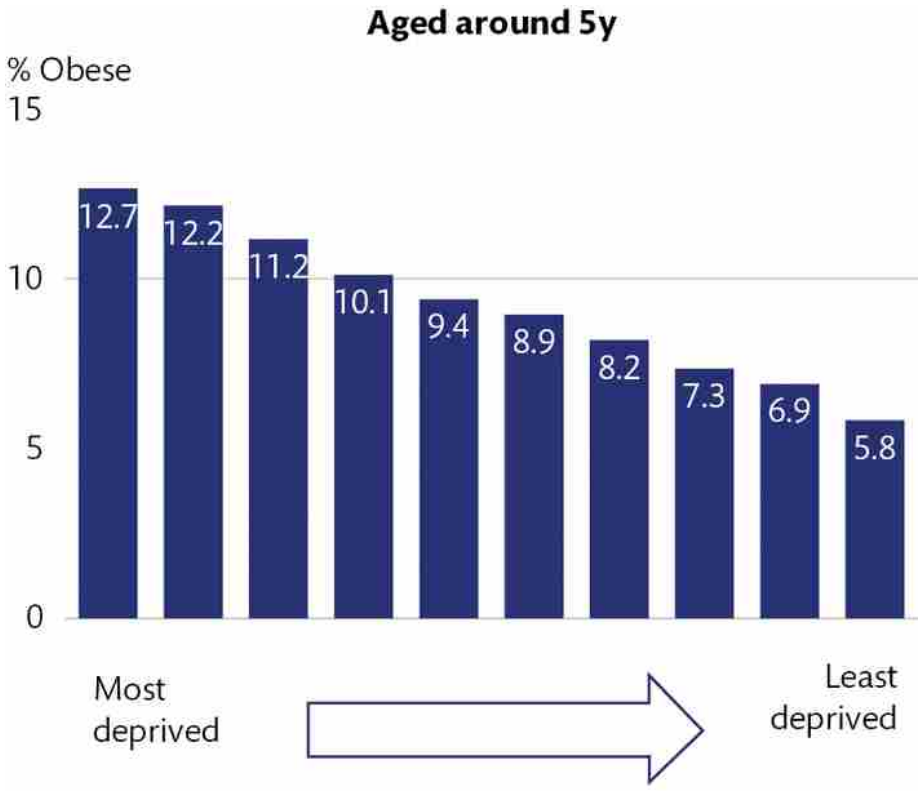


Fig. 8.4.2 Increasing BMI by age to late middle-age in the UK.
Source: data from NHS Digital (2017), 'Health Survey for England 2016: Adult Overweight and Obesity', (London: Health and Social Care Information Centre).

Note that these data reflect cross-sectional surveys in the years shown, and therefore the decline in average BMI after age 60y could be due to several effects apart from an early demise of those people suffering highest levels of overweight. For example, it could reflect the perpetuation of lower BMI among previous generations persisting through each decade, or a specific cohort effect for those who were growing up in the 1920s or 1940s. Equally, it could reflect a significant change in circumstances, for example, moving from earned income to a pension-based income, or to reduced muscle mass rather than reduced

adiposity. The decline applies to both genders, so it is unlikely to be due to a difference in gender survival rates.

Differences in obesity prevalence according to socioeconomic status (measured by level of educational achievement or household income) can be seen in many countries: in high-income countries the risk of developing obesity tends to be higher in less affluent or less educated population groups, with the gradient from highest to lowest socioeconomic (SES) group tending to be most clearly seen for women and for children, and less so for men (McLaren 2007). The data for children are very clearly shown in the UK's National Child Measurement Programme, a survey measuring virtually every child in the country when they enter and leave primary school (aged around 5y and 11y, respectively) (Fig. 8.4.3) (NHS Digital 2017b). Classified by levels of deprivation in the school environs, the data show a clear near-uniform decrease in obesity for each decile decrease in deprivation score.


Aged around 5y
% Obese
15
10
5
0
12.7
12.2
11.2
10.1
9.4
8.9
8.2
7.3
6.9
5.8
Most deprived
Least deprived

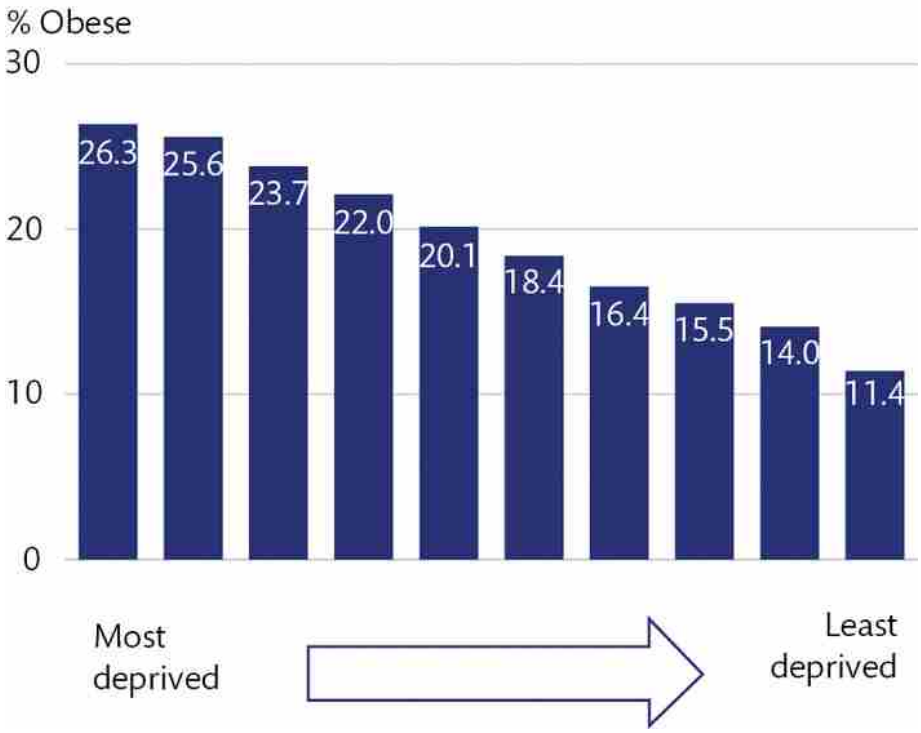

Aged around 11y
% Obese
30
20
10
0
26.3
25.6
23.7
22.0
20.1
18.4
16.4
15.5
14.0
11.4
Most deprived
Least deprived

Fig. 8.4.3 UK National Child Measurement Programme data on obesity according to decile of index of deprivation. Obesity defined by UK growth reference cut-offs.

Source: data from NHS Digital. National Child Measurement Programme—England, 2016–17. London: Health and Social Care Information Centre 2017.

In low income countries, the gradient tends in the opposite direction, with wealthier or better educated families experiencing a higher risk of developing obesity, possibly in part to their greater exposure to labour-saving, urbanized, processed food environments. This gradient has been shown to reverse as countries progress through the nutrition and epidemiologic transition. The classic papers by Monteiro et al. in 2004 (Monteiro et al. 2004a, 2004b) demonstrate the 'cross-over point' to be at around ($2,500 national per capita GDP at the time).

Other subpopulation groups may show disparities after adjusting for SES. Background racial or ethnic status can determine obesity risk status, and this may interact with gender. In the USA, obesity prevalence is lower for Hispanic males and higher for Hispanic females than it is for the overall population, while for groups of Asian origin prevalence is generally significantly lower than the average, but for black or African-American population groups obesity risk is higher than average for both men and especially women, with the obesity affecting more than 50% of women in this group (Ogden et al. 2014).

Significant disparities are found between ethnic groups in many other countries. In New Zealand, higher than average rates of obesity are found among Maori and especially Pacific Islander groups (Ministry of Health—Manatū Hauora) and the same is the case for Australia's indigenous Aboriginal and Torres Islander peoples (Australian Bureau of Statistics 2018).

While increasing urbanization appears linked to obesity prevalence generally, in some countries—e.g. the USA and Australia—residence in rural or periurban areas confers a greater risk of obesity than residence in inner-urban areas. It is not clear if this holds true after SES differences are taken into account. It should also be noted that there are other population groups likely to be at increased risk of obesity, including recent immigrant populations and refugees dependent on food aid (Grijalva-Eternod et al. 2012; Wandell 2013).

Lastly, nations differ in many respects, but one which opens interesting questions for public health policy is the degree of inequality manifest in a society. Measurements of disparities in wealth, e.g. using the Gini Index, have been found to have a close relationship with various measures of health status in the population, including obesity in adults and children (Pickett et al. 2005; Knai et al. 2012). Thus countries may differ in obesity prevalence in

association with their differences in wealth distribution despite having similar average per capita GDP.

Future prospects

Projecting future prevalence rates from past trends is fraught with difficulty. Past trends differ considerably between countries as we have shown. Changes in prevalence are non-linear, and presumably are self-limiting at some maximum level below 100%, although this may differ between countries and subpopulation groups. Modellers must attempt to distinguish age, cohort, and period effects alongside subgroup differences and changing demographics. While past trends can be helpful to predict future prospects, local and national responses to the obesity crisis may lead to policies which alter the trajectory in one direction, while economic crises, changing market dynamics, trade policies fuel prices, or the impact of global warming on food supplies may alter the trajectories in another.

Despite the problems, some attempt at predicting future trends over the coming decade are helpful, not least for authorities responsible for healthcare services facing rising demand for treatment for the chronically disabling diseases consequent on obesity. Interventions to treat obesity rather than waiting to treat its consequences can be cost-saving, and although guidelines vary between countries, a threshold for intervention most commonly occurs at BMI levels above 35 kg/m^2. Crude projections of the numbers of adults with BMI above 35 kg/m^2 have been made in Table 8.4.4 for some of the world's larger countries.

Table 8.4.4 Top 20 countries for numbers of adults projected to be living with severe obesity (BMI>35kg/m^2)

Millions of adults	2010	2025
Global	73.73	177.62
United States	16.57	25.48
China	1.42	16.90
Brazil	4.62	9.49
Mexico	4.43	8.09
Russian Federation	5.18	7.06
Egypt, Arab Rep.	2.94	5.73
Turkey	3.01	5.35
Iran, Islamic Rep.	2.78	4.79
United Kingdom	2.64	4.26
Nigeria	0.69	3.97
India	0.08	3.76
South Africa	1.74	3.43
France	2.21	3.39
Germany	2.33	3.25
Saudi Arabia	1.37	2.71
Argentina	1.43	2.60
Italy	1.78	2.55
Canada	1.50	2.34
Spain	1.67	2.33
Colombia	1.16	2.32

Source: data from CD Risk Factor Collaboration. Trends in adult body-mass index in 200 countries from 1975 to 2014: a pooled analysis of 1698 population-based measurement studies with 19.2 million participants. *The Lancet* 2016; 387: 1377–96.

The social and health costs of overweight and obesity

Language of ‘cost’

Before considering the costs of obesity we need to be clear what we mean. Firstly, there is a move to resist the phrase ‘cost of obesity’ (or ‘burden of obesity’) if it is used in the context of weight-biased language that implies people living with obesity are the cause of social and economic costs to the wider population. In the present discussion we use ‘direct costs of obesity’ to refer to the costs of either treating the disease, or treating the consequences of obesity such as diabetes, liver disease and cardiovascular disease; and ‘indirect costs of obesity’ to refer to the economic losses due to ill-health or premature death. There are also consequential costs to individuals and their families in terms of lost educational opportunity and lowered quality of life.

Adverse health outcomes

Obesity brings a significant change in body function, with alterations in hormonal activity involving insulin, growth hormone, prostaglandins, androgen and oestrogen, and other steroid hormones, changes in joint and muscular physiology, along with raised risks of liver damage, adverse cardiovascular function, poor control of blood sugar and blood lipids, and consequential risk of osteoarthritis, heart disease, stroke, diabetes, liver dysfunction and a range of cancers (WHO 2000).

Additionally, obesity is highly stigmatized in many countries in terms of both the perceived ‘undesirable’ bodily appearance and the apparent character defects it is supposed to indicate, leading to social discrimination. Compared to non-obese peers, people living with obesity are likely to complete fewer years of schooling, are less likely to gain tertiary education and, among women particularly, less likely to enter better-paid employment. Perhaps not surprisingly in these conditions, people living with obesity are more likely to see their own bodies as unattractive and to exclude themselves from social interaction.

Estimated costs and burden of disease

Higher health risks are reflected in greater need for healthcare. Estimates of the relative increase in the need for health services by BMI status have been made in several countries and are shown in Table 8.4.5.

Table 8.4.5 Relative use of health services or expenditure on healthcare

	Australia: use of hospital outpatients (Price Waterhouse Coopers 2015)	Australia: use of hospital inpatients (Price Waterhouse Coopers 2015)	Malta: use of GP services (Price Waterhouse Coopers 2017)	Malta: use of hospital services (Price Waterhouse Coopers 2017)	USA: per capita healthcare expenditure (Arterburn et al. 2005)	Multiple counties: per capita healthcare expenditure (Kent et al. 2017)
BMI 18.5–24.9 Normal weight	1.0	1.0	1.0	1.0	1.0	1.0
BMI 25–29.9 Overweight non-obese	1.14	1.07	1.03–1.15	1.01–1.38	1.099	0.85–1.65
BMI 30–34.9 Obese I	1.25	1.22	1.26–1.38	1.0–1.75	1.231	1.08–1.41
BMI 35–39.9 Obese II	1.39	1.49	1.23–1.57	1.01–1.57	1.448	1.21–1.64
BMI 40+ Obese III	1.07	1.76	1.84–1.85	1.56–4.03	1.815	1.44–2.10

The metrics are not identical across studies, and clearly do not include a number of relevant costs, but the relative increment according to weight status is clear. From these figures, the World Obesity Federation (World Obesity Federation 2017) estimated that the consequential healthcare expenditure resulting from raised body mass in a variety of countries, including those shown in Table 8.4.6.

Table 8.4.6 Examples of estimated healthcare expenditure associated with raised BMI

	2014	2025	2017–2025
United States	$325 bn	$555 bn	$4.2 trn
Egypt	$3.9 bn	$7.8 bn	$58 bn
Australia	$9.0 bn	$17.0 bn	$127 bn
Mexico	$10.2 bn	$20.7 bn	$152 bn
Malta	$115 m	$206 m	$1.6 bn
United Kingdom	$19.1 bn	$30.6 bn	$237 bn
South Africa	$4.4 bn	$8.6 bn	$63 bn
Germany	$31.3 bn	$50.2 bn	$390 bn
Brazil	$16.6 bn	$34.4 bn	$251 bn
Colombia	$2.7 bn	$5.6 bn	$42 bn
Worldwide	**US$600 bn**	**US$1.2 trillion**	**$9.1 trn**

Total (private and government) healthcare expenditure in US$ at 2014 value, adjusted for purchasing power parity, from WHO World Health Report 2014. Projections to 2025 assume World Bank estimates for population growth and linear trends in BMI from 2000–2014 data.

Source: data from World Obesity Federation. Global Data on Costs and Consequences. 2017 Available from: https://www.obesityday.worldobesity.org/ourdata2017

These estimates may prove conservative. An assessment by the Milken Institute in the USA suggested that the direct costs for medical treatment for health conditions causally related to obesity and overweight were valued at US$427.8bn in 2014, or 14% of US healthcare spending (Waters and DeVol 2016).

Besides direct healthcare costs, there are indirect economic costs due to a reduction in economic activity resulting from ill-health attributable to the consequences of obesity. These include costs from absenteeism, presenteeism, and reduced productivity. Examples of these estimates are shown in Table 8.4.7.

Table 8.4.7 Examples of recent estimates of indirect costs associated with obesity

Date	Country	Estimated costs	Basis
2015	Germany	€33.65 bn	Sickness absence, unemployment, nursing care, rehabilitation, early retirement pension, premature mortality, pension for widows and orphans.
2013	Canada	CA$16.4 bn	Unrealised earnings due to poor health and premature mortality (based on the ratio to direct costs).
2012	USA	US$8.65 bn	Loss of productivity due to sickness absence.
2009	Germany	€772 per capita	Loss of productivity due to sickness absence.
2009	Thailand	US$391.8 m	Loss of productivity due to hospital absence and premature mortality.

Sources: data from Tremmel M, Gerdtham UG, Nilsson PM, Saha S. Economic Burden of Obesity: A Systematic Literature Review. Int J Environ Res Public Health 2017; 14. Krueger H, Krueger J, Koot J. Variation across Canada in the economic burden attributable to excess weight, tobacco smoking and physical inactivity. *Can J Public Health* 2015; 106: e171–7.

The Milken Institute report cited here also estimated the indirect costs due to absence or underperformance at work due to ill-health associated with obesity and its consequences, which the Institute estimated at an extraordinary US$988.8bn for 2014. Added to the Institute's estimate of direct costs, the total of $1.4 trillion was equivalent to 8.2% of US gross domestic product for the year.

Other forms of indirect cost have been considered. Additional food has to be grown and transported to meet the needs of heavier populations. Over the three decades 1985–2016 the average adult's BMI in Australia increased by 10.1%, in the UK by 10.4% and in the USA by 12.6%.(NCD Risk Factor Collaboration) Simply to maintain body function at a stable level this implies a 10% increase in food intake for the average adult, implying a significant increase in food supplies beyond that needed for a rise in the numbers of adults. The majority of people do not accumulate excess adiposity until well into adulthood, so it is of concern that for children and adolescents the increases in average BMI over the same period for these three countries were 6.8%, 5.7%, and 8.1%, respectively.

Further costs include the additional costs to infrastructure, both in hospitals and broader industries. A calculation of airline fuel use in the USA suggested that the additional BMI of the USA population in the period 1990–2000 led to a 2.4% increase in fuel requirements for domestic services, or 350m gallons of jet fuel, and added 3.8m tonnes of CO_2 emissions to greenhouse gas emissions (Dannenberg et al. 2004). Calculations have been made for automobile transport in the USA (Jacobson and King 2009) which suggest that in 2008 around 1.1 bn gallons of fuel were used to transport the excess weight of the US population in private vehicles, releasing some 9 m tonnes of CO_2.

The double burden of nutrition

The last century has seen an extraordinary 'nutrition transition' which has led to energy-dense foods and beverages high in sugar and fat becoming widely available to developed and urbanized populations around the globe. A potential double burden of ill-health may occur in populations where the healthfulness of food supplies remains poor: surveys in Vietnam, Brazil, and South Africa have found evidence of overweight children showing lack of growth—stunting—and indeed children in lower-income households in the UK are more likely to be overweight and more likely to be below average height, compared with children in higher-income families (Lobstein et al. 2015).

We are not aware of estimates of the costs of this specific issue. However, there have been recent attempts to estimate the direct and indirect costs for whole populations failing to meet dietary recommendations. A UK study estimated the healthcare costs of poor diets at £5.8 bn for 2006–2007 (Scarborough et al. 2011). An analysis of the economic costs of failing to meet Canadian food recommendations in terms of disease consequences estimated that poor diets were responsible for CA$5.1 bn annually for direct healthcare and CA$8.7bn for indirect costs (Lieffers et al. 2018). This is in addition to the costs associated with overweight and obesity, which have been estimated at CA$23.5 bn annually (Krueger et al. 2015).

Causes of overweight and obesity

A complex system of determinants

For individuals, obesity results from an imbalance between energy intake and expenditure, whereby energy intake is consistently greater than energy

expenditure. Behind this apparently simple equation, obesity is best contextualized as resulting from a complex set of interactions between a wide range of drivers, reflective of the broader socioecological model (Dahlgren and Whitehead 1991; Swinburn et al. 2011; Ohri-Vachaspati et al. 2015).

Obesity results from complex interactions between an individual's physiology, psychology, physical activity, and food consumption, and the social and environmental factors that surround them, including the broader social psychology, the physical activity environment, and food production and trade processes (Vandenbroeck et al. 2007). This complexity was mapped in an obesity system map for the UK Foresight report in 2007 (Vandenbroeck et al. 2007)—which is colloquially referred to as the 'spaghetti map' (Allender et al. 2015) due to the number of interconnections within and across each of the factors (Fig. 8.4.4).

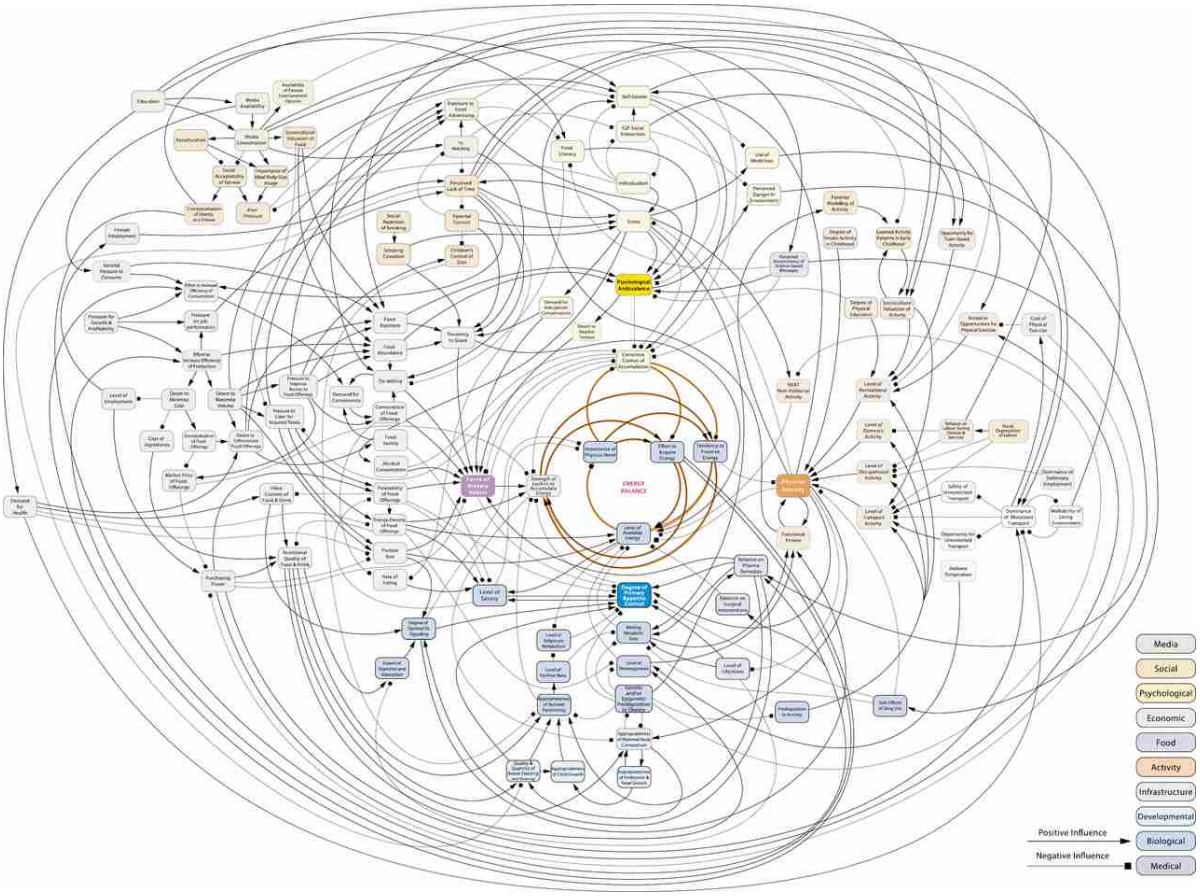


Fig. 8.4.4 The Foresight obesity system map.

Reproduced from Vandenbroeck, IP et. al. (2007), 'Foresight. Tackling Obesities: Future Choices—Obesity System Atlas', (Government Office for Science, UK Government's Foresight Programme) under the Open Government License.

In the centre of these models is the set of individual characteristics such as age, sex, race/ethnicity, and genes that influence an individual's propensity for weight gain. These individual characteristics are then influenced by a range of lifestyle behaviours, including diet, physical activity, and sleep. At the level of social and community networks and environments, influencers include sociocultural, socioeconomic, workplace, recreation, and transport environments, all of which amplify or attenuate the individual level drivers. These in turn are influenced by, and influence, the high-level systemic drivers such as food and transport systems and the broader socioeconomic, commercial, and political determinants of health.

The obesity epidemic

The obesity epidemic, now seen across the world, has resulted from a series of large changes to how we live, including greater urbanization, a food system providing more accessible, affordable, and heavily marketed products and automation leading to decreased activity in our daily lives. While, as indicated earlier, for an individual there are a number of factors contributing directly to the development of obesity at the individual level, the major cause of the global pandemic of obesity is thought to be the global changes in food and diet (Swinburn et al. 2011).

The increases in the prevalence of obesity first seen in high-income countries are now being observed in low- and middle-income countries as they rapidly urbanize and food companies increase their investment in these regions (Friel and Baker 2009; Baker and Friel 2014). China is an illustrative example having recently experienced a transition from a history of undernutrition to a rapid increase in obesity (Gordon-Larsen et al. 2014). Over the past two decades the national health survey in China documents rapid changes in diet, activity, and obesity, all reflective of large social change (Popkin 2014). Through analysis of the China Health and Nutrition Survey, Popkin et al. have documented changes to the food supply, cooking and eating behaviours, transportation modes, and physical activity patterns, all reflective of the tripling in the prevalence of overweight in adults observed between 1991 and 2009 (Popkin 2014).

Globally, these changes to our food and activity environments have been termed the development of ‘obesogenic’ environments (Swinburn et al. 1999) The recognition that our environments encourage and incentivise the very behaviours that contribute to obesity was pivotal in broadening the search for solutions from individual education and behavioural change to transformation of our social, economic, political, food, and physical activity environments.

Causes of causes

As well described by Friel et al. (2007), ‘Focusing only on direct action to make people eat more healthily and be more physically active misses the heart of the problem: the underlying unequal distribution of factors that support the opportunity to be a healthy weight. Unless this oversight is addressed the obesity epidemic and its inequities will persist and possibly increase.’ These causes of causes, or upstream determinants of health, set the context for our obesogenic environments. Housing, income, trade, and societal norms all affect the likelihood of overconsumption and sedentary behaviour.

The policy and economic context strongly influences both the level of growth and consumption, and the resources (social, economic, knowledge, and

skills) that citizens have at their disposal to maintain healthy lifestyles. While direct action will need to be taken at the level of food and activity environments, this will need to include consideration of the upstream 'causes of causes' in order to identify effective and equitable intervention targets.

Diet and physical activity

In 2011 Swinburn et al. postulated a 'tipping point' for the obesity epidemic, identifying diet to be the primary direct driver of current obesity trends (Swinburn et al. 2011). While acknowledging that all the factors identified in the Foresight map contribute both at an individual and societal level to weight gain and obesity, they illustrate that the obesity epidemic since the 1970s can be explained from two distinct phases: 'the so-called move less, stay lean phase (1910–60), characterized by decreasing physical activity levels and energy intake, and a population that remained lean; and the subsequent so-called eat more, gain weight phase, characterized by increasing energy intake and a concomitant rise in population weight' (Swinburn et al. 2011). Here we explore the characteristics of diet and physical activity that have been shown to affect the propensity of an individual, and a population, to gain weight.

The evidence demonstrates that an individual's diet is driven by factors across multiple levels of the socioecological model: the microlevel (interpersonal and intrapersonal factors)—such as an individual's knowledge and skills and their social networks and social norms; the mesolevel (food environment)—such as the marketing of foods and beverages and their perceived and actual availability and affordability; and the macrolevel (broader sociopolitical factors)—such as marketing regulations and agricultural policies (Peeters 2018).

There are many studies demonstrating links between specific foods, nutrients, and dietary patterns and obesity. A review in 2016 by Mozaffarian (Mozaffarian 2016) concludes that robust evidence exists to recommend increased fruits, non-starchy vegetables, nuts, legumes, fish, vegetable oils, yogurt, and minimally processed whole grains; and fewer red meats, processed (e.g. sodium-preserved) meats, and foods rich in refined grains, starch, added sugars, salt, and trans fat, when considering the outcomes of obesity, diabetes, and cardiovascular disease. Further studies have suggested prospective associations between factors including diet quality and weight maintenance (Romaguera et al. 2010; Bertoia et al. 2015; Fung et al. 2015; Schulze et al. 2018) and between soft drink consumption and weight gain (Malik et al. 2006; Malik et al. 2013; Singh et al. 2015; Bucher Della Torre et al. 2016).

Levels of physical activity and sedentary time are similarly driven by factors across multiple levels of the socioecological model: the microlevel (interpersonal and intrapersonal factors)—such as an individual's knowledge and skills, perceptions of safety, and their social networks, and social norms; the mesolevel (physical activity and transport environments)—such as the walkability of a neighbourhood, the availability of public transport and the availability and affordability of physical activities; and the macrolevel (broader sociopolitical factors)—such as urban planning and transport policies (Durand et al. 2011; Hesketh et al. 2017). While there are clear links between activity levels and weight gain for the individual, population studies demonstrate that levels of PA have not changed substantially since the 1980s in many countries (Swinburn et al. 2011; Rodgers et al. 2018).

A number of observational studies and reviews demonstrate the link between the number of hours spent sitting, the number of hours spent watching television, or screens in general, levels of moderate and vigorous physical activity, and transport mode with weight gain and obesity (Dunstan et al. 2005; Turrell et al. 2018).

Genes and biology

The recent World Obesity Federation position statement on obesity articulated a perspective of obesity as a chronic, relapsing disease process. In this process the primary driver of current weight gain, food (particularly foods that are high in energy density such as fat, or in sugar-sweetened beverages), interacts with the other drivers of weight gain, including low physical activity and inadequate sleep, and with the individual's genetic susceptibility.

There are a few known genes with a strong effect on the development of obesity, such as that for leptin, and there are a large number of genes with contributing effects to weight gain susceptibility. In addition to direct genetic effects, we also now know that both epigenetics and the body's weight maintenance physiology play a role in an individual's susceptibility to weight gain and capacity for weight maintenance and weight loss.

In the past decade, our understanding of the role of epigenetics in the development of obesity has increased rapidly (van Dijk et al. 2014). A recent review of studies investigating the association between obesity and either global, site-specific, or genome-wide methylation of DNA has demonstrated that, while there was evidence for a number of obesity-associated differentially methylated sites, there was no consistent evidence for a relationship between global methylation and obesity. Still an emerging field, the current evidence

suggests there are epigenetic modifications *in utero* that may both assist in predicting future obesity risk, and represent modifiable targets to reduce this risk (van Dijk et al. 2014; van Dijk et al. 2015).

For those who attempt weight loss, the physiology of the body is a powerful driver of weight regain. As described in detail by Sumithran et al. (Sumithran and Proietto 2013), in the hypothalamus, hormonal signals from the gastrointestinal tract, adipose tissue, and other peripheral sites are integrated to influence appetite and energy expenditure. In those who lose weight intentionally there are several physiological changes that encourage weight regain, including alterations in energy expenditure, metabolism, and hormone pathways involved in appetite regulation, many of which may persist long term.

Obesity in a 'syndemic' of epidemics

Obesity, together with undernutrition and other dietary risks, are currently by far the largest cause of poor health globally, but are likely to be overtaken by the consequences of climate change. These three pandemics—obesity, undernutrition, and climate change—represent a 'syndemic' or synergy of epidemics because they co-occur in time and place, interact with each other to produce complex sequelae, and share common underlying societal drivers. Although beyond the scope of the present chapter, this global syndemic represents the paramount public health challenge in the twenty-first century. It is discussed in greater detail in the forthcoming report of the Lancet Commission on Obesity (Lancet Commission on Obesity 2019).

Prevention of overweight and obesity

A range of public health strategies are needed to ensure reductions in risk of obesity throughout the life course (Baker et al. 2017). Specifically, strategies are needed to reduce intergenerational transfer and development of susceptibility, to reduce weight gain and support weight maintenance throughout life for those at a healthy weight, to support weight maintenance and moderate weight loss in those with a weight just outside the healthy weight range, and to support weight maintenance after weight loss for those with obesity.

In addition, public health strategies are needed to mitigate the harms of obesity. Specifically, to reduce the stigma associated with obesity (Puhl and Heuer 2010) and to reduce the impact of obesity on people's capacity to lead a

fulfilling life, including opportunities in employment, housing, and marriage (Black et al. 2018). The advantage of a well-constructed, multisectoral, and multifactorial public health approach is that it can address all these issues concurrently.

Current recommendations for a public health approach to obesity

There is currently significant coherence in the recommendations national and internationally for a successful public health approach to obesity. Many of the recommendations fall under the remit of government, but it is clear that to enable the cultural, political, and economic change required for reduced obesity rates into the future, action is required from the full range of societal stakeholders, from citizens through community and non-government organizations to the private sector (Huang et al. 2015). In addition, it is clear that there is a role for multiple levels of government, from local through state to national governments, as well as multiple sectors of government (Sacks et al. 2009).

Whether the recommendations focus on childhood or whole of society it is clear that we need to enact a range of levers to re-orient our food and activity environments, empower and upskill our citizens, and rethink our societal conversations about obesity (Fig. 8.4.5) (Moodie 2008; McGuire 2012; Commission on Ending Childhood Obesity 2017; WHO 2017b, 2018c). In 2012 WHO set a target of reducing premature mortality from non-communicable diseases by 25% by the year 2025 and for there to be no increase in adult and adolescent overweight and obesity (WHO 2013) In 2017, this was followed by a report from the Commission on Ending Childhood Obesity including the high-level policy actions governments will need to take to achieve this latter target (Commission on Ending Childhood Obesity 2017). WHO has also developed recommendations for action for the private sector (WHO 2004).

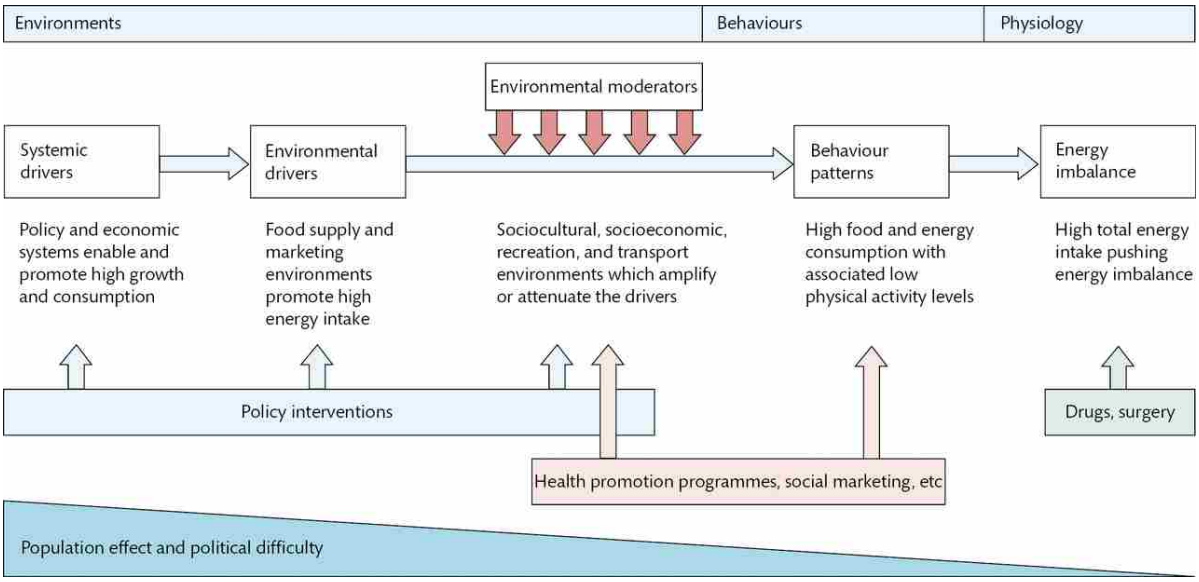


Fig. 8.4.5 A framework to categorize obesity determinants and solutions.
Reprinted from Swinburn, B. A., et al. (2011), 'The global obesity pandemic: shaped by global drivers and local environments', *Lancet*, 378 (9793), 804–14, with permission from Elsevier.

The government recommendations identified by the WHO as the best buys to reduce unhealthy diets and increase physical activity span the key public health levers from pricing, through labelling and guidelines to public education campaigns (Box 8.4.2). Obesity prevention is likely to require government intervention in further areas, such as international trade and food procurement policies (Friel et al. 2013; Hawkes et al. 2013) and is likely to require development of national nutrition, physical activity, and/or obesity plans (Commission on Ending Childhood Obesity 2017; Peeters 2018; WHO 2018c, d) accompanied by ongoing community consultation, research and monitoring (Swinburn et al. 2013).

Box 8.4.2 Key elements of a public health approach to obesity by government

Reduce unhealthy diet

- Implement the global strategy on diet, physical activity, and health
- Implement the WHO recommendations on the marketing of foods and non-alcoholic beverages to children
- Reduce sugar consumption through effective taxation on sugar-sweetened beverages
- Promote and support exclusive breastfeeding for the first 6 months of life, including promotion of breastfeeding
- Implement subsidies to increase the intake of fruits and vegetables
- Limiting portion and package size to reduce energy intake and the risk of overweight/obesity
- Implement nutrition education and counselling in different settings (e.g. in preschools, schools, workplaces, and hospitals) to increase the intake of fruits and vegetables
- Implement nutrition labelling to reduce total energy intake (kcal), sugars, sodium, and fats
- Implement mass media campaign on healthy diets, including social marketing to reduce the intake of total fat, saturated fats, sugars, and salt, and promote the intake of fruits and vegetables

Reduce physical inactivity

- Implement the global strategy on diet, physical activity, and health
- Implement community-wide public education and awareness campaign for physical activity which includes a mass media campaign combined with other community-based education, motivational and environmental programmes aimed at supporting behavioural change of physical activity levels
- Provide physical activity counselling and referral as part of routine primary healthcare services through the use of a brief intervention
- Ensure that macrolevel urban design incorporates the core elements of residential density, connected street networks that include sidewalks, easy access to a diversity of destinations and access to public transport
- Implement a whole-of-school programme that includes quality physical education, availability of adequate facilities and programmes to support physical activity for all children
- Provide convenient and safe access to quality public open space and adequate infrastructure to support walking and cycling
- Implement multicomponent workplace physical activity programmes
- Promotion of physical activity through organized sport groups and clubs, programmes, and events

Source: data from World Health Organization. (2017). Tackling NCDs: ‘best buys’ and other recommended interventions for the prevention and control of noncommunicable diseases. World Health Organization. https://apps.who.int/iris/handle/10665/259232.

A number of international bodies have also provided detailed recommendations to improve population diets and physical activity levels. The World Cancer Research Fund, International hosts the NOURISHING framework (Hawkes et al. 2013). This framework summarizes the policy domains required to improve population diets and an accompanying database collates information about implemented polices to promote healthy diets and reduce obesity. WHO’s Global database on the Implementation of Nutrition Action (GINA) serves a similar purpose, including how programmes are implemented, country adaptations, and lessons learnt (WHO 2018a). The Global Physical Activity Network hosts a knowledge database focusing on

physical activity policies and including summaries of related evidence and implementation (WHO 2018b).

Further, to see such recommendations enacted there will need to be an increase in the political and societal will for action (Huang et al. 2015; Baker et al. 2017). A recent review by Baker and colleagues (Baker et al. 2017) posits that it is possible to build and galvanize political commitment to obesity prevention. They identify a significant role for public health professionals in this process, suggesting that the catalysts will include 'strategic actions by cohesive, resourced and strongly led nutrition actor networks' (Baker et al. 2017). Such actions can be supported by evidence that supports a political ideology and feasible policy options (Clarke et al. 2016). Successful implementation of recommendations will also require a response to the substantial industry opposition to many of the recommended policies. For the first time the Director General of the WHO spoke about the critical role of regulation to contain the food and drink industries, stating 'it is not just Big Tobacco anymore. Public health must also contend with Big Food, Big Soda, and Big Alcohol. All of these industries fear regulation, and protect themselves by using the same tactics' (Chan 10 June 2013). One consequence of this recognition is the development of strong and transparent conflict of interest guidance documents (Buse et al. 2017; WHO 2017a).

Finally, it is important to recognize that while a public health approach to obesity is urgently needed, without an integrated and concurrent focus on greater accessibility and availability of appropriate treatment options, the burden of obesity will continue for many decades.

Evidence of impact

Many of the proposed population obesity prevention strategies have been implemented, evaluated, and modelled for population impact and cost-effectiveness. Comparative cost-effectiveness analyses have demonstrated that some of the regulatory population approaches, including a sugary drinks tax, restricting marketing of unhealthy food and drink and increasing availability of healthy food options relative to unhealthy food options, are likely to be cost-saving at a societal level, leading to reductions in obesity and its sequelae over the lifetime (Table 8.4.8) (Gortmaker et al. 1993; Carter et al. 2009; Vos et al. 2010; Dobbs et al. 2014). Large population benefit has also been identified as a consequence of strategies that change the behavioural defaults, such as food reformulation and decreases in standard portion sizes.

Table 8.4.8 Cost-effectiveness of population obesity prevention interventions

Intervention	Cost-effectiveness*	Other information
Portion size control	Cost-effective	High population impact Evidence of behaviour change
Reformulation	Cost-effective	High population impact Limited evidence for behaviour change
School curriculum	Cost-effective	High population impact Evidence of behaviour change
Labelling	Cost-effective	Moderate population impact Limited evidence for behaviour change
Price promotions	Cost-effective	Moderate population impact Logic model of behaviour change
Marketing restrictions	Cost-effective	Moderate population impact Limited evidence for behaviour change
Tax on high sugar high fat products	Cost-effective	Moderate population impact Logic model of behaviour change
Public health campaigns	Cost-effective	Low population impact Logic model of behaviour change

* Not cost-effective >=$USD50,000/DALY saved; Cost-effective <$USD50,000/ DALY saved; Cost-saving

Note: all analyses done in the context of the United Kingdom.

Source: data from Dobbs, R, et al. (2014), 'Overcoming obesity: An initial economic analysis': McKinsey Global Institute.

The impact of a population strategy to improve one element of the food or activity pathways currently driving obesity will necessarily take a number of years to effect weight gain and obesity. In addition, the effect of any given pathway may be difficult to identify among the range of societal changes affecting obesity. In order to support evidence-based decision making it will be important to ensure the full range of evidence is collected and integrated. Approaches integrating logic pathways and systems dynamics will facilitate conclusions on whether a given policy, intervention, or programme is having its intended effect on the direct target behaviours, any unintended effects, and eventually its intended effect on weight gain and obesity (Peeters 2018). Recent evaluations have begun to adapt approaches from systems science (Rutter et al.

2017; Owen et al. 2018). While these methods are still evolving, by identifying and integrating a variety of evidence sources, such approaches have the potential to provide a more comprehensive evaluation of obesity prevention policies.

One approach that has underpinned a number of intervention evaluations has been the quantification of the effect of energy imbalance on bodyweight by Hall and colleagues (Hall et al. 2011). This mathematical modelling approach simulates energy expenditure adaptations during weight loss using findings from human metabolism studies. In addition to informing multiple modelling studies, this approach identified an important characteristic of the obesity epidemic: that a very small average daily energy imbalance between intake and expenditure of about 30 kJ per day can explain the observed average weight gain. This is equivalent to less than a biscuit a day, demonstrating the potential of effective prevention strategies to impact on population obesity rates.

For long-term impact, implemented policies and programmes will need not only to be effective, but also to be sustained and to be equitable. Recent analyses of population obesity trends have shown no reduction in inequalities in obesity in high-income countries (Devaux and Sassi 2013; Chung et al. 2016). Further, there is some evidence that while obesity rates may be improving in some population groups, these improvements are occurring more often in less disadvantaged groups than more disadvantaged groups (Chung et al. 2016). To ensure that implemented policies benefit the most disadvantaged citizens, who are also likely to carry the highest burden of risk, a more comprehensive understanding of differential effectiveness across different population groups is required. A selection of population obesity prevention policies have been evaluated for such differential effectiveness, demonstrating that a sugary drinks tax and whole of community interventions are likely to be effective across the socioeconomic spectrum. However, the majority of recommended policies have not been evaluated for the equity impact. It will be important to continue to build evaluations of the equity effect of such policies (Backholer et al. 2014; Peeters and Backholer 2015). It will also be critical to continue to build the evidence regarding synergistic effects with other related and substantial population issues, including undernutrition climate change (Lancet Commission on Obesity 2019).

Conclusions

A successful public health approach to obesity will require strong political commitment and leadership, accompanied by community and organizational support. Re-orienting our food and physical activity environments alongside supporting and upskilling our citizens will drive the culture change required to reverse the obesity epidemic. A comprehensive, integrated multisectoral approach will be required to reduce intergenerational transfer and excess weight gain, as well as to support weight maintenance in those who have experienced intentional weight loss. Evaluation techniques will need to continue to evolve to ensure implemented interventions are effective, equitable, and sustainable.

ACKNOWLEDGEMENTS

The authors would like to thank Gary Sacks, Phil Baker, and Emma Gearon for initial inspirations, thoughtful review and critique of this chapter. They would also like to thank Professors W. Philip. T. James and Tim Marsh, the authors of the previous version of the Chapter.

REFERENCES

All-Party Parliamentary Group on Obesity (2018). *The Current Landscape of Obesity Services*. London: House of Commons.

Allender, S., et al. (2015). A community based systems diagram of obesity causes. *PLoS One*, 10(7), e0129683.

Anonymous (2017). Worldwide trends in body-mass index, underweight, overweight, and obesity from 1975 to 2016: a pooled analysis of 2416 population-based measurement studies in 128.9 million children, adolescents, and adults. *The Lancet*, 390(10113), 2627–42.

Arterburn, D.E., Maciejewski, M.L., and Tsevat, J. (2005). Impact of morbid obesity on medical expenditures in adults. *International Journal of Obesity (London)*, 29(3), 334–9.

Australian Bureau of Statistics (2018). *4727.0.55.001—Australian Aboriginal and Torres Strait Islander Health Survey: First Results*, Australia, 2012–13 (updated 27 November 2013). [Online] Available at: http://www.abs.gov.au/ausstats/abs@.nsf/0/A07BD8674C37D838CA257C2F001459FA?opendocument

Backholer, K., et al. (2014). A framework for evaluating the impact of obesity prevention strategies on socioeconomic inequalities in weight. *American Journal of Public Health*, 104(10), e43–50.

Baker, P. and Friel, S. (2014). Processed foods and the nutrition transition: evidence from Asia. *Obesity Review*, 15(7), 564–77.

Baker, P., et al. (2017). Generating political priority for regulatory interventions targeting obesity prevention: an Australian case study. *Social Science & Medicine*, **177**, 141–49.

Bertoia, M.L., et al. (2015). Changes in intake of fruits and vegetables and weight change in united states men and women followed for up to 24 years: analysis from three prospective cohort studies. *PLoS Medicine*, 12(9), e1001878.

Black, N., Kung, C.S.J., and Peeters, A. (2018). For richer, for poorer: the relationship between adolescent obesity and future household economic prosperity. *Preventive Medicine*, **111**, 142–50.

Bray, G.A., et al. (2017). Obesity: a chronic relapsing progressive disease process. A position statement of the World Obesity Federation. *Obesity Review*, 18(7), 715–23.

Bucher Della Torre, S., et al. (2016). Sugar-sweetened beverages and obesity risk in children and adolescents: a systematic analysis on how methodological quality may influence conclusions. *Journal of the Academy of Nutrition and Diabetes*, 116(4), 638–59.

Buse, K., Tanaka, S., and Hawkes, S. (2017). Healthy people and healthy profits? Elaborating a conceptual framework for governing the commercial determinants of non-communicable diseases and identifying options for reducing risk exposure. *Global Health*, 13(1), 34.

Carter, R., et al. (2009). Assessing cost-effectiveness in obesity (ACE-obesity): an overview of the ACE approach, economic methods and cost results. *BMC Public Health*, **9**, 419.

Chan, M. (10 June 2013). *WHO Director-General Addresses Health Promotion Conference: Opening Address at the 8th Global Conference on Health Promotion.* Finland: World Health Organization.

Chung, A., et al. (2016). Trends in child and adolescent obesity prevalence in economically advanced countries according to socioeconomic position: a systematic review. *Obesity Review*, 17(3), 276–95.

Clarke, B., Swinburn, B., and Sacks, G. (2016). The application of theories of the policy process to obesity prevention: a systematic review and meta-synthesis. *BMC Public Health*, 16(1), 1084.

Cole, T.J. and Lobstein, T. (2012). Extended international (IOTF) body mass index cut-offs for thinness, overweight and obesity. *Pediatric Obesity*, 7(4), 284–94.

Cole, T.J., et al. (2000). Establishing a standard definition for child overweight and obesity worldwide: international survey. *British Medical Journal*, 320(7244), 1240–3.

Commission on Ending Childhood Obesity (2017). *Report of the Commission on Ending Childhood Obesity.* Geneva: World Health Organization.

Dahlgren, G. and Whitehead, M. (1991). *Policies and Strategies to Promote Social Equity in Health.* Stockholm: Institute for Futures Studies.

Dannenberg, A.L., Burton, D.C., and Jackson, R.J. (2004). Economic and environmental costs of obesity: the impact on airlines. *American Journal Preventive Medicine*, 27(3), 264.

Devaux, M. and Sassi, F. (2013). Social inequalities in obesity and overweight in 11 OECD countries. *European Journal of Public Health*, 23(3), 464–9.

Dobbs, R., et al. (2014). *Overcoming Obesity: An Initial Economic Analysis*. McKinsey Global Institute.

Dunstan, D.W., et al. (2005). Associations of TV viewing and physical activity with the metabolic syndrome in Australian adults. *Diabetologia*, 48(11), 2254–61.

Durand, C.P., et al. (2011). A systematic review of built environment factors related to physical activity and obesity risk: implications for smart growth urban planning. *Obesity Review*, 12(5), e173–82.

Friel, S. and Baker, P.I. (2009). Equity, food security and health equity in the Asia Pacific region. *Asia Pacific Journal of Clinical Nutrition*, 18(4), 620–32.

Friel, S., Chopra, M., and Satcher, D. (2007). Unequal weight: equity oriented policy responses to the global obesity epidemic. *British Medical Journal*, 335(7632), 1241–3.

Friel, S., et al. (2013). Monitoring the impacts of trade agreements on food environments. *Obesity Reviews*, **14**, 120–34.

Fung, T.T., et al. (2015). Long-term change in diet quality is associated with body weight change in men and women. *Journal of Nutrition*, 145(8), 1850–6.

Garrow, J.S. (1974). *Energy Balance and Obesity in Man*. Amsterdam: North-Holland Publishing, p. 287.

Gearon, E., et al. (2018). Diet and physical activity as possible mediators of the association between educational attainment and body mass index gain among Australian adults. *International Journal of Public Health*, 63(7), 883–93.

Gordon-Larsen, P., Wang, H., and Popkin, B.M. (2014). Overweight dynamics in Chinese children and adults. *Obesity Review*, 15 Suppl 1, 37–48.

Gortmaker, S.L., et al. (1993). Social and economic consequences of overweight in adolescence and young adulthood. *New England Journal of Medicine*, 329(14), 1008–12.
Grijalva-Eternod, C.S., et al. (2012). The double burden of obesity and malnutrition in a protracted emergency setting: a cross-sectional study of Western Sahara refugees. *PLoS Medicine*, 9(10), e1001320.
Hall, K.D., et al. (2011). Quantification of the effect of energy imbalance on bodyweight. *The Lancet*, 378(9793), 826–37.
Hawkes, C., Jewell, J., and Allen, K. (2013). A food policy package for healthy diets and the prevention of obesity and diet-related non-communicable diseases: the NOURISHING framework. *Obesity Review*, 14 Suppl 2, 159–68.
Hesketh, K.R., Lakshman, R., and van Sluijs, E.M.F. (2017). Barriers and facilitators to young children's physical activity and sedentary behaviour: a systematic review and synthesis of qualitative literature. *Obesity Review*, 18(9), 987–1017.
Huang, T.T., et al. (2015). Mobilisation of public support for policy actions to prevent obesity. *The Lancet*, 385(9985), 2422–31.
Jacobson, S.H. and King, D.M. (2009). Measuring the potential for automobile fuel savings in the US: the impact of obesity. *Transportation Research Part D (Transport and Environment)*, 14(1), 6–13.
James, W.P.T. (1976), Research on obesity: a report of the DHSS/MRC group/compiled by James, W.P.T.; [for the] Department of Health and Social Security [and the] Medical Research Council. Waterlow, J.C (ed.) London: H.M.S.O.
Kannel, W.B. and Gordon, T. (1974). Some determinants of obesity and its impact as a cardiovascular risk factor. In: Howard, A. (ed.) *Recent Advances in Obesity Research: I. Proceedings of the First International Congress on Obesity*. Tunbridge Wells: Newman Publishing, p. 14.
Kent, S., et al. (2017). Body mass index and healthcare costs: a systematic literature review of individual participant data studies. *Obesity Review*, 18(8), 869–79.
Knai, C., et al. (2012). Socioeconomic patterning of childhood overweight status in Europe. *International Journal Environmental Research and Public Health*, 9(4), 1472–89.
Krueger, H., Krueger, J., and Koot, J. (2015). Variation across Canada in the economic burden attributable to excess weight, tobacco smoking and physical inactivity. *Canadian Journal of Public Health*, 106(4), e171–7.
Lancet Commission on Obesity (2019). *Report of the Lancet Commission on Obesity.* London: Lancet.
Lean, M.E., Han, T.S., and Morrison, C.E. (1995). Waist circumference as a measure for indicating need for weight management. *British Medical Journal*, 311(6998), 158–61.
Lean, M.E., Han, T.S., and Deurenberg, P. (1996). Predicting body composition by densitometry from simple anthropometric measurements. *American Journal of Clinical Nutrition*, 63(1), 4–14.
Lieffers, J.R.L., et al. (2018). The economic burden of not meeting food recommendations in Canada: the cost of doing nothing. *PLoS One*, 13(4), e0196333.
Lobstein, T., et al. (2017). Comment: obesity as a disease—some implications for the World Obesity Federation's advocacy and public health activities. *Obesity Review*, 18(7), 724–26.
Lobstein, T., et al. (2015). Child and adolescent obesity: part of a bigger picture. *The Lancet*, 385(9986), 2510–20.
Malik, V.S., Schulze, M.B., and Hu, F.B. (2006). Intake of sugar-sweetened beverages and weight gain: a systematic review. *American Journal of Clinical Nutrition*, 84(2), 274–88.
Malik, V.S., et al. (2013). Sugar-sweetened beverages and weight gain in children and adults: a systematic review and meta-analysis. *American Journal of Clinical Nutrition*, 98(4), 1084–102.
McGuire, S. (2012). Institute of Medicine. 2012. Accelerating progress in obesity prevention: solving the weight of the nation. Washington, DC: the National Academies Press. *Advanced Nutrition*, 3(5), 708–9.
McLaren, L. (2007). Socioeconomic status and obesity. *Epidemiology Review*, **29**, 29–48.
Ministry of Health—Manatū Hauora obesity statistics. (updated 01 August 2018) [Online]. Available at: https://www.health.govt.nz/nz-health-statistics/health-statistics-and-data-sets/obesity-statistics
Monteiro, C.A., et al. (2004a). Obesity and inequities in health in the developing world. *International Journal of Obesity and Related Metabolic Disorders*, 28(9), 1181–6.

Monteiro, C.A., et al. (2004b). Socioeconomic status and obesity in adult populations of developing countries: a review. *Bulletin of the World Health Organization*, 82(12), 940–6.
Moodie, A.R. (2008). Australia: the healthiest country by 2020. *Medical Journal of Australia*, 189(10), 588–90.
Mozaffarian, D. (2016). Dietary and policy priorities for cardiovascular disease, diabetes, and obesity: a comprehensive review. *Circulation*, 133(2), 187–225.
NCD Risk Factor Collaboration (2017). Worldwide trends in body-mass index, underweight, overweight, and obesity from 1975 to 2016: a pooled analysis of 2416 population-based measurement studies in 128.9 million children, adolescents, and adults. *The Lancet*, 390(10113), 2627–42.
NCD Risk Factor Collaboration. Adult body-mass index: evolution of BMI over time. [Online] Available at: http://ncdrisc.org/data-downloads.html
NCD Risk Factor Collaboration (2016). Trends in adult body-mass index in 200 countries from 1975 to 2014: a pooled analysis of 1698 population-based measurement studies with 19.2 million participants. *The Lancet*, 387(10026), 1377–96.
NHS Digital (2017a). *Health Survey for England 2016: Adult Overweight and Obesity.* London: Health and Social Care Information Centre.
NHS Digital (2017b). *National Child Measurement Programme—England, 2016–17. National Child Measurement Programme*. London: Health and Social Care Information Centre.
Ogden, C.L., et al. (2014). Prevalence of childhood and adult obesity in the United States, 2011–2012. *Journal of the American Medical Association*, 311(8), 806–14.
Ohri-Vachaspati, P., et al. (2015). The relative contribution of layers of the Social Ecological Model to childhood obesity. *Public Health Nutrition*, 18(11), 2055–66.
Owen, B., et al. (2018). Understanding a successful obesity prevention initiative in children under 5 from a systems perspective. *PLoS One*, 13(3), e0195141.
Peeters, A. and Backholer, K. (2015). Reducing socioeconomic inequalities in obesity: the role of population prevention. *Lancet Diabetes and Endocrinology*, 3(11), 838–40.
Peeters, A. (2018). Obesity and the future of food policies that promote healthy diets. *Nature Reviews Endocrinology*, **14**, 430–7.
Peeters, A., et al. (2016). Beyond BMI: how to capture influences from body composition in health surveys. *Current Nutrition Reports*, 1–9.
Pickett, K.E., et al. (2005). Wider income gaps, wider waistbands? An ecological study of obesity and income inequality. *Journal of Epidemiology and Community Health*, 59(8), 670–4.
Popkin, B.M. (2014). Synthesis and implications: China's nutrition transition in the context of changes across other low- and middle-income countries. *Obesity Review*, 15 Suppl 1, 60–7.
Price Waterhouse Coopers (2015). *Weighing the Costs of Obesity: A Case for Action*. Sydney: PwC.
Price Waterhouse Coopers (2017). *Weighing the Costs of Obesity in Malta.* Qurmi: PwC.
Puhl, R.M. and Heuer, C A. (2010). Obesity stigma: important considerations for public health. *American Journal of Public Health*, 100(6), 1019–28.
Rodgers, A., et al. (2018). Prevalence trends tell us what did not precipitate the US obesity epidemic. *Lancet Public Health*, 3(4), e162–3.
Romaguera, D., et al. (2010). Mediterranean dietary patterns and prospective weight change in participants of the EPIC-PANACEA project. *American Journal of Clinical Nutrition*, 92(4), 912–21.
Rutter, H., et al. (2017). The need for a complex systems model of evidence for public health. *The Lancet*, 390(10112), 2602–4.
Sacks, G., Swinburn, B., and Lawrence, M. (2009). Obesity Policy Action framework and analysis grids for a comprehensive policy approach to reducing obesity. *Obesity Review*, 10(1), 76–86.
Scarborough, P., et al. (2011). The economic burden of ill health due to diet, physical inactivity, smoking, alcohol and obesity in the UK: an update to 2006–07 NHS costs. *Journal of Public Health (Oxford)*, 33(4), 527–35.
Schulze, M.B., et al. (2018). Food based dietary patterns and chronic disease prevention. *British Medical Journal*, 361, k2396.

Singh, G.M., et al. (2015). Estimated global, regional, and national disease burdens related to sugar-sweetened beverage consumption in 2010. *Circulation*, 132(8), 639–66.
Stenholm, S., et al. (2008). Sarcopenic obesity: definition, cause and consequences. *Current Opinion in Clinical Nutrition & Metabolic Care*, 11(6), 693–700.
Sumithran, P. and Proietto, J. (2013). The defence of body weight: a physiological basis for weight regain after weight loss. *Clinical Science (London)*, 124(4), 231–41.
Swinburn, B., Egger, G., and Raza, F. (1999). Dissecting obesogenic environments: the development and application of a framework for identifying and prioritizing environmental interventions for obesity. *Preventive Medicine*, 29(6 Pt 1), 563–70.
Swinburn, B., et al. (2013). INFORMAS (International Network for Food and Obesity/non-communicable diseases Research, Monitoring and Action Support): overview and key principles. *Obesity Reviews*, **14**, 1–12.
Swinburn, B.A., et al. (2011). The global obesity pandemic: shaped by global drivers and local environments. *The Lancet*, 378(9793), 804–14.
Tanamas, S.K., et al. (2016). Changing guards: time to move beyond body mass index for population monitoring of excess adiposity. *QJM: An International Journal of Medicine*, 109(7), 443–46.
Tremmel, M., et al. (2017). Economic burden of obesity: a systematic literature review. *International Journal of Environmental Research and Public Health*, 14(4), 435.
Turrell, G., et al. (2018). Do active modes of transport cause lower body mass index? Findings from the HABITAT longitudinal study. *Journal of Epidemiology and Community Health*, 72(4), 294–301.
van Dijk, S. J., et al. (2015). Recent developments on the role of epigenetics in obesity and metabolic disease. *Clinical Epigenetics*, **7**, 66.
van Dijk, S. J., et al. (2014). Epigenetics and human obesity. *International Journal of Obesity*, **39**, 85.
Vandenbroeck, I.P., Goossens, J., and Clemens, M. (2007). Foresight. Tackling obesities: future choices — obesity system atlas. In: Vandenbroeck, P., Goossens, J., and Clemens, M. (eds.), *Tackling Obesities: Future Choices—Obesity System Atlas.* London: Government Office for Science, UK Government's Foresight Programme.
Vos, T., et al. (2010). *Assessing Cost-Effectiveness in Prevention (ACE–Prevention): Final Report.* Melbourne: University of Queensland, Brisbane and Deakin University.
Wandell, P.E. (2013). Population groups in dietary transition. *Food Nutrition Research*, 57. Doi: 10.3402/fnr.v57i0.21668.
Waters, H. and DeVol, R. (2016). *Weighing Down America: The Health and Economic Impact of Obesity*. Santa Monica: Milken Institute.
WHO Expert Consultation (2004). Appropriate body-mass index for Asian populations and its implications for policy and intervention strategies. *The Lancet*, 363(9403), 157–63.
World Health Organization (WHO) (1995). *Physical Status: The Use and Interpretation of Anthropometry. Report of a WHO Expert Committee*. WHO Technical Report Series 854. Geneva: WHO.
World Health Organization (WHO) (2000). *Obesity: Preventing and Managing the Global Epidemic. Report of a WHO Consultation*. WHO technical report series: 894. Geneva: WHO.
World Health Organization (WHO) (2004). *Global Strategy on Diet, Physical Activity and Health.* Geneva: WHO.
World Health Organization (WHO) (2011). *Obesity and Overweight. Fact Sheet No. 311.* [Online] Available at: http://www.who.int/mediacentre/factsheets/fs311/en/
World Health Organization (WHO) (2013). *Global Action Plan for the Prevention and Control of NCDs 2013–2020.* Geneva: WHO.
World Health Organization (WHO) (2017a). Safeguarding against possible conflicts of interest in nutrition programmes: approach for the prevention and management of conflicts of interest in the policy development and implementation of nutrition programmes at country level. [Online] Available at: http://www.who.int/nutrition/consultation-doi/comments/en/
World Health Organization (WHO) (2017b). *'Best Buys' and Other Recommended Interventions for the Prevention and Control of Noncommunicable Diseases: Updated (2017) Appendix 3 of the*

Global Action Plan for the Prevention and Control of Noncommunicable Diseases 2013–2020. Geneva: World Health Organization.

World Health Organization (WHO) (2018a). Global database on the Implementation of Nutrition Action (GINA). [Online] Available at: https://extranet.who.int/nutrition/gina/en

World Health Organization (WHO) (2018b). *Global Action Plan on Physical Activity 2018–2030: More Active People for a Healthier World.* Geneva: World Health Organization.

World Health Organization (WHO) (2018c). *Growth Reference 5–19 Years.* [Online] Available at: http://www.who.int/growthref/en/

World Health Organization (WHO) (2018d). *Child Growth Standards.* [Online] Available at: http://www.who.int/childgrowth/en/

World Obesity Federation (2017). Global data on costs and consequences. [Online] Available at: https://www.obesityday.worldobesity.org/ourdata2017

8.5

Physical activity and public health

Nyssa T. Hadgraft, Neville Owen, and Paddy C. Dempsey

Introduction

Physical activity has well-established health benefits. However, large proportions of the adult populations of developed countries, and rapidly urbanizing developing countries are insufficiently physically active for health benefits. Additionally, many people's lives are characterized by large amounts of time spent sitting. This represents a formidable set of public health challenges.

Although the relationships of exercise and physical activity with health have been topics of interest for millennia (Tipton 2014; Myers et al. 2015), the modern era of physical activity and public health arguably dates back to the early 1950s with the seminal epidemiological studies of Morris and colleagues. These involved employees of the London Transport Executive (bus drivers compared to conductors) and Post Office (civil servants compared to postal workers). The main findings were that the largely sedentary bus drivers had almost double the age-adjusted rate of fatal coronary heart disease, when compared with conductors who spent much of their workday climbing stairs, walking, and standing. Similarly, the incidence of coronary heart disease was lower in the active postal workers compared to sedentary civil service clerks and executives (Morris et al. 1966, 1953). From a broader public health perspective, it is interesting to note that one of the initial hypotheses considered by Morris and colleagues was the potential effects of stress on the cardiovascular health of bus drivers, whom it would be assumed would be more socially isolated in their jobs than would the conductors.

Some 70 years on from the Morris studies, residents of developed and developing countries have been subject to a broad set of demographic,

epidemiological, economic, social, and technological transitions. Widespread urbanization, particularly in low- and middle-income countries, has meant large segments of traditionally rural populations are moving into cities. Many previously manual tasks in the occupational and household sectors have become automated, making life easier and safer in many respects. A pervasive consequence of these changes is that the living environments of large numbers of people globally are becoming less supportive of structured and incidental physical activity, and more conducive to large amounts of time spent sitting. For this reason, physical inactivity has been called a 'global pandemic' (Kohl et al. 2012); one requiring multidisciplinary and multisectoral approaches (Reis et al. 2016).

In this chapter, we focus on physical activity and sedentary behaviour (put simply, too little exercise and too much sitting), in the contemporary public health context, highlighting:

the key definitions and distinctions around physical activity, exercise, and sedentary behaviour
current evidence on relationships with health outcomes, and current physical activity recommendations
current evidence on prevalence and trends in physical activity and sedentary behaviour, and some of the key issues for surveillance and measurement
how physical activity and sedentary behaviour may be understood in ways that will inform public health approaches

We emphasize the importance of an interdisciplinary and intersectoral strategy—working with constituencies beyond the public health field, such as urban planners, architects, and policymakers, in order to create activity-friendly built environments that provide socially equitable opportunities for physical activity to be part of everyday living.

Physical activity and sedentary behaviour: definitions, health consequences, and guidelines

Definitions

When thinking about what constitutes physical activity and where it takes place, images that might come to mind are people working out in a gym, or

playing sport. However, physical activity is broader than structured exercise. Although often used interchangeably, ‘physical activity’, ‘exercise’, ‘physical fitness’, and ‘sedentary behaviour’ are terms that describe related but distinct concepts. Clear definitions and distinctions are important for research and practice in this field (see Table 8.5.1) (Caspersen et al. 1985; Pate et al. 2008; Brown et al. 2012; Tremblay et al. 2017).

Table 8.5.1 Key terms and definitions: physical activity and sedentary behaviour

Physical activity	Any bodily movement produced by skeletal muscles that expends energy. The term ‘physical activity’ usually refers to the subset of activity that produces substantial health benefits—e.g. moderate-intensity to vigorous intensity activity in bouts of at least 10 minutes at a time. This can be from leisure-time physical activity, transportation (e.g. walking or cycling), occupational or household work, play, games, sports, or planned exercise.
Exercise	A*subset* of physical activity that is purposeful, planned, and structured—with an objective more geared towards improving facets of physical fitness (see next); the term ‘exercise’ is at times used synonymously with ‘physical activity’.
Physical fitness	A set of attributes that people have or achieve that are health or skill-related, such as cardiorespiratory endurance, muscular strength and endurance, flexibility, balance, and body composition. The degree to which people have these attributes can be measured with specific tests.
Baseline activity	All light-intensity activities of daily life, such as standing, walking slowly, and lifting light weight objects. It may also include short episodes of moderate or vigorous activity, such as climbing a few flights of stairs.
Sedentary behaviours	Any waking behaviour characterized by an energy expenditure ≤1.5 metabolic equivalents (METs), while in a sitting, reclining, or lying posture. Note that this is different from ‘physical inactivity’ (see earlier).

As noted earlier, there is now increased interest in the distinct health effects and determinants of sedentary behaviour—sitting. While ‘sedentary’ was often historically used to describe someone who did not engage in sufficient physical activity (i.e. physically inactive), more recently it has been specifically defined to refer to *any waking behaviour characterized by an energy expenditure ≤1.5 metabolic equivalents (METs), while in a sitting, reclining or lying posture* (Sedentary Behavior Research Network 2012; Tremblay et al. 2017).

Physical activity and exercise can be quantified using the constructs of *intensity* (how hard?), *duration* (how long?), *frequency* (how often?), and *mode* (what type?), for example, walking, running, swimming, cycling, etc. (Powell et al. 2004). These constructs are used to determine whether someone meets physical activity guidelines (see Box 8.5.2).

Health consequences: why physical inactivity and sedentary behaviour are crucial public health issues

Strong evidence demonstrates that physical inactivity is a modifiable risk factor for a wide spectrum of health conditions, including non-communicable diseases that are a rapidly growing public health concern in developed and developing countries. It has been estimated that 6–10% of all deaths from non-communicable diseases (including 7% of deaths associated with type 2 diabetes, 6% of coronary heart disease deaths, and 10% of breast/colon cancer deaths) worldwide can be attributed to physical inactivity, making it the fourth leading cause of death worldwide (Lee et al. 2012).

Physical inactivity is related not only to increased risk of illness and premature mortality, it also results in a substantial economic burden worldwide. In the first global estimate for the economic burden of physical inactivity (based on data from 142 countries, representing 93.2% of the world's population), it was conservatively estimated that the effect of physical inactivity cost the world economy more than $67.5 billion through healthcare expenditure and productivity losses in 2013. Low- and middle-income countries had the largest disease burden from physical inactivity (Ding et al. 2016).

There are numerous benefits to health from physical activity; including reducing the risk of excessive weight gain, type 2 diabetes, injurious falls, depression, and loss of cognitive function (Warburton et al. 2010; WHO 2010; Warburton and Bredin 2016; Physical Activity Guidelines Advisory Committee 2018); see Box 8.5.1).

There is also moderately strong evidence that physical activity maintains functional ability in older adults, helps maintain weight loss, improves sleep quality, and reduces the risk of hip fracture and osteoporosis.

Different types of physical activity acting through multiple physiological pathways influence a broad array of health outcomes. The acute and chronic

physiological adaptations that can lead to ill-health largely depend on the type of activity performed. For example, aerobic activities tend to cause many changes to the heart and vascular system (e.g. increased stroke volume, increased capillary density, reduced peripheral resistance). Resistance activities (e.g. weight training or working against body weight) have modest effects on the cardiovascular system, but do improve muscle strength and endurance (Powell et al. 2004; Garber et al. 2011). This highlights that any single activity is unlikely to be optimal for comprehensive health benefits and will depend upon the health outcome of most interest to the population subgroup of concern.

At the lower end of the activity spectrum, accumulating research also indicates that reducing and replacing sedentary behaviours with light-intensity activities is beneficial in a variety of populations, on a smaller scale, for metabolic and cardiovascular health outcomes (Dempsey et al. 2016; Chastin et al. 2017; Chastin et al. 2018). Consistent evidence from prospective observational studies across numerous countries now also indicates that excessive sedentary behaviour (sitting) is also associated with an increased risk for all-cause mortality, heart disease, diabetes and some cancers (Wilmot et al. 2012; Biswas et al. 2015), and may be further exacerbated in those who are insufficiently active or unfit (Biswas et al. 2015; Bouchard et al. 2015; Ekelund et al. 2016). A recent meta-analysis (Chau et al. 2013) estimated that some 6% of all deaths could be attributed to daily self-reported sitting time.

Since much of the world's population is physically inactive (Hallal et al. 2012), and time spent sedentary (with a high proportion of that time spent in sitting in prolonged bouts) accounts for a majority of adults' waking hours (Matthews et al. 2008; Healy et al. 2015), many of the aforementioned figures are likely underestimating the global economic and disease burden associated with physical inactivity and excessive sitting time. From a public health perspective, whether there is a minimum volume of physical activity that may reduce some of the health risks associated with prolonged sitting, or a threshold above which the benefits of physical activity are attenuated, is a topic of contemporary research interest (Powell et al. 2004).

Box 8.5.1 Strength of the evidence on the health benefits of physical activity

Strong evidence of reduced rates of:

- all-cause mortality
- coronary heart disease
- high blood pressure
- stroke
- metabolic syndrome
- type 2 diabetes
- breast cancer
- colon cancer
- depression
- falling

Strong evidence of:

- increased cardiorespiratory and muscular fitness
- healthier body mass and composition
- improved bone health
- increased functional health
- improved cognitive function

Moderately strong evidence that physical activity:

- helps maintain functional ability in older adults, helps maintain weight loss, improves sleep quality, and reduces the risk of hip fracture and osteoporosis

Physical activity and sedentary behaviour guidelines

Evidence supporting the importance of physical activity for population health may be found in a number of landmark documents, including:

the 1996 US Department of Health and Human Services' *Surgeon General's Report on Physical Activity and Health* (US Department of Health and Human Services 1996)
the 2008 US Department of Health and Human Services *Physical Activity Guidelines for Americans* (US Department of Health and Human Services 2008; Physical Activity Guidelines Advisory Committee 2018)
the 2010 World Health Organization (WHO) *Global Recommendations on Physical Activity for Health* (WHO 2010)
and, in 2010, the *Toronto Charter for Physical Activity* was ratified at the 3rd International Congress for Physical Activity and Health (Bull et al. 2010), calling for global action to prioritize physical activity for all.

Based on extensive evidence syntheses and reviews compiled by national and international bodies, there have been generally consistent recommendations developed in relation to adults' physical activity. The WHO's recommendations for adults aged 18–64 are to perform at least 150 min/week of moderate-intensity aerobic physical activity (e.g. brisk walking, gentle bicycling), or at least 75 min/week of vigorous intensity aerobic physical activity (e.g. jogging, running, hard bicycling), or an equivalent combination of moderate and vigorous. In addition, it is recommended that individuals should engage in muscle-strengthening activities at least 2 days/week (see Box 8.5.2) (WHO 2010). More than the suggested minimum total of 150 min/week of activities (up to 300 min/week are suggested) is likely to provide additional benefits. The WHO recommendations were developed with the goal of primary prevention of non-communicable diseases (Vuori 2018).

For those who are inactive or highly sedentary, any physical activity is beneficial. There are health benefits from volumes of physical activity less than half the recommended amount (Powell et al. 2004; WHO 2010). Indeed, a curvilinear reduction in risk occurs for a variety of diseases and conditions across volume of activity, with the greatest improvements in health and mortality risk tending to occur when those who are very inactive/sedentary become more physically active (Powell et al. 2004; Haskell et al. 2007; Garber et al. 2011).

Recommendations for older adults (aged 65 or over) are generally the same as for adults, as much as abilities allow, with particular emphasis of avoiding excessive inactivity and sedentary time. For children/adolescents, the recommended amount of moderate-vigorous physical activity is at least

60 min/day, with muscle/bone strengthening exercises recommended at least three times per week.

Recommendations on sedentary behaviour have recently been introduced into some country-specific guidelines, but are less specific than are those for physical activity. For example, Australia's *Physical Activity and Sedentary Behaviour Guidelines* recommend that adults 'minimise the amount of time spent in prolonged sitting' and 'break up long periods of sitting as often as possible' (DoH 2017).

Box 8.5.2 Recommended guidelines for physical activity and sedentary behaviour

- Adults should accumulate at least 150 minutes of moderate-intensity physical activity, or at least 75 minutes of vigorous intensity physical activity throughout the week, or an equivalent combination of moderate and vigorous intensity activity (e.g. running, brisk walking, bicycling, playing tennis, dancing, and swimming).
- For additional health benefits, adults should increase their moderate-intensity physical activity to 300 minutes per week, or equivalent.
- Muscle-strengthening activities involving major muscle groups should be done on 2 or more days a week, e.g. weight training, calisthenics (body weight resistance such as half-squats, push ups, pull ups, sit ups), carrying heavier loads, heavy gardening (such as lifting, digging, or shovelling).
- Avoid inactivity and minimise the total amount of time that is spent sitting, particularly in sitting prolonged bouts; e.g. replace prolonged periods (~30–60 minutes) of sitting (e.g. at work, watching television) with more standing and more incidental activity.

The WHO has provided leadership in recent decades in advocating for action on physical inactivity by member states (Vuori 2018). The *WHO Global Action Plan for the Prevention and Control of Non-Communicable Diseases 2013–2020*, endorsed by the World Health Assembly in 2013, identifies physical inactivity as one of four risk factors for non-communicable diseases, alongside tobacco use, alcohol consumption and unhealthy diets (WHO 2013). In 2018, WHO launched the *Global Action Plan on Physical Activity 2018–2030: More Active People for a Healthier World*, which aims to provide countries with up-to-date guidance on physical activity and a framework for policy responses (WHO 2018). The plan sets a target of a 15% relative reduction in the prevalence of physical inactivity in adults and adolescents by 2030 (WHO 2018).

Physical activity and sedentary behaviour: prevalence and trends

Physical activity

Country comparisons provide insight into variations in physical activity and sedentary behaviour across the world. When assessing self-reported physical activity against the guidelines outlined earlier, the proportion of adults globally who are not meeting guidelines has been estimated at 31.1%. However, country- and region-specific averages vary substantially. For example, only 4.7% of the population are considered inactive in Bangladesh, through to a high of 71.9% in Malta (Hallal et al. 2012). Higher levels of inactivity have been found in the eastern Mediterranean region (43.2%) and the Americas (43.3%), compared to Southeast Asia (17.0%) and Africa (27.5%) (Hallal et al. 2012). In general, higher levels of inactivity are found in higher-income compared to lower-income countries (Dumith et al. 2011). It has been suggested that this may be because the contribution of physical activity from work tends to be lower in higher-income countries (Hallal et al. 2012).

Within individual countries, average physical activity levels also vary across subpopulations—differences that have implications for health inequities and developing appropriately targeted health promotion programmes. One contributing factor may be that safe and affordable spaces to be physically active are less accessible to some population groups, including those of low socioeconomic position, those with disabilities, and girls, women, and older adults (WHO 2018).

Gender differences in physical activity levels have been observed across most countries, with women more likely to be physically inactive than men (Hallal et al. 2012; Bauman et al. 2009, 2012). These differences can be attributed to a range of economic, social, and cultural factors. Women often spend more time in household activities, which are not always captured by physical activity questionnaires (Hallal et al. 2003). However, women across the world also tend to have lower levels of leisure-time physical activity than men (Brown et al. 2016). This may be due to social and cultural factors, such as in countries where women's participation in physical activity can be restricted (e.g. those in the Gulf Cooperation Council) (Mabry et al. 2010). However, gender differences also exist in high-income countries where

cultural barriers to participation are less apparent, suggesting that social norms may also play a role. For this reason, gender-specific strategies may be needed to address the physical activity gap (Brown et al. 2016).

Physical activity levels also appear to vary by socioeconomic status. Leisure-related physical activity levels tend to be greater among those in higher socioeconomic positions, while occupational physical activity is higher for those with a lower socioeconomic position (Gidlow et al. 2006; Beenackers et al. 2012). There is some evidence to suggest that occupational physical activity (which often involves heavy lifting and prolonged, static postures) may not be as beneficial for improving health outcomes as volitional physical activity performed during leisure (Holtermann et al. 2012; Coenen et al. 2018), which has been suggested as a possible pathway through which health inequities manifest (Coenen et al. 2018). This emphasizes the need to consider the *mode* and *setting* when assessing population levels of physical activity and designing interventions, as discussed later in this chapter.

Sedentary behaviour

Recent findings highlight the extent to which adults around the world engage in high volumes of sitting time, but less is known about international prevalence of sedentary behaviour, relative to what is known about physical activity. Cross-country studies have found that self-reported daily sitting time can range from 2.5 to 3 hours per day in countries such as Portugal, Brazil, and Columbia, through to 6 hours/day in Denmark and the Netherlands, and seven hours/day in Japan and Saudi Arabia (Bauman et al. 2011; Loyen et al. 2016). However, data obtained from studies using accelerometers (a type of activity monitor) suggest that these self-report figures may be underestimates of actual time spent in sedentary behaviour, with some reporting adults spend 55–70% of their waking hours engaged in sedentary behaviours (Matthews et al. 2008; Hagstromer et al. 2010) or over 8 hours/day (Bauman et al. 2018).

Physical activity and sedentary behaviour trends

Evidence from high-income countries suggest that the direction of physical activity trends over time vary depending on the setting in which the activity

is being assessed. While leisure-related physical activity levels have tended to remain steady or slightly increased over time, physical activity during work, the domestic environment and transportation has decreased in recent decades (Hallal et al. 2012; Ng and Popkin 2012). Across the globe, fewer people now work in occupations that require them to be physically active (Ng and Popkin 2012). In the United States, the proportion of workers in manufacturing jobs has declined sharply over the last 50 years, while service-related jobs have become more prevalent. In 2008, only 20% of occupations required at least moderate intensity physical activity, down from 48% in 1960 (Church et al. 2011). Many adults now work in desk-based occupations, where over two-thirds of the workday can be spent sitting (Clemes et al. 2014; van Dommelen et al. 2016). Figure 8.5.1, adapted from Ng and Popkin (Ng and Popkin 2012) provides compelling estimates of past, recent, and projected trends across domains of physical activity and sedentary behaviour.

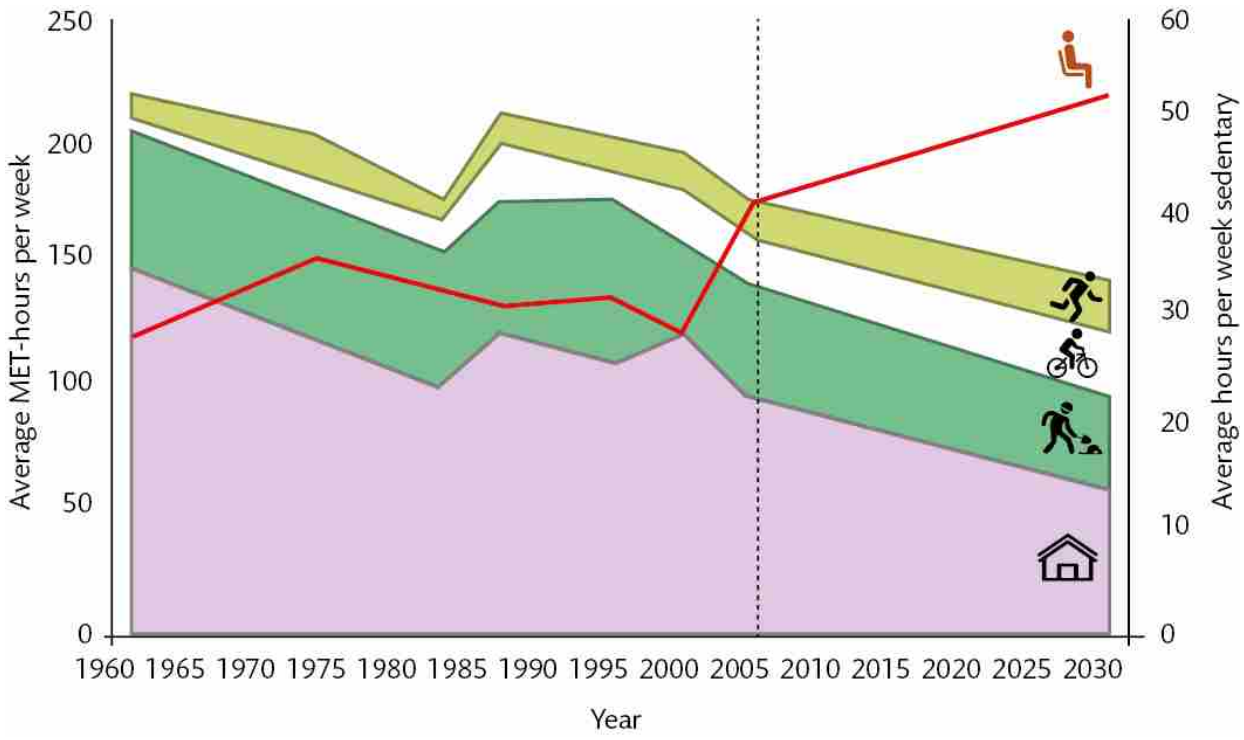


Fig. 8.5.1 UK adults' metabolic equivalents of task (MET)-hours per week of time in sedentary behaviour, and hours per week of all physical activity in different domains (active-leisure, travel, occupational and domestic physical activity): based on time-use surveys measured for 1961–2005 and forecasted for 2006–2030.

Reproduced from Ng SW, Popkin BM. Time use and physical activity: a shift away from movement across the globe. *Obes Rev*. 2012;13(8):659–80, with permission from John Wiley and Sons.

While much of the evidence on physical activity trends has come from high-income countries evidence from China and Brazil suggests that these countries have also experienced reductions in occupational and domestic physical activity in recent decades (Ng and Popkin 2012; Ng et al. 2014). For example, national surveys in China reveal that physical activity levels dropped by 32% between 1991 and 2006 (Ng et al. 2009), with much of this decline due to falling levels of occupational and domestic physical activity (Ng et al. 2014). At the same time, leisure-time physical activity levels have generally remained stable for the majority of the population (Ng et al. 2014).

New technologies are changing, shifting, and refining our perspectives on physically active and sedentary behaviours and creating new research agendas. The rapidly growing use of consumable wearable technologies and smart phones/watches (which have inbuilt accelerometers) is also providing unique opportunities to monitor, and potentially intervene, on physical activity and sedentary behaviour (Althoff et al. 2017; Phillips et al. 2018).

Until recently, surveillance studies examining the population prevalence of physically active and sedentary time (such as those described earlier) have typically relied on time-use and self-reported moderate-vigorous physical activity during leisure time, or people's estimates of overall daily sitting or television/screen viewing time. However, activity monitoring using accelerometers (which can record time, duration, frequency, and intensity of movements) and inclinometer-accelerometers (which can provide more accurate data on sitting, standing, and stepping) are starting to provide more accurate and precise measurement of activity levels across the day. It is important to note that both self-report and sensor-assessed methods have their strengths and limitations. Self-report is prone to recall and response bias, social desirability, or under- or over-reporting, while sensor-assessed measurements can be expensive in larger studies and are typically unable to differentiate between the different contexts of physically active and sedentary behaviours (Sallis and Saelens 2000; Prince et al. 2008; Healy et al. 2011).

A challenge for monitoring global physical activity levels is the variability between countries in the availability of current and representative data. Measures used to access physical activity can differ across countries, which creates challenges in comparisons (Varela et al. 2017). Regular, national surveillance of physical activity levels is generally concentrated in highly developed countries, while a number of low-and middle-income countries in Africa, parts of Latin America, and central Asia have limited or no data on their population's physical activity levels (Hallal et al. 2012). The Global Observatory for Physical Activity (GoPA!) recently developed a series of Country Cards to amalgamate information on country-specific indicators for physical activity surveillance, policy, and research data (Varela et al. 2017). The cards are intended to be used as an advocacy tool to promote greater awareness of, and investment in public health policy, research, and programmes relating to physical activity, which may be particularly valuable for low-and middle-income countries (Ramirez Varela et al. 2018).

Understanding and influencing physical activity and sedentary behaviour: an ecological perspective

Physical activity is broader than activities that constitute structured exercise or sport. It can also, and predominantly does, take place in a variety of different settings—including during leisure, transportation, and in the workplace or school environment. Similarly, when we think about sedentary behaviour, sitting occurs at desks in the workplace, watching television in the home, or driving a car for transport. Understanding the settings in which physical activity and sedentary behaviour take place, and the factors that influence these behaviours, is particularly important for identifying opportunities for public health interventions that do not exacerbate existing social disadvantage. Box 8.5.3 outlines three major settings in which physical activity and sedentary behaviour typically take place.

Box 8.5.3 Physical activity and sedentary behaviour settings

Leisure: Physical activity and sedentary behaviour performed during leisure is generally performed with a degree of volitional control relative to other settings such as the workplace or school environment. For many adults, physical activity comprises a small proportion of their total leisure time, if any. Instead, the majority of leisure time is spent in sedentary pursuits, such as watching television, using a computer, or socializing with friends.

Work: Physical activity performed at work is often termed 'occupational physical activity'. Jobs which require heavy lifting and other manual tasks have higher levels of occupational physical activity. With increasing automation within many industries and a large number of the workforce 'desk-bound', occupational physical activity has been decreasing over recent decades. Many workers (such as those in desk-based jobs, parts of manufacturing, transport) spend much of the day sitting. The workplace has therefore become a key target setting for reducing sedentary behaviour.

Transportation: Physical activity performed during transportation is often termed 'active transport'. Examples of this include walking or cycling to get to and from destinations, including public transport stops. The rates of active transport vary substantially across the world. For example, over a third of the population walks or cycles to work in China and the Netherlands, while fewer than 5% do so in Australia or the United States (Hallal et al. 2012). Cycling in particular, is very common in China, the Netherlands, and Denmark (Hallal et al. 2012), while the vast majority of people in Australia and the United States use cars to get to and from work. In addition to less time spent in physical activity, high car dependency means more time spent in prolonged bouts of sitting.

Conceptual models and theories can be helpful for public health practitioners who are interested in understanding why a behaviour does, or does not occur, and therefore how to best intervene for positive health outcomes. Ecological models are helpful frameworks that have been applied to assist our understanding of a variety of health behaviours (Owen et al. 2011). A strength of ecological models is that they can incorporate multiple theories within an overarching framework, and thus allow for a broader

approach to understanding potential determinants of major health risks and multilevel approaches for change (Sallis and Owen 2015).

Rather than focusing primarily on changing behaviour at the individual level—which poses limitations beyond the individual's control, ecological approaches also consider extrinsic influences on behaviour (see Fig. 8.5.2). The ultimate goal being to create environments and policies that make it safe, convenient, attractive, and economical to make health-promoting choices, and then motivate and educate people to take up those healthy options.

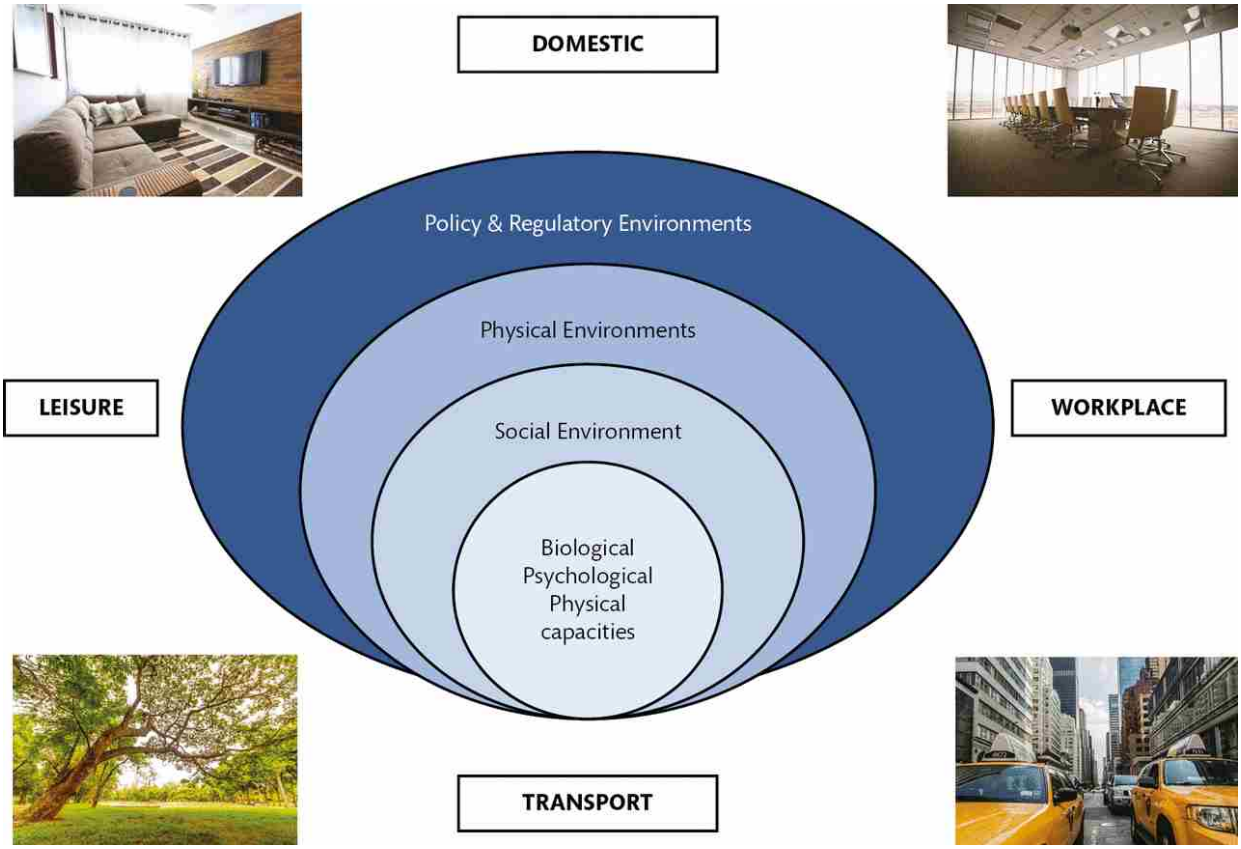


Fig. 8.5.2 A summary version of ecological models of physical activity and sedentary behaviour. Reprinted from Owen N et al (2011) 'Adults' Sedentary Behavior Determinants and Interventions' *Am. J. Preventive Medicine* 41(2):189–96, with permission from Elsevier.

Ecological models have five key principles as outlined to follow (Sallis and Owen 2015), which have relevance to approaches for influencing both physical activity and sedentary behaviour:

There are multiple levels of influence on health behaviours
Environmental contexts are significant determinants of health behaviours
Influences on behaviours interact across levels
Ecological models should be behaviour specific
Multilevel interventions should be most effective in changing behaviours

Considering these principles in the context of physical activity and sedentary behaviour, the influences on these behaviours are complex and multifaceted. These might include individual factors (such as preferences, abilities or motivation), social factors (e.g. extent of social support for activity, social norms), environmental factors (e.g. availability of equipment or proximity to pleasant surroundings for activity), and policy factors (e.g. workplace policies around active transport) (Owen et al. 2011). As these influences are hypothesized to interact across levels, interventions or strategies that focus only on one level of influence (such as individual, behavioural factors) are likely to be limited in their success at changing behaviour (Hadgraft et al. 2018).

A public health approach to addressing physical inactivity and sedentary behaviour recognizes that strategies premised on the idea of 'individual choice' will only go so far, and it is important to recognize and address the broader structural barriers many in the population face to being physically active. This includes the broader social and environmental conditions in which we live and work (Owen et al. 2014). The WHO emphasizes the importance of a systems-based approach for addressing physical inactivity; one that focuses on addressing both upstream policy approaches and downstream individual-focused strategies (WHO 2018).

Developing public health strategies to increase physical activity and reduce sedentary time

Following the release of the Toronto Charter for Physical Activity, an evidence review identified seven 'best investments' for physical activity that are considered to be applicable across population groups (see Box 8.5.4) (GAPA-ISPAH 2012). Importantly, these approaches target the multiple levels of influence of physical activity, incorporating upstream social and environmental approaches, with downstream individually focused approaches (WHO 2018).

Box 8.5.4 Seven best investments for physical activity

1. **Supportive school environments:** Prioritizing and resourcing physical activity (structured and unstructured) across the entire school day, through approaches that engage the entire school community.
2. **Transport policies and systems:** Promoting active transport by prioritizing walking, cycling, and public transport.
3. **Urban design infrastructure and regulations that support physical activity:** the built environment should support transport and recreational walking, cycling, and recreation, across different age groups.
4. **Physical activity integrated into primary healthcare systems:** physical activity should be promoted within primary healthcare as an important modifiable risk factor for non-communicable diseases.
5. **Education and awareness raising:** multiple approaches, such as traditional media, social media, and public education campaigns should be used to promote physical activity awareness and knowledge and shift societal norms.
6. **Whole-of-community approaches to physical activity:** Physical activity should be promoted across the life course, and programmes should be integrated across the multiple settings in which people spend their time (e.g. workplaces, schools, community settings).
7. **Making sport programmes accessible for all:** promoting physical activity through sport by making sure it is enjoyable and accessible for people of different ages, gender identities, abilities, and across the socioeconomic spectrum.

Many of the best investments for physical activity may also have a positive effect on reducing sedentary behaviour; for example, mass media campaigns, infrastructure, and policies that promote walking and cycling, rather than car use (Scheepers et al. 2014). While there is a smaller body of knowledge about what works for reducing sedentary behaviour relative to physical activity, evidence suggests that environmental modifications in the workplace environment, such as the implementation of sit-stand workstations, can lead to significant reductions in sitting time (Neuhaus et al. 2014; Chu et al. 2016). In line with the ecological model of health behaviours, these types of interventions appear to be more effective when included as part of a multicomponent approach that targets multiple levels of influence (Chu et al. 2016).

In accordance with a key principle of ecological models, it is important that public health strategies for increasing physical activity and reducing sedentary behaviour are specific to the context in which these behaviours occur. As outlined earlier, physical activity and sedentary behaviour can

occur in multiple settings and contexts. The main factors that influence time spent in physical activity and sedentary behaviour in each of these settings are not necessarily the same, and understanding these differences is important for intervention development (Owen et al. 2011). Designing effective interventions requires identifying the settings-specific determinants or influences on these behaviours, to enable the development of appropriately targeted strategies. Importantly, consideration should also be given to the equity of physical activity initiatives, including providing no- or low-cost strategies where possible (Boxes 8.5.4 and 8.5.5).

Box 8.5.5 Examples of settings-based approaches for increasing physical activity and/or reducing sedentary behaviour

Workplace

- Introduction of sit-stand workstations to address the significant influence of desk-based work on sitting time (Neuhaus et al. 2014)
- Modifying the internal physical environment to promote stair use (e.g. through attractive design, point of decision prompts) and providing end-of-trip facilities to support active transport (Oldenburg et al. 2002)

School

- Walking School Bus programmes to encourage children to walk to school rather than being driven (Mendoza et al. 2011).
- Incorporating physical activity into academic classes and promoting physical activity breaks during lessons (Watson et al. 2017)

Transport

- Building safe and separated bicycle lanes to encourage active transport (Pucher and Buehler 2008)
- Reducing the availability of free car parking to promote use of active and public transport (Scheepers et al. 2014).

Community

- Organized activities in public spaces (e.g. tai chi, dance class, yoga) (Heath et al. 2012)
- Community walking groups (Kassavou et al. 2013)

It is worth noting that much of the evidence on ‘what works’ for increasing physical activity and reducing sedentary behaviour has arisen from controlled research studies in high-income countries, with relatively short-term follow up durations (Foster et al. 2005). Evidence for the effectiveness of community-wide interventions is somewhat inconsistent, potentially due to methodological issues of studies and the challenges inherent with population-level approaches (Baker et al. 2015). To increase our knowledge about the types of population-level physical activity interventions that are both effective and cost-effective, funding and resources should be prioritized towards conducting longer-term and more rigorous evaluations of these types of approaches (Laine et al. 2014; Reis et al. 2016). In particular, it is important to access the efficacy of approaches in a diversity of populations, including those from low-middle-income countries, less privileged backgrounds, those with disabilities, and those from culturally and linguistically diverse backgrounds.

Cities and transport systems, physical activity, and sedentary time

One significant set of challenges (and potential opportunities) for public health has been the rapid urbanization of developing countries around the world. As greater proportions of the population transition from rural to urban living, there can be benefits to public health through better access to healthcare and reduced incidence of infection (Leon 2008). However, these benefits are counteracted by a rise in non-communicable diseases, such as diabetes and cardiovascular disease (Yusuf et al. 2001), brought about by risk factors such as poor nutrition, insufficient physical activity, and high levels of sedentary time (Popkin 2004).

While traditional approaches to urban transport have considered how we can design safer roads for cars, there is now increasing interest in the ‘walkability’ of urban environments, or how features of the built environment may promote or discourage physical activity (Giles-Corti et al. 2016). Evidence from large cross-country studies has shown that particular urban design features can significantly influence the amount of time people spend in physical activity (Sallis et al. 2016). Higher connectedness of streets, higher density (to support mixed development and public transport), and the availability of local destinations are associated with higher levels of

walking for transport (Giles-Corti et al. 2016). Cycling is also more common in areas where there is supportive infrastructure—such as networks of separated bicycles lanes and bicycle parking (Pucher and Buehler 2008; Hallal et al. 2012). An advantage of walking and cycling is that they are low to no cost, meaning that they are broadly accessible to a large proportion of the population (i.e. those who have the physical capacity to walk or cycle). This makes walking in particular a feasible and accessible strategy for people as they age.

Increasing the number of trips that are conducted by active transport and reducing the amount of time that people spend in cars may also have broad-ranging benefits for public health beyond impacts on physical activity levels. Car time, for example, has been associated with a number of risk factors for chronic disease, including higher waist circumference, BMI, fasting plasma glucose and a clustered cardiometabolic risk score (Sugiyama et al. 2016). Reducing car emissions will also have broader environmental and health benefits (Woodcock et al. 2009).

The International Physical Activity and the Environment Network

In addressing key elements of the environmental and policy agenda advocated earlier, the International Physical Activity and the Environment Network (IPEN) project was established in 2004. IPEN has been working as an international team from 12 countries (Australia, Belgium, Brazil, Columbia, Czech Republic, Denmark, China, Mexico, New Zealand, Spain, the United Kingdom, and the United States), carrying out studies to identify neighbourhood environmental factors associated with physical activity and sedentary behaviour.

Several findings with important public health implications have emerged. The IPEN project has measured physical activity using accelerometer devices, has examined variations by country in activity levels and participants' perceptions of a range of built-and natural-environment attributes of their neighbourhoods, and has employed geographic information system databases to provide objective indices of environmental factors (see Fig. 8.5.3 for country-specific examples).

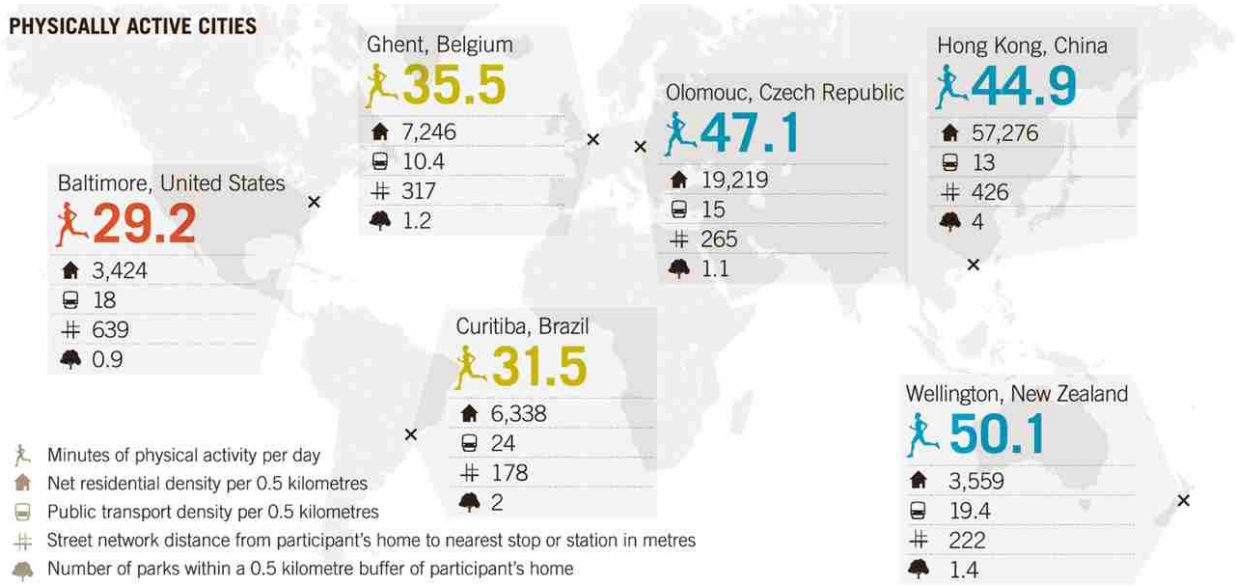


Fig. 8.5.3 Examples of country-specific activity levels and related environmental attributes, from the IPEN project.
Source: J. F. Sallis et al. The Lancet 387, 2207–2217 (2016).

In identifying associations of neighbourhood environmental attributes with physical activity and health, using the 'walkability' construct (a composite measure employing indices of net residential density, intersection density, land use mix, and net retail area ratio) has been particularly informative.

Through the IPEN project, interesting findings have emerged on the role that the built environment may play in promoting or discouraging physical activity. Based on data from 14 cities in 10 countries, several neighbourhood environment attributes were related consistently to accelerometer-assessed moderate to vigorous physical activity. These included:

residential density
number of public transport stops
number of street intersections
and, number of parks within walking distance

The relationships were linear; for example, the higher the level of residential density was related to a correspondingly higher level of physical activity. Those living in the most 'activity-friendly' neighbourhoods did 68–89 minutes more of physical activity per week than those in the least activity-friendly neighbourhoods (Sallis et al. 2016).

The IPEN project has identified considerable similarity of findings across the cities investigated, adding strong evidence to support previous calls for policy changes in urban planning, public transport, and development of parks and recreational facilities as ways to increase physical activity. It is of considerable interest that 'activity-friendly' neighbourhood characteristics were found to be related to physical activity in a linear fashion—suggesting that every improvement in the built environment of a city can be expected to help to increase physical activity, whether or not residents are starting at a low or high level. These findings, however, are from a set of developed countries. Comparable findings have emerged from studies in Japan (Liao et al. 2016). Whether such environmental attributes are similarly important in low- and middle-income countries remains to be determined.

Creating environments in cities that promote, rather than discourage, activity will require the involvement of (among others) urban planners, engineers, and policy makers at different levels of government, alongside public health practitioners. Shared decision making and building trust with stakeholders from the variety of sectors involved is important for effective intersectoral action to create healthy cities (Barr et al. 2008).

Conclusions

While efforts to understand and influence physical inactivity have been a public health priority for several decades, progress in improving the physical activity levels of populations has been modest (Heath et al. 2012). Understanding how to address low levels of physical activity in populations that are increasingly urbanized, automobile dependent, and captured by screens in their domestic and workplace environments is a complex issue, like many public health challenges.

Public health initiatives focused on creating healthy and enabling environments, such as through targeting urban planning and the built environment attributes of cities, offer great potential to improve health outcomes. Such approaches would not only have positive benefits on and through physical activity, but also through influencing other adverse exposures such as air pollution, noise, stress, and limited access to good-quality food (Giles-Corti et al. 2016).

The causes of low levels of physical activity and high levels of sedentary behaviour are multifactorial, and different approaches are likely to be

necessary depending on the setting and the population group of interest. To achieve large-scale change, interventions and policies will need to be developed using intersectoral and systems-based approaches, as well as perspectives that incorporate broad-ranging expertise and recognize the influence of traditional non-health sectors on health risk behaviours and outcomes.

REFERENCES

Althoff, T., Sosič, R., Hicks, J.L., King, A.C., Delp, S.L., and Leskovec, J. (2017). Large-scale physical activity data reveal worldwide activity inequality. *Nature*, 547(7663), 336–9.

Baker, P.R., Francis, D.P., Soares, J., Weightman, A.L., and Foster, C. (2015). Community wide interventions for increasing physical activity. *Cochrane Database of Systemic Reviews*, 1, CD008366.

Barr, V., Pedersen, S., Pennock, M., and Rootman, I. (2008). *Health Equity Through Intersectoral Action: An Analysis of 18 Country Case Studies.* Ontario: Public Health Agency of Canada and WHO.

Bauman, A., et al. (2011). The descriptive epidemiology of sitting. A 20-country comparison using the International Physical Activity Questionnaire (IPAQ). *American Journal of Preventive Medicine*, 41(2), 228–35.

Bauman, A., et al. (2009). The International Prevalence Study on Physical Activity: results from 20 countries. *International Journal of Behavioral Nutrition and Physical Activity*, **6**, 21.

Bauman, A.E., Petersen, C.B., Blond, K., Rangul, V., and Hardy, L.L. (2018). The descriptive epidemiology of sedentary behaviour. In: *Sedentary Behaviour Epidemiology*. Cham: Springer, pp. 73–106.

Bauman, A.E., Reis, R.S., Sallis, J.F., Wells, J.C., Loos, R.J.F., and Martin, B.W. (2012). Correlates of physical activity: why are some people physically active and others not? *The Lancet*, 380(9838), 258–71.

Beenackers, M.A., et al. (2012). Socioeconomic inequalities in occupational, leisure-time, and transport related physical activity among European adults: a systematic review. *International Journal of Behavioral Nutrition and Physical Activity*, 9(1), 116.

Biswas, A., et al. (2015). Sedentary time and its association with risk for disease incidence, mortality, and hospitalization in adults: a systematic review and meta-analysis. *Annals of Internal Medicine*, 162(2), 123–32.

Bouchard, C., Blair, S.N., and Katzmarzyk, P.T. (2015). Less sitting, more physical activity, or higher fitness? *Mayo Clinic Proceedings*, 90(11), 1533–40.

Brown, W.J., Bauman, A.E., Bull, F.C., and Burton, N.W. (2012). *Development of Evidence-Based Physical Activity Recommendations for Adults (18–64 Years). Report prepared for the Australian Government Department of Health*. [Online] Available at: https://www1.health.gov.au/internet/main/publishing.nsf/Content/F01F92328EDADA5BCA257BF0001E720D/$File/DEB-PAR-Adults-18-64years.pdf

Brown, W.J., Mielke, G.I., and Kolbe-Alexander, T.L. (2016). Gender equality in sport for improved public health. *The Lancet*, 388(10051), 1257–8.

Bull, F.C., Gauvin, L., Bauman, A., Shilton, T., Kohl, H., and Salmon, A. (2010). The Toronto Charter for Physical Activity: a global call for action. *Journal of Physical Activity and Health*, 7(4), 421–2.

Caspersen, C.J., Powell, K.E., and Christenson, G.M. (1985). Physical activity, exercise, and physical fitness: definitions and distinctions for health-related research. *Public Health Reports*, 100(2), 126–31.

Chastin, S., et al. (2017). Interventions for reducing sedentary behaviour in community-dwelling older adults (Protocol). *Cochrane Database of Systematic Reviews*, 9, CD012784.

Chastin, S.F.M, et al. (2018). How does light-intensity physical activity associate with adult cardiometabolic health and mortality? Systematic review with meta-analysis of experimental and observational studies. *British Journal of Sports Medicine*, 53(6), 370–6.

Chau, J.Y., et al. (2013). Daily sitting time and all-cause mortality: a meta-analysis. *PLoS One*, 8(11), e80000.

Chu, A.H., Ng, S.H., Tan, C.S., Win, A.M., Koh, D., and Muller-Riemenschneider, F. (2016). A systematic review and meta-analysis of workplace intervention strategies to reduce sedentary time in white-collar workers. *Obesity Review*, 17(5), 467–81.

Church, T.S., et al. (2011). Trends over 5 decades in U.S. occupation-related physical activity and their associations with obesity. *PLoS One*, 6(5), e19657.

Clemes, S.A., O'Connell, S.E., and Edwardson, C.L. (2014). Office workers objectively measured sedentary behavior and physical activity during and outside working hours. *Journal of Occupational and Environmental Medicine*, 56(3), 298–303.

Coenen, P., et al. (2018). Do highly physically active workers die early? A systematic review with meta-analysis of data from 193 696 participants. *British Journal of Sports Medicine*, 52(20), 1320–6.

Dempsey, P.C., Owen, N., Yates, T.E., Kingwell, B.A., and Dunstan, D.W. (2016). Sitting less and moving more: improved glycaemic control for type 2 diabetes prevention and management. *Current Diabetes Reports*, 16(11), 114.

Department of Health (DoH) (2017). Australia's Physical Activity and Sedentary Behaviour Guidelines 2017 [updated 21 November]. [Online] Available at: http://www.health.gov.au/internet/main/publishing.nsf/content/health-pubhlth-strateg-phys-act-guidelines

Ding, D., et al. (2016). The economic burden of physical inactivity: a global analysis of major non-communicable diseases. *The Lancet*, 388(10051), 1311–24.

Dumith, S.C., Hallal, P.C., Reis, R.S., and Kohl, H.W. (2011). Worldwide prevalence of physical inactivity and its association with human development index in 76 countries. *Preventive Medicine*, 53(1–2), 24–8.

Ekelund, U., et al. (2016). Does physical activity attenuate, or even eliminate, the detrimental association of sitting time with mortality? A harmonised meta-analysis of data from more than 1 million men and women. *The Lancet*, 388(10051), 1302–10.

Foster, C., Hillsdon, M., Thorogood, M., Kaur, A., and Wedatilake, T. (2005). Interventions for promoting physical activity. *Cochrane Database of Systematic Reviews*, 1, CD003180.

Garber, C.E., et al. (2011). American College of Sports Medicine position stand. Quantity and quality of exercise for developing and maintaining cardiorespiratory, musculoskeletal, and neuromotor fitness in apparently healthy adults: guidance for prescribing exercise. *Medicine & Science in Sports & Exercise*, 43(7), 1334–59.

Gidlow, C., Johnston, L.H., Crone, D., Ellis, N., and James, D. (2006). A systematic review of the relationship between socio-economic position and physical activity. *Health Education Journal*, 65(4), 338–67.

Giles-Corti, B., et al. (2016). City planning and population health: a global challenge. *The Lancet*, 388(10062), 2912–24.

Global Advocacy for Physical Activity the Advocacy Council of the International Society for Physical Activity and Health (GAPA-ISPAH) (2012). NCD prevention: investments that work for physical activity. *British Journal of Sports Medicine*, 46(10), 709–12.

Hadgraft, N.T., Dunstan, D.W., and Owen, N. (2018). Models for understanding sedentary behaviour. In: Leitzmann, M., Jochem, C., and Schmid, D. (eds.) *Sedentary Behaviour Epidemiology*. Cham: Springer, pp. 381–403.

Hagstromer M, Troiano RP, Sjostrom M, and Berrigan D. (2010). Levels and patterns of objectively assessed physical activity--a comparison between Sweden and the United States. *American Journal of Epidemiology*, 171(10), 1055–64.

Hallal, P.C., Andersen, L.B., Bull, F.C., Guthold, R., Haskell, W., and Ekelund, U. (2012). Global physical activity levels: surveillance progress, pitfalls, and prospects. *The Lancet*, 380(9838), 247–57.

Hallal, P.C., Victora, C.G., Wells, J.C., and Lima, R.C. (2003). Physical inactivity: prevalence and associated variables in Brazilian adults. *Medical Science and Sports Exercise*, 35(11), 1894–900.

Haskell, W.L., et al. (2007). Physical activity and public health: updated recommendation for adults from the American College of Sports Medicine and the American Heart Association. *Circulation*, 116(9), 1081–93.

Healy, G.N., Clark, B.K., Winkler, E.A., Gardiner, P.A., Brown, W.J., and Matthews, C.E. (2011). Measurement of adults' sedentary time in population-based studies. *American Journal of Preventive Medicine*, 41(2), 216–27.

Healy, G.N., Winkler, E.A., Owen, N., Anuradha, S., and Dunstan, D.W. (2015). Replacing sitting time with standing or stepping: associations with cardio-metabolic risk biomarkers. *European Heart Journal*, 36(39), 2643–9.

Heath, G.W., et al. (2012). Evidence-based physical activity intervention: lessons from around the globe. *The Lancet*, **380**, 272–81.

Holtermann, A., Hansen, J.V., Burr, H., Sogaard, K., and Sjogaard, G. (2012). The health paradox of occupational and leisure-time physical activity. *British Journal of Sports Medicine*, 46(4), 291–5.

Kassavou, A., Turner, A., and French, D.P. (2013). Do interventions to promote walking in groups increase physical activity? A meta-analysis. *International Journal of Behavioral Nutrition and Physical Activity*, 10(1), 18.

Kohl, H.W., et al. (2012). The pandemic of physical inactivity: global action for public health. *The Lancet*, 380(9838), 294–305.

Laine, J., Kuvaja-Kollner, V., Pietila, E., Koivuneva, M., Valtonen, H., and Kankaanpaa, E. (2014). Cost-effectiveness of population-level physical activity interventions: a systematic review. *American Journal of Health Promotion*, 29(2), 71–80.

Lee, I.-M., et al. (2012). Effect of physical inactivity on major non-communicable diseases worldwide: an analysis of burden of disease and life expectancy. *The Lancet*, 380(9838), 219–29.

Leon, D.A. (2008). Cities, urbanization and health. *International Journal of Epidemiology*, 37(1), 4–8.

Liao, Y., et al. (2016). Associations of perceived and objectively-measured neighborhood environment attributes with leisure-time sitting for transport. *Journal of Physical Activity and Health*, 13(12), 1372–7.

Loyen, A., van der Ploeg, H.P., Bauman, A., Brug, J., and Lakerveld, J. (2016). European sitting championship: prevalence and correlates of self-reported sitting time in the 28 European Union member states. *PLoS One*, 11(3), e0149320.

Mabry, R.M., Reeves, M.M., Eakin, E.G., and Owen, N. (2010). Evidence of physical activity participation among men and women in the countries of the Gulf cooperation council: a review. *Obesity Review*, 11(6), 457–64.

Matthews, C.E., et al. (2008). Amount of time spent in sedentary behaviors in the United States, 2003–2004. *American Journal of Epidemiology*, 167(7), 875–81.

Mendoza, J.A., Watson, K., Baranowski, T., Nicklas, T.A., Uscanga, D.K., and Hanfling, M.J. (2011). The walking school bus and children's physical activity: a pilot cluster randomized

controlled trial. *Pediatrics*, 128(3), e537–44.
Morris, J.N., Heady, J.A., Raffle, P.A., Roberts, C.G., and Parks, J.W. (1953). Coronary heart-disease and physical activity of work. *The Lancet*, 265(6796), 1111–20.
Morris, J.N., Kagan, A., Pattison, D.C., and Gardner, M.J. (1966). Incidence and prediction of ischaemic heart-disease in London busmen. *The Lancet*, 2(7463), 553–9.
Myers, J., McAuley, P., Lavie, C.J., Despres, J.P., Arena, R., and Kokkinos, P. (2015). Physical activity and cardiorespiratory fitness as major markers of cardiovascular risk: their independent and interwoven importance to health status. *Progressive Cardiovascular Disease*, 57(4), 306–14.
Neuhaus, M., et al. (2014). Reducing occupational sedentary time: a systematic review and meta-analysis of evidence on activity-permissive workstations. *Obesity Review*, 15(10), 822–38.
Ng, S.W., Howard, A.G., Wang, H.J., Su, C., and Zhang, B. (2014). The physical activity transition among adults in China: 1991–2011. *Obesity Review*, **15** (Suppl 1), 27–36.
Ng, S.W., Norton, E.C., and Popkin, B.M. (2009). Why have physical activity levels declined among Chinese adults? Findings from the 1991–2006 China Health and Nutrition Surveys. *Social Science and Medicine*, 68(7), 1305–14.
Ng, S.W. and Popkin, B.M. (2012). Time use and physical activity: a shift away from movement across the globe. *Obesity Review*, 13(8), 659–80.
Oldenburg, B., Sallis, J.F., Harris, D., and Owen, N. (2002). Checklist of Health Promotion Environments at Worksites (CHEW): development and measurement characteristics. *American Journal of Health Promotion*, 16(5), 288–99.
Owen, N., Salmon, J., Koohsari, M.J., Turrell, G., and Giles-Corti, B. (2014). Sedentary behaviour and health: mapping environmental and social contexts to underpin chronic disease prevention. *British Journal of Sports Medicine*, 48(3), 174–7.
Owen, N., Sugiyama, T., Eakin, E.E., Gardiner, P.A., Tremblay, M.S., and Sallis, J.F. (2011). Adults' sedentary behavior determinants and interventions. *American Journal of Preventive Medicine*, 41(2), 189–96.
Pate, R.R., O'Neill, J.R., and Lobelo, F. (2008). The evolving definition of 'sedentary'. *Exercise and Sport Sciences Reviews*, 36(4), 173–8.
Phillips, S.M., Cadmus-Bertram, L., Rosenberg, D., Buman, M.P., and Lynch, B.M. (2018). Wearable technology and physical activity in chronic disease: opportunities and challenges. *American Journal of Preventive Medicine*, 54(1), 144–50.
Physical Activity Guidelines Advisory Committee (2018). *Physical Activity Guidelines Advisory Committee Scientific Report*. Washington, DC: US Department of Health and Human Services.
Popkin, B.M. (2004). The nutrition transition: an overview of world patterns of change. *Nutrition Review*, 62(Suppl_2), S140–3.
Powell, K.E., Paluch, A.E., and Blair, S.N. (2004). Physical activity for health: what kind? How much? How intense? On top of what? *Annual Review of Public Health*, **32**, 349–65.
Prince, S.A., Adamo, K.B., Hamel, M.E., Hardt, J., Connor Gorber, S., and Tremblay, M. (2008). A comparison of direct versus self-report measures for assessing physical activity in adults: a systematic review. *International Journal of Behavior and Nutrition and Physical Activity*, **5**, 56.
Pucher, J. and Buehler, R. (2008). Making cycling irresistible: lessons from the Netherlands, Denmark and Germany. *Transport Reviews*, 28(4), 495–528.
Ramirez Varela, A., et al. (2018). Worldwide use of the first set of physical activity country cards: the Global Observatory for Physical Activity—GoPA! *International Journal of Behavioural Nutrition and Physical Activity*, 15(1), 29.
Reis, R.S., Salvo, D., Ogilvie, D., Lambert, E.V., Goenka, S., and Brownson, R.C. (2016). Scaling up physical activity interventions worldwide: stepping up to larger and smarter approaches to get people moving. *The Lancet*, 388(10051), 1337–48.
Sallis, J.F., et al. (2016). Physical activity in relation to urban environments in 14 cities worldwide: a cross-sectional study. *The Lancet*, 387(10034), 2207–17.

Sallis, J.F. and Owen, N. (2015). Ecological models of health behavior. In: Glanz, K., Rimer, B.K., Viswanath, K. (eds.) *Health Behavior: Theory, Research, and Practice*. (5th ed.). San Francisco, CA: Jossey-Bass, pp. 43–64.

Sallis, J.F. and Saelens, B.E. (2000). Assessment of physical activity by self-report: status, limitations, and future directions. *Research Quarterly for Exercise and Sport*, 71 Suppl 2, 1–14.

Scheepers, C.E., Wendel-Vos, G.C.W., den Broeder, J.M., van Kempen, E.E.M.M., van Wesemael, P.J.V., and Schuit, A.J. (2014). Shifting from car to active transport: a systematic review of the effectiveness of interventions. *Transportation Research Part A: Policy and Practice*, **70**, 264–80.

Sedentary Behavior Research Network (2012). Standardized use of the terms 'sedentary' and 'sedentary behaviours'. *Applied Physiology, Nutrition, & Metabolism*, **37**, 540–2.

Sugiyama, T., Wijndaele, K., Koohsari, M.J., Tanamas, S.K., Dunstan, D.W., and Owen N. (2016). Adverse associations of car time with markers of cardio-metabolic risk. *Preventive Medicine*, **83**, 26–30.

Tipton, C.M. (2014). The history of 'Exercise Is Medicine' in ancient civilizations. *Advances in Physiology Education*, 38(2), 109–17.

Tremblay, M.S., et al. (2017). Sedentary Behavior Research Network (SBRN)—Terminology Consensus Project process and outcome. *International Journal of Behavioral Nutrition and Physical Activity*, 14(1), 75.

US Department of Health and Human Services (2008). *2008 Physical Activity Guidelines for Americans: Be Active, Healthy, and Happy!* [Online] Available at: https://health.gov/sites/default/files/2019-09/paguide.pdf

US Department of Health and Human Services (1996). *Physical Activity and Health: A Report of the Surgeon General*. Atlanta, GA: Centers for Disease Control and Prevention.

van Dommelen, P., Coffeng, J.K., van der Ploeg, H.P., van der Beek, A.J., Boot, C.R., and Hendriksen, I.J. (2016). Objectively measured total and occupational sedentary time in three work settings. *PLoS One*, 11(3), e0149951.

Varela, A.R., et al. (2017). Worldwide surveillance, policy, and research on physical activity and health: the global observatory for physical activity. *Journal of Physical Activity and Health*, 14(9), 701–9.

Vuori, I. (2018). World Health Organization and physical activity. *Progress in Preventive Medicine*, 3(1), e0012.

Warburton, D.E. and Bredin, S.S. (2016). Reflections on physical activity and health: what should we recommend? *Canadian Journal of Cardiology*, 32(4), 495–504.

Warburton, D.E., Charlesworth, S., Ivey, A., Nettlefold, L., and Bredin, S.S. (2010). A systematic review of the evidence for Canada's Physical Activity Guidelines for Adults. *International Journal of Behavioral Nutrition and Physical Activity*, **7**, 39.

Watson, A., Timperio, A., Brown, H., Best, K., and Hesketh, K.D. (2017). Effect of classroom-based physical activity interventions on academic and physical activity outcomes: a systematic review and meta-analysis. *International Journal of Behavioral Nutrition and Physical Activity*, 14(1), 114.

Wilmot, E.G., et al. (2012). Sedentary time in adults and the association with diabetes, cardiovascular disease and death: systematic review and meta-analysis. *Diabetologia*, 55(11), 2895–905.

Woodcock, J., et al. (2009). Public health benefits of strategies to reduce greenhouse-gas emissions: urban land transport. *The Lancet*, 374(9705), 1930–43.

World Health Organization (WHO) (2010). Global recommendations on physical activity for health. Geneva: WHO. Available from: http://whqlibdoc.who.int/publications/2010/9789241599979_eng.pdf

World Health Organization (WHO) (2013). *Global Action Plan for the Prevention and Control of Noncommunicable Diseases 2013–2020*. Geneva: WHO.

World Health Organization (WHO) (2018). *Global Action Plan on Physical Activity 2018–2030: More Active People for a Healthier World*. Geneva: WHO.

Yusuf, S., Reddy, S., Ôunpuu, S., and Anand, S. (2001). Global burden of cardiovascular diseases: part I: general considerations, the epidemiologic transition, risk factors, and impact of urbanization. *Circulation*, 104(22), 2746–53.

8.6

Diabetes mellitus

Farah Naz Khan, Nida Izhar Shaikh, K.M. Venkat Narayan, and Mohammed K. Ali

Introduction

Diabetes: definition, classification, and diagnosis

Definition and classification

Diabetes mellitus is a complex chronic disease characterized by elevated blood sugars. The term *diabetes* has Greek origins and was meant to describe the passing of urine, while *mellitus* stems from the Latin word for sweet. Elevated blood sugars or hyperglycaemia seen in diabetes can be attributed to either impaired insulin production, impaired insulin action, or both (American Diabetes Association 2018). Produced by the beta cells in the pancreatic islets of Langerhans, insulin is the main hormone responsible for regulating glucose. In response to elevated blood sugars after food intake, the pancreatic beta cells secrete insulin, thereby facilitating the utilization of glucose and keeping blood glucose within an acceptable normal range. Prolonged hyperglycaemia over many years can cause damage to both large (macrovascular) and small (microvascular) blood vessels. The macrovascular complications from diabetes include coronary artery disease, cerebrovascular disease, and peripheral arterial disease, and microvascular complications encompass retinopathy (which can lead to vision loss), neuropathy (which can contribute to foot ulcers), and nephropathy (which can culminate in end-stage renal disease). Epidemiological studies examining the sugar levels at which the risk of macrovascular and microvascular complications increase have influenced the thresholds that we use to define and classify hyperglycaemia and diabetes.

The American Diabetes Association (ADA) and the World Health Organization (WHO) provide widely recognized recommendations on the

classification and diagnosis of diabetes (WHO 2006); however, it is important to note that the distinctions between the types are somewhat arbitrary and factors separating them exist on a spectrum as shown in Figure 8.6.1. Notably, those with type 1 diabetes will require insulin for survival, whereas other classifications may potentially be managed with non-insulin therapies. The main categories of diabetes classification that will be discussed are type 1 diabetes, type 2 diabetes, and gestational diabetes.

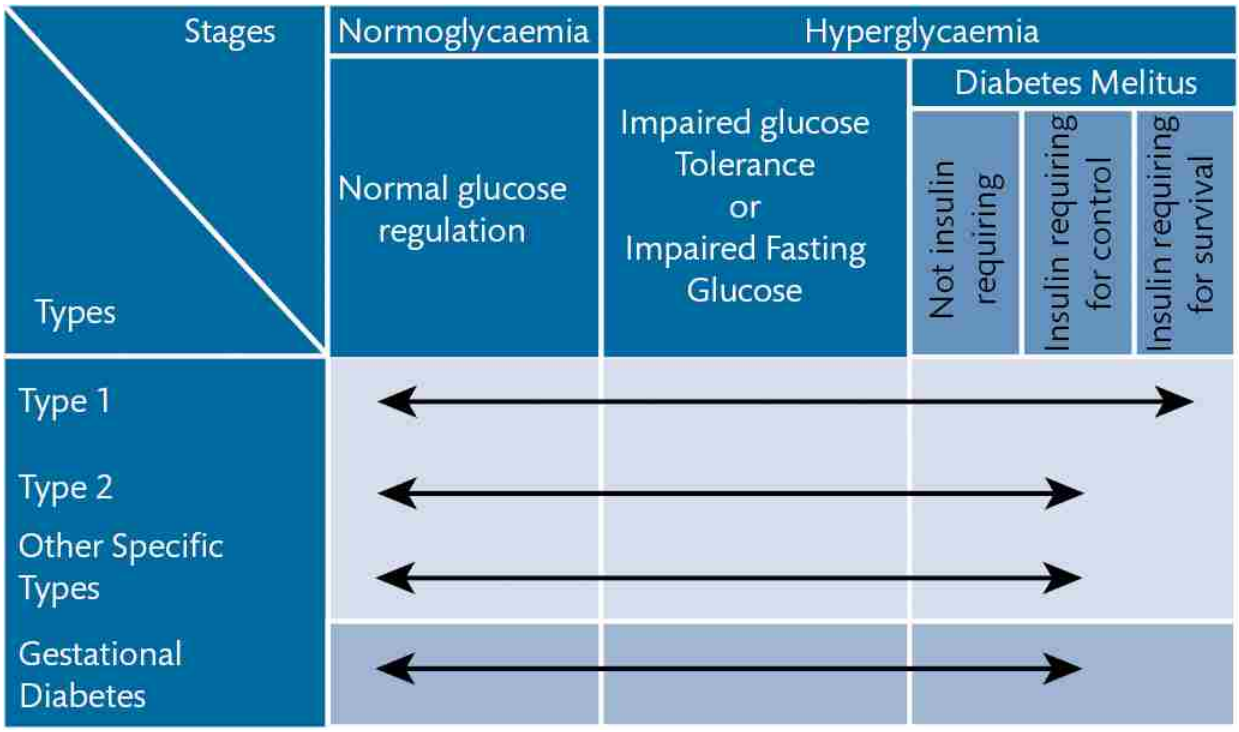


Fig. 8.6.1 Definition and diagnosis of diabetes mellitus and intermediate hyperglycaemia.
Reprinted from World Health Organization and International Diabetes Federation (2006). 'Definition and diagnosis of diabetes mellitus and intermediate hyperglycemia. A report of a WHO/IDF Consultation.' With permission from the WHO. Available from http://www.who.int/diabetes/publications/diagnosis_diabetes2006/en/.

Type 1 diabetes is caused by the destruction of insulin-producing pancreatic beta cells. Most commonly, this is due to autoantibodies against the beta cells, but can been seen in post-pancreatectomy patients and in patients with chronic pancreatic inflammation or injury. Type 1 diabetes accounts for only a minority of diabetes across the world.

A milieu of impaired metabolic functions contribute to the development of type 2 diabetes including: impaired pancreatic insulin secretion, increased hepatic gluconeogenesis, decreased incretin effects, decreased skeletal muscle glucose uptake, increased lipolysis, neurotransmitter dysfunction, and increased renal glucose reabsorption (Defronzo 2009).

Gestational diabetes is similar to type 2 diabetes in that it occurs as a result of insulin resistance and impaired insulin secretion as opposed to the beta cell destruction seen in type 1 diabetes. Diagnosed either in the second or third trimester of pregnancy, gestational diabetes usually resolves after delivery. However, women with gestational diabetes have over 7 times higher risk of developing type 2 diabetes compared to women with normoglycemic pregnancies (Bellamy et al. 2009).

There are also other (relatively) rare and specific forms of diabetes that include: (i) exocrine pancreatic insufficiency related diabetes such a cystic fibrosis-related diabetes; (ii) monogenic diabetes syndromes such as neonatal diabetes and maturity-onset diabetes of the young; and (iii) medication-related diabetes such as steroid-induced diabetes.

Diagnosis

Diagnosis of diabetes relies on blood glucose levels, which can be done to reflect fasting serum glucose levels, 2-hour post-load glucose levels, or a combination of the two. These can be done as part of an oral glucose tolerance test (OGTT). OGTTs detect impaired glucose metabolism as the test measures blood glucose before and then 2 hours after the ingestion of a glucose load. The OGTT is often considered too cumbersome and more costly than using a fasting glucose alone, which is why the ADA prefers the use of fasting plasma glucose or glycated haemoglobin (HbA1c) for diabetes diagnoses. Glycated haemoglobin is a non-fasted test that reflects the average blood glucose over the previous 8–12 weeks (Nathan et al. 2007). This being said, the OGTT remains the best test to detect those with impaired glucose tolerance who have normal fasting blood sugars (Balkau 2000), so it is still favoured as the diagnostic method of choice by the WHO. The latest ADA and WHO criteria for diagnosis of diabetes are summarized in Table 8.6.1 (WHO 2006; American Diabetes Association 2018), which includes the WHO 2011 special consultative report that also endorsed the HbA1c as a diagnostic test (WHO 2011).

Table 8.6.1 ADA and WHO diabetes and non-diabetes hyperglycaemia diagnostic criteria

	ADA	WHO
Diabetes		
Fasting glucose	≥7.0 mmol/L	
2-hour glucose	≥11.1 mmol/L	
HbA1c	> 6.5%	
Impaired fasting glucose		
Fasting glucose	5.6 to 6.9 mmol/L	6.1 to 6.9 mmol/L
2-hour glucose	n/a	
Impaired glucose tolerance		
Fasting glucose	<7.0 mmol/L AND	
2-hour glucose	≥7.8 and <11.0 mmol/L	
Prediabetes		
HbA1c	5.7% to 6.4%	n/a

The diagnostic criteria noted in Table 8.6.1 could all be used individually to diagnose diabetes, impaired fasting glucose, or impaired glucose tolerance. The largest discrepancy is in the criteria for impaired fasting glucose (IFG), where the ADA and WHO differ slightly. The ADA also uses HbA1c values of 5.7% to 6.4% to diagnose prediabetes, which encompasses those who are at increased risk of developing diabetes. While the HbA1c test may seem to be the simplest tool given its ability to give average blood glucose over time with a simple non-fasting blood test, it was initially not recommended by the WHO as a diagnostic test (WHO 2006). The WHO's main concerns involved cost and availability, international laboratory consistency, and HbA1c's variability with health conditions that affect red blood cell turnover, as is the case for sickle cell anaemia. These concerns remain valid, and should be kept in mind when using HbA1c as diagnostic tool across populations.

Diabetes: worldwide prevalence and trends

The burden of diabetes across the world has been steadily increasing in both men and women since 1980 (NCD Risk Collaborators 2016). Over time, diabetes prevalence has increased in every country studied. This increase in diabetes prevalence was most strongly associated with ageing and growth of populations, and to some extent, to changes in lifestyles (NCD Risk Collaborators 2016).

The International Diabetes Federation (IDF) has tracked diabetes prevalence for nearly two decades. In the most recent Diabetes Atlas, the IDF estimates that there are 463 million adults age 20–79 years who have diabetes. Unfortunately, almost half of those with diabetes are thought to be unaware of their diagnosis. Figure 8.6.2 illustrates the millions of adults across the world living with diabetes. This population is expected to grow to 700 million by the year 2045. Additionally, hyperglycaemia in pregnancy is estimated to affect 16.2% or 21.3 million live births across the world (Cho et al. 2018).

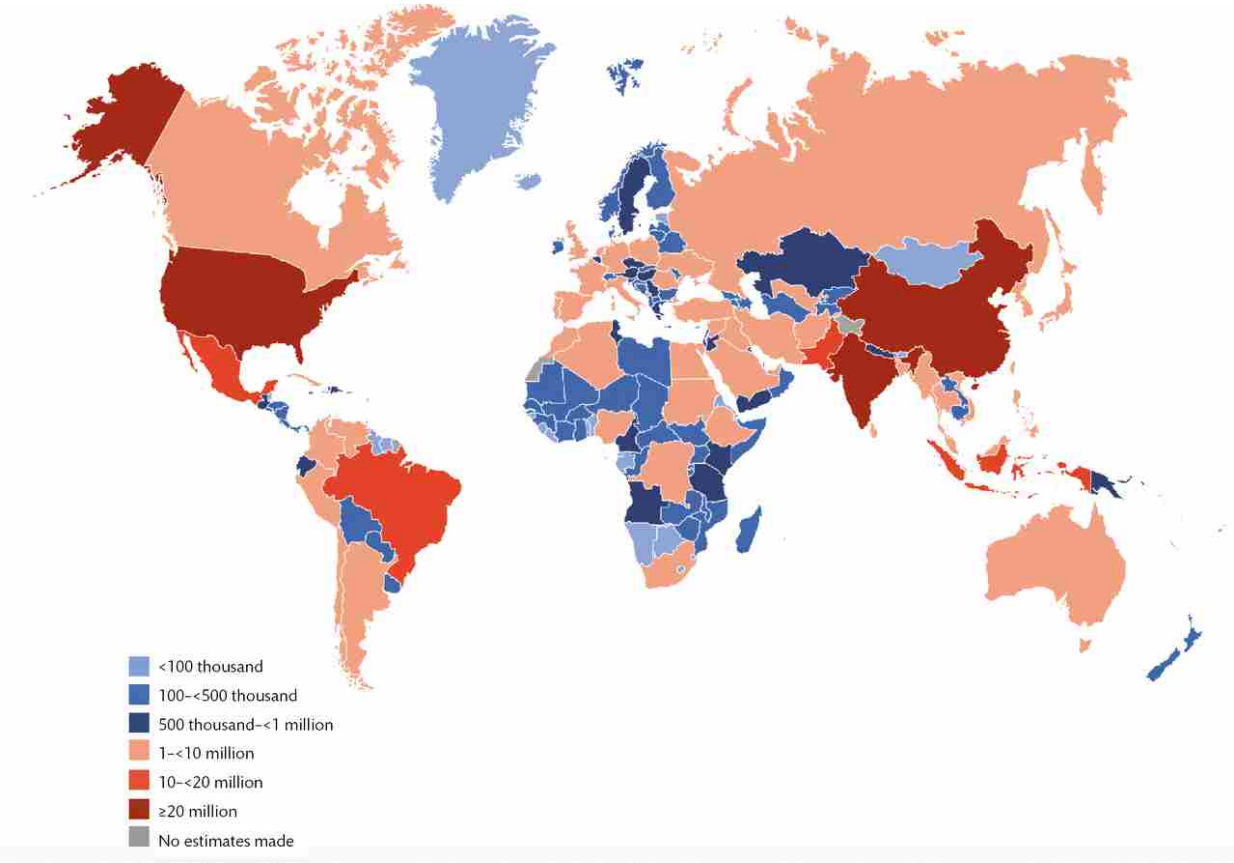


Fig. 8.6.2 Estimated total number of adults living with diabetes, 2017.
Reproduced from International Diabetes Federation, Diabetes Atlas, Ninth Edition, Map 3.1, International Diabetes Federation, Brussels, Copyright © 2019.

Of note, women have slightly lower prevalence than men and the burden of diabetes is most prevalent in the 65–79 year age group (Fig. 8.6.3). The prevalence of diabetes is higher in urban compared to rural environments (10.2% vs. 6.9%). North America, the Middle East and North Africa, and Southeast Asia have the highest diabetes prevalence currently. These trends and findings are also corroborated by other large comprehensive analyses as shown in Figure 8.6.4 (NCD Risk Collaborators 2016).

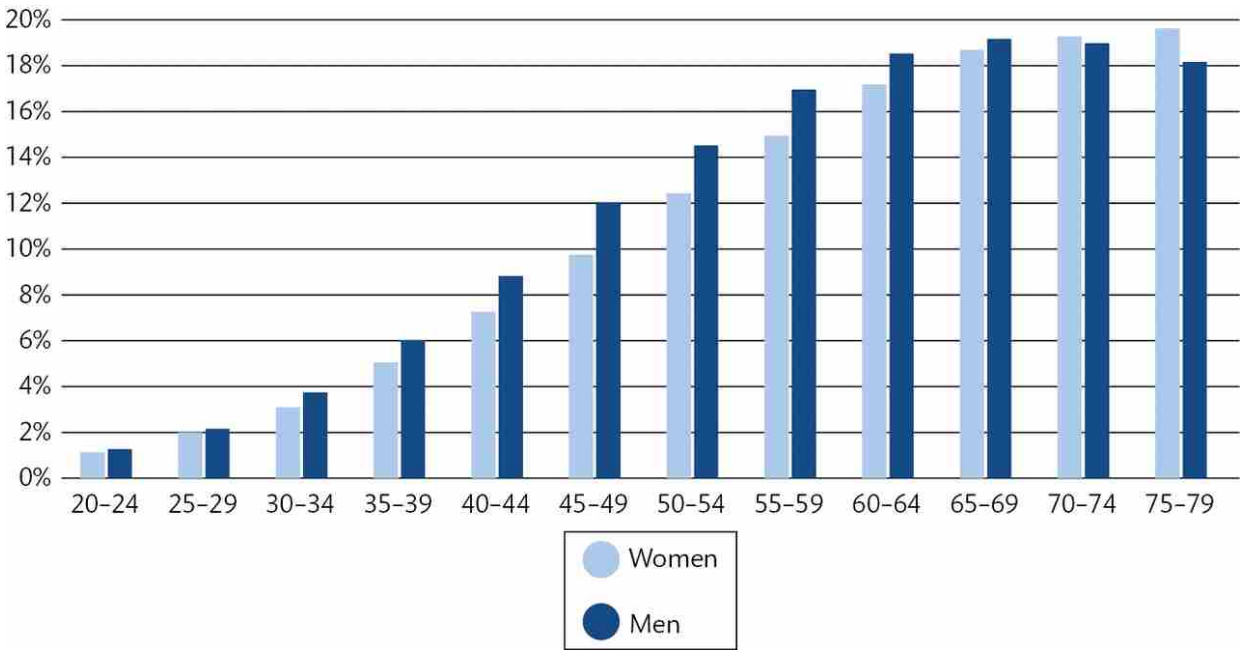


Fig. 8.6.3 Prevalence of people with diabetes by age and sex, 2017.
Reproduced from International Diabetes Federation, Diabetes Atlas, Eighth Edition, Figure 3.1, International Diabetes Federation, Brussels, Copyright © 2017, available from http://diabetesatlas.org/resources/2017-atlas.html.

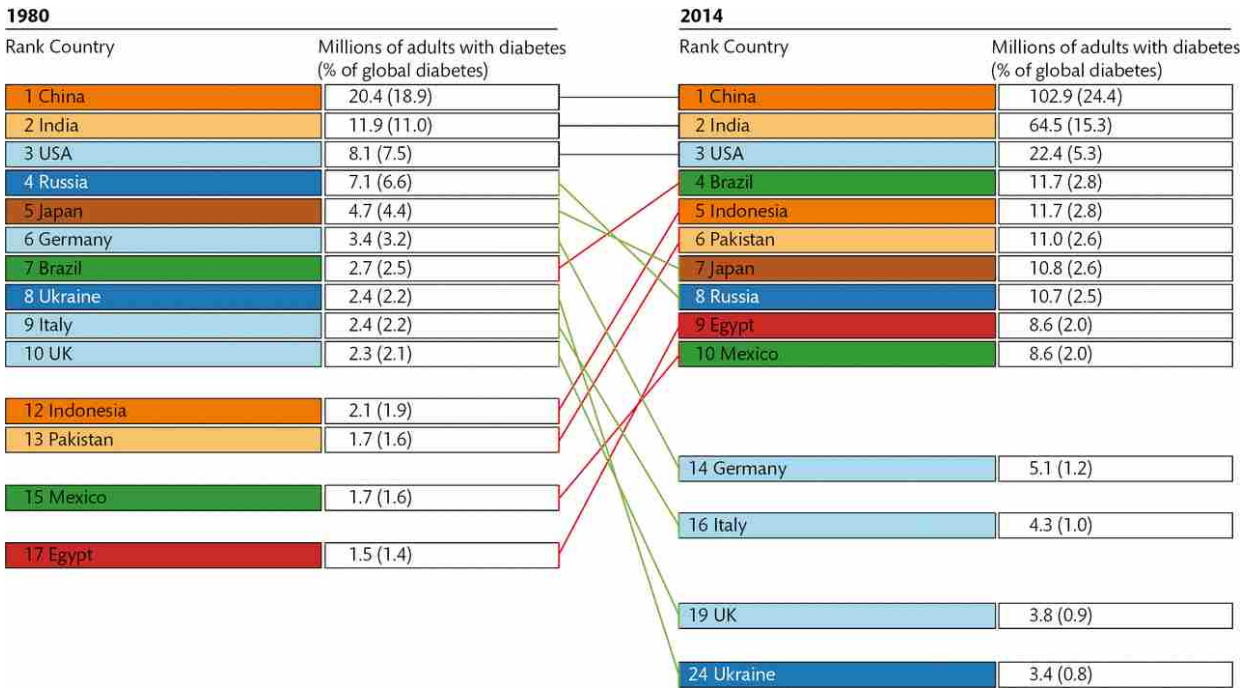


Fig. 8.6.4 Ten countries with the largest number of adults with diabetes in 1980 and 2014.
Reprinted from Zhou B. et al (NCD Risk Factor Collaboration, NCD-RisC) (2016) Worldwide trends in diabetes since 1980: a pooled analysis of 751 population-based studies with 4·4 million participants. *The Lancet*. 387:10027, 9–15 Pages 1513–30 under a CC BY 4.0 license.

Over 75% of the IDF's estimated 463 million people with diabetes live in low- and- middle-income countries (LMICs). The IDF and NCD-RiSC Collaboration show that the largest and fastest increases in diabetes prevalence in the coming years will occur in LMICs (NCD Risk Collaborators 2016; Cho et al. 2018).

Contributions of diabetes to global burdens of disease

Disease burdens are most commonly evaluated in terms of health, economic, and social impacts. In terms of health burdens, diabetes-related complications can be acute where extreme fluctuations in blood glucose require urgent medical care to sustain life, or can be chronic when diabetes has been poorly controlled for long periods of time. Chronic complications are most commonly characterized by disruptions in blood flow (due to damage to the vasculature) to the heart, brain, eyes, kidneys, or limbs, but can also include mental health and infectious disease outcomes.

The risk of cardiovascular disease in people with diabetes is two- to fourfold higher than in people without diabetes (Stamler et al. 1993; Folsom et al. 1997), and this accounts for much of the increased mortality associated with diabetes (Roglic et al. 2005). Diabetes is the leading cause of renal failure in high-income countries (HICs), responsible for 40–50% of all new patients requiring dialysis in United States and Canada, and 15–33% in Europe and Australia (Atkins 2005). The nerve damage associated with diabetes can affect both peripheral and autonomic nerves. Peripheral nerve damage and disruption in peripheral vascular blood flow lead to diabetic limb and foot disease, which can culminate in amputations if untreated (Edmonds et al. 1996). In HICs, diabetes is the single most important cause of non-traumatic lower limb amputation, accounting for 40–60% of all amputations (Unwin 2000), and people with diabetes have a 15-fold higher risk of amputation compared to people without diabetes (Edmonds et al. 1996). Individually and together, these conditions result in considerable morbidity, decreased quality of life, higher use and costs of healthcare, and increased mortality.

Notably, while most other causes of mortality are declining worldwide, diabetes-related mortality has remained flat and/or increased (Ali et al. 2015). Annually, diabetes contributes to approximately 5 million (8.4%) of all deaths, and LMICs appear to shoulder nearly four-fifths of these deaths (Cho et al. 2018).

In addition to premature mortality leading to lost productivity, diabetes is a costly condition requiring long-standing and progressively more healthcare visits and medical treatments. Currently, it is estimated that diabetes results in nearly US$1 trillion direct medical expenditures annually. Diabetes also results in higher and growing non-medical costs for people with diabetes and their caregivers worldwide. At an individual level, the costs of care increase dramatically with the presence of complications; being two- to threefold higher with the presence of either micro- or macrovascular complications, and five- to sixfold higher with the presence of both micro- and macrovascular complications than in people with diabetes without complications (Williams 2005). Diabetes, as with any other chronic illness, may limit employment and impact upon the income of people with it and the people caring for them. In 2007, the impact of diabetes on gross domestic product (GDP) was estimated for several countries; the United Kingdom lost

0.4% of GDP, the US 1.2%, and India 2.1% (Economist Intelligence Unit 2007).

Efforts to characterize the social impacts of diabetes have noted that this disease interferes with societal roles and functioning, increases the need for long-term care (at home or in facilities) to support people with physical deficits of diabetes like amputations or need for dialysis, and leads to mental health manifestations such as depression, cognitive dysfunction, impaired social integration, and poorer quality of relationships. These costs to society, along with their associated psychological harms, are extremely hard to quantify tangibly.

Challenges in different settings—LMICs vs. HICs

Compared to HICs, LMICs often have worse health outcomes in the face of limited access to routine care, medications (e.g. insulin), and infrastructural resources (e.g. refrigeration for insulin) that can be lifesaving (Sobngwi et al. 2012). With regard to diabetes-related complications, there have been large reductions in classic complications of type 2 diabetes, predominantly reductions in myocardial infarctions, stroke, amputations, and mortality in HICs (Gregg et al. 2016). However, trends in burdens of diabetes-related complications in LMICs have been unclear (Gregg et al. 2016). Diabetes and diabetes-related complications are seen more commonly among youth and adults ages 15–69 years in LMICs versus among adults ≥65 years in HICs.

While LMICs shoulder disproportionate burdens of diabetes, research into diabetes remains concentrated in a few HICs (Narayan 2016). The discordance in research outputs and research capacity between HICs and LMICs has implications for policies and decision-making. Research capacity in LMICs is insufficient and often fragmented. Locally led diabetes research in LMICs is critical to the proposal of culturally suitable and cost-effective interventions. Additionally, local research in LMICs will help better investigate obstacles that prevent recommended strategies from being implemented.

Phenotypic differences

People with either IFG (higher-than-normal fasting blood glucose levels) or impaired glucose tolerance (IGT, higher-than-normal post-prandial blood

glucose levels) or both are all considered to have prediabetes despite their potential phenotypic and physiologic differences. The IDF estimates that 318 million people (6.7% of the adult population) worldwide have some form of IGT, but the IDF estimations do not include individuals with isolated IFG (iIFG) or fully capture those with combined IFG-IGT (International Diabetes Federation 2013). Estimates regarding the distribution of the subtypes of prediabetes worldwide are not available because the blood tests required to diagnose prediabetes are relatively expensive and logistically inconvenient to administer (Ali et al. 2017). This lack of data has major implications for understanding the natural history of different types of diabetes and implementing appropriate interventions.

The aetiologies, clinical presentations, diagnoses, treatments, and prevention of these different phenotypes of diabetes and prediabetes differ (Ahlqvist et al. 2018). As an example, when comparing Pima Indians vs. Asian Indians (Narayan 2016), the Pimas represent an obese, highly insulin-resistant phenotype, unlike South Asians who exhibit a relatively thin insulin secretory deficit phenotype. Studies examining the South Asian phenotype suggest that generations of undernutrition may underlie a greater propensity to fat storage, insulin resistance, and early problems with insulin secretion (Enas et al. 2007; Unnikrishnan et al. 2014; Narayan 2016). This geno- or epigenetic vulnerability is then triggered by lifestyle choices—of note, South Asians exhibit a high prevalence of smoking and heavy alcohol use, high intake of refined carbohydrates (e.g. white rice), and decreased physical activity levels (Hu 2011). Despite the influence of lifestyle factors, studies examining the effectiveness of lifestyle modification among people with thin, metabolically unhealthy profiles have shown modest benefits at reducing diabetes incidence (Absetz et al. 2007; Weber et al. 2016). This may be because early β-cell dysfunction is a more dominant pathophysiology in these subgroups, and this phenomenon is quite common in populations of emerging economies (Chan et al. 2004; Maskarinec et al. 2009; Gujral et al. 2013; Mohan et al. 2013; Gujral et al. 2014; Staimez et al. 2013; Sattar and Gill 2015). South Asian populations appear to have a higher susceptibility for poor β-cell function compared to their counterparts in Europe (Ikehara et al. 2015), US (Kanaya et al. 2014), and Netherlands (Jainandunsing et al. 2015). However, most of the published data are from cross-sectional studies, and investments in longitudinal multiethnic studies are needed to establish causality and mechanism (Narayan 2016).

The impacts of diabetes also vary by ethnicity. For example, South Asians are five to nine times more likely than other racial/ethnic groups to exhibit dyslipidaemia (characterized by low HDL cholesterol and high triglycerides), even at a healthy BMI (Patel et al. 2016; Gujral et al. 2018). In the United States, people with diabetes of African, Asian (Chinese and Japanese), and Hispanic origin have substantially lower rates of myocardial infarction than those of European origin, and those of the latter two groups have lower rates of stroke. Similar findings were seen for people of African origin in the United Kingdom (Davis 2008). People with diabetes of South Asian origin are at similar or higher risk of cardiovascular disease than those of European origin (Gholap et al. 2011). These differences are not fully explained by differences in conventional risk factors.

Studies have reported marked differences by ethnic group in the risk of end-stage renal disease, with people of African and Asian origin in particular being at greater risk compared to those of European origin (Karter et al. 2002; Lanting et al. 2005). A high risk of diabetes-related renal disease has also been described in Asia (Chan et al. 2009). Differences exist in the risk of lower extremity amputation by ethnic group, for example: people of Asian origin in the United States and United Kingdom having a markedly lower risk than the European population (Lanting et al. 2005; Chan et al. 2009).

Unfortunately, the relative contributions of healthcare access, health behaviours, and genetic susceptibility to these ethnic group differences are not completely understood (Karter et al. 2002; Lanting et al. 2005).

Differences in access and care

In HICs, people typically access diabetes care through insurance-financed health care, whereas in LMICs, the majority self-pay to get access to care, thereby often impoverishing individuals and families. Varied access and low-resource settings affect care and outcomes for diabetes. In LMICs, resources are scarce in the face of rapidly increasing need. Low-resource settings face numerous challenges in delivering diabetes care and education such as limited access to care; cultural, social, and financial barriers; and lower literacy levels.

There are large gaps in care in both HICs and LMICs. With the growing availability of mobile phones and access to internet worldwide, not only do HICs, but also LMICs have the opportunity to benefit from SMS-based services that can address the problems of access to diabetes care. Even still,

data from Africa, Asia, Europe, Latin America, and the United States show that more than half of the people with diabetes are not meeting recommended targets to prevent complications (Gakidou et al. 2011; Sobngwi et al. 2012; Ali et al. 2013; Stone et al. 2013). Another study from eight countries in Europe showed that although 80–98% of people with diabetes had recorded values for glycated haemoglobin, blood pressure, and lipid values in the past year, few had achieved the recommended values (54%, 19%, and 55%, respectively) (Stone et al. 2013). In LMICs, there are almost no data representative of the entire population (Narayan et al. 2012). A study of six African countries reported that 47% of all respondents with diabetes had glycated haemoglobin levels recorded in the past year, 29% had glycated haemoglobin <6.5%, and only 13% and 65% were treated with lipid-lowering or antihypertensive medications, respectively (Sobngwi et al. 2012). In general, there is wide variation in the way different healthcare facilities, regions, and countries monitor, record, and report quality of care for diabetes across the world.

There are also large gaps in prevention of diabetes worldwide. In the United States, it is estimated the 90% of 84 million people with prediabetes who would be eligible for preventive services (Li et al. 2012) are not aware of their status and therefore not engaged in diabetes prevention programmes. Data from LMICs are lacking and we do not know the proportion of people who are at high risk for diabetes and are aware of their prediabetes status as glycated haemoglobin tests are costly and OGTTs are time consuming and disliked by patients.

Socioeconomic gradients

The prevalence and impacts of diabetes vary across geographical and socioeconomic contexts defined by rural versus urban, education levels, occupation, and income. Importantly, the patterns of these gradients appear to be at different stages in their evolutions in HIC and LMIC settings.

In LMICs, populations with higher education, those experiencing rapid industrialization, urbanization, international trade, migration from rural to urban centres, and lifestyle changes including diets high in calories, reduced physical activity, and tobacco use are experiencing faster and higher diabetes burdens (Robbins et al. 2005; Willi et al. 2007; Ramachandran et al. 2010; Anjana et al. 2011; Hu 2011; Lee et al. 2012; Xu et al. 2013; Dagenais et al. 2016).

In HICs, the relationship between socioeconomic status and diabetes appears to be reversed, with the poor at greatest risk and most burdened. For instance, strong associations have been observed between poverty, low education, and type 2 diabetes (Robbins et al. 2001; Krishnan et al. 2010; Sims et al. 2011). A recent meta-analysis of 23 case–control and cohort studies that examined SES–diabetes associations using 43 measures found an overall higher risk for type 2 diabetes among low SES groups based on education, occupation, and income (Agardh et al. 2011). The strength of the association, however, was less consistent in LMICs, where few studies have been conducted (Corsi and Subramanian 2012).

The changing face of diabetes globally

Diabetes remains a vast and significant global health problem. However, the epidemiology and burdens associated with diabetes are evolving as we learn more, implement prevention programmes, modify our approach to care delivery, and as other influential contextual factors change (e.g. in social, political, and economic circumstances). We will focus on a few key trends that are changing the face of diabetes globally.

First, the prevalence and incidence of type 2 diabetes in younger adults have been on the rise. Increasing incidence of type 2 diabetes in youth has been noticed over the past several decades, and this trend follows a growing burden of obesity in younger generations (Dabelea 2018), particularly so in minority and indigenous groups. The risk factors for young-onset type 2 diabetes share similarities with the predisposing risk factors found in those with older-onset—namely, adiposity and metabolic factors and family history, but additional factors, such as maternal obesity and diabetes during pregnancy or undernutrition *in utero* may play a strong role. The implication of earlier onset, prior to the age of 40, is that this increased exposure to diabetes leads to increased risk of long-term microvascular and macrovascular complications. Some evidence also suggests that younger onset type 2 diabetes may have a more aggressive presentation that results in earlier development of diabetes complications (Lascar et al. 2018); this has been shown in South Asian populations as well (Amutha et al. 2017). This calls for focused research in young type 2 patients to identify more effective means of preventing and treating early onset of diabetes.

Second, with greater access to effective diabetes care, lifespans of those with diabetes are also increasing (Gregg et al. 2016). Living longer with this chronic disease has three important implications:

The profile of diabetes complications is changing. Diabetes complications have traditionally been dominated by vascular diseases of the coronary, cerebrovascular, and peripheral vasculature; however, with successes in preventing and treating these macrovascular diseases and expanding life expectancy, there is now emergence of additional complications like cancers, infections, and mental health disorders in the ageing diabetes patient population. Furthermore, successes have not extended to declines in renal disease, and growth of this complication expands global financial and health impacts of diabetes.

There is a growth of multimorbidity in people with diabetes. Most adults with diabetes have at least one other chronic disease, and some estimates suggest, with longer life expectancy, as many as 40% of adults with diabetes would have accrued at least three other chronic diseases (Piette and Kerr 2006; Teljeur et al. 2013), which is often defined as ‘multimorbidity’. Multimorbidity complicates management of diabetes. For those caring for people with diabetes, additional comorbidities can often complicate the choice of treatments as some medications have deleterious effects on other organ systems. Furthermore, additional comorbidities come with added financial costs, as well as physical, mental, and emotional tolls that drastically affect self-care for many adults with diabetes. Together, more complex management, polypharmacy, and poorer self-care all increase risks of disability and mortality (Ekundayo et al. 2009), and increase the economic and social impacts of diabetes.

With continued differences in access to, and quality of delivery of effective prevention and management interventions, diabetes and its associated comorbidities have the potential to widen disparities in health even more, both across and within countries. This calls for policy action to identify ways to scale up effective prevention and treatments for diabetes, or suffer downstream population impacts of ‘unnatural’ selection.

Can the curve be bent? What can be done?

Once a disease of HICs, type 2 diabetes is now a global public health crisis that has spread to every country in the world. The dynamics of the diabetes epidemic are changing rapidly. We have identified three areas to address the diabetes epidemic.

Invest in research to understand and treat the different phenotypes of diabetes

Recent advances have substantially improved our understanding of the pathophysiology of diabetes, but the currently identified genetic susceptibility loci are insufficient to explain differences in diabetes risk across different ethnic groups or the rapid rise in diabetes prevalence over the past several decades (Hu 2011). An important area for future investigation of phenotypes relevant to type 2 diabetes prevention and treatment is the role of innate or early problems with insulin secretion. It is time to consider the pathophysiology of type 2 diabetes across the spectrum, from those dominantly affected by insulin resistance (type 2A) to those dominantly affected by insulin secretion (type 2B) and to tailor measures of prevention for different subpopulations based on the primary problem (Narayan 2016). For instance, the Asian Indian type 2 diabetes phenotype may involve greater innate insulin secretion problems, both in the fasting state and in response to glucose challenges. This has implications for both prevention and treatment.

Another area for future investigation is in diabetes translational research that is aimed at translating results from trials into everyday healthcare and community settings (Ali et al. 2012; Johnson et al. 2013). While studies including epidemiologic and randomized clinical trials have shown that type 2 diabetes is preventable through lifestyle (diet and physical activity) modifications, translating these findings into practice requires fundamental changes in public policies, social environments, and health systems. Diabetes translation issues are diverse and complicated, but identifying and understanding barriers and facilitators to diabetes translation in different settings is needed (Garfield et al. 2003). Another issue is the need for sustainable interventions—in particular how to sustain an intervention once the intensity and excitement of the study are over (Boothroyd et al. 1994; Haw et al. 2017).

More research is needed to understand how to better implement and sustain evidence-based diabetes care in the real world and that addresses important knowledge gaps, especially regarding implementation and scalability (Ali et al. 2017). There is a need to identify highest priority interventions, particularly to detect, prevent, and manage diabetes in LMICs (Ali et al. 2017). We need to identify research that explores the best life stage for diabetes prevention, better understand early life factors that predispose individuals and subpopulations to obesity and type 2 diabetes (Cunningham et al. 2014), and is directed at improved understanding of the potential role of early life determinants (e.g. the maternal environment) to help shape disease prevention programmes (Zimmet et al. 2014).

Lastly, it is also time for research that addresses the disconnect between where most of the research is done (HICs) and where most of the diabetes burden lies (LMICs). Overall, longitudinal data from LMICs are needed to understand the natural history of diabetes and whether different ethnic, genotypic, or phenotypic variants of prediabetes and diabetes progress differently.

Implement what we know

Prevention: To curb the escalating diabetes epidemic, investing in research to understand primary prevention and phenotypic treatment differences should be a global policy priority (Hu 2011). Efforts at primary prevention of diabetes can include targeted screening of people with risk factors to identify prediabetes and diabetes (Ali et al. 2017) and engaging in diabetes prevention strategies for those at risk for diabetes. Evidence suggests lifestyle modification and medications successfully reduced diabetes incidence and comorbidities of diabetes. And although scale up efforts are underway in several countries including Finland (Saaristo et al. 2010), United States (Albright and Gregg 2013), India (Ramachandran et al. 2006), and Australia (Dunbar et al. 2014), more needs to be done to recruit at-risk persons to join and complete these programmes.

Management: To improve the management of prediabetes and diabetes, and lower the risk of complications, we need to address patient-level, provider-level, and system-level barriers that block achievement of care goals (Ali et al. 2017). One proposed idea is to use metformin for people

with prediabetes to prevent diabetes incidence but this is not yet an approved indication for the drug (Moin et al. 2015).

Collaborative care to treat multimorbidities: Collaborative care models have shown some promise in reducing the physical and mental burdens associated with this multimorbidity. Guidelines often focus on individualized treatment of diabetes while efforts to scale tend to focus on standardized approaches. More research is needed to determine whether standardized versus individualized care can most help those with diabetes across the world (Coventry et al. 2015), but likely a careful approach that marries these two entities will be key in assisting people with diabetes to age healthily.

Invest in policies and models to narrow disparities

Studies to date have noted obesity-associated type 2 diabetes in youth, often in minority and indigenous groups (Cunningham et al. 2017b). The only way to curb this growing problem of younger onset type 2 diabetes may be to prevent obesity through lifestyle promotion and integrated cultural change. Research, programmes, and policies aimed at increasing physical activity and making healthier food options more affordable than calorie-dense fast foods are much needed. These cultural changes are of particular importance given the notion of social networks. The concept of the impact of social relationships on health and the so-called social spread of chronic disease was first described in early social network conceptual modelling work that was done in the 1970s (Berkman 2000). Over the years, the impact of social networks on health and mortality have been further analysed and examined, and the impact of social networks on the spread of obesity was described in 2007 (Christakis and Fowler 2007). A subsequent study published several years later found an association between incident type 2 diabetes and social contacts with diabetes or obesity. Further stratification of the results found that incident type 2 diabetes was associated with diabetes in siblings and obesity in spouses—and these associations led to higher individual incident diabetes risk in those individuals with high diabetes genetic risk scores (Raghavan et al. 2016; Cunningham et al. 2017a). More research into social networks and their contribution to the spread of type 2 diabetes and obesity may help elucidate solutions to help narrow disparities across patient populations. Ultimately, all programmes and policies need to recognize that

there are large disparities due to poverty, growing worldwide urbanization, and variable access to healthcare.

Conclusions

Diabetes is a burdensome, chronic, costly, and common disease. Diabetes burdens are felt in every corner of the globe and are evolving. The medical, scientific, and public health communities have made important strides towards improving the detection, prevention, and management of diabetes, but the agenda is still incomplete. There are still many unknowns in our understanding of what works for different phenotypes of diabetes, large gaps in implementation of what we know, and wide disparities in access and engagement with known interventions. Ultimately, given the multifactorial nature of diabetes, addressing and preventing its burdens requires concerted efforts at policy, programme, clinical, and individual levels (Hill et al. 2013).

REFERENCES

Absetz, P., et al. (2007). Type 2 diabetes prevention in the 'real world': one-year results of the GOAL Implementation Trial. *Diabetes Care*, 30(10), 2465–70.

Agardh, E., Allebeck, P., Hallqvist, J., Moradi, T., and Sidorchuk, A. (2011). Type 2 diabetes incidence and socio-economic position: a systematic review and meta-analysis. *International Journal of Epidemiology*, 40(3), 804–18.

Ahlqvist, E., et al. (2018). Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables. *The Lancet Diabetes & Endocrinology*, 6(5), 361–9.

Albright, A.L., and Gregg, E.W. (2013). Preventing type 2 diabetes in communities across the U.S.: the National Diabetes Prevention Program. *American Journal of Preventive Medicine*, 44(4 Suppl 4), S346–51.

Ali, M., et al. (2017). Chapter 12. Diabetes: an update on the pandemic and potential solutions cardiovascular, respiratory, and related disorders. In: Prabhakaran, D., et al. (eds.) *Disease Control Priorities* (3rd ed.) (Vol. 5). Washington, DC: World Bank.

Ali, M.K., Bullard, K.M., Saaddine, J.B., Cowie, C.C., Imperatore, G., and Gregg, E.W. (2013). Achievement of goals in U.S. diabetes care, 1999–2010. *New England Journal of Medicine*, 368(17), 1613–24.

Ali, M.K., Echouffo-Tcheugui, J., and Williamson, D.F. (2012). How effective were lifestyle interventions in real-world settings that were modeled on the Diabetes Prevention Program? *Health Affairs (Millwood)*, 31(1), 67–75.

Ali, M.K., Jaacks, L.M., Kowalski, A.J., Siegel, K.R., and Ezzati, M. (2015). Noncommunicable diseases: three decades of global data show a mixture of increases and decreases in mortality rates. *Health Affairs*, 34(9), 1444–55.

American Diabetes Association (2018). 2. Classification and diagnosis of diabetes: standards of medical care in Diabetes—2018. *Diabetes Care*, 41(Suppl 1), S13–S27.

Amutha, A., et al. (2017). Incidence of complications in young-onset diabetes: Comparing type 2 with type 1 (the young diab study). *Diabetes Research and Clinical Practice*, **123**, 1–8.

Anjana, R.M., et al. (2011). Prevalence of diabetes and prediabetes (impaired fasting glucose and/or impaired glucose tolerance) in urban and rural India: phase I results of the Indian Council of Medical Research-INdia DIABetes (ICMR-INDIAB) study. *Diabetologia*, 54(12), 3022–7.

Atkins, R.C. (2005). The epidemiology of chronic kidney disease. *Kidney International Supplement*, (*94*), S14–18.

Balkau, B. (2000). The DECODE study. Diabetes epidemiology: collaborative analysis of diagnostic criteria in Europe. *Diabetes & Metabolism*, *26*(4), 282–286.

Bellamy, L., Casas, J.P., Hingorani, A.D., and Williams, D. (2009). Type 2 diabetes mellitus after gestational diabetes: a systematic review and meta-analysis. *The Lancet*, 373(9677), 1773–9.

Berkman, L.F. (2000). Social support, social networks, social cohesion and health. *Social Work in Health Care*, 31(2), 3–14.

Boothroyd, P., Green, L.W., Hertzman, C., Lynam, J., and McIntosh, J. (1994). *Tools for sustainability: Iteration and implementation Ecological Public Health: From Vision to Practice*. Toronto: Centre for Health Promotion, University of Toronto, pp. 111–21.

Chan, J.C., et al. (2009). Diabetes in Asia: epidemiology, risk factors, and pathophysiology. *Journal of the American Medical Association*, 301(20), 2129–40.

Chan, W.B., et al. (2004). The associations of body mass index, C-peptide and metabolic status in Chinese Type 2 diabetic patients. *Diabetes Medicine*, 21(4), 349–53.

Cho, N., et al. (2018). IDF Diabetes Atlas: global estimates of diabetes prevalence for 2017 and projections for 2045. *Diabetes Research and Clinical Practice*, **138**, 271–81.

Christakis, N.A., and Fowler, J.H. (2007). The spread of obesity in a large social network over 32 years. *New England Journal of Medicine*, 357(4), 370–9.

Corsi, D.J., and Subramanian, S.V. (2012). Association between socioeconomic status and self-reported diabetes in India: a cross-sectional multilevel analysis. *BMJ Open*, 2(4), e000895.

Coventry, P., et al. (2015). Integrated primary care for patients with mental and physical multimorbidity: cluster randomised controlled trial of collaborative care for patients with depression comorbid with diabetes or cardiovascular disease. *British Medical Journal*, 350, h638.

Cunningham, S.A., Adams, S.R., Schmittdiel, J.A., and Ali, M.K. (2017a). Incidence of diabetes after a partner's diagnosis. *Preventive Medicine*, **105**, 52–7.

Cunningham, S.A., Kramer, M.R., and Narayan, K. (2014). Incidence of childhood obesity in the United States. *New England Journal of Medicine*, **370**, 401–9.

Cunningham, S.A., et al. (2017b). County-level contextual factors associated with diabetes incidence in the United States. *Annals of Epidemiology*, 28(1), 20–5.e2.

Dabelea, D. (2018). Diabetes in youth-looking backwards to inform the future: Kelly West Award Lecture 2017. *Diabetes Care*, 41(2), 233–40.

Dagenais, G.R., et al. (2016). Variations in diabetes prevalence in low-, middle-, and high-income countries: results from the prospective urban and rural epidemiological study. *Diabetes Care*, 39(5), 780–7.

Davis, T.M. (2008). Ethnic diversity in type 2 diabetes. *Diabetes Medicine*, 25 Suppl 2, 52–6.

Defronzo, R.A. (2009). Banting lecture. From the triumvirate to the ominous octet: a new paradigm for the treatment of type 2 diabetes mellitus. *Diabetes*, 58(4), 773–95.

Dunbar, J.A., et al. (2014). Scaling up diabetes prevention in Victoria, Australia: policy development, implementation, and evaluation. *Diabetes Care*, 37(4), 934–42.

Economist Intelligence Unit (2007). *The Silent Epidemic: An Economic Study of Diabetes in Developed and Developing Countries*. [Online] Available at: https://fliphtml5.com/htbs/kmtr

Edmonds, M., et al. (1996). Report of the Diabetic Foot and Amputation Group. *Diabetes Medicine*, 13(9 Suppl 4), S27–42.

Ekundayo, O.J., Muchimba, M., Aban, I.B., Ritchie, C., Campbell, R.C., and Ahmed, A. (2009). Multimorbidity due to diabetes mellitus and chronic kidney disease and outcomes in chronic heart failure. *American Journal of Cardiology*, 103(1), 88–92.

Enas, E. A., Mohan, V., Deepa, M., Farooq, S., Pazhoor, S., and Chennikkara, H. (2007). The metabolic syndrome and dyslipidemia among Asian Indians: a population with high rates of diabetes and premature coronary artery disease. *Journal of Cardiometabolic Syndromes*, 2(4), 267–75.

Folsom, A.R., Szklo, M., Stevens, J., Liao, F., Smith, R., and Eckfeldt, J.H. (1997). A prospective study of coronary heart disease in relation to fasting insulin, glucose, and diabetes. The Atherosclerosis Risk in Communities (ARIC) Study. *Diabetes Care*, 20(6), 935–42.

Gakidou, E., et al. (2011). Management of diabetes and associated cardiovascular risk factors in seven countries: a comparison of data from national health examination surveys. *Bulletin of the World Health Organization*, 89(3), 172–83.

Garfield, S.A., et al. (2003). Considerations for Diabetes Translational Research in Real-World Settings. *Diabetes Care*, 26(9), 2670.

Gholap, N., Davies, M., Patel, K., Sattar, N., and Khunti, K. (2011). Type 2 diabetes and cardiovascular disease in South Asians. *Primary Care Diabetes*, 5(1), 45–56.

Gregg, E.W., Sattar, N., and Ali, M.K. (2016). The changing face of diabetes complications. *The Lancet Diabetes & Endocrinology*, 4(6), 537–47.

Gujral, U.P., Narayan, K.M., Kahn, S.E., and Kanaya, A.M. (2014). The relative associations of beta-cell function and insulin sensitivity with glycemic status and incident glycemic progression in migrant Asian Indians in the United States: the MASALA study. *Journal of Diabetes Complications*, 28(1), 45–50.

Gujral, U.P., Pradeepa, R., Weber, M.B., Narayan, K.M., and Mohan, V. (2013). Type 2 diabetes in South Asians: similarities and differences with white Caucasian and other populations. *Annals of the New York Academy of Sciences*, **1281**, 51–63.

Gujral, U.P., Weber, M.B., Staimez, L.R., and Narayan, K.M.V. (2018). Diabetes among non-overweight individuals: an emerging public health challenge. *Current Diabetes Reports*, *18*(8), 60.

Haw, J. S., et al. (2017). Long-term sustainability of diabetes prevention approaches: a systematic review and meta-analysis of randomized clinical trials. *JAMA Internal Medicine*, 177(12), 1808–17.

Hill, J.O., et al. (2013). Scientific statement: socioecological determinants of prediabetes and type 2 diabetes. *Diabetes Care*, 36(8), 2430–9.

Hu, F.B. (2011). Globalization of diabetes. *Diabetes Care*, 34(6), 1249.

Ikehara, S., et al. (2015). Age trajectories of glycaemic traits in non-diabetic South Asian and white individuals: the Whitehall II cohort study. *Diabetologia*, 58(3), 534–42.

International Diabetes Federation (2013). *Clinical Practice Guidelines*. [Online] Available at: http://www.idf.org/guidelines

Jainandunsing, S., et al. (2015). Failing beta-cell adaptation in South Asian families with a high risk of type 2 diabetes. *Acta Diabetologica, 52*(1), 11–19.

Johnson, M., et al. (2013). Can diabetes prevention programmes be translated effectively into real-world settings and still deliver improved outcomes? A synthesis of evidence. *Diabetes Medicine*, 30(1), 3–15.

Kanaya, A.M., et al. (2014). Understanding the high prevalence of diabetes in U.S. south Asians compared with four racial/ethnic groups: the MASALA and MESA studies. *Diabetes Care*, 37(6), 1621–8.

Karter, A.J., Ferrara, A., Liu, J.Y., Moffet, H.H., Ackerson, L.M., and Selby, J.V. (2002). Ethnic disparities in diabetic complications in an insured population. *Journal of the American Medical Association*, 287(19), 2519–27.

Krishnan, S., Cozier, Y.C., Rosenberg, L., and Palmer, J.R. (2010). Socioeconomic status and incidence of type 2 diabetes: results from the Black Women's Health Study. *American Journal of Epidemiology*, 171(5), 564–70.

Lanting, L.C., Joung, I.M., Mackenbach, J.P., Lamberts, S.W., and Bootsma, A.H. (2005). Ethnic differences in mortality, end-stage complications, and quality of care among diabetic patients: a review. *Diabetes Care*, 28(9), 2280–8.

Lascar, N., Brown, J., Pattison, H., Barnett, A. H., Bailey, C. J., and Bellary, S. (2018). Type 2 diabetes in adolescents and young adults. *Lancet Diabetes Endocrinol*, 6(1), 69–80.

Lee, I.M., Shiroma, E.J., Lobelo, F., Puska, P., Blair, S.N., and Katzmarzyk, P.T. (2012). Effect of physical inactivity on major non-communicable diseases worldwide: an analysis of burden of disease and life expectancy. *The Lancet*, 380(9838), 219–29.

Li, Y., Geiss, L.S., Burrows, N.R., Rolka, D.B., and Albright, A. (2012). Awareness of prediabetes: United States 2005–2010. [Press release] *Morbidity and Mortality Weekly Report (MMWR)*, *62*(11), 209–12.

Maskarinec, G., et al. (2009). Diabetes prevalence and body mass index differ by ethnicity: the multiethnic cohort. *Ethnicity & Disease*, *19*(1), 49–55.

Mohan, V., et al. (2013). Associations of beta-cell function and insulin resistance with youth-onset type 2 diabetes and prediabetes among Asian Indians. *Diabetes Technology & Therapeutics*, 15(4), 315–22.

Moin, T., et al. (2015). Metformin prescription for insured adults with prediabetes from 2010 to 2012: a retrospective cohort study. *Annals of Internal Medicine*, 162(8), 542–8.

Narayan, K.M., Echouffo-Tcheugui, J.B., Mohan, V., and Ali, M.K. (2012). Analysis & commentary: global prevention and control of type 2 diabetes will require paradigm shifts in policies within and among countries. *Health Affairs (Millwood)*, 31(1), 84–92.

Narayan, K.M.V. (2016). Type 2 diabetes: why we are winning the battle but losing the war? 2015 Kelly West award lecture. *Diabetes Care*, *39*(5), 653.

Nathan, D.M., Turgeon, H., and Regan, S. (2007). Relationship between glycated haemoglobin levels and mean glucose levels over time. *Diabetologia*, *50*(11), 2239–44.

NCD Risk Collaborators (2016). Worldwide trends in diabetes since 1980: a pooled analysis of 751 population-based studies with 4.4 million participants. *The Lancet*, 387(10027), 1513–30.

Patel, S.A., et al. (2016). Is the ‘South Asian phenotype’ unique to South Asians?: Comparing cardiometabolic risk factors in the CARRS and NHANES Studies. *Global Heart*, *11*(1), 89–96.e83.

Piette, J.D., and Kerr, E.A. (2006). The impact of comorbid chronic conditions on diabetes care. *Diabetes Care*, 29(3), 725–31.

Raghavan, S., et al. (2016). Incident type 2 diabetes risk is influenced by obesity and diabetes in social contacts: a social network analysis. *Journal of General Internal Medicine*, 31(10), 1127–33.

Ramachandran, A., et al. (2006). The Indian Diabetes Prevention Programme shows that lifestyle modification and metformin prevent type 2 diabetes in Asian Indian subjects with impaired glucose tolerance (IDPP-1). *Diabetologia*, 49(2), 289–97.

Ramachandran, A., Ma, R.C., and Snehalatha, C. (2010). Diabetes in Asia. *The Lancet*, 375(9712), 408–18.

Robbins, J.M., Vaccarino, V., Zhang, H., and Kasl, S.V. (2001). Socioeconomic status and type 2 diabetes in African American and non-Hispanic white women and men: evidence from the Third National Health and Nutrition Examination Survey. *American Journal of Public Health*, 91(1), 76–83.

Robbins, J.M., Vaccarino, V., Zhang, H., and Kasl, S.V. (2005). Socioeconomic status and diagnosed diabetes incidence. *Diabetes Research in Clinical Practice*, 68(3), 230–6.

Roglic, G., et al. (2005). The burden of mortality attributable to diabetes: realistic estimates for the year 2000. *Diabetes Care*, 28(9), 2130–5.

Saaristo, T., et al. (2010). Lifestyle intervention for prevention of type 2 diabetes in primary health care: one-year follow-up of the Finnish National Diabetes Prevention Program (FIN-D2D). *Diabetes Care*, 33(10), 2146–51.

Sattar, N., and Gill, J.M. (2015). Type 2 diabetes in migrant south Asians: mechanisms, mitigation, and management. *Lancet Diabetes Endocrinology*, 3(12), 1004–16.

Sims, M., et al. (2011). The socioeconomic gradient of diabetes prevalence, awareness, treatment, and control among African Americans in the Jackson Heart Study. *Annals of Epidemiology*, 21(12), 892–8.

Sobngwi, E., et al. (2012). Type 2 diabetes control and complications in specialised diabetes care centres of six sub-Saharan African countries: the Diabcare Africa study. *Diabetes Research in Clinical Practice*, 95(1), 30–6.

Staimez, L.R., et al. (2013). Evidence of reduced beta-cell function in Asian Indians with mild dysglycemia. *Diabetes Care*, 36(9), 2772–8.

Stamler, J., Vaccaro, O., Neaton, J.D., and Wentworth, D. (1993). Diabetes, other risk factors, and 12-yr cardiovascular mortality for men screened in the Multiple Risk Factor Intervention Trial. *Diabetes Care*, 16(2), 434–44.

Stone, M.A., et al. (2013). Quality of care of people with type 2 diabetes in eight European countries: findings from the Guideline Adherence to Enhance Care (GUIDANCE) study. *Diabetes Care*, 36(9), 2628–38.

Teljeur, C., Smith, S.M., Paul, G., Kelly, A., and O'Dowd, T. (2013). Multimorbidity in a cohort of patients with type 2 diabetes. *European Journal of General Practice*, 19(1), 17–22.

Unnikrishnan, R., Anjana, R.M., and Mohan, V. (2014). Diabetes in South Asians: is the phenotype different? *Diabetes*, 63(1), 53–5.

Unwin, N. (2000). Epidemiology of lower extremity amputation in centres in Europe, North America and East Asia. *British Journal of Surgery*, 87(3), 328–37.

Weber, M.B., et al. (2016). The stepwise approach to diabetes prevention: results from the D-CLIP randomized controlled trial. *Diabetes Care*, 39(10), 1760–7.

Willi, C., Bodenmann, P., Ghali, W.A., Faris, P.D., and Cornuz, J. (2007). Active smoking and the risk of type 2 diabetes: a systematic review and meta-analysis. *Journal of the American Medical Association*, 298(22), 2654–64.

Williams, R. (2005). Medical and economic case for prevention of type 2 diabetes and cardiovascular disease. *European Heart Journal Supplements*, *7*(Suppl D), D14–D17.

World Health Organization (WHO) (2006). *Definition and Diagnosis of Diabetes Mellitus and Intermediate Hyperglycemia: Report of a WHO/IDF Consultation*. Geneva: WHO.

World Health Organization (WHO) (2011). *Use of Glycated Haemoglobin (HbA1c) in Diagnosis of Diabetes Mellitus: Abbreviated Report of a WHO Consultation*. Geneva: WHO. [Online] Available at: https://www.who.int/diabetes/publications/report-hba1c_2011.pdf

Xu, Y., et al. (2013). Prevalence and control of diabetes in Chinese adults. *Journal of the American Medical Association*, 310(9), 948–59.

Zimmet, P.Z., Magliano, D.J., Herman, W.H., and Shaw, J.E. (2014). Diabetes: a 21st century challenge. *The Lancet Diabetes & Endocrinology*, 2(1), 56–64.

8.7

Public mental health and suicide

Danuta Wasserman and Kristian Wahlbeck

Introduction to public mental health and suicide

As defined by the World Health Organization (WHO), mental health is not just the absence of illness, rather it is conceptualized as a state of well-being, in which the individual realizes his or her own abilities, can cope with the normal stresses of life, can work productively and fruitfully, and is able to make a contribution to his or her community (WHO 2013). Consequently, public mental health is not just about the occurrence and prevention of mental disorders in the population, but it also concerns the promotion of mental health and well-being. Thus, public mental health can be defined as actions aimed at developing the mental health of populations and producing healthy societies.

Public mental health research encompasses: research that describes collective experience, occurrence, distribution, and trajectories of positive mental health, mental health problems, and their determinants; research on mental health promotion and prevention of mental disorders; research on mental health system policies and governance; service delivery; and the organization of mental health services.

Mental health is more crucial today than it has ever been, due to the societal transition into the information society era. The population's mental capital (i.e. cognitive, emotional, and social-skill resources required for role functioning) is a prerequisite for the prosperity of individuals, companies, and societies. Mental health becomes more valuable and more vulnerable, due to the many stressors in information-driven economies. At the same time, in low-income countries, poverty, discrimination, and poor mental health form a vicious circle which becomes an obstacle to economic development (Patel and Thornicroft 2009; Patel et al. 2015).

The huge burden and costs of impaired mental health for individuals, families, society, and the economy call for actions to prevent mental ill-health and promote positive mental health and well-being. This necessitates public mental health actions, not only to treat, but also to prevent impaired mental health and promote positive mental well-being.

Magnitude and costs of mental health problems

The magnitude and costs of mental health and mental disorders is relatively neglected (Jané-Llopis et al. 2011; Wahlbeck 2015), in spite of data showing that mental health is a growing health challenge. The share of burden attributable to mental and behavioural disorders, which is already substantial, will probably increase steadily in the future. The rising burden from mental and behavioural disorders will impose new challenges on health systems worldwide.

The recent update of the Global Burden of Diseases study (Kassebaum et al. 2016) demonstrates the magnitude of the burden of mental health disorders. Disability-adjusted life-years (DALYs) include potential years of life lost, due to premature death and equivalent years of 'healthy' life lost by virtue of being in states of poor health or disability. DALY data indicate that mental disorders constitute a growing share of the global disease burden. From 2005 to 2015, mental and substance abuse disorders showed a worldwide increase in DALYs by 14.9%. A single major source of disability and premature death is major depressive disorder (MDD). From 1990 to 2015, depressive disorders increased in ranking from nineteenth to fifteenth place among leading causes of DALYs worldwide. In order of importance, other main causes of DALYs among mental disorders were anxiety disorders (1.0% of DALYs worldwide), drug use disorders (0.7%), schizophrenia (0.6%), and alcohol use disorders (0.5%) (Kassebaum et al. 2016). Self-harm, which in 2015 ranked number 21 among worldwide causes of DALYs, adds to the total impact of mental and behavioural disorders on population health.

The distribution of DALYs varies considerably between geographical regions. The relative contribution of mental and behavioural disorders is as a rule higher in developed economies. According to the Global Burden of Diseases study, depressive disorders accounted for the eighth leading cause

of DALYs in high-income regions, such as Australasia, Europe, and North America. In many Western economies, up to half of the disability pensions are granted due to mental disorders, especially depression. It is, thus, foreseeable that a continued increase in the standard of living globally will contribute to a continued epidemiological transition towards increasing public health impact of mental disorders. In many of the developed economies, depressive disorders rank third, fourth, or fifth among causes of DALYs (Murray et al. 2012; Kassebaum et al. 2016). In regions with high suicide mortality rates, such as Eastern Europe, Japan, and South Korea, self-harm ranks fifth or sixth among causes of DALYs.

The disabling burden of mental and behavioural disorders is almost the same for males and females, but the major contributing causes are different. The burden of depression is 50% higher for females than males. In contrast, the occurrence of alcohol and drug use disorders in males is nearly seven times higher than that for females, and accounts for almost one-third of the male burden of mental and behavioural disorders. In both low- and middle-income countries, and high-income countries, alcohol use disorders are among the ten leading causes of years lived with disability (YLD). This includes only the direct burden of alcohol dependence and problem use. The total attributable burden of disability, due to alcohol use, is much larger.

The Sustainable Development Goals (SDGs) have the primary target to reduce economic poverty and inequalities, improve environmental stability, as well as reassure education, good health, and well-being, for all genders (United Nations 2017). The SDGs have demonstrated the epidemiological shift from communicable diseases towards non-communicable psychosocial disorders, which constitute a global health challenge. Promotion of population mental health is important, as mental health has vast and global well-being consequences. The WHO Action Plan 2013–2020 on mental health, approved by Member States at the WHO World Health Assembly, is a milestone in acknowledging the importance of mental health and the well-being of entire populations (WHO 2013).

Available data confirm the immense disability burden of mental disorders, not only in terms of DALYs, but also in terms of other indicators, such as loss of work days, work life productivity, and poor quality of life (Wittchen et al. 2011). This is due to the critical combination of high prevalence and high level of associated impairments and disabilities of

mental disorders. Many mental disorders—anxiety, substance use, and somatoform disorders in particular—start early, which typically has adverse effects on further psychological and social development, thereby reducing school and academic achievement, social functioning, and social integration.

Compared to many somatic diseases, costs for mental disorders differ with disproportionately high indirect cost items (e.g. sick days, disability, early retirement, and deaths due to suicide) and relatively low direct costs of healthcare (i.e. low expenditures for diagnostic measures, treatment, and care). Mental disorders are dominant contributors to the global economic burden of diseases. The global cost of mental health conditions in 2010 was estimated at US$2.5 trillion, with the cost projected to surge to US$6 trillion by 2030. About two-thirds of the total cost comes from indirect costs, and the remainder from direct costs. Currently, high-income countries shoulder about 65% of the burden, which is not expected to change over the next 20 years (Bloom et al. 2011).

For instance, in Europe, depressive disorder is a major cause of lost productivity (Wahlbeck 2009). In 2004, economic costs of depression were estimated to be €250 per inhabitant, or €118 billion. In the United Sates, the economic cost of MDD was estimated to be $210.5 billion in 2010 (Greenberg et al. 2015). As mentioned earlier, direct costs account for only a minor part of the total economic burden (Sobocki et al. 2006). A majority of costs, between 65 and 85%, are indirect. Data from Europe indicate that costs for depression are increasing. Moreover, depression causes significant reductions of well-being and in health-related quality of life (Sobocki et al. 2007).

People with mental disorders, including substance use disorders, are at risk of suicide. Ninety% of suicides are estimated to be associated with mental disorders, mostly with mood disorders, like depression, which is associated with 60% of suicides (Mann et al. 2005), and also with alcohol use disorders. However, suicidal behaviour is usually an outcome of complex interactions of socioenvironmental, behavioural, and psychiatric factors. High-risk groups include those with severe somatic illness, the socially disadvantaged, those with recent loss, especially through suicide (Nordentoft 2007), and some migrant groups (Garssen et al. 2006; Westman et al. 2006; Gilliver et al. 2014). People with a history of suicide attempts are especially at higher risk of dying by suicide (Fleischmann et al. 2008).

According to WHO data, suicides (i.e. deaths caused from self-inflicted intentional injuries) amounted to 804,000 in 2012, which was 1.4% of the total world mortality (WHO 2014), ranging from 0.5% in the African region to 1.9% in the Southeast Asian region (Värnik 2012). However, the WHO estimate includes neither injury deaths of undetermined intent nor deaths from unknown causes, which are generally believed to conceal unrecognized suicides. Moreover, the data from all WHO Member States are not reported.

The estimated global suicide rate is 11.4 suicides per 100,000 inhabitants, including 15 suicides per 100,000 for males and 8 suicides per 100,000 for females (WHO 2014). For both males and females, the highest suicide rates are found in Europe, predominantly in Eastern Europe, that is, Lithuania, the Russian Federation, Belarus and, to a lesser extent, Finland, Hungary, and Latvia: a group of countries that share similar historical and sociocultural characteristics, as well as similar alcohol use patterns. Nevertheless, some similarly high rates are found in countries that are quite distinct in relation to these characteristics, such as Cuba, Japan, Sri Lanka, and South Korea. As a whole, the lowest rates are found in Islamic countries (Bertolote and Fleischmann 2009; WHO 2017b). Due to stigmatization and legal sanctions against suicide, underreporting of suicide rates in Muslim countries may partly contribute to the low suicide mortality rates reported (Khan and Syed 2011).

The male suicide rate is higher than the rate of females, with the exception of some Asian countries, where the sex differences for suicide are reversed, and young women are more at risk for suicide than men (WHO 2017a), mainly due to impulsive attempts using pesticides as the mode of attempt (Law and Liu 2008).

Suicide in later life is a global public health problem. Those aged 65 years and above are the demographic group that constitutes the highest suicide rate in most countries reporting suicide statistics to the WHO (Conejero et al. 2018). Non-fatal suicidal acts tend to be less common in this age group (Fässberg et al. 2012).

The epidemiology of adolescent suicide has shown striking changes over the last 100 years, with a steady decline in recent decades. One of the suggested factors explaining this trend is the growing use of antidepressants, especially selective serotonin reuptake inhibitors, in the adolescent population (Bursztein and Apter 2009). Even within countries,

there is significant variability in suicidal ideation and behaviours among adolescents of different ethnic backgrounds. In the United States, American Indian/Alaska Native youth tend to have the highest rates of fatal suicidal behaviour among all ethnic groups (Amitai and Apter 2012; Hern et al. 2014). There are similar findings for Inuit populations in Canada, and the Ethiopian population in Israel, with regards to fatal suicide. As well, the Amerindian population in French Guiana has very high suicide rates, in contrast with other French territories, with the younger age groups of both genders affected (Pacot et al. 2018). All of these youth show a similar pattern of failure to integrate traditional cultures with a modern Western culture (Amitai and Apter 2012). However, it is important to note the problems colonization has created for indigenous peoples. As those who have maintained control over their lands and culture have shown much lower rates of youth suicide, as found in British Columbia, Canada (Chandler and Lalonde 2008).

Suicide rates are higher in many disadvantaged populations groups, such as ethnic or sexual minorities. Family rejection or negative family reaction to an adolescent who is gay, lesbian, or bisexual is associated with an eightfold greater likelihood of attempting suicide compared to adolescents who experienced minimal or no family rejection (Ryan et al. 2009; Cotter et al. 2014). The suicide risk of gay and lesbian youth is amendable; lesbian and gay youths living in areas with school antibullying policies that did not specifically mention sexual orientation as a protected group were nearly two times more likely to have attempted suicide in the past year compared with those living in counties where more districts had these policies (Hatzenbuehler and Keyes 2013).

School-based suicide preventive interventions have shown to be effective in significantly reducing suicide attempts (Zalsman et al. 2016). Emerging research is showing that these interventions are not only efficacious, but can also be cost-effective, such as the Youth Aware of Mental Health (YAM) intervention (Ahern et al. 2018), which was presented as part of the Saving and Empowering Young Lives in Europe (SEYLE) study (Wasserman et al. 2015).

A highly increased risk of suicide after a suicide attempt can be found in people with mental disorders. To reduce risk of attempting further suicides, people with mental disorders need aftercare, especially during the first 2

years after the suicide attempt (Fleischmann et al. 2008; Tidemalm et al. 2008).

During the last 50 years, the epicentre of suicide mortality has moved eastwards from Western Europe to Eastern Europe and now to Asia (Värnik 2012). The biggest contributors to the number of suicides in the world are China and India, where half of the total suicides in the world occur, while South Korea has experienced enormous growth in suicide rates during the past decade (Vijayakumar 2004; Värnik 2012; Jeon et al. 2016).

Mental health policy

In spite of their public health impact, mental health issues were largely neglected in the international public health agenda until the 1990s. In order to successfully introduce mental health issues on the political agenda during that time, a new approach has been developed in Europe. It has been stressed that mental health cannot be reduced to an issue about mental disorders; neither the high prevalence of mental disorders, nor the need for more resources in psychiatry was used as entry points, thus, avoiding the stigma associated with mental disorders. Instead, it has been stressed that mental health is an indivisible part of public health, and that it has a significant impact on countries and their human, social, and economic capital. The aim has been to raise mental health from its professional, organizational, and even political isolation to a broader sphere of public health, that is, to shift the focus from the individual level towards strengthening the mental health approach at the population level. Acknowledging the psychosocial determinants of mental health opens the door for horizontal actions to promote population mental health (Wahlbeck 2011b, Wahlbeck 2015). Durkheim's view that social structures, promoting or hindering integration and regulation of society, have influenced the lives of individuals, and the variation in suicide rates between different groups and time periods (Condorelli 2016). Consequently, Durkheim's empirical observations and theory should be taken into consideration when shaping social policies.

It is now widely accepted that major mental health determinants are modifiable by policy actions. During the last decades, many governments and intergovernmental organizations have developed mental health policies,

as a response to increasing mental health needs in the population. Typical objectives of the policies are increased parity between mental and physical disorders, reduction of stigma, respect for human rights, a shift from hospital-based to community-based service provision and strengthening of mental healthcare capacity in primary care.

The stigma associated with mental disorders may have a negative impact on the willingness of decision-makers to commit to a high-level mental health policy. A clear public health approach, and making mental health promotion one of the core objectives, will highlight the productivity links of mental health, and offers a route to successfully avoiding the stigma and discrimination attached to mental health issues.

Mental disorders are inextricably linked to human rights issues. The stigma, discrimination, and human rights violations that individuals and families affected by mental disorders suffer are intense and pervasive. Most international mental health policy documents build on the human rights principles and stress the need for respectful treatment and empowerment of mental health service users. The United Nations Convention on the Rights of People with Disabilities (CRPD), adopted in 2006, affirms that people with mental health disabilities have the right to full participation and inclusion in society, including the right to live independently, the right to education, and the right to work. The CRPD, and related pressure from the international community, will increasingly put human and fundamental rights issues at the forefront of new regional and national mental health policies across the globe.

In May 2013, the WHO World Health Assembly approved the first WHO Action Plan on mental health, covering the period from 2013 to 2020 (WHO 2013). The plan's four major objectives are to: strengthen effective leadership and governance for mental health; provide comprehensive, integrated, and responsive mental health and social care services in community-based settings; implement strategies for promotion and prevention in mental health; and strengthen information systems, evidence, and research for mental health. The plan sets new directions for mental health, including a central role for provision of community-based care, and a greater emphasis on human rights. It also emphasizes the empowerment of people with mental disabilities, the need to develop strong civil society, and implement health promotion and prevention activities. The document proposes specific actions for Member States, the WHO, and partners. It also

proposes indicators and targets, such as a 20% increase in service coverage for severe mental disorders, and a 10% reduction of the suicide rate in countries by the year 2020, which can be used to evaluate levels of implementation, progress, and impact.

Setting mental health targets in policy action plans may help focus attention and support implementation. Targets should, however, be formulated as health equity targets to support mental health improvement and suicide reduction among the most vulnerable groups or in the most deprived areas. To be meaningful, measurement of targets should avoid masking deterioration in one segment of the population by an improvement in another segment (e.g. an increase of suicides in young people by a decrease in older people) (Wahlbeck 2009).

Stigma and public attitudes

People with mental health problems have to cope with a double problem: first, the symptoms of the mental health problem, and second, with the stigma of having a mental health problem (Rüsch et al. 2005). Stigma is a core concept in understanding the field of public mental health and provision of health services. Stigma has a strong influence on the status of mental health services, their resource allocation, and attractiveness to the workforce. It constitutes a barrier that greatly contributes to low help-seeking behaviour among people with mental health problems, and affects provision of services negatively.

Stigma and stereotypes form the public attitude towards people with mental health problems and psychiatry. Stigma related to mental health problems can be divided into: (1) perceived public stigma, that is, stereotype awareness; (2) personal stigma, that is, stereotype agreement (people's personal beliefs about mental illness); and (3) self-stigma (people's view of their own mental health problem) (Rüsch et al. 2005). Discrimination of people with mental disorders is a common manifestation of stigma.

Stigma studies across the world have generally shown consistent findings. There is a lack of parity between mental and physical disorders; people with mental disorders, as well as the services they are provided, are less valued. This is extremely unfortunate, because there are clear links

between popular understandings of mental illness and whether people in mental distress seek help or feel able to disclose their problems (Littlewood 1998; Clement et al. 2015). The core experiences of shame (to oneself or to one's family) and blame (from others) are common, although they vary between cultures. However, international studies have shown that stigma and discrimination of people with mental disorders is consistently common across cultures (Thornicroft et al. 2009; Lasalvia et al. 2013; Seeman et al. 2016). Some studies have reported higher levels of stigma tend to be found within developing countries than in developed countries (Barke et al. 2011; Semrau et al. 2015). Within Asia, compared to Western countries, people who suffer from mental illnesses are feared, considered dangerous, and should be excluded from the community. Furthermore, many studies from Asia report that mental illness symptoms are believed to have derived from supernatural forces (Lauber and Rössler 2007). Thus, the experience of mental disorders is variable between cultures.

The stigma concept can be traced back to Goffman (1963) and social interactionists, who focused on the role of social interaction in stigma, including labelling and stereotyping. This view has been challenged by theorists who focus on structural stigma and discrimination; they argue that legislation and empowering people with mental disorders will reduce stigma. Link and Phelan reconciled these different approaches in 2001 in their modified labelling theory (Link and Phelan 2001). They identify stigma as a process involving labelling, stereotyping, separation, status loss, and discrimination. They highlight structural discrimination and disadvantage, and that the following is necessary to stigmatize; social, economic, and political power.

Unfortunately, not even large-scale and expensive antistigma campaigns have shown much promise in achieving changes in public attitudes (Sartorius 2010; Henderson and Thornicroft 2013). While the Beyond Blue campaign in Australia may have led to a reduction in discrimination and stigma, it is not only difficult to determine the extent to which it has achieved, but that is has correspondingly not achieved the same degree of change as the awareness raising activities (Nous Group 2014). A combination of positive social contact with people with mental disorders, protest against stigmatizing messages and measures, and education seems to be most effective in fighting stigma. Social contact has shown to be the most promising evidence-based intervention method, including 'proxy'

contact, for example, a narrative through a film (Quinn et al. 2011; Stuart 2016). It seems to be more productive to focus on promoting positive messages and strengths rather than dispelling negative beliefs. Programmes should offer a chance to revise stereotypes through identification with ordinary people and peers. Positive models of extraordinary people with mental disorders may not have the intended effect, because they can be dismissed as exceptions to the norm. At best, antistigma activities providing social contacts are mainstreamed in school curricula and training of professionals. Awareness of mental health among young Europeans in combination with role-play activities stimulating social interactions in the classroom and increasing understanding that mental health problems can be prevented are well perceived and attractive to young people (Wasserman C. et al. 2012).

Sensational reporting and reporting with a negative undertone about mental health problems is common in the media. Many national campaigns have a network of contacts who protest against negative stories on the radio, newspapers, or television by using letters, email, and the telephone. This may be effective against reporting that people with severe mental illness are dangerous. Public protests have played a limited role in mental health campaigns compared to other civil rights movements, perhaps due to the very deeply held prejudice in society and potential for ridicule (Goldie et al. 2012).

Mental health determinants

Social, behavioural, and biological sciences have provided substantial insight into the role of risk and protective factors in the developmental pathways to mental health and mental disorders. Biological, psychological, social, and societal risk and protective factors and their interactions have been identified across the lifespan from as early as fetal life. Many of these factors are amendable and, therefore, potential targets for promotion and prevention measures. Other socioeconomic and environmental determinants for mental health problems and disorders are related to macro-issues such as poverty, lack of freedom, war, and inequity (WHO 2004).

The roots of mental health and mental disorders often lay in early life. It is well documented that poor nutrition, exposure to toxic substances such as

alcohol during pregnancy, trauma during labour, maternal depression, parental neglect, sexual abuse, other forms of trauma, and lack of stimulation can impact a child's cognitive development and socioemotional status (Richter et al. 2010; Liu et al. 2017).

Mental health and mental illnesses are determined by multiple and interacting social, psychological, and biological factors. Individual, familial, and societal determinants of mental health often lie in non-health domains such as social policy, education, and urban planning. Promising evidence has, however, indicated that effective interventions exist (Doughty 2005; Das et al. 2016): interventions in local communities (Hawkins et al. 2002), child support programmes, home visiting programmes (Olds 2002), and school programmes (Weare and Nind 2011) are some examples of effective interventions for improving mental health.

Research has, so far, mainly focused on determinants of mental ill-health, largely neglecting the need to identify both risk and protective factors. Lately, there has been a heightened interest in the determinants of positive mental health and well-being. Evidence indicates that social relationships are critical for promoting well-being and buffering against mental ill-health (Diener and Seligman 2002). This seems to be the case for people across all ages (Forsman et al. 2013; Chang et al. 2017). Although food choices are often linked to well-being in popular press, the evidence on the role of different nutritional factors is complex; thus, the direct links between eating well and feeling good remain ambiguous until more research is completed (Kirkwood et al. 2008; Lai et al. 2014; O'Neil et al. 2014). There is an increasing amount of evidence about the effects of physical exercise on mental health (Firth et al. 2015; Kvam et al. 2016). Some evidence shows that more frequent physical activity and participation in sport among adolescence contribute to greater well-being and lower levels of anxiety and depressive symptoms in both sexes (McMahon et al. 2017). Longitudinal studies provide some evidence to indicate that physical activity protects against cognitive decline in later life and the onset of depressive symptoms and anxiety (Kirkwood et al. 2008).

Mental disorders are, in the developed and developing world alike, associated with disadvantage. Many people living in vulnerable circumstances are more affected by mental health problems and die more often from suicide. The greater vulnerability of disadvantaged people to mental illnesses may be explained by such factors as the experience of

insecurity and hopelessness, rapid social change, and the risks of violence and physical ill-health.

Major individual socioeconomic risk factors for mental health problems and suicide are poverty, poor education, unemployment, high debt, social isolation, and major negative life events. Socially excluded and deprived people are at a higher risk of developing mental health problems, especially depression.

On the community level, suicides are linked with socioeconomic deprivation and unemployment (Rehkopf and Buka 2005; Middleton et al. 2006; Milner at al. 2012). In Europe, between 1970 and 2007, a more than 3% increase in national unemployment rate was linked to a 4.5% increase in suicides (Stuckler et al. 2009). Moreover, suicide attempt at a young age is a strong indicator of subsequent labour market marginalization with respect to long-term unemployment, sickness absence, and disability pension (Niederkrotenthaler et al. 2014). Immigration status has also shown to affect suicide rates; in Scandinavian studies, second- and intermediate-generation immigrants show to be at higher risk of suicide (Di Thiene et al. 2015).

The effects of paid employment on mental health are complex. On the one hand, work is a source of personal satisfaction and accomplishment, interpersonal contacts, and financial security. These are all prerequisites for good mental health. The workplace social capital (i.e. trusting relationships) has been shown to protect against mental ill-health (Oksanen et al. 2010; Gao et al. 2014). Those who become unemployed are twice as likely to have increased depressive symptoms and to be diagnosed with clinical depression compared to those who remain employed (Dooley et al. 1994). On the other hand, there is evidence indicating that a high workload, precarious work, and high emotional demand, as well as school or workplace bullying and violence, are linked with depression (Theorell et al. 2015).

Prenatal maternal stress is a risk factor for behavioural and mental disorders, including depression. Those born small, due to fetal stressors, have an increased risk of mental disorders (Wahlbeck et al. 2001; Räikkönen et al. 2008; Lahti et al. 2010), as well as of suicide completion and attempt (Mittendorfer-Rutz et al. 2004). Hostile, unstable, and unsupportive parent–child relationships can lead to depression later in life (Berg et al. 2017). Corporal punishment, harsh parenting and child abuse (physical, sexual, and emotional), temporary separation (Räikkönen et al.

2011) from parents, and interparental conflict are associated with adverse psychological outcomes (Sarchiapone et al. 2007; Afifi et al. 2014) and disorders in childhood (Madigan et al. 2007) and adolescence (Allen et al. 2007; Mills et al. 2013). Children of parents with depression are a high-risk group: six in ten will develop a mental disorder before the age of 25 (Beardslee et al. 1993). Weissman et al. (2006) found that children of depressed parents have a 3 times higher risk of developing anxiety, depression, and substance abuse disorders in comparison with non-depressed parents. Promoting a nurturing early interaction between caregivers and the child increases resilience in the face of adverse life events and promotes life-long mental health and well-being.

Substance use is a risk factor for mental disorders. Alcohol use, for instance, causes depression (Rehm et al. 2003). Alcohol problems lead to a more serious course of depression, including earlier onset of the disorder, more episodes of depression, and more suicide attempts (Sher et al. 2008). A rise in per capita alcohol consumption has been linked to a post-war rise in suicide mortality in many European countries (e.g. Denmark, France, Hungary, Norway, and Sweden), but not in southern Europe (Norström and Ramstedt 2005). The link seems to be more pronounced in countries where strong spirits dominate the consumption. There is strong evidence supporting the effectiveness of policies that regulate the alcohol market, for example, by taxation and restricting access, in reducing the harm done by alcohol (Anderson 2006; Anderson et al. 2013). A natural experiment in the former USSR during the Gorbachev rule in the second half of the 1980s, characterized not only by political reforms, but also by strict limits on the sale of alcohol and restrictive attitude towards alcohol consumption, resulted in history's most effective suicide prevention programme for men. In all 15 republics of the USSR, there was a 40% decrease in the number of suicides for men between 1984 and 1986. In comparison to a 3% decrease in 22 European countries studied during the same time (Wasserman et al. 1998). Promotion of a healthy lifestyle and avoidance of harmful drinking are cornerstones in promoting good mental health and preventing suicides.

Mental, social, and behavioural health problems interact, which may intensify their effects on behaviour and well-being. Substance abuse, violence, and abuses of women and children, on the one hand, and health problems such as heart disease, depression, and anxiety on the other, are more prevalent and more difficult to cope with in conditions of high

unemployment, low income, limited education, stressful work conditions, gender discrimination, unhealthy lifestyle, and human rights violations.

Well-being and mental health promotion

In recent years, there has been a discernible shift of focus in the field of public mental health from illness to well-being. The emergence of positive psychology, and the science of well-being, has brought with it a greater interest in establishing the underlying causes of positive mental health and well-being. Standardized measures of well-being have been used in population-based surveys (Bech et al. 2003; Tennant et al. 2007, Vaingankar et al. 2018) and there are several countries in Northwestern Europe (e.g. England, Iceland, and Scotland) that perform repeated measures of mental well-being in the population.

Mental health promotion aims to improve mental health of the population by strengthening well-being. It implies the creation of individual, social, and environmental conditions that are empowering and enable optimal health and development. Such initiatives involve individuals in the process of achieving positive mental health, well-being, and quality of life. Mental health promotion is an enabling process, done by, with, and for the people.

Common principles and recommendations for modern mental health promotion were laid by The Melbourne Charter in 2008. The Charter provides a framework that recognizes the influence of social and economic determinants on mental health and mental illness, and it identifies the contribution that diverse sectors (including but not exclusive to health) make in influencing those conditions that create or ameliorate positive mental health. The Charter stresses that mental health promotion is everybody's concern and responsibility; that mental well-being is best achieved in equitable, just, and non-violent societies; and that mental health is best promoted through respectful, participatory means where culture and cultural heritage and diversity are acknowledged and valued (Global Consortium for the Advancement of Promotion and Prevention in Mental Health 2009).

In many aspects, the promotion of mental health overlaps with prevention, yet both are distinct. The emphasis in mental health promotion is on positive mental health and well-being (what can be done to keep

people healthy or to become even healthier), rather than illness prevention (what can be done to avoid illness).

Population-based approaches for promoting mental health and well-being utilize principles of public participation, engagement, and empowerment and are implemented in everyday contexts such as in families, schools, and workplaces. Effective mental health promotion builds on cross-sectoral collaboration with non-health sectors, including education, housing, employment and industry, transport, arts, sports, urban planning, and justice.

The lessons of the successful control of infectious and cardiovascular diseases indicate that the road to improved mental health among populations lies not in the investment in of mental health services, but in promotion and prevention activities. Thus, lately many policy documents have taken the 'Health in All Policies' (HiAP) approach, which is targeting determinants of mental health across policy areas in the whole population, reaching out to public policymakers to mainstream mental health promotion and prevention of mental health problems in non-health policies. HiAP is a relatively new approach to the development, implementation, and assessment of public policies across sectors that systematically take into account the health implications of decisions. It can be seen as a strategy to recognize that mental health is greatly affected by policies in other sectors than in the health sector (Wahlbeck and Taipale 2006; Mäkinen 2009; Wahlbeck 2015). For instance, the policies in the employment sector influence unemployment and suicide rates (Dooley et al. 1994; Norström and Grönqvist 2015).

Disadvantaged and vulnerable populations have a significant need for mental health promotion action. Strengthening mental health adds to the human capital asset of vulnerable populations. The promotion of mental health can play an important role in breaking the intergenerational cycle of poverty and mental ill-health through promoting positive mental health outcomes (Petersen et al. 2010).

The concept of well-being comprises two main elements: feeling good (hedonic well-being) and functioning well (eudaimonic well-being). Happiness and enjoyment are aspects of hedonic well-being. Resilience (the capacity to cope with adversity), sense of mastery of one's life, and sense of coherence and optimism are characteristics of eudaimonic well-being (Huppert 2008).

Being in a state known as mindfulness, ‘the state of being attentive to and aware of what is taking place in the present’, has also been shown to predict positive mental states, self-regulated behaviour, and heightened self-knowledge (Brown and Ryan 2003) and may significantly improve symptoms for various mental health conditions (Gotink et al. 2015). Self-determination theory suggests that an open awareness is particularly valuable for choosing behaviours that are consistent with one’s needs, values, and interests (Ryan and Deci 2000). This self-regulatory behaviour is thought to be important for well-being.

Although the intervention research-base in mental health promotion is still developing, several effective population-level interventions have been identified (Mrazek and Haggerty 1994; Sandler at al. 2014). A large range of multifaceted outcomes from mental health promotion programmes have impacted other sectors such as education, labour and employment, and family cohesion; however, effects have often been seen after a longer follow-up time than that usually used in most standard intervention studies.

An important target for mental health promotion interventions is parenting, including early parent–child interaction and approaches to discipline in child upbringing. Programmes should address the intergenerational transmission of poor parenting; abused and neglected children are more likely to become adults who are hostile, reject parenting, and display negative discipline practices towards their own children. Parenting support practices need to focus on the positive, taking an empowering approach, and enhancing positive mother–infant and father–infant interaction and enjoyment (Stewart-Brown and Schrader-McMillan 2011).

Promoting a nurturing early interaction between caregivers and the child increases the resilience of children in the face of adverse life events and promotes life-long mental health and well-being. Home visitation programmes that provide counselling, as well as a specific intervention to strengthen parent–child interaction, have been shown to be effective when delivered by trained lay women in low- and middle-income countries (Cooper et al. 2009), and by trained nurses in high-income settings (Olds 2002). Such programmes have shown good effects, improving maternal sensitivity, and reducing intrusiveness and improving attachment of children (Cooper et al. 2009).

The past two decades have seen a significant growth of research and good practice on mental health prevention and promotion in schools (Wasserman et al. 2015; Zalsman et al. 2016; Ahern et al. 2018). Schools are an important setting for mental health promotion, through their role in helping to establish identity, interpersonal relationships, and other transferable skills. Across the world, an increasing number of schools are engaging in a wide range of mental health-related initiatives and policies, which, in many places, are showing promising results. Activities operate under a variety of headings, not only ‘mental health’, but also ‘social and emotional learning’, ‘emotional literacy’, ‘emotional intelligence’, ‘resilience’, ‘life skills’, and ‘character education’. There is growing evidence showing the effectiveness of mental health promotion in schools. Positive impacts include the reduction of depression, aggression, impulsiveness, and antisocial behaviour, as well as the development of proficiencies that promote mental health such as cooperation, resilience, a sense of optimism, empathy, and a positive and realistic self-concept. Programmes have also been shown to help prevent and reduce early sexual experience, alcohol and drug use, violence and bullying in and outside schools, to promote pro-social behaviour and, in some cases, reduce juvenile crime (Weare and Nind 2011; Sandler et al. 2014). Furthermore, mental health promotion programmes significantly improve academic performance. A meta-analysis summarized research on 213 social and emotional learning programmes and found that schools with effective programmes had an 11-percentile improvement in achievement tests, a 25-percentile improvement in social and emotional skills, and a 10 percentile decrease in classroom misbehaviour, anxiety, and depression. Data indicates that successful school programmes include those with a sequential and integrated skills curriculum, active forms of learning to promote skills, and a focus on skill development and explicit learning goals (Durlak et al. 2011). Research has also indicated that during evaluations of school-based mental health programmes, when considering student feedback, it is imperative to be conscious of power dynamics that are present between mental health professionals and pupils, as the opinions of more reserved students may become unintentionally overlooked (Wasserman C. et al. 2018). Teacher well-being is an important component of the whole-school approach to mental well-being; results of an European study (Wasserman D. et al. 2010; Carli et al. 2013) showed that psychological well-being of

teachers is linked to teachers' readiness to help pupils with mental health problems (Sisask et al. 2014).

For the adult population, the workplace is an important setting for mental health promotion. Actions can be implemented at both an organizational level within the workplace and targeted at specific individuals. The former can target managers and include measures to promote awareness of mental health and well-being in the workplace and improve managers' skills in risk-management of stress and poor mental health (Nyberg et al. 2011). This can be achieved by examining job content, working conditions, terms of employment, social relations at work, modification of the physical working environment, flexible working hours, improved employer–employee communication, and opportunities for career progression. Actions targeted at individuals can include modifying workloads, providing cognitive behavioural therapy, relaxation, and meditation training, time management training, exercise programmes, journaling, biofeedback, and goal-setting. Structuring employment to create 'good work' brings health benefits to the individual, financial benefits to the corporation, and both direct and indirect improvements to the fabric of society. Effective programmes address work context, work content, and individual resilience. Effective management styles are those that promote organizational justice, workplace support, control possibilities of the work demand, active participation of employees, and provide clear and consistent communication. Effective work content is that which protects employees from unreasonable job demands, promotes employee control and autonomy, flexible working schedules, and job stability. For individuals, effective comprehensive programmes are those that promote training in both resilience and stress management (Czabała et al. 2011).

For older people, the most promising interventions promoting mental health include social activities. Studies have shown that associations exist between social capital in the ageing population and mental health (Forsman et al. 2011b, 2012). Crucial components of the individual-level social capital concept, such as low social support and small social network size, are associated with depressive symptoms and depression, while loneliness is linked to depressive symptoms and depression. Research has highlighted that civic mistrust and lack of reciprocity or social participation (i.e. low individual-level social capital) are associated with depressive symptoms among older adults. Mental health of older adults can be strengthened by

addressing social capital. With an increasing retirement age in many countries, work may well have a protective effect for older people, but only if it is flexible enough to take into account older people's differing capabilities and needs. Data highlight the effectiveness and subjective importance of social activities in maintaining mental health and well-being among older adults. The social activities are an important mental health resource among older adults, because of the accompanied sense of belonging to a social group, as well as feelings of purpose, with regard to everyday life and hope for the future. Psychosocial interventions aiming to increase the social contacts of older participants tend to improve mental well-being, as this reduces feelings of loneliness. Systematic review and meta-analysis show that social activities among older people significantly reduced depressive symptoms when compared to no-intervention controls (Forsman et al. 2011a) and support use of social capital interventions in older population (Coll-Planas et al. 2017).

Prevention of mental disorders

Prevention of mental health problems and disorders has a long history. The early ideas of the mental hygiene movement, at the beginning of the twentieth century, were first translated into experimental activities in primary healthcare, schools, and public health practices. However, the systematic development of science-based prevention programmes and controlled studies to test the effectiveness of preventive interventions did not emerge until around 1980. Since then, the multidisciplinary field of prevention science in mental health has developed at a rapid pace, facilitated by increasing knowledge on malleable risk and protective factors, generating evidence showing that preventive interventions can influence risk and protective factors and reduce the incidence and prevalence of some mental disorders (WHO 2004) and self-destructive behaviours (Zalsman et al. 2016).

Mental disorder prevention aims at reducing occurrence, frequency, and relapses, the time spent with symptoms, or the risk for a mental illness, preventing or delaying their occurrence and decreasing their impact in the affected person, their families, and society. Recently, there has been a wider

acceptance and recognition of the importance of prevention in mental disorders.

Many of the effective preventive measures are harmonious with principles of social equity, equal opportunity, and care of vulnerable groups in society. Examples of these interventions include improving nutrition, ensuring primary education, access to the labour market, removing discrimination based on race and gender, and ensuring basic economic security. A particularly potent and unfortunately common threat to mental health is conflict and violence, both between individuals and between communities and countries. The resulting mental distress and disorders are substantial. Preventing violence requires larger societal efforts, but mental health professionals may be able to ameliorate the negative impact of these phenomena by implementing some specific preventive efforts and making humanitarian assistance more mental health friendly (WHO 2004; Wahlbeck 2015).

Prevention can be categorized in a number of ways: primary prevention focuses on addressing wider determinants across whole populations. Depending on the target group, primary prevention can be universal, selective, or indicated. Universal prevention is aimed at the general population and is beneficial for everyone in the total population. Selective prevention focuses on targeting groups at higher risk of developing a disorder. Indicated prevention targets high-risk people who are identified as having minimal, but detectable, signs or symptoms foreshadowing mental disorder or biological markers indicating predisposition for mental disorder, but who do not meet all diagnostic criteria for a disorder at that time. Secondary prevention involves early detection and intervention. Tertiary prevention involves working with those with an established disease in order to promote recovery and reduce the risk of relapse (Mrazek and Haggerty 1994; Campion et al. 2012).

It is increasingly accepted that prevention efforts will have a broader spectrum impact, not only on mental health, but also on employability, educational achievement, and social adjustment. This is due to common risk factors for mental disorders and opportunities in life. By acting on common risk factors, instead of trying to prevent specific disorders, a range of positive outcomes will be achieved.

The postnatal period is important for the mental health in mother and child. It has been repeatedly shown that children of families with parental

mental disorder have an increased risk of developing mental disorders. A range of interventions have been developed aiming to prevent transgenerational transfer by addressing risk and protective factors in children and their families. Issues addressed include the family's knowledge about mental illness, fostering psychosocial resilience in children, improving parent–child and family interactions, diminishing stigma, and increasing social network support. Some interventions target early parent–child interaction; others use a whole-family approach during childhood and early adolescence or focus on the children at risk themselves (Siegenthaler et al. 2012; Sandler et al. 2014; Lauritzen and Reedts 2015). Systematic reviews and meta-analyses indicated that the risk of mental disorders in the offspring can be reduced by 40% by preventive interventions (Siegenthaler et al. 2012).

There is strong evidence showing that improving nutrition and development in socioeconomically disadvantaged children can lead to healthy cognitive development, improved educational outcomes, and reduced risk for mental ill-health, especially for those at risk or who are living in impoverished communities. The most effective intervention models combine nutritional interventions (such as food supplementation) and growth charts with counselling and psychosocial care (e.g. warmth, attentive listening) (WHO 2004). In addition, iodine, which secures an adequate function of the thyroid, plays a key role in preventing mental and physical retardation and impairment in learning ability. A too low iodine intake during early gestation has been linked to lower verbal IQ and poorer reading ability of offspring at age 8–9 years. Iodine supplementation programmes ensure that children obtain adequate levels of iodine through iodized salt or water. Pregnant women may in addition need targeted prevention actions to ensure sufficient iodine intake (Bath et al. 2013; Velasco et al. 2018).

Bullying among youth is a significant public health problem; it is prevalent and frequently has detrimental effects. Being a victim of bullying is associated with being depressed or anxious, poor social and emotional adjustment, lower academic achievement (Hertz et al. 2013) and alcohol use (Marschall-Levesque et al. 2016). Data indicates a causal association between being a bully-victim and subsequent depression and panic disorder as well as subsequent suicide-related behaviours (Klomek et al. 2010; Copeland et al. 2013; Holt et al. 2015; Brunstein Klomek et al. 2016;

Barzilay et al. 2017). Perpetrators of bullying have also been shown to be at risk of depression and suicidal ideation (Kaltiala-Heino et al. 1999; Saluja et al. 2004) Prevention of bullying would improve mental health outcomes for many young people, and effective antibullying prevention programmes have been developed. A meta-analysis of school bullying strategies concluded that whole-school approaches that included multiple disciplines and complementary components directed at different levels of school organization more often reduced victimization and bullying than the interventions that only included classroom-level curricula or social skills groups (Vreeman and Carroll 2007).

Community systems-strengthening (Fund 2010) has its roots in the action that communities have always taken to protect and support their members. Modern approaches to community healthcare are reflected in the WHO Alma Ata declaration of 1978, and in more recent WHO work on the social determinants of health (WHO n.d.). Community systems-strengthening interventions focus on developing empowering processes and building a sense of ownership and social responsibility within community members. An example of such an intervention is the *Communities that Care* (CTC) programme, which has been implemented successfully in the United States and is currently being adopted and replicated in several developed countries. The CTC intervention activates communities to implement community violence and aggression prevention systems (Hawkins et al. 2002). The strategy helps communities use local data on risk and protective factors to identify risks and develop actions. These include interventions that operate simultaneously at multiple ecological levels: the community (e.g. mobilization, media, and policy change), the school (e.g. changing school management structures, curricula, and teaching practices), the family (e.g. parent training strategies), and the individual (e.g. social competence strategies). Evaluations at various CTC sites have indicated improvements in youth outcomes such as reduction in school problems, weapons charges, burglary, drug offences, and assault charges.

There is rich evidence showing that conduct disorders, aggression, and violence of young people can be prevented. The most successful preventive interventions to reduce the risk of aggressive behaviour and conduct disorders focus on improving the social competence and pro-social behaviour of children, parents, peers, and teachers. Universal interventions that have been shown to have a successful impact on conduct problems are

all school-based and include classroom behaviour management, enhancing child social skills, and multimodal strategies, including the involvement of parents. Classroom behaviour management programmes (e.g. *Good Behaviour Game*) attempt to help children better meet the social demands of the classroom through the overt encouragement of desired behaviours, and the discouragement of undesired behaviours. Social skills programmes (e.g. *I Can Problem Solve* or *Promoting Alternative THinking Strategies* (PATHS)) attempt to provide children with cognitive skills that may help them cope better with difficult social situations. Skills related to listening, empathy, interpersonal problem-solving, and conflict and anger management are taught in a classroom setting. Also, selective interventions designed for a variety of settings have been found to be effective in preventing conduct problems, including prenatal and/or early childhood programmes and school-or community-based programmes. Prenatal and/or early childhood programmes usually attempt to improve the skills of parents to nurture, support, and teach their children pro-social behaviour patterns and/or to develop the social skills of children. These programmes have been shown to decrease child conduct problems during adolescence, including reductions in violence and police arrests. School- or community-based programmes for selective child populations at risk have successfully targeted child social and problem-solving skills and/or parent management skills, resulting in a decrease in negative parent–child interactions and teacher ratings of conduct problems at school (Powell et al. 2007).

Recent research demonstrates that depressive episodes and anxiety disorders can be prevented in a cost-effective way and even in a cost-saving way (Cuijpers et al. 2012; Chisholm et al. 2016). Preventive interventions can reduce the incidence of new episodes of MDD by about 25%. Adding a stepped-care model to the preventive intervention may reduce the number of new episodes even more (Van't Veer-Tazelaar et al. 2009). However, it seems that a stepped-care model may only be cost-effective in some settings, depending on the willingness-to-pay (Van der Aa et al. 2017). Methods with proven effectiveness involve educational, psychotherapeutic, pharmacological, lifestyle, and nutritional interventions. School-based programmes targeting cognitive, problem-solving, and social skills of children and adolescents (e.g. *The Resourceful Adolescent Programme*) have reported a reduction in high depressive symptom levels of 50% or more a year after the intervention. Several selective interventions targeted at

coping with major life events have shown significant and long-term reduction in high levels of depressive symptoms, such as programmes for children with parental death or divorce or for unemployed adults. Indicated programmes for those with elevated levels of depressive symptoms, but no depressive disorder, have shown significant effects in reducing high levels of depressive symptoms and preventing depressive episodes. Such programmes mainly use a group format to educate people at risk in positive thinking, challenge negative thinking styles and improve problem-solving skills and have been shown to be effective in school (Clarke et al. 1995; Weersing et al. 2016; Garber et al. 2018) or primary care settings.

Anxiety disorders can successfully be prevented by strengthening emotional resilience, self-confidence, and cognitive problem-solving skills in school-aged children. Cognitive behavioural programmes (e.g. the *FRIENDS* programme) have significantly reduced the first onset of anxiety disorders in school children (WHO 2004; Stallard et al. 2014).

The evidence for the effectiveness of interventions in preventing post-traumatic stress disorder (PTSD) is limited. However, some studies show that PTSD severity can be reduced by brief preventive trauma-focused cognitive behavioural therapy (CBT) and modified prolonged exposure delivered within weeks of potential traumatic events (Birur et al. 2017). Psychological debriefing does not reduce the incidence or severity of PTSD or related psychological symptoms in civilian victims of crime, assault, or accidental trauma and should not be routinely offered (Gartlehner et al. 2013).

Several eating disorder prevention programmes have been developed and tried among children and adolescents. There is currently limited evidence in the published literature to suggest that any particular type of programme is effective in preventing eating disorders. Eating disorder prevention programmes, based on challenging the thin ideal (*The Body Project*) or promoting energy balance (the *Healthy Weight* intervention) have reduced the risk of future onset of eating disorders, but more research on their effectiveness is needed before large-scale implementation of eating disorder prevention programmes can be initiated (Stice et al. 2013; Stice et al. 2015; Akers et al. 2017).

Substance use disorders can be prevented by universal policy actions by reducing the availability of alcohol and drugs. Effective regulatory interventions include taxation, restrictions on availability, and total bans on

all forms of direct and indirect advertising. Policy interventions aimed at reducing the harm from addictive substances have led to the prevention of substance use disorders. When applied to alcohol, education and persuasion strategies usually deal with less alcohol consumption, the hazards of driving under the influence of alcohol, and related topics. Despite their good intentions, public service announcements are considered an ineffective antidote to the high-quality pro-drinking messages that appear much more frequently through paid advertisements in the mass media.

Psychoses are severe mental disorders with a peak incidence in late adolescence and early adulthood. Universal and selective interventions are not yet viable strategies in the prevention of psychoses. The indicated prevention approach and early identification and intervention present promising possibilities to reduce the burden of schizophrenia and other psychoses. Typically, there is a delay of 1–2 years between the onset of schizophrenia and initiation of treatment, due to failure in identifying psychosis, which, in many cases, may have a slow and gradual onset. A prolonged duration of untreated psychosis has been linked to worse outcome. Several population-based preventive programmes have been developed to reduce the duration of untreated psychosis. Improving community awareness and mental health literacy of the general population reduced the delays into treatment in the Norwegian *Treatment and Identification of Psychosis Study* (TIPS) and subsequent studies in Australia (Correll et al. 2010). Results in the prevention and delay of transition to psychotic disorder from high-risk state in adults, mostly involving second-generation antipsychotics and CBT, are promising and have become part of public health practice recommendations; however, there is still a call for more research to investigate the specific treatment effects (Schmidt et al. 2015).

Prevention of suicides

Public health approaches to suicide prevention have to integrate societal and cultural viewpoints with medical and psychological viewpoints to develop strategies that will save most lives in an effective and measurable way (Wasserman D. 2016).

Suicides can be prevented by public health actions and suicide prevention has consistently been shown to be highly cost-effective. Considerable evidence is available for the effectiveness of broadly applied population-level interventions, such as restriction of access to lethal means of suicide, and responsible media coverage of suicide issues (Zalsman et al. 2016).

Restricting access to lethal means has been supported by strong evidence. Firearm control legislations have been associated with reduced suicide rates. Control of analgesics, particularly with regards to distributions in smaller packaging, has shown to be beneficial in reducing suicide deaths. Suicide incidences were also reduced through restricting prescription and sale of barbiturates. Restriction on the availability of pesticides has been effective in countries where this mean of suicide is common. Construction of barriers at sites popular for jumping has been linked to reduce overall suicide by jumping by 86%, with little evidence of major substitution to other sites. Detoxification of domestic gas and the introduction of catalytic converters in cars have also shown to be effective suicide prevention methods. Part 13—Public health strategies: Controlling the access to means of suicide, ch. 74–81, from the Oxford Textbook of Suicidology and Suicide Prevention, Second edition, edited by Danuta Wasserman, 2020.

The probability of individuals attempting suicide decreases when they are precluded from implementing a preferred method. Although some individuals might seek other methods, many do not; when they do, the means chosen are less lethal and are associated with fewer deaths than when more dangerous ones are available. When a lethal method is unavailable at the moment of potential action, suicide attempts might be delayed so that, in some cases, suicidal impulses will pass without fatal effects (Yip et al. 2012; Barber and Miller 2014).

As a public health measure, means restriction has a long history; removal of the pump handle in Broad Street, London, United Kingdom, by John Snow in 1854 was an early example and a historic landmark in public health practice (Johnson 2006). Means restriction entails a community or societal action that does not depend on an individual’s intention or volition. Applied to the population as a whole, it typically affects people whose suicide risk is otherwise undetected, and who do not seek therapeutic assistance to prevent their crisis or for life-saving interventions when necessary. Removal or restriction of access to a lethal method changes the

context of a potential suicide by precluding potentially fatal actions or forcing the use of a less lethal method.

Application of universal measures for means restriction might be considered more or less intrusive by the general public, depending on the country. The resistance to restrictions and safety measures is underpinned by the common misperception that, despite data showing powerful population level effects, a seriously suicidal person will inevitably find a way to die and that all methods have roughly equal case fatalities. In order to implement means restriction, it is important to address these misunderstandings and gain community support.

Choice of suicide mean varies according to the country and even inside one country, and by age and gender (Ajdacic-Gross et al. 2008; Mars et al. 2014; Qi et al. 2014). Suicide means used also vary over time. These patterns suggest a close link to differences in the availability and lethality of specific suicide means. Due to availability of pesticides, suicides by pesticide poisoning are common in agrarian populations in Asia (Phillips and Gunnell 2009; Mew et al. 2017), whereas access to high buildings and bridges make jumping from high places a common mean of suicide in urban settings (Beautrais and Gibb 2009; Chen and Yip 2009; Chen et al. 2012).

In times of peace, most firearm-related deaths are suicides (Resnick et al. 2017). Enforcement of gun control policies (e.g. purchase restrictions, waiting times for gun purchase, higher age limits, licensing of firearm owners, safe storage precautions) lowers numbers of firearm suicides (WHO 2010). Empirical data suggest that firearm regulations, which function to reduce overall gun availability, have a significant deterrent effect on male suicide, while regulations that seek to prohibit high-risk individuals from owning firearms have a lesser effect (Rodriguez Andrés and Hempstead 2011; Anestis et al. 2015). Having a firearm in the home increases the risk of a violent death in the home; data from the United States of America showed that the risk of dying from a suicide in the home was tenfold greater for males in homes with guns than for males without guns in the home. Persons with guns in the home were 30-fold more likely to have died from suicide committed with a firearm than from one committed by using a different method (Dahlberg et al. 2004). Guns are highly lethal, require little preparation, and may be chosen over less lethal methods to commit suicide, particularly when the suicide is impulsive. Removal of firearms from people’s homes by regulatory means is desirable

and would result in a reduction of suicides, but may be difficult to achieve in countries with a strong pro-gun lobby.

Safe environments (e.g. safety doors on railway platforms, barriers on bridges, locked storage of firearms) contribute to suicide prevention. Prevention of suicide can be taken into account during the planning process or after an environment (e.g. bridges and railways) has been identified as a suicide hot-spot (Beautrais and Gibb 2009; Ladwig et al. 2009; Pirkis et al. 2015; Ueda et al. 2015; Sinyor et al. 2017).

Media depiction of suicide and suicidal behaviour can have protective effects on the general population, but detrimental effects on vulnerable populations. Decreased suicidal behaviour is associated with media blackouts and better quality reporting. There is significant evidence from around the world to indicate that highly sensationalized reporting of suicides, providing detailed descriptions of the method used, can and does lead to 'copycat' suicides, particularly if the news reports about celebrity suicides are mirrored with more attention in social media (Ueda et al. 2017). Suicide contagion is a concept from the infective disease model and assumes that suicidal behaviour may facilitate the occurrence of subsequent, similar behaviours. Imitation refers to the process that explains the occurrence of contagion. Suicide contagion has been investigated primarily in adolescent populations, revealing that up to 5% of all adolescent suicides may be the result of suicide clustering (Westerlund et al. 2009). On the other hand, responsible reporting on suicide reduces copycat suicide and has protective effects, especially among adolescents (Niederkrotenthaler and Sonneck 2007; Niederkrotenthaler et al. 2010). Liaising with media as well as using the educational internet-based systems has an important role in educating the public about depression, suicide preventive methods and responsible coverage of suicide (Westerlund and Wasserman 2009). Media guidelines for reporting suicides and monitoring of stigmatizing media reports have been linked with reduced stigmatization in the press and reduction of suicides. Such reporting guidelines can be based on recommendations set forth by the WHO (2017b).

More evidence is needed for community-based simultaneous multilevel interventions targeting primary care providers, gatekeepers, general populations, and patients with their relatives (Zalsman et al. 2016). However, improving the awareness and coping skills of the general public, gatekeepers (such as youth workers, police officers, and teachers), general

practitioners (GPs), and other healthcare professionals to identify people at risk of suicide has been shown to be an effective way to prevent suicide (Hegerl et al. 2009; Hoven et al. 2009; Zalsman et al. 2016). Responsiveness of services to the needs of individuals with suicidal behaviours is an essential component of suicide prevention. Training of GPs and other healthcare staff to identify depression and substance use problems, and provide adequate treatment has been successful in reducing suicidality and suicides (Appleby et al. 2000; Rutz 2001; Roskar et al. 2010; Zalsman et al. 2016). This method is especially useful for preventing suicides in females, as in a number of cultures females have a greater tendency than males to seek help from healthcare professionals (Hadlaczky and Wasserman 2009).

Over the last decade, evaluation studies involving school-based programmes have been of better quality. Systematic reviews indicate improved knowledge and attitudes towards suicide and three large randomized control trials showed significant effects on reducing suicide attempts and ideation (Zalsman et al. 2016). The Saving and Empowering Young Lives in Europe (SEYLE)—a multicentre, cluster-randomized controlled trial from 168 schools with 11,110 pupils in 10 European Union countries—demonstrated that Youth Aware of Mental Health Programme (YAM) was associated with a significant reduction of incident suicide attempts and severe suicidal ideation, compared with the control group (Wasserman D. et al. 2015).

Community, family, and school-based mental health interventions and programmes increase awareness about treatment possibilities, and may promote treatment acceptance, factors that are important in suicide prevention.

The evidence for selected strategies, which target subgroups with increased risk of suicide, such as people bereaved by suicide, victims of violence, physical or sexual abuse, is growing (Wasserman D. et al. 2012; Zalsman et al. 2016). The focus on high-risk groups is important, as suicide rates are higher in these groups compared with the general population.

Indicated prevention using various psychiatric treatments has very well documented effects in suicide prevention (Wasserman D. et al. 2012; Zalsman et al. 2016). The pharmacological treatment of psychiatric disorders, such as depressive disorders, psychotic disorders, and substance use disorders, is proven to be substantial in the prevention of suicide.

Antidepressants have been associated with reduced suicide risk; however, in children and adolescence, antidepressants have been linked to increased suicide thoughts, yet they need to be considered, since there is an increased risk of suicide with untreated depression. Psychotherapies for individuals with mood disorders have also shown to be efficacious in reducing the risk of suicide.

Combination of pharmacological treatment with psychotherapies (such as cognitive behaviour therapy, dialectical behavioural therapy, interpersonal therapy, problem-solving therapy, etc.), as well as outreach, are effective and recommended in suicide prevention. Follow-up of people who attempt suicide has found strong support (Fleischmann et al. 2008), as well has the availability of mental health services (Pitt et al. 2013).

In spite of the fact that universal approaches hold the potential to prevent a greater number of deaths (Pitman and Caine 2012), it is important that they go hand-in-hand with selected and indicated strategies, which diminish suffering for the individuals and their families.

Mental health services

Design, management, and evaluation of mental health services are a core task of public mental health. Today, mental health service provision is in a global transition from hospital-based systems to community-based systems. The change reflects the growing evidence of what constitutes cost-effective care; it also acknowledges the failures concerning issues of social inclusion and human rights of the care systems, which were based on the old-fashioned and remote institutions (asylums).

The recent history of mental health services in economically developed countries can be seen in terms of three periods: first, the rise of the asylum; second, the decline of the asylum; and third, balancing mental health services (Table 8.7.1).

Table 8.7.1 The key characteristics of the three periods in the historical development of mental health systems of care

Period 1: the rise of the asylum	Period 2: the decline of the asylum	Period 3: balancing mental services
Asylums built	Asylums neglected	Asylums replaced by smaller facilities
Increasing number of hospital beds	Decreasing number of hospital beds	Decrease in the number of beds slows down
Reduced role for the family	Increasing but not fully recognized role of the family	Importance of families increasingly recognized, in terms of care given, therapeutic potential, the burden carried, and as a political lobbying group
Public investment in institutions	Public disinvestment in mental health services	Increasing private investment in treatment and care and focus in public sector on cost-effectiveness and cost containment
Staff: doctors and nurses only	Clinical psychologist, occupational therapist, and social worker disciplines evolve	More community-based staff and emphasis on multidisciplinary team working
	Effective treatments emerge, beginning of treatment evaluation and of standardized diagnostic systems, growing influence of individual and group psychotherapy	Emergence of 'evidence-based' psychiatry in relation to pharmacological, social, and psychological treatments
Primacy of containment overtreatment	Focus on pharmacological control and social rehabilitation, less disabled patients discharged from asylums	Emergence of concern about balance between control of patients and their independence

Source: data from Thornicroft G, and Tansella M (2002). 'Balancing community-based and hospital-based mental health care.' *World Psychiatry*, 1(2):, pp. 84–90, Copyright © 2002.

A modern mental healthcare system is based on balanced care, that is, it includes both modern community-based and modern hospital-based care. In balanced care, the focus is on services that are provided in normal community settings, as close to the population served as possible, and to which admissions to psychiatric wards in a general hospital can be arranged

promptly, but only when necessary (Thornicroft and Tansella 1999, 2002; Gaebel et al. 2012).

A health policy supporting the integration of health and social services, and the mainstreaming of mental health services into primary care, improves access to care. Integrating mental health services into general healthcare is the most viable way of implementing balanced care, closing the treatment gap, and ensuring that people get the care they need. Because of the high prevalence of mental disorders, only complex and/or severe cases should be treated by specialized mental healthcare. To achieve this consistently, the services need to identify priority groups who should receive access to specialist care from among the 25% of the whole population who suffer from a mental disorder in any given year. Well targeted services are those in which specialist care concentrates on providing direct services to people with the most severe degree of symptoms and disability. Primary care responsibility for common mental health disorders should be supported by accessible referral systems and specialist supervision (Tansella and Thornicroft 1999). Providing even minimal psychotherapy in primary care can prevent full-blown depression (Smit et al. 2006). Programmes aimed at education of primary care physicians have improved the detection of depression and even led to a decrease in suicides due to depression (Rutz 2001). The use of new media, such as e-mental health and smart phone technologies, and the use of lay health counsellors, may boost dissemination of mental health interventions, especially in low- and middle-income countries.

Another cornerstone of a balanced care system is the provision of safe and high-quality home and community-based specialist services. Any developed mental healthcare system encompasses specialized community services for children and adolescents. A modern mental health system includes crisis resolution teams, home treatment, assertive outreach, and early intervention teams, day treatment facilities, and supported work. Collated evidence suggests that supported employment schemes, which consist of arranging early placement in normal work with variable support from staff, may offer better outcomes than sheltered or transitional employment approaches. People suffering from mental disorders who want to work should be offered the option of supported employment as part of their treatment package. Current evidence indicates that supported work improves clinical outcomes and supports social inclusion of people with

severe mental disorders (Crowther et al. 2001; Bruns et al. 2007; Luciano et al. 2014).

In diversified modern mental healthcare systems, particular attention needs to be paid to maintaining high-quality communication between all parts of the care system. There is a need to address three types of interfaces: (1) those within the mental health service, between its components; (2) those within the health service, between mental health and other services (both primary and secondary care); and (3) those between health and other public services, including social services and housing departments.

Available evidence indicates that community-based and diversified mental health systems, with a wide range of services, are superior to hospital-centred mental health systems, according to a range of outcomes. For instance, community-based, well-developed, and multifaceted mental health services have been linked with lower suicide rates than hospital-based traditional services (Pirkola et al. 2009). Discharged patients benefit from well-developed community care; community follow-up within 7 days of psychiatric discharge has shown a significant reduction in suicides among recently discharged psychiatric patients (While et al. 2012). Additionally, research has demonstrated that brief information interventions combined with long-term follow-up after suicide attempts have a significant influence on preventing subsequent deaths from suicide, in comparison with treatment as usual (Fleischmann et al. 2008).

A core element of modern mental healthcare is the empowerment of service users and informal carers. Historically, people with mental health problems have lacked a voice. Empowerment translates into being treated with dignity and respect in mental health services and participation of users and carers in decisions. Data suggest that the possibility to exercise control and influence leads to positive mental health outcomes. These outcomes include increased emotional well-being, independence, motivation to participate, and more effective coping strategies. Key issues that users and carers have expressed to be important to advocate for are the rights to autonomy and self-determination, to acceptable and accessible services, user-led evaluation of services, the right for everyone to be recognized as a person before the law without discrimination, the de-stigmatization of mental disorders, and more inclusive and respectful services with user and carer involvement (WHO 2005). Increased used of peer support and ‘experts by experience’ in the provision of mental health services will

support empowerment of service users and improve services (Pitt et al. 2013).

Development of mental health service systems requires access to reliable data on both mental health services and trends in mental health service provisions. International benchmarking, based on comparable data, is an important moving force for the development of mental health services in countries. Unfortunately, mental health information systems, in most countries, are geared towards hospital data, which are of less interest when developing a community-based mental health service provision system. Many highly relevant aspects of modern service provision, such as patient choice, service user empowerment, and respect for human rights, are hardly ever covered by health information systems (Wahlbeck 2011a). Additionally, in many low- and middle-income countries, there are limited mental health resources and activities, which are unequally distributed and inefficiently used (WHO 2009).

In conclusion, sufficient evidence supports effective public mental health measures to prevent suicidal behaviours. Therefore, the implementation of evidence-based strategies should be the responsibility of public health policymakers and organizations responsible for implementation.

ACKNOWLEDGEMENTS

The authors thank Daniel Ventus and Tony Durkee for assistance in preparing the text of the previous version of this chapter, along with Iva Tadic and Danielle Stone for assisting with updating the current version.

REFERENCES

Afifi, T.O., MacMillan, H.L., Boyle, M., Taillieu, T., Cheung, K., and Sareen, J. (2014). Child abuse and mental disorders in Canada. *Canadian Medical Association Journal*, 186(9), E324–32.

Ahern, S., et al. (2018). A cost-effectiveness analysis of school-based suicide prevention programmes. *European Child & Adolescent Psychiatry*

Ajdacic-Gross, V., et al. (2008). Methods of suicide: international suicide patterns derived from the WHO mortality database. *Bulletin of the World Health Organization*, **86**, 726–32.

Akers, L., et al. (2017). Cost-effectiveness of achieving clinical improvement with a dissonance-based eating disorder prevention program. *Eating Disorders*, 25(3), 263–72.

Allen, J.P., Porter, M., and McFarland, C. (2007). The relation of attachment security to adolescents' paternal and peer relationships, depression, and externalizing behavior. *Child*

Development, **78**, 1222–39.
Amitai, M. and Apter, A. (2012). Social aspects of suicidal behavior and prevention in early life: a review. *International Journal of Environmental Research and Public Health*, 9(3), 985–94.
Anderson, P. (2006). *Alcohol in Europe: A Public Health Perspective*. Luxembourg: European Commission.
Anderson, P., Moller, L., Galea, G., et al. (2013). *Alcohol in the European Union: Consumption, Harm and Policy Approaches*. Copenhagen: WHO Europe.
Anestis, M.D., Khazem, L.R., Law, K.C., et al. (2015). The association between state laws regulating handgun ownership and statewide suicide rates. *American Journal of Public Health*, **105**, 2059–67.
Appleby, L., et al. (2000). An educational intervention for front-line health professionals in the assessment and management of suicidal patients (the STORM project). *Psychological Medicine*, 30(4), 805–12.
Barber, C.W., and Miller, M.J. (2014). Reducing a suicidal person's access to lethal means of suicide: a research agenda. *American Journal of Preventive Medicine*, 47(3; Suppl 2), S264–72.
Barke, A., Nyarko, S., and Klecha, D. (2011). The stigma of mental illness in Southern Ghana: attitudes of the urban population and patients' views. *Social Psychiatry & Epidemiology*, **46**, 1191–202.
Barzilay S., et al. (2017). Bullying victimization and suicide ideation and behavior among adolescents in Europe: a 10-country study. *Journal of Adolescent Health*, 61(2), 179–86.
Bath, S.C., Steer, C.D., Golding, J., Emmett, P., and Rayman, M.P. (2013). Effect of inadequate iodine status in UK pregnant women on cognitive outcomes in their children: results from the Avon Longitudinal Study of Parents and Children (ALSPAC). *The Lancet*, 382(9889), 331–7.
Beardslee, W.R., et al. (1993). The impact of parental affective disorder on depression in offspring: a longitudinal follow-up in a nonreferred sample. *Journal of the American Academy of Child & Adolescent Psychiatry*, **32**, 723–30.
Beautrais, A. and Gibb, S. (2009). Protecting bridges and high buildings in suicide prevention. In: Wasserman, D. and Wasserman, C. (eds.) *The Oxford Textbook of Suicidology and Suicide Prevention: A Global Perspective*. Oxford: Oxford University Press, pp. 563–8.
Bech, P., Olsen, L.R., Kjoller, M., and Rasmussen, N.K. (2003). Measuring well-being rather than the absence of distress symptoms: a comparison of the SF-36 Mental Health subscale and the WHO-Five Well-Being Scale. *International Journal of Methods in Psychiatric Research*, 12(2), 85–91.
Berg, N., Kiviruusu, O., Rahkonen, O., and Huurre, T. (2017) Pathways from problems in adolescent family relationships to midlife mental health via early adulthood disadvantages ± a 26-year longitudinal study. *PLoS One*, 12(5), e0178136.
Bertolote, J.M. and Fleischmann, A. (2009). A global perspective of the magnitude of suicide mortality. In: Wasserman, D. and Wasserman, C. (eds.) *The Oxford Textbook of Suicidology and Suicide Prevention: A Global Perspective*. Oxford: Oxford University Press, pp. 91–8.
Birur, B., Moore, N.C., and Davis, L.L. (2017). An evidence-based review of early intervention and prevention of posttraumatic stress disorder. *Community Mental Health Journal*, 53(2), 182–201.
Bloom, D.E., et al. (2011). *The Global Economic Burden of Noncommunicable Diseases*. Geneva: World Economic Forum.
Brown, K.W. and Ryan, R.M. (2003). The benefits of being present: mindfulness and its role in psychological well-being. *Journal of Personality and Social Psychology*, **84**, 822–48.
Bruns, T., et al. (2007). The effectiveness of supported employment for people with severe mental illness: a randomised controlled trial. *The Lancet*, **370**, 1146–52.
Brunstein Klomek, A., et al. (2016). Association between victimization by bullying and direct self-injurious behavior among adolescence in Europe: a ten-country study. *European Child &*

Adolescent Psychiatry, 25(11), 1183–93.
Bursztein, C. and Apter, A. (2009). Adolescent suicide. *Current Opinion in Psychiatry*, **22**, 1–6.
Campion, J., Bhui, K., and Bhugra, D. (2012). European Psychiatric Association (EPA) guidance on prevention of mental disorders. *European Psychiatry*, **27**, 68–80.
Carli, V., et al. (2013). The Saving and Empowering Young Lives in Europe (SEYLE) randomized controlled trial (RCT): methodological issues and participant characteristics. *BMC Public Health*, 13(1), 479.
Chandler, M.J. and Lalonde, C. (2008). Cultural continuity as a moderator of suicide risk among Canada's First Nations. In: Kirmayer, L. and Valaskaskis, G. (eds.) *Healing Traditions: The Mental Health of Aboriginal Peoples in Canada*. Vancouver: University of British Columbia Press, pp. 221–48.
Chang, Q., Chan, C.H., and Yip, P.S.F. (2017). A meta-analytic review on social relationships and suicidal ideation among older adults. *Social Science & Medicine*, **191**, 65e76.
Chen, Y.-Y. and Yip, P. (2009). Prevention of suicide by jumping: experiences from Taipei City (Taiwan), Hong Kong and Singapore. In: Wasserman, D. and Wasserman, C. (eds.) *The Oxford Textbook of Suicidology and Suicide Prevention: A Global Perspective*. Oxford: Oxford University Press, pp. 569–72.
Chen, Y.-Y., Wu, K. C.-C., Yousuf, S., and Yip, P.S.F. (2012). Suicide in Asia: opportunities and challenges. *Epidemiological Review*, 43(1), 129–44.
Chisholm, D., et al. (2016). Scaling-up treatment of depression and anxiety: a global return on investment analysis. *Lancet Psychiatry*, **3**, 415–24.
Clarke, G.N., et al. (1995). Targeted prevention of unipolar depressive disorder in an at-risk sample of high school adolescents: a randomized trial of group cognitive intervention. *Journal of the American Academy of Child & Adolescent Psychiatry*, **34**, 312–21.
Clement, S., et al. (2015). What is the impact of mental health-related stigma on help-seeking? A systematic review of quantitative and qualitative studies. *Psychological Medicine*, **45**, 11–27.
Coll-Planas, L., Nyqvist, F., Puig, T., Urrutia, G., Sola, I., and Monteserin, R. (2017). Social capital interventions targeting older people and their impact on health: a systematic review. *Journal of Epidemiology and Community Health*, 71(7), 663–72.
Condorelli, R. (2016). Social complexity, modernity and suicide: an assessment of Durkheim's suicide from the perspective of a non-linear analysis of complex social systems. *SpringerPlus*, **5**, 374.
Conejero, I., Olie, E., Courtet, P., and Calati, R. (2018). Suicide in older adults: current perspectives. *Clinical Interventions in Aging*, **13**, 691–9.
Cooper, P.J., et al. (2009). Improving the quality of the mother–infant relationship and infant attachment in a socio-economically deprived community in a South African context: a randomised controlled trial. *British Medical Journal*, 338, b974.
Copeland, W.E., Wolke, D., Angold, A., and Costello, E.J. (2013). Adult psychiatric outcomes of bullying and being bullied by peers in childhood and adolescence. *JAMA Psychiatry*, **70**, 419–26.
Correll, C.U., Hauser, M., Auther, A.M., and Cornblatt, B.A. (2010). Research in people with psychosis risk syndrome: a review of the current evidence and future directions. *Journal of Child Psychology and Psychiatry*, 51(4), 390–431.
Cotter, P., et al. (2014). Victimisation and psychosocial difficulties associated with sexual orientation concerns: a school-based study of adolescents. *Irish Medical Journal*, 107(10), 310–13.
Crowther, R., Marshall, M., Bond, G.R., and Huxley, P. (2001). Vocational rehabilitation for people with severe mental illness. *Cochrane Database of Systematic Reviews*, 2, CD003080.
Cuijpers, P., Beekman, A.T., and Reynolds, C.F. 3rd. (2012). Preventing depression: a global priority. *Journal of the American Medical Association*, 307(10), 1033–4.

Czabała, C., Charzyńska, K., and Mroziak, B. (2011). Psychosocial interventions in workplace mental health promotion: an overview. *Health Promotion International*, *26(Suppl.* 1), i70–84.
Dahlberg, L.L., Ikeda, R.M., and Kresnow, M. (2004). Guns in the home and risk of a violent death in the home: findings from a national study. *American Journal of Epidemiology*, **160**, 929–36.
Das, J.K., et al. (2016). Interventions for adolescent mental health: an overview of systematic reviews. *Journal of Adolescent Health*, 59, S49–S60.
Di Thiene, D., Alexanderson, K., Tinghög, P., La Torre, G., and Mittendorfer-Rutz, E. (2015). Suicide among first-generation and second-generation immigrants in Sweden: association with labour market marginalisation and morbidity. *Journal of Epidemiology and Community Health*, 69(5), 467–73.
Diener, E. and Seligman, M.E.P. (2002). Very happy people. *Psychological Science*, **13**, 81–4.
Dooley, D., Catalano, R., and Wilson, G. (1994). Depression and unemployment: panel findings from the Epidemiologic Catchment Area Study. *American Journal of Community Psychology*, 22(6), 745–65.
Doughty, C. (2005). The effectiveness of mental health promotion, prevention and early intervention in children, adolescents and adults. A critical appraisal of the literature. *New Zealand Health Technology Assessment (NZHTA) Report*, 8(2). Christchurch: NZHTA.
Durlak, J.A., Weissberg, R.P., Dymnicki, A.B., Taylor, R.D., and Schellinger, K.B. (2011). The impact of enhancing students' social and emotional learning: a meta-analysis of school-based universal interventions. *Child Development*, 82(1), 405–32.
Fässberg, M.M., et al. (2012). A systematic review of social factors and suicidal behavior in older adulthood. *International Journal of Environmental Research and Public Health*, 9(3), 722–45.
Firth, J., Cotter, J., Elliott, R., French, P., and Yung, A.R. (2015). A systematic review and meta-analysis of exercise interventions in schizophrenia patients. *Psychological Medicine*, 45(7), 1343–61.
Fleischmann, A., et al. (2008). Effectiveness of brief intervention and contact for suicide attempters: a randomized controlled trial in five countries. *Bulletin of the World Health Organization*, 86(9), 703–9.
Forsman, A.K., Nordmyr, J., and Wahlbeck, K. (2011a). Psychosocial interventions for the promotion of mental health and the prevention of depression among older adults. *Health Promotion International*, *26*(Suppl. 1), i85–107.
Forsman, A., Nyqvist, F., Herberts, C., Wahlbeck, K., and Schierenbeck, I. (2013). Understanding the role of social capital for mental well-being among older adults. *Ageing & Society*, 33(5), 804–25.
Forsman, A., Nyqvist, F., Schierenbeck, I., Gustafson, Y., and Wahlbeck, K. (2012). Structural and cognitive social capital and depression among older adults in two Nordic regions. *Aging & Mental Health*, 16(6), 771–9.
Forsman, A.K., Nyqvist, F., and Wahlbeck, K. (2011b). Cognitive components of social capital and mental health status among older adults: a population-based cross-sectional study. *Scandinavian Journal of Public Health*, **39**, 757–65.
Fund, G. (2010). *Community Systems Strengthening Framework*. Geneva: The Global Fund.
Gaebel, W., et al. (2012). EPA guidance on the quality of mental health services. *European Psychiatry*, 27(2), 87–113.
Gao, J., et al. (2014). Workplace social capital and mental health among Chinese employees: a multi-level, cross-sectional study. *PLoS ONE*, 9(1), e85005.
Garber, J., et al. (2018). Prevention of depression in at-risk adolescents: moderators of long-term response. *Prevention Science*, 19(1), 6–15.
Garssen, M.J., Hoogenboezem, J., and Kerkhof, A.J. (2006). [Suicide among migrant populations and native Dutch in The Netherlands]. *Nederlands Tijdschrift voor Geneeskunde*, 150(39),

2143–9.
Gartlehner, G., et al. (2013). Interventions for the Prevention of Posttraumatic Stress Disorder (PTSD) in Adults after Exposure to Psychological Trauma. *Comparative Effectiveness Review*, 109. Rockville, MD: Agency for Healthcare Research and Quality. [Online] Available at: http://www.effectivehealthcare.ahrq.gov/reports/final.cfm
Gilliver, S.C., Sundquist, J., Li X., and Sundquist, K (2014). Recent research on the mental health of immigrants to Sweden: a literature review. *European Journal of Public Health*, 24(1), 72–9.
Global Consortium for the Advancement of Promotion and Prevention in Mental Health (From Margins to Mainstream: 5th World Conference on the Promotion of Mental Health and the Prevention of Mental and Behavioural Disorders, Melbourne, 2008) (2009). *The Melbourne Charter for Promoting Mental Health and Preventing Mental and Behavioural Disorders.* [Online] Available at: http://www.vichealth.vic.gov.au/Publications/Mental-health-promotion/Melbourne-Charter.aspx
Goffman, E. (1963). *Stigma: Notes on Management of Spoiled Identity*. Englewood Cliffs, NJ: Prentice-Hall.
Goldie, I., Quinn, N., and Knifton, L. (2012). Best practice challenging stigma and discrimination against people with depression: best practice guidelines, values and resources. The ASPEN Project. [Online] Available at: http://www.antistigma.eu/sites/default/files/ASPEN_WP4_BEST_PRACTICE_TOOLKIT.pdf
Gotink, R.A., Chu, P., Busschbach, J.J.V., Benson, H., Fricchione, G.L., and Hunink, M.G.M. (2015). Standardised mindfulness-based interventions in healthcare: an overview of systematic reviews and meta-analyses of RCTs. *PloS One*, 10(4), e0124344.
Greenberg, P.E., Fournier, A-A., Sisitsky, T., Pike, C.T., and Kessler, R.C (2015). The economic burden of adults with major depressive disorder in the United States (2005 and 2010). *Journal of Clinical Psychiatry*, 72(2), 155–62.
Hadlaczky, G. and Wasserman, D. (2009). Suicidality in women. In: Chandra, P.S., et al. (eds.) *Contemporary Topics in Women's Mental Health: Global Perspectives in a Changing Society*. Chichester: Wiley, pp. 117–37.
Hatzenbuehler, M.L. and Keyes, K.M. (2013). Inclusive anti-bullying policies and reduced risk of suicide attempts in lesbian and gay youth. *Journal of Adolescent Health*, 53(Suppl. 1), S21.e6.
Hawkins, J.D., Catalano, R.F., and Arthur, M.W. (2002). Promoting science-based prevention in communities. *Addictive Behaviors*, 27(6), 951–76.
Hegerl, U., Dietrich, S., Pfeiffer-Gerschel, T., Wittenburg, L., and Althaus, D. (2009). Education and awareness programmes for adults: selected and multilevel approaches in suicide prevention. In: Wasserman, D. and Wasserman, C. (eds.) *The Oxford Textbook of Suicidology and Suicide Prevention: A Global Perspective*. Oxford: Oxford University Press, pp. 495–500.
Henderson, C. and Thornicroft, G. (2013). Evaluation of the Time to Change programme in England 2008–2011. *British Journal of Psychiatry*, 202, s45–8.
Hern, M., Bartholomew, M.L., and Weahkee, R.L. (2014). Suicide mortality among American Indians and Alaska Natives, 1999–2009. *American Journal of Public Health*, 104(3), s336–42.
Hertz, M.F., Donato, I., and Wright, J. (2013). Bullying and suicide: a public health approach. *Journal of Adolescent Health*, 53(Suppl. 1), S1–S3.
Holt, M.K., et al. (2015). Bullying and suicidal ideation and behaviors: a meta-analysis. *Pediatrics*, 135(2), e496–509.
Hoven, C.W., Wasserman, D., Wasserman, C., and Mandell, D.J. (2009). Awareness in nine countries: a public health approach to suicide prevention. *Legal Medicine*, 11(Suppl. 1), S13–17.
Huppert, F. (2008). *Psychological Well-Being: Evidence Regarding its Causes and its Consequences*. London: Foresight Mental Capital and Wellbeing Project.

Jané-Llopis, E., et al. (2011). Reducing the silent burden of impaired mental health. *Journal of Health Communication*, 16(Suppl. 2), 59–74.
Jeon, S. Y., Reither, E.N., and Masters, R.K. (2016). A population-based analysis of increasing rates of suicide mortality in Japan and South Korea, 1985–2010. *BMC Public Health*, **16**, 356.
Johnson, S. (2006). *The Ghost Map: The Story of London's Most Terrifying Epidemic: And How it Changed Science, Cities and the Modern World*. London: Riverhead Books.
Kaltiala-Heino, R., Rimpelä, M., Marttunen, M., Rimpelä, A., and Rantanen, P. (1999). Bullying, depression, and suicidal ideation in Finnish adolescents: school survey. *British Medical Journal*, 319(7206), 348–51.
Kassebaum N., et al (2016). Global, regional, and national disability-adjusted life-years (DALYs) for 315 diseases and injuries and healthy life expectancy (HALE), 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015. *The Lancet*, **388**, 1603–58.
Khan, M.M. and Syed, E.U. (2011). Suicide in Asia: epidemiology, risk factors, and prevention. In: O'Connor, R.C., Platt, S., and Gordon, J. (eds.) *International Handbook of Suicide Prevention: Research, Policy and Practice*. Oxford: Wiley-Blackwell, pp. 487–506.
Kirkwood, T., Bond, J., May, C., McKeith, I., and Teh, M. (2008). *Mental Capital Through Life Challenge Report*. London: Foresight Mental Capital and Wellbeing Project.
Klomek, A.B., Sourander, A., and Gould, A. (2010). The association of suicide and bullying in childhood to young adulthood: a review of cross-sectional and longitudinal research findings. *Canadian Journal of Psychiatry*, **55**, 282.e8.
Kvam, S., Kleppe, C.L., Nordhus, I.H., and Hovland, A. (2016). Exercise as a treatment for depression: a meta-analysis. *Journal of Affective Disorders*, **202**, 67–86.
Ladwig, K.-H., Ruf, E., Baumert, J., and Erazo, N. (2009). Prevention of metropolitan and railway suicide. In: Wasserman, D. and Wasserman, C. (eds.) *The Oxford Textbook of Suicidology and Suicide Prevention: A Global Perspective*. Oxford: Oxford University Press, pp. 589–94.
Lahti, M., et al. (2010). Prenatal origins of hospitalization for personality disorders: the Helsinki Birth Cohort Study. *Psychiatry Research*, 179(2), 226–30.
Lai, S.J., et al. (2014). A systematic review and meta-analysis of dietary patterns and depression in community-dwelling adults. *The American Journal of Clinical Nutrition*, 99(1), 181–97
Lasalvia, A., et al. (2013). Global pattern of experienced and anticipated discrimination reported by people with major depressive disorder: a cross-sectional survey. *The Lancet*, 381(9860), 55–62.
Lauber, C., and Rössler, W. (2007). Stigma towards people with mental illness in developing countries in Asia. *International Review of Psychiatry*, 19(2), 157–178.
Lauritzen, C., and Reedts, C. (2015). Knowledge transfer in the field of parental mental illness: objectives, effective strategies, indicators of success, and sustainability. *International Journal of Mental Health Systems*, **9**, 6.
Law, S. and Liu, P. (2008). Suicide in China: unique demographic patterns and relationship to depressive disorder. *Current Psychiatry Reports*, **10**, 80–6.
Link, B.G. and Phelan, J.C. (2001). Conceptualising stigma. *American Sociological Review*, **27**, 363–85.
Littlewood, R. (1998). Cultural variation in the stigmatisation of mental illness. *The Lancet*, **352**, 1056–7.
Liu, Y., et al. (2017). Maternal depressive symptoms and early childhood cognitive development: a meta-analysis. *Psychological Medicine*, **47**, 680–689.
Luciano, A., et al. (2014). Evidence-based supported employment for people with severe mental illness: past, current, and future research. *Journal of Vocational Rehabilitation*, **40**, 1–13.
Madigan, S., Moran, G., and Schuengel, C. (2007). Unresolved maternal attachment representations, disrupted maternal behaviour and disorganized attachment in infancy: links to

toddler behaviour problems. *Journal of Child Psychology and Psychiatry*, **48**, 1042–50.
Mann, J., et al. (2005). Suicide prevention strategies—a systematic review. *Journal of the American Medical Association*, 294(16), 2064–74.
Mars, B., Burrows, S., Hjelmeland, H., and Gunnell, D., (2014). Suicidal behaviour across the African continent: a review of the literature. *BMC Public Health*, **14**, 606.
Marschall-Levesque, B., et al. (2016). Victimization, suicidal ideation, and alcohol use from age 13 to 15 years: support for the self-medication model. *Journal of Adolescent Health*, **60**, 380–7.
McMahon, E.M., et al. (2017). Physical activity in European adolescents and associations with anxiety, depression and well-being. *European Child & Adolescent Psychiatry*, 26(1), 111–22.
Mew, E.J., et al. (2017). The global burden of fatal self-poisoning with pesticides 2006–15: systematic review. *Journal of Affective Disorders*, **219**, 93–104.
Middleton, N., Sterne, J.A.C., and Gunnell, D. (2006). The geography of despair among 15–44-year-old men in England and Wales: putting suicide on the map. *Journal of Epidemiology and Community Health*, **60**, 1040–7.
Mills, R., Scott, J., Alati, R., O'Callaghan, M., Najman, J.M., and Strathearn L. (2013). Child maltreatment and adolescent mental health problems in a large birth cohort. *Child Abuse & Neglect*, **37**, 292–302.
Milner, A., McClure, R., and De Leo, D. (2012). Socio-economic determinants of suicide: an ecological analysis of 35 countries. *Social Psychiatry and Psychiatric Epidemiology*. 47(1), 19–27
Mittendorfer-Rutz, E., Rasmussen, F., and Wasserman, D. (2004). Restricted fetal growth and adverse maternal psychosocial and socioeconomic conditions as risk factors for suicidal behaviour of offspring: a cohort study. *The Lancet*, 364(9440), 1135–40.
Mrazek, P. and Haggerty, R. (1994). *Reducing Risks of Mental Disorder: Frontiers for Preventive Intervention Research*. Washington, DC: National Academy Press.
Murray, C.J.L., Vos, T., Lozano, R., et al. (2012). Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010. *The Lancet*, **380**, 2197–223.
Mäkinen, I.H. (2009). Social theories of suicide. In: Wasserman, D. and Wasserman, C. (eds.) *Oxford Textbook of Suicidology and Suicide Prevention*. United Kingdom: Oxford University Press, pp. 139–47.
Niederkrotenthaler, T. and Sonneck, G. (2007). Assessing the impact of media guidelines for reporting on suicides in Austria: interrupted time series analysis. *Australian & New Zealand Journal of Psychiatry*, 41(5), 419–28.
Niederkrotenthaler, T., et al. (2010). Role of media reports in completed and prevented suicide. Werther v. Papageno effect. *British Journal of Psychiatry*, 197(3), 234–43.
Niederkrotenthaler, T., et al. (2014). Future risk of labour market marginalization in young suicide attempters—a population-based prospective cohort study. *International Journal of Epidemiology*, 43(5), 1520–30.
Nordentoft, M. (2007). Prevention of suicide and attempted suicide in Denmark. *Danish Medical Bulletin*, 54(2), 306–69.
Norström, T. and Ramstedt, M. (2005). Mortality and population drinking: a review of the literature. *Drug and Alcohol Review*, 24(6), 537–47.
Norstöm, T., and Grönqvist, H. (2015). The Great Recession, unemployment and suicide. *Journal of Epidemiology and Community Health*, 69(2), 110–16.
Nous Group (2014). Independent evaluation of beyondblue. [Online] Available at: https://www.beyondblue.org.au/docs/default-source/research-project-files/bw0265.pdf?sfvrsn=33739de9_0
Nyberg, A., et al. (2011). Destructive managerial leadership and psychological well-being among employees in Swedish, Polish, and Italian hotels. *Work*, 39(3), 267–281.

Oksanen, T., et al. (2010). Prospective study of workplace social capital and depression: are vertical and horizontal components equally important? *Journal of Epidemiology and Community Health*, 64(8), 684–9.

Olds, D.L. (2002). Prenatal and infancy home visiting by nurses: from randomized trials to community replication. *Prevention Science*, **3**, 1153–72.

O'Neil, A., et al. (2014). Relationship between diet and mental health in children and adolescents: a systematic review. *American Journal of Public Health*, 104, e31–e42.

Pacot, R., Garmit, B., Pradem, M., Nacher, M., and Brousse, P. (2018). The problem of suicide among Amerindians in Camopi-Trois Sauts, French Guiana 2008–2015. *BMC Psychiatry*, **18**, 99.

Patel, V., Chisholm, D., Dua T., Laxminarayan R., and Medina-Mora, M.E. (2015). *Mental, Neurological, and Substance Use: Disease Control Priorities* (3rd ed., *Volume 4)*. Washington, DC: The International Bank for Reconstruction and Development/The World Bank.

Patel, V. and Thornicroft, G. (2009). Packages of care for mental, neurological, and substance use disorders in low- and middle-income countries. *PLoS Medicine*, 6(10), e1000160.

Petersen, I., Swartz, L., Bhana, A., and Flisher, A.J. (2010). Mental health promotion initiatives for children and youth in contexts of poverty: the case of South Africa. *Health Promotion International*, 25(3), 331–41.

Phillips, M.R. and Gunnell, D. (2009). Restrictions of access to pesticides in suicide prevention. In: Wasserman, D. and Wasserman, C. (eds.) *The Oxford Textbook of Suicidology and Suicide Prevention: A Global Perspective*, pp. 583–8. Oxford: Oxford University Press.

Pirkis, J., Too, L.S., Spittal, M.J., Krysinska, K., Robinson, L., and Cheung, Y.T.D. (2015). Interventions to reduce suicides at suicide hotspots: a systematic review and meta-analysis. *Lancet Psychiatry*, **2**, 994–1001.

Pirkola, S., Sund, R., Sailas, E., and Wahlbeck, K. (2009). Community mental health services and suicide rate in Finland: a nationwide small area analysis. *The Lancet*, 373(9658), 147–53.

Pitman, A. and Caine, E. (2012). The role of the high-risk approach in suicide prevention. *British Journal of Psychiatry*, **201**, 175–7.

Pitt, V., Lowe, D., Pictor, M., Hetrick, S.E., Ryan, R., and Berends, L. (2013). Consumer-providers of care for adult clients of statutory mental health services (Review). *Cochrane Database of Systematic Reviews*, 3, CD004807.

Powell, N.R., Lochman, J.E., and Boxmeyer, C.L. (2007). The prevention of conduct problems. *International Review of Psychiatry*, 19(6), 597–605.

Qi, X., Hu, W., Page, A., and Tong, S. (2014). Dynamic pattern of suicide in Australia, 1986–2005: a descriptive-analytic study. *BMJ Open*, 4, e005311.

Quinn, N., Shulman, A., Knifton, L., and Byrne, P. (2011). The impact of a national mental health arts and film festival on stigma and recovery. *Acta Psychiatrica Scandinavica*, 123(1), 71–81.

Räikkönen, K., et al. (2011). Risk of severe mental disorders in adults separated temporarily from their parents in childhood: the Helsinki Birth Cohort Study. *Journal of Psychiatric Research*, 45(3), 332–8.

Räikkönen, K., et al. (2008). Depression in young adults with very low birth weight: the Helsinki study of very low-birth-weight adults. *Archives of General Psychiatry*, **65**, 290–6.

Rehkopf, D.H. and Buka, S.L. (2005). The association between suicide and the socio-economic characteristics of geographical areas: a systematic review. *Psychological Medicine*, 36(2), 145–57.

Rehm, J., et al. (2003). The relationship of average volume of alcohol consumption and patterns of drinking to burden of disease: an overview. *Addiction*, **98**, 1209–28.

Resnick, S., et al. (2017). Firearm deaths in America: can we learn from 462,000 lives lost?. *Annals of Surgery*, 266(3), 432–40.

Richter, L., Dawes, A., and de Kadt, J. (2010). Early childhood. In: Petersen, I., Bhana, A., Swartz, L., Flisher, A., and Richter, L. (eds.) *Mental Health Promotion and Prevention for Poorly Resourced Contexts: Emerging Evidence and Practice*. Pretoria: HSRC Press, pp. 91–123.

Rodriguez Andrés, A. and Hempstead, K. (2011). Gun control and suicide: the impact of state firearm regulations in the United States 1995– 2004. *Health Policy*, **101**, 95–103.

Roskar, S., et al. (2010). Effects of training program on recognition and management of depression and suicide risk evaluation for Slovenian primary-care physicians: follow-up study. *Croatian Medical Journal*, 51(3), 273–42.

Rüsch, N., Angermeyer, M.C., and Corrigan, P.W. (2005). Mental illness stigma: concepts, consequences, and initiatives to reduce stigma. *European Psychiatry*, **20**, 529–39.

Rutz, W. (2001). Preventing suicide and premature death by education and treatment. *Journal of Affective Disorders*, **62**, 123–9.

Ryan, C., Huebner, D., Diaz, R.M., and Sanchez, J. (2009). Family rejection as a predictor of negative health outcomes in white and Latino lesbian, gay, and bisexual young adults. *Pediatrics*, **123**, 346–52.

Ryan, R.M. and Deci, E.L. (2000). Self-determination theory and facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, **55**, 68–78.

Saluja, G., Iachan, R., Scheidt, P.C., Overpeck, M.D., Sun, W., and Geidd, J.N. (2004). Prevalence and risk factors for depressive symptoms among young adolescents. *Archives of Pediatric & Adolescent Medicine*, 158(8), 760–5.

Sandler, I., et al. (2014). Overview of meta-analyses of the prevention of mental health, substance use and conduct problems. *Annual Reviews of Clinical Psychology*, **10**, 243–73.

Sarchiapone, M., et al. (2007). Childhood trauma and suicide attempts in patients with unipolar depression. *Depression and Anxiety*, 24(4), 268–72.

Sartorius, N. (2010). Short-lived campaigns are not enough. *Nature*, **468**, 163–5.

Schmidt, S.J., et al. (2015). EPA guidance on the early intervention in clinical high risk states of psychoses. *European Psychiatry*, **30**, 388–404.

Seeman, N., Tang, S., Brown, A.D., and Ing, A. (2016). World survey of mental illness stigma. *Journal of Affective Disorders*, **190**, 115–21.

Semrau, M., Evans-Lacko, S., Koschorke, M., Ashenafi, L., and Thornicroft, G. (2015). Stigma and discrimination related to mental illness in low- and middle-income countries. *Epidemiology and Psychiatric Sciences*, **24**, 382–94.

Sher, L., et al. (2008). Depressed patients with co-occurring alcohol use disorders: a unique patient population. *Journal of Clinical Psychiatry*, 69(6), 907–15.

Siegenthaler, E., Munder, T., and Egger, M. (2012). Effect of preventive interventions in mentally ill parents on the mental health of the offspring: systematic review and meta-analysis. *Journal of the American Academy of Child & Adolescent Psychiatry*, 51(1), 8–17.

Sinyor, M., et al. (2017). Did the suicide barrier work after all? Revisiting the Bloor Viaduct natural experiment and its impact on suicide rates in Toronto. *BMJ Open*, 7, e015299.

Sisask, M., et al. (2014). Teacher satisfaction with school and psychological well-being affects their readiness to help children with mental health problems. *Health Education Journal*, **73**, 382–93.

Smit, F., et al. (2006). Cost-effectiveness of preventing depression in primary care patients: randomised trial. *British Journal of Psychiatry*, **188**, 330–6.

Sobocki, P., Jönsson, B., Angst, J., and Rehnberg, C. (2006). Cost of depression in Europe. *Journal of Mental Health Policy and Economics*, 9(2), 87–98.

Sobocki, P., et al. (2007). The economic burden of depression in Sweden from 1997 to 2005. *European Psychiatry*, 22(3), 146–52.

Stallard, P., et al. (2014). Classroom-based cognitive behaviour therapy (FRIENDS): a cluster randomised controlled trial to Prevent Anxiety in Children through Education in Schools (PACES). *Lancet Psychiatry*, **1**, 185–192.
Stewart-Brown, S.L. and Schrader-McMillan, A. (2011). Parenting for mental health: what does the evidence say we need to do? Report of Workpackage 2 of the DataPrev project. *Health Promotion International*, 26(Suppl. 1), i10–28.
Stice, E., Becker, C.B., and Yokum, S. (2013). Eating disorder prevention: current evidence-base and future directions. *International Journal of Eating Disorders*, 46(5), 478–85.
Stice, E., Yokum, S., Burger, K., Rohde, P., Shaw, H., and Gau, J.M. (2015). A pilot randomized trial of a cognitive reappraisal obesity prevention program. *Physiology & Behavior*, **138**, 124–32.
Stuart, H. (2016) Reducing the stigma of mental illness. *Global Mental Health*, 3(e17), 1–14.
Stuckler, D., et al. (2009). The public health effect of economic crises and alternative policy responses in Europe: an empirical analysis. *The Lancet*, 374(9686), 315–23.
Tansella, M. and Thornicroft, G. (1999). *Common Mental Disorders in Primary Care*. London: Routledge.
Tennant, R., et al. (2007). The Warwick–Edinburgh Mental Well-being Scale (WEMWBS): development and UK validation. *Health and Quality of Life Outcomes*, **5**, 63.
Theorell, T., et al. (2015). A systematic review including meta-analysis of work environment and depressive symptoms. *BMC Public Health*, **15**, 738.
Thornicroft, G., Brohan, E., Rose, D., Sartorius, N., Leese, M.; INDIGO Study Group (2009). Global pattern of experienced and anticipated discrimination against people with schizophrenia: a cross-sectional survey. *The Lancet*, 373(9661), 408–15.
Thornicroft, G. and Tansella, M. (1999). *The Mental Health Matrix. A Manual to Improve Services*. Cambridge: Cambridge University Press.
Thornicroft, G. and Tansella, M. (2002). Balancing community-based and hospital-based mental health care. *World Psychiatry*, 1(2), 84–90.
Tidemalm, D., et al. (2008). Risk of suicide after suicide attempt according to coexisting psychiatric disorder: Swedish cohort study with long term follow-up. *British Medical Journal*, **337**, 1328–30.
Ueda, M., Mori, K., Matsubayashi, T., and Sawada, Y. (2017). Tweeting celebrity suicides: users' reaction to prominent suicide deaths on Twitter and subsequent increases in actual suicides. *Social Science & Medicine*, **189**, 158–66.
Ueda, M., Sawada, Y., and Matsubayashi, T. (2015). The effectiveness of installing physical barriers for preventing railway suicides and accidents: evidence from Japan. *Journal of Affective Disorders*, **178**, 1–4.
United Nations (2017). *Sustainable Development Goals Report 2017*. New York: United Nations. [Online] Available at: https://unstats.un.org/sdgs/files/report/2017/TheSustainableDevelopmentGoalsReport2017.pdf.
Vaingankar, J.A., et al. (2018). Psychometric properties and population norms of the positive mental health instrument in a representative multi-ethnic Asian population. *BMC Medical Research Methodology*, **18**, 29.
Van der Aa, H.P.A., van Rens, G.H.M.B., Bosmans, J.E., Comijs, H.C., and van Nispen, R.M.A. (2017). Economic evaluation of stepped-care versus usual care for depression and anxiety in older adults with vision impairment: randomized controlled trial. *BMC Psychiatry*, **17**, 280.
Van't Veer-Tazelaar, P.J., et al. (2009). Stepped-care prevention of anxiety and depression in late life. A randomized controlled trial. *Archives of General Psychiatry*, 66(3), 297–304.
Värnik, P. (2012). Suicide in the world. *International Journal of Environmental Research and Public Health*, **9**, 1.

Velasco, I., Bath, S.C., and Rayman, M.P. (2018). Iodine as essential nutrient during the first 1000 days of life. *Nutrients*, 10(3), 290.
Vijayakumar, L. (2004). Suicide prevention: the urgent need in developing countries. *World Psychiatry*, 3(3), 158–9.
Vreeman, R.C. and Carroll, A.E. (2007). A systematic review of school-based interventions to prevent bullying. *Archives of Pediatrics & Adolescent Medicine*, **161**, 78–88.
Wahlbeck, K. (2009). *Background Document for the Thematic Conference on Prevention of Depression and Suicide*. Luxembourg: European Communities.
Wahlbeck, K. (2011a). European comparisons between mental health services. *Epidemiology and Psychiatric Sciences*, 20(1), 15–18.
Wahlbeck, K. (2011b). European mental health policy should target everybody. *European Journal of Public Health*, 21(5), 551–3.
Wahlbeck, K. (2015). Public mental health: the time is ripe for translation of evidence into practice. *World Psychiatry*, 14(1), 36–42.
Wahlbeck, K., et al. (2001). Association of schizophrenia with low maternal body mass index, small size at birth, and thinness during childhood. *Archives of General Psychiatry*, **58**, 48–52.
Wahlbeck, K. and Taipale, V. (2006). Europe's mental health strategy. *BMJ*, **333**, 210–11.
Wasserman, C., et al. (2012). Suicide prevention for youth—a mental health awareness program: lessons learned from the Saving and Empowering Young Lives in Europe (SEYLE) intervention study. *BMC Public Health*, 12(1), 776.
Wasserman, C., Postuvan, V., Herta, D., Iosue, M., Värnik, P., and Carli, V. (2018). Interactions between youth and mental health professionals: the Youth Aware of Mental health (YAM) program experience. *PLoS One*, 13(2), e0191843.
Wasserman, D. (2016). *Suicide: An Unnecessary Death* (2nd ed.). New York: Oxford University Press.
Wasserman, D., et al. (2010). Saving and Empowering Young Lives in Europe (SEYLE): a randomized controlled trial. *BMC Public Health*, **10**, 192.
Wasserman, D., et al. (2015). School-based suicide prevention programmes: the SEYLE cluster-randomised, controlled trial. *The Lancet*, 385(9977), 1536–44.
Wasserman, D., et al. (2012). The European Psychiatric Association (EPA) guidance on suicide treatment and prevention. *European Psychiatry*, 27(2), 129–41.
Wasserman, D., Värnik, A., Dankowicz, M., and Eklund, G. (1998). Suicide-preventive effects of perestroika in the former USSR: the role of alcohol restriction. *Acta Psychiatrica Scandinavica Supplementum*, **394**, 1–44.
Weare, K. and Nind, M. (2011). Mental health promotion and problem prevention in schools: what does the evidence say? *Health Promotion International*, 26(Suppl. 1), i29–69.
Weersing, V.R., et al. (2016). Prevention of depression in at-risk adolescents: predictors and moderators of acute effects. *Journal of the American Academy of Child & Adolescent Psychiatry*, 55(3), 219–26.
Weissman, M.M., Wickramaratne, P., Nomura, Y., Warner, V., Pilowsky, D., and Verdeli, H. (2006). Offspring of Depressed Parents: 20 Years Later. *American Journal of Psychiatry*, **163**, 1001–8.
Westerlund, M., Sylvia, S., and Schmidtke, A. (2009). The role of mass-media reporting and suicide prevention. In: Wasserman, D. and Wasserman, C. (eds.) *The Oxford Textbook of Suicidology and Suicide Prevention: A Global Perspective*. Oxford: Oxford University Press, pp. 515–24.
Westerlund, M., and Wasserman, D. (2009). The role of the internet in suicide prevention. In: Wasserman, D. and Wasserman, C. (eds.) *The Oxford Textbook of Suicidology and Suicide Prevention: A Global Perspective*. Oxford: Oxford University Press, pp. 525–32.

Westman, J., et al. (2006). Country of birth and suicide: a follow-up study of a national cohort in Sweden. *Archives of Suicide Research*, 10(3), 239–48.

While, D., et al. (2012). Implementation of mental health service recommendations in England and Wales and suicide rates, 1997–2006: a cross-sectional and before-and-after observational study. *The Lancet*, **379**, 1005–12.

Wittchen, H.U., et al. (2011). The size and burden of mental disorders and other disorders of the brain in Europe 2010. *European Neuropsychopharmacology*, **21**, 655–79.

World Health Organization (WHO) (2004). *Prevention of Mental Disorders: Effective Interventions and Policy Options: Summary Report*. Geneva: WHO.

World Health Organization (WHO) (2005). *Empowerment and Mental Health Advocacy. Briefing Paper for the WHO European Ministerial Conference on Mental Health: Facing the Challenges, Building Solutions*. Copenhagen: WHO.

World Health Organization (WHO) (2009). *Mental Health Systems in Selected Low- and Middle-Income Countries: A WHO-AIMS Cross-Nation Analysis*. Geneva: WHO.

World Health Organization (WHO) (2010). *Guns, Knives and Pesticides: Reducing Access to Lethal Means*. Geneva: WHO. [Online] Available at: https://www.who.int/mental_health/prevention/suicide/vip_pesticides.pdf

World Health Organization (WHO) (2013). *Mental Health Action Plan 2013–2020*. Geneva: WHO. [Online] Available at: http://apps.who.int/iris/bitstream/handle/10665/89966/9789241506021_eng.pdf?sequence=1

World Health Organization (WHO) (2014). Preventing suicide: a global imperative. Geneva: WHO. [Online] Available at: http://apps.who.int/iris/bitstream/handle/10665/131056/9789241564878_eng.pdf;jsessionid=E9D04176FDB088AF4D135ACDAC1CF3D1?sequence=8

World Health Organizations (WHO) (2017a). Age-standardized suicide rates (per 100 000 population), both sexes, 2015. Geneva: WHO. [Online] Available at: http://gamapserver.who.int/mapLibrary/Files/Maps/Global_AS_suicide_rates_bothsexes_2015.png?ua=1

World Health Organization (WHO) (2017b). *Preventing Suicide: A Resource for Media Professionals, Update 2017*. Geneva: WHO.

World Health Organization (WHO) (n.d.). *Social Determinants of Health*. [Online] Available at: http://www.who.int/hia/evidence/doh/en/index.html.

Yip, P.S.F., Caine, E., Yousuf, S., Chang, S.-S., Wu, K.C.-C., and Chen, Y.-Y. (2012). Means restriction for suicide prevention. *The Lancet*, **379**, 2393–9.

Zalsman, G., et al. (2016). Suicide prevention strategies revised: 10-year systematic review. *Lancet Psychiatry*, **3**, 646–59.

8.8

Dental public health

Amira S. Mohamed and Peter G. Robinson

Introduction to dental public health

Poor oral health is a major public health problem. The considerable burden of oral disease at a societal level reflects the cumulative impact of many people affected to a modest degree. Oral diseases are exceedingly widespread. Their impact on individuals and the community are significant, ranging from mortality to effects on general health and quality of life, and in terms of direct costs and lost productivity.

The scope of dental public health

Dental public health is the science and art of preventing oral disease, promoting oral health, and improving the quality of life through the organized efforts of society.

Major areas of dental public health activity include:

Oral health surveillance.
Assessing the evidence on oral health and dental interventions, programmes, and services.
Policy and strategy development, and implementation.
Strategic leadership and collaborative working for health.
Oral health improvement.
Health and public protection.
Developing and monitoring quality dental services.
Dental public health intelligence.
Academic dental public health.
Role within health services.

In some countries the specialty also involves providing services to special population subgroups.

This chapter reviews the epidemiology, aetiology, and management of the four oral conditions of greatest public health importance: dental caries, periodontal diseases, oral cancer, and orofacial trauma. Approaches to oral health improvement are then outlined using the framework of the Commission on the Social Determinants of Health. Finally, we consider future developments in population oral health.

Oral diseases of public health importance

Dental caries

Dental caries is the demineralization of tooth substance by acid metabolites of oral bacteria. In the very early stages, the lesion appears as a chalky white spot. If the lesion progresses, the surface of the tooth breaks down leading to cavitation. If the caries reaches the underlying dentine, it can spread more readily through this porous and less mineralized tissue towards the pulp. Infection of the pulp may allow the passage of bacteria along the root canals to the alveolar bone.

The direct consequences of this process are destruction of the tooth, pain, and a possible dental abscess. When dentine is exposed by cavitation, there may be transient pain associated with hot or cold drinks or sweet foods. Later, as the pulp becomes inflamed, the discomfort may be spontaneous, exquisitely painful, and of longer duration. In a dental abscess, pressure to the tooth is transmitted to the infected alveolus and the affected person avoids biting or knocking the tooth.

Risk factors

Four factors are necessary for caries development: dietary sugars, a susceptible tooth surface, the microflora of dental plaque, and adequate time. Dental plaque that forms on the tooth surface comprises mainly bacteria, particularly *Streptococcus mutans*, which metabolize sugars to produce acids. The acids cause demineralization of the tooth structure. With each exposure to sugar, the plaque pH falls sharply and rises slowly back to

normal levels over the following hour. Hence, caries incidence is related to the frequency of sugar intake.

At high pH, there is remineralization of the tooth, especially in the presence of fluoride. Saliva plays a crucial protective role against caries by simple dilution, buffering plaque acid, and acting as a source of minerals and immunological plaque inhibitory factors. For these reasons, dental caries is more frequent in the sites less accessible to saliva as the pits and fissures of posterior teeth and between these teeth and also in people with restricted salivary flow.

Epidemiology

Caries of the permanent dentition is traditionally measured by the number of decayed, missing, and filled teeth (the DMFT index). While, the dft index is used for deciduous teeth due to difficulty in differentiation between the naturally exfoliated teeth and missing teeth due to caries. For more precision, the number of surfaces affected can be recorded (DMFS, dfs). The index aggregates both disease and treatment experience, hence it is affected by treatment decisions of dentists and is less valid with increasing age. Since each of the categories is equally weighted, it is insensitive to both the severity of the disease and outcomes of treatment. Nonetheless, the DMFT index, having been used for almost 80 years, will continue to be used for some time. In addition to the assessment of the prevalence of caries, its impact is now routinely measured, most frequently by measures of oral health-related quality of life.

Typically, the pattern has been of high caries levels in developed countries associated with exposure to sugars. In the mid-1970s, DMF levels in 12 years old began to fall dramatically in many developed countries such as the United States and European high-income countries, with less obvious decreases in low-income countries (Kassebaum et al. 2017). National surveys in United Kingdom over 40 years (1973–2013) showed that the prevalence of obvious cavitation (caries in dentin) decreased gradually from 72% to 26% from 1973 till 2013 in 5-year-old children. Similarly, 15-year-old children showed a fall in obvious caries prevalence from 97% in 1973 to 42% in 2013 (Murray et al. 2015). The National Health and Nutrition Examination Survey (NHANES) in 2011–2012 also showed a decrease in caries in primary teeth and to a lower extend in permanent teeth in

comparison to 1999–2004 (Dye et al. 2007). This decrease can be attributed to greater availability of fluoride, changing diagnostic behaviours, and greater access to care in some countries.

Aggregate national data on caries levels are useful, but mask important trends, particularly inequalities in oral health. When disease prevalence was high, disease was almost universal and inequalities were manifest as differences in the numbers of teeth affected in individuals. With lower disease prevalence, the distribution is highly skewed with typically 20–25% of the population having moderate or severe caries (more than three teeth affected) and the majority having no obvious caries. This affected minority tends to be those living in the most deprived areas. Furthermore, dental caries has less impact on the lives of children and older adults from high-income environments than those from low-income environments. These findings reinforce the importance of measuring both the disease impact and prevalence. Based on the World Health Organization (WHO) databank data from 2000 to 2015, the median of mean DMF in 12-year-old children ranged from 0.9 in low-income countries to 2.1 in upper-middle countries (Frencken et al. 2017).

Although data are scarce, there are concerns with the rising caries levels in children in some countries undergoing economic growth and nutritional transition. For example, dental caries is very common and severe in 12-year-old children from several Asian and South American countries but remains relatively low in most African countries where poor economic growth restricts sugar consumption. Here again, national data may mask local variations, particularly high caries levels in urban areas.

When estimating the burden of dental caries, untreated caries is more relevant than treated caries. In 2015, the age-standardized prevalence of untreated caries was estimated to be 34.1% in permanent teeth, affecting 2.5 billion people and 7.8% in deciduous teeth, affecting 573 million people worldwide (Kassebaum et al. 2017).

Management

Until the nineteenth century, tooth extraction was the only useful treatment for dental caries. Treatment evolved to restorative care in which the infected parts of the tooth are removed and replaced with an inert obturating filling, for example, dental amalgam. During the latter half of the twentieth century,

in developed countries, operative treatment for adults has become increasingly complex and technology intensive. Badly decayed teeth can now be restored with a range of adhesive tooth-coloured materials that are either formed in the mouth or prepared in laboratories and then fitted. Originally, missing teeth could only be replaced with removable dentures. Now they can be replaced with bridges or prostheses supported by osseointegrated implants that project out through the gingivae.

In developing countries, the necessary infrastructure is often not available for complex treatment of carious teeth. In some areas, atraumatic restorative treatment (ART) is the only sustainable method because the decay is removed with hand instruments and the cavities filled with glass-ionomer cements. The instruments are easily portable and treatment can be provided painlessly, at low cost, without local anaesthesia, electricity, or expensive dental equipment. Systematic reviews have shown the effectiveness of the ART approach (Duangthip et al. 2017).

Implications of changes in caries prevalence

At the low disease prevalence in the developed world, proportionately more caries affects the accessible occlusal surfaces of the teeth. Fissure sealants can be applied to these surfaces to prevent caries and are cost-effective at low caries levels. Only simple restorations are needed to treat existing disease at these sites. The disease also progresses more slowly, so allowing deferred operative treatment. Many lesions are detected at an earlier stage so that new dental materials can be used in minimally invasive techniques. Minimal intervention dentistry is a new approach that aims to preserve dental tissue by integrating caries prevention, remineralization of demineralized enamel and dentine, and conservative placement and replacement of restorations.

Increasing the intervals between dental examinations is safe and effective for children and adults with low disease incidence. There is greater scope to employ dental care professionals (DCPs), other than dentists, who can also deliver effective preventive interventions such as fluoride varnish. However, because of the skewed distribution of dental caries, there remains a minority with high levels of caries to be managed.

There is also a large cohort of people more than 50 years old who have experienced severe dental caries and cycles of re-restoration. This so-called

heavy metal generation (Fig. 8.8.1) will continue to need complex and specialized treatment for decades to come, with high expectations of retaining their natural dentition into older age. By contrast, the needs of younger people are likely to be more straightforward. The proportion of dental treatment that falls within this straightforward category will continue to grow and can be delegated to other DCPs, especially in deprived communities, hence liberating dentists' time to undertake more complex treatments.

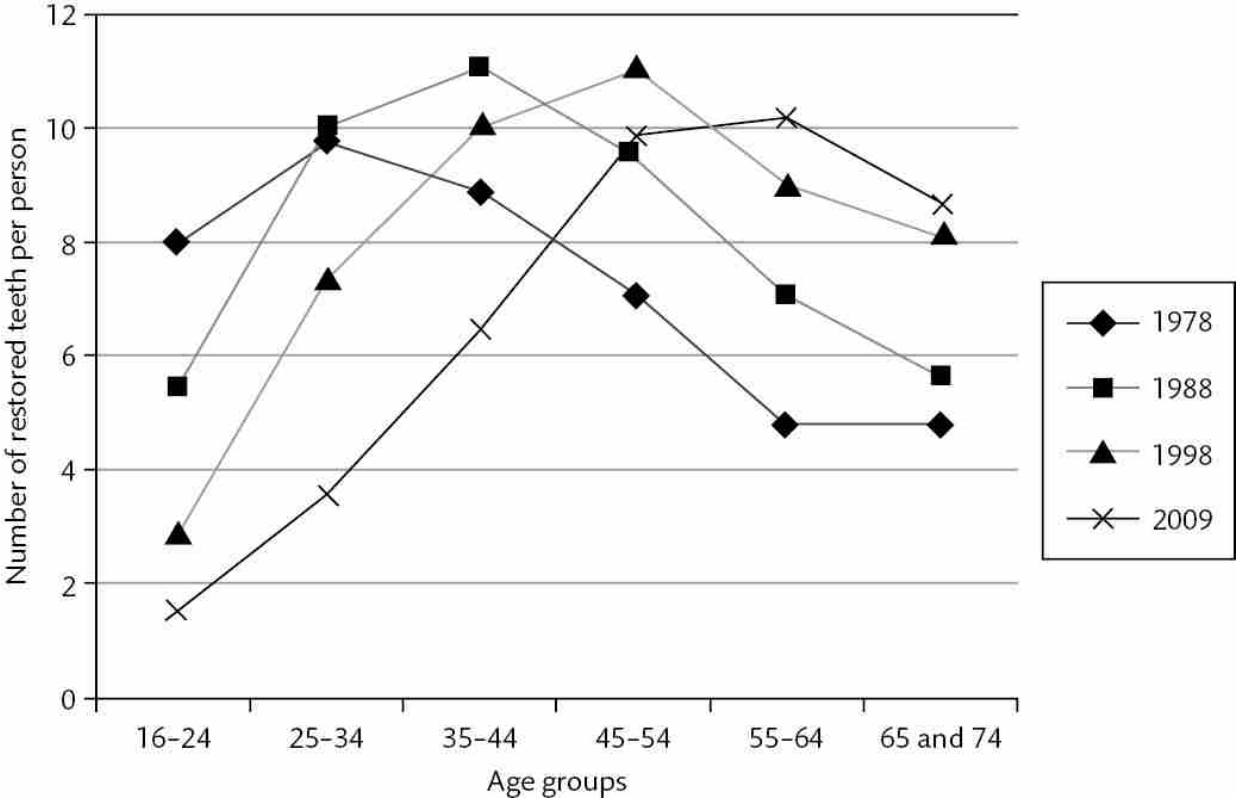


Fig. 8.8.1 The 'heavy metal' cohort.

Reproduced from Steele, J. et al., Complexity and maintenance—a report from the Adult Dental Health Survey 2009, The Health and Social Care Information Centre, London, UK, Copyright © 2011, Re-used with the permission of the Health and Social Care Information Centre (now NHS Digital). All rights reserved.

Periodontal diseases

Periodontal diseases comprise a range of inflammatory diseases of the periodontium: the attachment between gingivae and tooth. In gingivitis, the attachment remains in its original position, whereas periodontitis is defined by migration of the epithelial attachment which reduces the amount of periodontal ligament and bone supporting the tooth.

Gingivitis, which is exceedingly common, is an inflammatory response to plaque. Along with redness and swelling, the gums may bleed on gentle provocation, such as tooth brushing. Pain is uncommon. Systemic involvement including hormonal changes, skin diseases, and medication use may modify these diseases or cause other gingival changes. In periodontitis, the loss of periodontal attachment deepens of the pockets between the gingivae and teeth and by recession of the gingivae. In severe cases, the supporting structures are so depleted that the teeth become loose. The disease is rarely painful unless an acute infection complicates a periodontal pocket (‘a lateral periodontal abscess’) or if the exposed root surfaces are temperature sensitive.

Periodontitis has been consistently linked with psychological and social consequences on everyday life, which intervention studies indicate is reduced by treatment. More severe and untreated disease results in the loss of teeth and the need for additional treatment to replace them.

Another periodontal disease of public health importance is acute necrotizing ulcerative gingivitis (ANUG or ‘trenchmouth’), which causes necrosis, ulceration, soreness, and bleeding of the gingivae. The ulcerated papillae may have a grey slough, and there may be a characteristic foetor. Lymphadenopathy, and mild fever are variable findings. In severe cases, necrosis may extend over adjacent tissues to cause gross destruction of oral, and facial tissues (known as cancrum oris or noma).

Pathogenesis

The pathogenesis of periodontitis involves the interaction of plaque pathogens with the host’s immune system. Considerable research is devoted to determining which, if any, specific pathogens are responsible for periodontal destruction (the ‘specific plaque hypothesis’). Dental plaque is ubiquitous, but destructive disease occurs only in a minority of people. Apart from plaque, there are important determinants of periodontal disease susceptibility.

It is now clear that tobacco exerts a powerful independent deleterious effect. Periodontal treatment is also less effective in smokers, meanwhile smoking cession can have a positive impact on the periodontium (Alexandridi et al. 2018). Stress is another risk factor: for example, greater occupational stress is associated with progression of periodontitis. ANUG has been noted among soldiers on difficult postings, students during exam terms, and in people experiencing negative life events.

Periodontitis often takes decades to become clinically detectable. While it is more common and severe with advanced years, periodontitis is not a consequence of age. It is associated with poor oral hygiene and does not progress in adults with good oral hygiene.

Epidemiology

Gingivitis affects almost half of the 15-year old children and its prevalence has remained stable through the UK national surveys over 40 years (Murray et al. 2015). Mild periodontal pocketing is common, being seen in approximately half of UK and 63% of US adults (White et al. 2011; Eke et al. 2012). Severe periodontitis is much less frequent. Lost attachment or pockets of 6 mm or more (thought to be sufficient to threaten tooth survival) are seen in 9% of UK and 25% of US adults. These apparent differences between countries may be in part attributed to different survey methods. In 2015, the age-standardized prevalence of severe periodontitis globally was 7.4% (affecting 538 million), with no sex difference in incidence (Kassebaum et al. 2017). The disease is more frequent and severe in countries where tooth cleaning practices are less sophisticated

In many developed countries, ANUG is a disease of young adults. There are no good incidence data, but anecdotally, it has become less frequent among developed populations in recent years. A variant of the disease is associated with HIV infection. ANUG is also seen in children in developing countries where it can progress in the absence of treatment.

Management

For most people, the progression of periodontal destruction is compatible with the retention of a natural dentition into old age. A significant minority of people may lose teeth from periodontal diseases and considerable effort is

spent by dental professionals to prevent and treat them, primarily by mechanically removing plaque through regular toothbrushing and flossing of teeth. Adjunctive treatments include removal of calcified plaque (calculus), planing away the superficial layers of the roots and periodontal surgery in severe cases.

Systematic reviews are beginning to provide an evidence base for periodontal therapy although there may be considerable Hawthorne effects reflecting differences between the effectiveness of treatment provided routinely in primary care and its efficacy in clinical trials conducted in university departments. There are also few long-term studies showing effects on tooth retention. A metanalysis showed that powered toothbrushes caused more reduction in plaque and gingivitis on both short term and long term (more than 3 months) if compared with the manual brushes (Yaacob et al. 2014). The most common professional procedure to prevent and treat periodontal diseases is removal of calculus by scaling and polishing. Evidence of low to moderate strength shows its effectiveness in improving measures of dental plaque and gingivitis especially if combined with oral hygiene instruction. There is no compelling evidence of benefits from professional mechanical plaque removal alone over simple oral hygiene instruction to prevent periodontal diseases. However, repeated oral hygiene instructions appear as beneficial as repeated professional mechanical plaque removal (Needleman et al. 2005, 2015). Non-surgical periodontal debridement is the gold standard for management of periodontal disease and provides similar levels of reduction and maintenance of pocket depth as surgical treatment especially in case of pockets of shallow or moderate depths. While it is less effective in the initial pocket reduction than surgical treatment in case of deep pockets. The use of systemic antibiotics can add a small benefit to the treatment outcomes (Drisko 2014).

Oral cancer

Benign oral cavity neoplasms include papillomas, polyps, and various types of granuloma. This subsection will deal principally with malignant oral neoplasms, of which approximately 90% are squamous cell carcinomas. They may occur on the lip, tongue, gingivae, oral floor, or elsewhere in the mouth (Some statistics on mouth cancer also include the pharynx). The site is often related to the aetiological factors. Lesions may present as swellings,

ulcers, or red, white patches, and many are painless until they become large. Malignant change can occur in a number of lesions which precede tumour development and present as white or red patches of unknown origin. Malignant change is also occasionally seen in oral lichen planus and hyperplastic candidiasis.

Risk factors

Cancer of the lower lip is strongly associated with sunlight exposure, especially in fair-skinned people. Tobacco use, whether chewed or smoked, predisposes to intraoral cancer. The high incidence of oral cancer among southern Asians is in part due to the addition of tobacco to betel quid or 'paan'. There are dose–response relationships for the duration of use and type of tobacco inhaled. Oral cancer risk in current smokers is three times higher than non-smokers. The risk in-ex smokers is lower than in current smokers (Gandini et al. 2008). Alcohol has additive effects with tobacco use while it may have an inverse association with oral cancer in absence of smoking and betel quid chewing (Globocan 2010; Petti et al. 2013). Other risk factors include human papillomavirus (HR-HPV) types (HPV16 and HPV18) and nutritional deficiency (Shield et al. 2017). HPV-associated tumours show a predilection for the tonsils and oropharynx and although apparently aggressive, carry a better prognosis.

In the past, oral cancer occurred mainly in older people (over 60 years). However, the incidence is increasing in developed countries because of a rising incidence in people younger than 45 years, where oral cancer may be a distinct form related to human papilloma virus (HPV) and associated with vaginal and oral sex. Higher risk of head and neck cancer is associated with lower levels of income and education, a difference not totally explained by smoking, alcohol drinking, and diet (Conway et al. 2015).

Epidemiology

Oral cancer is the eighth most common cancer worldwide. More than 450,000 people globally are diagnosed each year with oral and/or pharyngeal cancers (Crall and Forrest 2018). Lip, oral, and pharyngeal cancers represent 3.8% of all cancer cases and cause 3.6% of cancer deaths (Shield et al.

2017). Oral cancer is characterized by late diagnosis and a low 5-year survival rate of 50% worldwide (Gupta et al. 2016)

The incidence is increasing in some Western European countries. The highest reported incidence rates are in South Asian Countries like India, Srilanka, Pakistan, and Bangladesh, where oral cancer is the most prevalent cancer, representing one-quarter of all new cases of cancer (Gupta et al. 2016)

In 2015, the estimated age-standardized incidence rate of mouth and lip cancer per 100,000 population worldwide is 8.1 for males and 4.1 for females. The age-standardized death rate per 100 000 population among males is 3.1 which is also higher than in females (1.4) (Global Burden of Disease Cancer Collaboration 2017).

Management

As early intervention improves oral cancer survival and many cases are preceded by premalignant lesions, case finding would seem useful for disease control. In a 15-year follow-up study in India, oral cancer mortality rate was not significantly different between screened and unscreened groups while among tobacco and alcohol drinkers, the screened individuals had a significantly lower mortality rate than unscreened (Sankaranarayanan et al. 2013). However, there is still no strong evidence to justify active screening in countries with a lower prevalence and available dental services. Opportunistic screening as part of routine dental visits may be viable.

Trauma

Trauma to the teeth is common and frequently causes fracture of the tooth or supporting bone or bodily movement of the tooth including complete avulsion. In many cases, the long-term survival of the tooth is threatened. As the anterior and most visible teeth are most often involved, the result can be disfiguring.

Risk factors

Main causes are falls (31.7 to 64.2%), sport activities (up to 40.2%), bicycling accidents (up to 19.5%), traffic accidents (up to 7.8%), and

physical violence (up to 6.6%) Protruding upper front teeth and inadequate lip coverage also put individuals at risk (Zaleckiene et al. 2014).

Epidemiology

Worldwide, the prevalence of dentofacial trauma experience is approximately 10–15%. The experience of traumatic dental injuries among schoolchildren aged 12–15 years varies between 14.4 and 33.8%. The peak age for trauma is early adolescence, with boys experiencing more injuries than girls. The prevalence of dental injuries in permanent teeth (58.6%) is higher than that in primary teeth (36.8%). Tooth displacement (Luxation) is more common in primary teeth as the supporting bone is less mineralized while tooth fracture is more common in permanent teeth (Zaleckiene et al. 2014).

Management

Trauma to the deciduous teeth is managed by monitoring, in case any consequent infection threatens the permanent tooth developing beneath it. In permanent teeth, adhesive fillings can be used to protect sensitive fractured teeth and dressings (usually calcium hydroxide) to allow continued root development in immature teeth. Permanent teeth that have been knocked out, should be replaced in the socket as soon as possible or placed in a suitable container of milk or normal saline. Dental care should be sought immediately as tooth replantation within 30 minutes provides the best long-term survival of avulsed teeth. The tooth should be splinted to adjacent teeth for 7–14 days and antibiotic and antitetanus treatments given.

The treatment of dental trauma, especially luxation injuries is expensive, time-consuming, and involves different specialists. In the United Kingdom, the average total cost of treating one patient with a dental traumatic injury has been estimated at US$1,665 (Wong and Kolokotsa 2004). In Denmark, annual cost of treatment is estimated to be US $2–$5 million per 1 million inhabitants (Andersson 2013).

Impacts of oral disease

In addition to mortality, oral disease affects other aspects of general health. Limited dietary choice and calorific and micronutrient intake are direct consequences of conditions such as dry mouth, poorly fitting dentures, and loss of teeth. Early childhood caries limits growth and development at the most important stage of life.

Oral diseases also directly affect quality of life. Dental pain is very common. In the United Kingdom, 12-month prevalence of oral pain ranges from 8% to 40.4% which leads to functional limitation and disability (Joury et al. 2018). Even the appearance of the mouth is hugely important, affecting self-esteem and willingness to interact with others. People with visible dental disease are judged by others to be less socially and intellectually competent and less well psychologically adjusted. Good dental appearance is regarded as a requirement for some prestigious occupations.

Considerable effort has been made to assess the extent to which oral disorders compromise aspects of daily life. This research is beginning to reveal the determinants of these impacts. In turn, that information is being used to prioritize resources, to suggest avenues for action, and to evaluate interventions.

Global burden of oral diseases

Despite the relative stability in age-standardized prevalence of oral disease from 1990 to 2015, the number of untreated cases of oral disease increased from 2.5 billion to 3.5 billion due to population growth and ageing (Kassebaum et al. 2017).

Direct costs are incurred from the treatment of oral disease as well as the costs of other conditions caused or exacerbated by oral disease. Globally, the direct cost for treating oral diseases in 2010 was around US$298 billion, which constituted about 4.6% of global health expenditure (Kassebaum et al. 2017). In 2015, the direct treatment costs reached $356.80 billion. The highest expenditures were in high-income north America followed by Western Europe. The cost of the Western model of dental treatment would exceed total national healthcare expenditure in many developing countries.

Indirect costs might include reduced employment or promotion expectations and opportunities, limitation of academic achievement, and the loss of economic productivity. In Australia, 1.1 million work days and 600,000 school days are lost yearly (Richardson and Richardson 2011).

Low-income families are more likely to lose time from work and school because of dental disease, hence compounding the inequalities that already exist in health, income, and educational attainment. Worldwide, the total productivity loss associated with oral diseases was $187.61 billion in 2015. Severe tooth loss was the main reason responsible for $126.67 billion (67%) followed by severe periodontitis ($38.85 billion, 21%). Untreated caries in permanent and deciduous teeth was responsible for $21.19 billion (11%) and $0.90 billion (0.5%), respectively. Western Europe and North America had the highest productivity losses. Cumulatively, the direct and indirect costs of dental diseases form a total global cost of $544.41 billion in 2015 (Righolt et al. 2018).

Links with other diseases

Many associations between oral conditions and general health have been proposed in the last two decades. Periodontal diseases have been linked to cardiovascular diseases, diabetes, stroke, pre-term birth, and low birth weight. Some authors have even suggested that periodontal diseases are independent risk factors for these diseases or claimed that specific cardiovascular risk factors such as tobacco use, obesity, and lower serum high-density lipoprotein cholesterol are more common among people with high dental disease experience. Some, such as tobacco smoking, are independent risk factors for both cardiovascular and periodontal diseases.

Poor oral hygiene predicted cardiovascular outcomes among Scottish adults, although the nature of the relationship was not determined. Having fewer teeth and not wearing dentures prospectively predicted the incidence of falls among Japanese elders. Relationships between oral health and all-cause mortality are partly explained by covariates but persist after many of these covariates have been accounted for (Sabbah et al. 2013).

Systematic and narrative reviews of these associations have not yielded firm conclusions because of methodological complexity. For the time being, we assume these relationships are confounded by environmental and lifestyle factors common to both sets of diseases. The existing evidence suggests that interventions improving periodontal health do not modify these other health problems. However, these epidemiological associations between oral and general health are compatible with the strategy of targeted interventions which could adopt a common risk factor approach (CRFA),

tackling risk factors that are common to many conditions. The CRFA was anticipated to provide multiple benefits, reduce duplication, save resources, improve effectiveness, and reduce social inequalities. However, a recent concern with the CRFA is 'lifestyle drift' where the focus has been on behaviours rather than their upstream structural, social, and psychological determinants. This recognition of the social context in which personal choices are made avoids the social iatrogenesis of describing oral health in individual terms.

Oral health programmes should be closely integrated into programmes to improve general health for the whole population and for specific groups such as older people (Satur et al. 2010). Similar observations have been made about global health policy where a focus on single diseases has ignored the problems of people experiencing multiple health problems and reinforced health inequalities.

Inequalities in oral health

Inequalities occur in the distribution of almost all oral conditions and this area of research has grown in parallel with the interest in health inequalities, led by the Commission for the Social Determinants of Health (CSDH 2008). The resultant formation of the Global Oral Health Inequalities Research Agenda (Sgan-Cohen et al. 2013) of the International Association of Dental Research will add impetus and hopefully, practicable policies.

Oral health inequalities are related to socioeconomic status, ethnicity, or region, both in gradients and in fine grain. Whether socioeconomic status is measured by education, occupation, or income, these inequalities consistently relate to the degree of social inequality and not just the distribution of resources. Differences in oral health by ethnicity and family deprivation were found among British children but were less evident among adolescents. However, differences by residential deprivation were greater among adolescents, which indicates the importance of investigating the effect of socioeconomic disadvantage using multiple measures (Rouxel and Chandola 2018). Furthermore, levels of total tooth loss across the United States are related to the level of income inequality within each state and not just absolute levels of poverty (Bernabé and Marcenes 2011).

There are social gradients for oral cancer incidence (Johnson et al. 2011), tooth loss (Tsakos et al. 2011), total absence of teeth, presence of untreated

decay or the presence of a filled tooth (Elani et al. 2012), severe periodontitis (Eke et al. 2012), pain caused by untreated oral conditions (Joury et al. 2018), self-rated oral health, and oral impacts (Tsakos et al. 2011). Periodontal diseases are unusual in childhood and yet have been linked to the structural determinants of health more strongly than to individual factors among Brazilian 12-year-olds (Jordao et al. 2012).

The life course approach indicates how people's circumstances in early life, accumulate to influence their health throughout life. Long-term follow-up of birth cohorts confirms the importance of this accumulated effect. In one study, 5-year-olds of low socioeconomic status had more decay, more fillings, and worse periodontitis and were more likely to have had a tooth extracted at the age of 26 years, even after controlling for childhood oral health. Furthermore, changes in socioeconomic status throughout life were associated with differing levels of oral health in adulthood (Thomson et al. 2004).

Reducing oral health inequalities

Greater concern about inequalities in health has had a number of broader implications. Earlier understanding had seen such factors as general socioeconomic, cultural, and environmental conditions, and consequently people's living and working conditions as the upstream determinants of *health*. More radical models look even further these upstream factors to consider the social determinants of health inequalities (Krieger 2008). From this perspective, the social determinants of health arise and are sustained by political ecologies, power, and social relations (Krieger et al. 2012).

The CSDH adopted a conceptual framework in which a toxic combination of poor social policies and programmes, unfair economic arrangements, and bad politics creates a socioeconomic and political context. These structural factors shape intermediary determinants including the social position of individuals and groups (as indicated by education, occupation, income, gender, or race/ethnicity). This social position then determines the material circumstances of people, their behaviours, social cohesion, psychosocial and biological factors, and their use of the healthcare system. It is this cascade of factors that shapes the distribution of health and well-being (see Chapter 2.1).

For those concerned with improving health and reducing health inequalities, the CSDH provides a comprehensive framework for planning interventions, which although not specific to any particular set of diseases, is directly applicable to oral health (Kwan and Petersen 2010; Marmot and Bell 2011). The commission recommended three principles of action:

To tackle the inequitable distribution of power, money, and resources and create a more equitable socioeconomic and political context.
To improve daily living conditions so that the material circumstances, social cohesion, psychosocial and behavioural factors, and use of health services are more equitably distributed.
To understand the problem and assess the impact of action.

These three activities are neither discrete nor mutually exclusive. For example, action on the distribution of power and resources will improve living conditions and work at either level will involve the same actors to a greater or lesser degree. The emphasis should be to work across all levels of the model (CSDH 2008; Kwan and Petersen 2010) and hence, at all levels of society.

Systematic reviews indicate that downstream, behaviourally based interventions do not appear to reduce inequalities and may increase them (Lorenc et al. 2013). In oral health, it has long been recognized that health education carries its own dangers of disempowerment and victim-blaming. It may increase inequalities in oral health by being more effective among higher than lower socioeconomic groups (Qadri et al. 2018). Conversely, upstream approaches such as structural workplace interventions and pricing policies have the potential to reduce inequalities (Satur et al. 2010). Conway and colleagues' (2008) systematic review associated oral cancer risk with low socioeconomic status but provided evidence to steer health policies which focus on lifestyles factors towards an integrated approach to tackle the root causes of disadvantage.

One refinement of the CSDH approach is the adoption of the principle of proportionate universalism. Marmot (2010) suggested that the gradient of health inequality should be reduced by working to improve the health of all portions of society, but the scale and intensity of the action should be in proportion to the level of disadvantage, for example, by targeting particular interventions at groups at greatest risk within a whole population strategy.

To public health practitioners concerned about social injustice, the growing attention given to health inequalities is exciting. However, the likelihood of reaching a consensus and prevailing upon the power imbalances that structure the social determinants of health and consequent health inequalities has been reduced during the current global economic crisis. Moreover, inequalities appear to be increasing (Thomson et al. 2004). Health inequalities may persist because the increased social mobility of recent generations has left the lower socioeconomic groups more homogeneous in terms of the individual characteristics that predispose to ill health and the difference in the consumption behaviour between socioeconomic groups due to difference in material and immaterial resources such as cultural capital (Mackenbach 2012).

Actions to improve health and reduce health inequality will have similar effects on oral health but some aspects are especially relevant (Box 8.8.1). Therefore, the CSDH framework and these principles will be used in the remainder of this section.

Box 8.8.1 Improving oral health and reducing inequalities within the Commission on Social Determinants of Health framework

1. Tackling the inequitable distribution of power, money, and resources: changing the structural drivers of conditions of daily life
    - Action on health and health equity at the highest level of government with coherent policies across all departments.
    - Redistributive direct taxation systems.
    - Publicly funded action on child development, education, working and living conditions, and healthcare.
    - Market responsibility.
    - Regulation of goods and services with a direct role in health.
    - Political empowerment and fair representation of all.
    - Oral health impact and oral health equity audit of all policies.
2. Improving daily living conditions: changing the circumstances in which people are born, grow, live, work, and age
    - Comprehensive approaches to early life, child survival, and early life interventions.
    - Comprehensive universal social protection.
    - Urban planning and fairness between areas to prevent rural poverty.
    - Full and fair employment with acceptable, safe work, and a healthy work–life balance.
    - Access to universal healthcare.
    - Increasing the availability of fluoride:
        - Water
        - Milk
        - Salt
        - Tablets/drops
        - Toothpastes
        - Mouth rinses
        - Varnishes.
    - Dietary sugars.
    - Oral cleaning.
    - Smoking cessation.
    - Prevention and management of trauma.
    - Use of oral health services.
3. Measuring and understanding the problem and assessing the impact of action
    - Routine systems with relevant measures of oral health, its determinants, and health equity.
    - Dentists and dental public health specialists as advocates.

Tackle the inequitable distribution of power, money, and resources

This part of the CSDH (2008) report makes recommendations which are the most upstream and radical, to change the way that society is organized. Action on health and health equity is advocated at the highest level of

government to ensure effective and coherent policies across all departments. To ensure adequate financing for publicly funded action, the CSDH proposed redistributive direct taxation systems, which demonstrably reduce poverty.

The provision of basic human and social needs including clean water, healthcare, and education are seen as a primary role of the state rather than the market. The commission also proposed that health equity be made a global development goal and that the social determinants of health be placed as a guiding principle across the WHO.

Oral health can be fitted into these broad actions on health inequalities and into all health policies. For example, in many countries dental caries is now a disease of poverty and particularly of children, where redistributive taxation might reduce oral health inequalities. Work with the agricultural sector, especially in relation to the production of cash crops like sugar and tobacco, has obvious applicability to oral health. Further regulation of the imports and marketing of tobacco, alcohol, and cariogenic foods although the evidence base for the effectiveness of the use of such policies is weak.

Oral diseases may contribute to social stratification as they are often visible. The appearance of the mouth is used by others to form social judgements which may limit life opportunities and employment. Seen this way, oral health is not only a goal but a means towards greater equality. However, to reduce the social impact of oral disease, oral health services need to be organized to meet the needs of the population. Where possible there should be universal access to primary oral healthcare services.

Improve daily living conditions

The CSDH recognized that the factors that enable good health are not distributed equally throughout life and made a series of recommendations to improve daily living conditions.

The accumulation of adversity during the life course means there must be comprehensive approaches to early life for child survival and development and comprehensive social mechanisms to support an income sufficient for healthy living. Secure, and fairly paid employment and access to universal healthcare are vital. These actions will have direct and indirect effects on oral health. One well-evaluated strategy for improving daily living conditions is increasing the availability of fluoride. Other strategies focus on

oral health behaviours such as oral cleaning, smoking cessation, and trauma prevention.

Increasing the availability of fluoride

Fluoride prevents or reverses dental caries mainly through topical effects, including the inhibition of demineralization of the tooth structure, enhancement of remineralization with a stronger crystal structure, and inhibition of bacterial enzymes. Consequently, fluoride has been used to prevent caries by adding it to water supplies, toothpastes, foods, drinks, and in professionally applied products.

Water fluoridation

Beneficial effects of fluoride on dental health were inadvertently discovered during investigations of developmental staining of children's teeth in Colorado, United States, due to chronic ingestion of fluoride during tooth development.

Water fluoridation has been designated one of the ten most important public health measures currently available (Centers for Disease Control and Prevention 2010). Dean originally suggested that the optimal concentration for water fluoridation was 1 part per million (ppm) (1 mg/L) (Dean et al. 1950). Recently, the US Department of Health and Human Services (US HHS) suggested 0.7 mg fluoride per litre as the optimum concentration (Spencer and Do. 2016)

The effectiveness and safety of water fluoridation in caries prevention is consistently supported in authoritative narrative and systematic reviews. A systematic review found that water fluoridation was associated with an increase in percentage of caries free children of 15% in deciduous dentition and 14% in permanent dentition, with a 35% reduction in dmft and 26% reduction in DMFT (Iheozor-Ejiofor et al.2015). A national study in Australia found significant association between lifetime exposure to fluoridated water and childhood caries while adjusting for socioeconomic status (Spencer et al. 2018). Water fluoridation was also found to have an effect over and above that of other sources of fluoride, particularly toothpaste.

A direct dose–response relationship was found between water fluoridation and dental fluorosis. At 1 ppm F, the prevalence of fluorosis was estimated

to be 48% and for fluorosis of aesthetic concern 12.5%. Other possible negative effects considered included bone fracture, cancer, Down's syndrome, senile dementia, and goitre. No clear evidence of these potential effects was found.

The cost-effectiveness of water fluoridation depends on the baseline level of caries and the capital costs of the necessary equipment. Fluoridation is most cost-effective in areas where water is supplied to large numbers of people via a single source. Further high-quality research into the safety, efficacy, cost-effectiveness, and impact of fluoridation on quality of life has been recommended.

The introduction of water fluoridation schemes is invariably opposed by active antifluoridationist lobbies. The main arguments against water fluoridation include the safety concerns outlined earlier and ethical objections related to the four principles; respect for autonomy, beneficence, non-maleficence, and justice (Jones and Lennon 1997).

With regard to the claims of loss of autonomy, those proposing fluoridation suggest some reduction of individual freedom of choice is accepted for the overall good of the community. On the argument that only dentate members of society will potentially benefit from water fluoridation, those in favour acknowledge this but argue that fluoridation will 'do no harm' to any members of society. On the principle of justice, some claim that imposing water fluoridation is not fair on those against it. Proponents counter this argument by suggesting that the reduction in inequalities in dental health that fluoridation potentially provides is a just way of helping those least able to help themselves.

A more recent alternative view sees water fluoridation as modestly effective in developed countries with low caries prevalence (as demonstrated by the relatively high number needed to be treated to prevent a case), yet the political battle for new fluoridation schemes may make it very expensive to introduce.

Other vehicles for administering fluoride include the following:

Fluoride milk

Fluoride has been added to milk as an alternative to water since the 1950s in many countries. These schemes vary in the age at which children start drinking the milk, the number of years over which it is administered, the number of days a year the milk is consumed, and the concentration of fluoride added (Marino et al. 2006). There is low quality evidence for effectiveness of fluoridated milk (Yeung et al. 2015).

Salt fluoridation

Fluoridated salt was first used in Switzerland in the 1950s. In certain cantons and in some Latin America countries, all salt for human consumption is fluoridated. Other European countries including France only sell fluoridated salt for domestic use. The typical concentration is 250 ppm F. Several cross-sectional studies report fluoride salt to be effective in reducing caries. However, the reduction of salt consumption is also a public health goal.

Fluoride tablets/drops

Fluoride tablets/drops use also began in the 1950s. They have been used as dietary supplements with different countries having different dosing regimens (based on the child's age, weight, caries-risk, and fluoride in the water). Fluoride supplements resulted in 24% reduction in DMFS compared with no fluoride supplements while the effect on deciduous teeth less clear (Marinho 2014). There is some evidence that when sucked to maximize the topical dose, a preventive effect is observed in deciduous teeth. Two key problems with fluoride supplementation are adherence to dosing regimens and the risk of fluorosis. These concerns have led to a reduction in the use of tablets/drops as a public health measure.

Topically applied fluoride

There are several delivery systems where fluoride is applied to tooth surfaces at high concentrations for a local protective effect.

a) Fluoride toothpaste:

Toothpaste is the most widespread source of fluoride and the decline in caries experience in children in some countries has been attributed to its regular use. The fluoride compounds and concentrations found in toothpastes vary between brands and between countries. The usual concentration is 1,000–1,500 ppm F with higher (over 2,000 ppm F) formulations available. They present an excellent vehicle for the frequent low-dose application of fluoride and their effectiveness in reducing caries in children has been confirmed with a prevented fraction (PF) of 23% with additional prevented fraction if other type of topical fluoride treatment is used (pooled prevented fraction of 10%) (Marinho 2014). The risk of fluorosis from toothpaste can be minimized by using a smear of paste and supervising children under 6 years of age.

Fluoride toothpaste has a distinct disadvantage over water fluoridation as a broad preventive strategy as it relies on people brushing their teeth. In developed countries infrequent brushing and high caries incidence are associated and so the people who have most to benefit from the use of fluoride toothpastes are less likely to use them frequently. In developing countries, there may not be a tradition of tooth cleaning with toothpastes and Western proprietary brands are likely to be expensive. However, cheap locally made pastes can be profoundly effective.

b) Fluoride mouth rinses:

Fluoride mouth rinses have been used extensively for the past 40 years. School-based programmes were common in certain countries although individual home use now predominates. With the general decline in dental caries and the spread of fluoride toothpastes, population-based rinsing programmes are now recommended only for high risk children. Mouth rinsing is not recommended for children under 6 years of age, due to the risk of fluoride ingestion. The two main concentrations available are 0.05% (230 ppm F) sodium fluoride used daily and 0.2% (900 ppm F) sodium fluoride used weekly.

c) Fluoride varnishes/gels:

Topically applied fluoride varnishes and gels have been used widely for over two decades, both as part of community-based programmes and on an individual basis. Fluoride varnishes are professionally applied, with the two most commonly used varnishes containing 22,600 ppm F or 7,000 ppm F (Marinho et al. 2015). Fluoride gels can be professionally or self-applied under supervision. The most commonly used gel is a professionally applied 12,300 ppm F acidulated phosphate fluoride. Due to the risk of excessive ingestion, this gel is not recommended for young children.

A series of systematic reviews of trials have investigated the effectiveness of various topical fluorides expressed as the preventive fraction (the difference in caries increments between the intervention and control groups) (Table 8.8.1).

Table 8.8.1 Relative effectiveness of sources of topical fluoride

Source	ppm F	Prevented fraction (%)
Toothpaste	1,000–1,250	23
	2,400–2,800	36
	1,450–1,500	30
Mouthrinse	230–900	26
Varnish	7000–22,000	43(permanent) 37 (deciduous)
Gel	12,300	28 (permanent) 20 (primary)

Data from Marinho, V. C. 2014. Cochrane fluoride reviews: an overview of the evidence on caries prevention with fluoride treatments. *Faculty Dental Journal*, 5(2), 78–83 and Marinho, V. C., et al. 2015. Fluoride gels for preventing dental caries in children and adolescents. The Cochrane Library.

Dietary sugars

Dietary sugars are a necessary cause of clinically relevant dental decay. Unfortunately, repeated systematic reviews indicate that dietary advice neither reduces their intake nor the incidence of caries (Kay and Locker 1997). Health education programmes only appear to be effective if they also involve the recommendation or prescription of fluoride vehicles.

However, the data suggest that restriction of dietary sugars by other means is preventive. For example, per capita sugar supplies and caries experience data correlate significantly in national ecological comparisons. A children's home in Australia had a dietary regimen with almost no sugar and the children had very low caries levels until they were allowed to make their own food choices at the age of 12. Likewise, caries levels fell in parallel with the decreased consumption of sugars during the Second World War.

It is difficult for people to restrict their consumption of sugars. There are evolutionary reasons for sweet foods and drinks to be pleasant. Consequently 'hidden' sugars are used to make processed foods nicer. In addition, products used to sweeten foods may not be labelled as such, even if they carry similar risks of harm (e.g. fructose corn syrup). To add to this confusion, ecological and animal studies indicate that sugars in whole fruit (sometimes called intrinsic sugars) do not cause caries. Furthermore,

intrinsic sugars have a relatively low glycaemic index and are accompanied by fibre, and so are not usually associated with obesity.

The WHO recommend a daily intake of sugars of less than 5% of the daily energy (Moynihan et al. 2018). In addition, the frequency of consumption of foods containing free sugars should be limited to a maximum of four times per day. A health-directed food policy focusing on the supply of products containing unhealthy sugars is likely to impact on dental diseases as well as broader health problems, given the evidence implicating them as a cause of obesity, myocardial infarction, dyslipidaemia, pancreatitis, and hepatic dysfunction (Lustig et al. 2012).

Such policies should form part of a broader strategy to reduce dietary intakes of refined grain products and potatoes. Possible strategies fall within the framework of education, substitution, regulation, pricing, or provision. Multilevelled fiscal policies can be used to discourage the manufacture and sale of sugar-containing products, including taxing sugar-containing foods and drinks and restricting advertising to children. Beyond approaches aimed at individuals, education can take the form of authoritative dietary guidelines to inform national policies, community initiatives, and caterers. Recommendations by the Finnish National Board of Education reduced the sale of sweet products in Finnish Schools (Kankaanpää et al. 2012). Healthcare workers can routinely ask their patients about sugar consumption (dentists do this all the time!). All of these approaches and direct consumer pressure can help convince caterers and retail outlets to provide food in a way that makes the healthy choices the easier choices.

Dietary sugars can be substituted with artificial sweeteners to reduce caries incidence. Sales of sugar-free carbonated drinks in Europe and North America demonstrate the compatibility of this measure with commercial interests. However, substitution of dietary sugars has only limited potential because the manufacture of many foodstuffs relies on other specific properties of sugars. In addition, some sweeteners have side effects and resistance to the extended use of artificial sweeteners persists.

Oral cleaning

As plaque is a necessary cause of periodontal diseases it seems logical that tooth cleaning (principally by toothbrushing) should be the cornerstone of their management and prevention. Interventions aimed at improving oral hygiene can achieve a commensurate reduction in gingival inflammation.

However, most studies have had short follow-up periods and the effectiveness of even the best interventions diminishes with time. Therefore, few data show that attempts to improve oral hygiene to prevent destructive disease are effective. Nonetheless, the consensus is that the public health approaches to improve periodontal health should include improved oral hygiene.

The relationship between plaque removal and tooth decay is much more contentious. In a carefully designed trial, professional oral cleaning did not demonstrate any additional preventive effect above a standard preventive programme of fissure sealants and locally applied topical fluoride. Reviews of the effect of research methods on the observed relationship between oral cleaning and dental caries concluded that there was ‘no unequivocal evidence that good oral cleanliness reduces caries experience’. This area of research is fraught with difficulty, as apart from the difficulties of measuring dental disease, studies are susceptible to selection bias, leakage of intervention, and the likely confounding effects between self-reported behaviours, diet, and oral hygiene. Studies where professional cleaning has been effective have used toothpastes containing fluoride. What is known is that brushing with fluoride toothpaste is effective in preventing caries. Therefore, brushing as it is currently practised in most developed countries combats both caries and periodontal diseases and is to be encouraged.

It is difficult to achieve sustainable changes in oral hygiene behaviour. Toothbrushing is a habit learned as a young child and is difficult to change later in life, hence interventions via healthcare workers and social agencies working with young children and their parents may be useful.

Toothbrushing often clusters with other health behaviours such as avoiding tobacco use, exercise, healthy eating, and other aspects of hygiene such as hand washing. In turn, these behaviours may relate to personal factors including greater self-efficacy, self-esteem, having confidence in one’s family, life satisfaction, managing in school, and not having other health problems. Unsurprisingly then, toothbrushing is determined by structural and resource-related factors and living conditions. These factors include family structure, social support, external barriers, domestic arrangements, social norms, cultural factors, socioeconomic status, and socioeconomic development. Therefore, actions across all these levels are most likely to improve periodontal health and periodontal health equity.

Smoking cessation

Cessation of smoking decreases the associated elevated risk of oral cancer within 5–10 years. Numerous other oral conditions are directly or indirectly associated with tobacco smoking. Many of these, such as stained teeth, receding gums, and altered taste are readily perceptible to the individual and may encourage or reinforce the desire to stop smoking. The dental team are also often aware of the personal and social circumstances (e.g. pregnancy or a new job) that prompt people to give up. The dental team can hence support smoking cessation by providing advice and signposting dedicated cessation services. Social inequality is evident in smoking rates and is effectively reduced through taxation polices and well-designed mass media campaigns (Purcell et al. 2015).

Prevention of trauma

Maxillofacial trauma is related to environmental and structural factors such as socioeconomic status, alcohol use, and road planning and use. In children, dental trauma is closely related to area-based measures of socioeconomic status and overcrowding. Many of these factors can be alleviated by improving living conditions.

Local and individual strategies can also help reduce trauma to the face and teeth. Playground surfaces can be made of impact-absorbing materials that cushion against trauma. The use of mouthguards is compulsory for some sports in some countries. Mouthguards not only prevent dental injuries, but also prevent laceration of the facial soft tissues against the teeth, reduce the risk of mandibular fracture, and may protect the cranial cavity. The most basic type may be obtained prefabricated in a range of sizes. A more sophisticated type may be adapted to fit the mouth, typically by softening it in hot water first. Custom-made devices are the most comfortable and can be made to support the lower teeth and mandible during trauma

The need for urgent care (particularly teeth that have been knocked out) means that skilled emergency dental services should ideally be available around the clock. Informing athletes and their teachers and trainers of the need for immediate action can reduce the impact of the trauma. Unfortunately, orthodontic treatment of protruding front teeth is complex and prolonged, but can be justified in children of 8 or 9 years to reduce the risk of trauma

Use of health services

Internationally, the emphasis and resources within dentistry remain focused on downstream approaches to ‘cure’ patients of their dental diseases, rather than evidence-based strategies to improve the oral health of populations (Baelum 2011).

Similar problems exist with medical, and dental care. Resources are constrained and may be poorly allocated. Services still cluster around people who can afford them, while those who cannot, have little control over their own health and healthcare. In every country of the world, there are people who cannot attend and/or afford dental treatment and even in well-developed socialized systems of dental care major inequalities exist.

Specific changes to ensure access to dental services can be categorized in the framework used by Penchansky and Thomas (1981). Services must be available, accessible, affordable, acceptable, and accommodating. This framework is broadly compatible with the characteristics of dental services as seen within Andersen’s (1995) behavioural model of access.

Increasing availability has a direct effect on service use. A number of countries including Australia, New Zealand, Canada, and the United Kingdom employ DCPs with a limited repertoire of treatment options (variously called school dental nurses, dental auxiliaries, and dental therapists). The scope of practice of dental therapists varies from country to country. In general, they may undertake diagnostic procedures, treatment planning, and preventive services, place fillings in primary and permanent teeth, and extract primary teeth. They may also be permitted to extract permanent teeth. Therapists can be used to increase access to care by supplementing the volume and reach of services, which enhances equity (Dyer et al. 2013). Therapists are socially acceptable where they are integrated into existing dental systems. Available data suggest that therapists offer care that is similarly effective to that provided by dentists. In developing countries with a growing incidence of dental caries which may not be able to afford sufficient dentists, dental professionals may be an appropriate solution.

One particular aspect of dentistry in many countries that could be revised is the direct fee-for-service system of payment to dentists. Fee-per-item service payments encourage dentists to work quickly and may incentivize overtreatment rather than prevention unless there is a specific fee for this. On the other hand, services based on capitation per patient enrolled reduce

restorative treatment and increase preventive care with no evidence of 'supervised neglect'.

The acceptability of dental services was highlighted in a qualitative study of people who did not go to the dentist. As well as being influenced by the perceived accessibility of care, participants were swayed by their trust in dentistry and their perceptions of oral health as a commodity. For these people at least, greater marketing may have been counterproductive in encouraging use of dental services.

Dental services in middle and low-income countries

Non-industrialized countries usually suffer from resource constraints including workforce, appropriate technology, and reliable power supplies. For example, the dentist: population ratio in many countries in Africa is around 1:150,000 compared to 1:2,000 in most industrialized countries. The ratio recommended by the WHO is 1:7500 (Kandelman et al. 2012). Access inequalities are exacerbated by the unequal distribution of dentists and their migration from middle and low-income countries to more developed countries.

The limitations of the traditional curative approach to dental health are more extreme in the context of the developing world. The surgical approach to dentistry is technology intensive and requires the availability of continuous power and expensive equipment. Dentists, therefore, need to treat patients who can help them recoup their costs. These pressures limit the availability of services and contribute to the inequalities in their provision. Large parts of the Western model of dental care may be inappropriate in developing countries, including an overemphasis of clinical surveys in healthcare planning. Services based on normative assessment limit community participation in healthcare and ignore the sociodental implications of oral disease. In one notorious example, survey data used to calculate the periodontal treatment needs of children in Kenya would have used the entire national dental workforce for up to 21 years. Services could concentrate on the relatively few conditions that comprise the bulk of oral health problems: toothache (not *tooth decay*), trauma, oral infections, and neoplasms.

The primary oral healthcare approach (PHCA) is still relevant in all countries but particularly so in the developing world. It has five principles:

Equitable distribution of services: The 'inverse care law' between the availability of services and the need for them also occurs in dentistry. It is particularly extreme in countries where there are wide disparities between rich and poor. In Africa, 80% of the trained professional personnel live and work in affluent neighbourhoods in cities although the same proportion of the population lives in rural areas. The scope for other DCPs in developing countries may be greater since political lobbies of dentists may be less well established than in industrialized nations. DCPs can provide simple, but essential treatments to extend the availability of services and reduce inequalities in access.

Community involvement in health: this means that people are allowed to take control of their own health and is necessary if programmes are to thrive. It is perhaps the most difficult aspect of the PHCA since it requires health professionals to relinquish their traditional hierarchical role. Individuals and communities often also regard health as beyond their control and may not consider oral health a priority. There are isolated examples of wide involvement in oral health. In Glasgow, Scotland, community-based Oral Health Action Teams (partnerships of parents, teachers, nursery nurses, health visitors, and dentists) have reduced caries levels in young children in deprived areas of the city (Blair et al. 2004).

Focus on prevention: prevention of oral diseases is universally accepted as an essential component of care, but in the past has involved people as passive recipients of information rather than active participants in health promotion.

Use of appropriate technology: 'appropriate technology' does not connote 'cheap' and 'second rate', but an approach which recognizes the needs and resources of the local community. For example, ART combines these requirements with knowledge of dental caries progression and developments in dental materials science.

Multisectoral approach: we have seen how an effective health strategy might involve a number of departments of both national and local governments, water providers, the educational system, community members, and healthcare workers. All of this needs to be integrated coherently to go beyond just using the resources of other sectors to promote oral health.

Measure and understand the problem, and assess the impact of action

The third aspect of the CSDH framework would involve monitoring the social determinants of oral health. This requires routine systems to record relevant measures of health, its determinants, and health equity. Such data can be collected during dental treatment or through surveys.

New evidence on the role of social determinants in oral health equity and evaluations of interventions to reduce inequality should emerge from the international emphasis given to the topic. For example, WHO attention on the social determinants of health should make funding for such work a priority. One area where the global oral health inequalities research agenda can make an immediate impact is to provide a forum for sharing and interpreting new evidence.

Training on this topic is available through undergraduate and postgraduate courses for dentistry, although their extent and focus varies. The WHO Oral Health Programme provides guidance on tackling oral health inequalities but could benefit from greater resources (Kwan and Petersen 2010).

Future developments

Several trends in oral health have been observed. First, some countries, particularly developing ones, may experience increases in dental caries due to the adoption of diets high in sugars. These trends are worrying even if they are limited to more affluent city dwellers as the increase in treatment and the burden on developing economies while this burden will be moderated by the low levels of perceived need in communities unused to receiving dental treatment.

Second, the fall in caries prevalence seen in developed countries over the past 40 years appears to have slowed, with evidence of wide inequalities in dental caries and its impact on people’s daily lives. Third, in many developed countries the combination of demographic changes, changed attitudes towards oral health, and the preservation of teeth produces an interesting pattern. The growing number of older people who will live longer and retain many heavily restored teeth will increase the need for more care, some of it more complex, which in turn will drive more specialization.

The deprofessionalization of dentistry continues to gather pace and is manifest in several different forms. In many developed countries patients are

demanding ‘rights’ as consumers of care, even as marketing theory is being applied to dentistry, which places consumer satisfaction as being essential for business success. Similar principles are cornerstones of the new public health. Health promotion takes public involvement in oral health well beyond clinical dentistry. Even within clinical care, satisfaction is an integral part of the process of care rather than just an outcome. Other agencies, such as governments and insurance companies, are increasingly involved in healthcare. Externally applied measures to minimize the costs of care and increase the accountability of healthcare organizations while assuring the quality of care all serve to reduce professional power within dentistry.

Gigerenzer and Gray (2011) have named the twenty-first century as the *century of the patient*. They suggest widespread stakeholder participation to solve common healthcare problems using all the theoretical and practical knowledge available.

This trend of deprofessionalization is likely to continue and may help to make oral care more relevant to the needs of the people it serves. Deprofessionalization and greater lay involvement in dentistry, both at the individual and community level, will reduce the power structures and hierarchies that underpin the social determinants of health. This might be particularly important where perceived social position is a cause of health inequality independent of objective assessments of socioeconomic status. Given that professions are likely to resist any tendency to undermine their power, there is an opportunity for dental public health to facilitate and manage the deprofessionalization of dentistry.

Evidence base

Only a minority of practitioners and academics contribute to the growth in the evidence base of dentistry through systematic reviews. Unfortunately, these reviews repeatedly conclude that there is a dearth of good primary data on which decisions can be based. While a greater evidence base for individual care is needed, more research is also needed on the benefits of interventions at a population level, not only to improve health, but also to reduce inequalities.

Other chapters in this text will consider the evaluation of health promotion in some detail (see Chapter 2.1), but the debate about controlled trials in public health continues. Health promotion takes place in the ‘real world’ and

other factors and time intervene between the intervention, and health outcomes. For example, implementing a life course approach to reduce inequalities may take decades before its benefits are manifest. These factors have led the WHO (1998) to conclude that 'The use of RCTs to evaluate health promotion initiatives is, in most cases, inappropriate, misleading, and unnecessarily expensive'. Even the keenest advocates of RCTs acknowledge that they are unsuitable for evaluating legislative and policy changes. However, many of the problems often cited against conducting trials in public health can be overcome.

A further problem in both clinical and industry-sponsored trials is the widespread failure to identify treatment effects. Further consideration of methodology may be required ahead of new evaluative research.

Technology

Technological developments also influence oral health and care. In developed countries, osseointegrated implants are increasingly used to support dental prostheses. At present, however, implant treatment demands specialist expertise, requires regular maintenance, and is therefore costly. Implants treatment is hence limited to those who can afford it and thus contributes to inequalities in oral health. Other more long-term technological possibilities include a vaccine for dental caries, growing teeth from stem cells, the use of nanotechnology and laser drills.

Advances in 3D imaging and printing have the potential to simplify and streamline the production of prostheses such as crowns and dentures. However, even as these technologies are introduced, it remains to be seen if they will have public health benefits.

Cosmetic dentistry

Cosmetic dentistry procedures include bleaching teeth and the use of veneers to change their appearance.

Cosmetic dentistry can be seen as the medicalization of beauty and as a form of iatrogenesis. Supplier-induced demand is evident when dentists draw patients' attention to minor dental anomalies (that may have no effect on their lives) and suggest they have them corrected. Attractiveness is

perhaps more important now than ever before and society is said to be suffering from a normative discontent with personal appearance. Dentistry may contribute to these problems if the availability of treatment to enhance dental appearance makes deviations from the norm less acceptable. This message is amplified in printed and broadcast media that expose us to enhanced images. Photographs in magazines have whiter and straighter teeth than would normally be encountered and exposure to these images increases our dissatisfaction with our own appearance. In addition, cosmetic dentistry will only be available to those who can afford it. It may be that the visible consumption of cosmetic dentistry will further distinguish socioeconomic groups and so contribute to the stratification of society and social inequalities. From this perspective, cosmetic dentistry might have adverse effects both on the people who receive it and on wider society. Consideration may be required about whether its growth should continue unchecked.

Consumerism and the role of industry in oral health

Consumerism is widely criticized from the perspective of overconsumption, its links with media culture, and from socialist and feminist perspectives. Its role in oral health is not so clear cut. The addition of fluoride to toothpastes has been one of the most successful dental public health interventions of all time. Moreover, its effect is cumulative and is staggering when seen over decades. The first fluoride toothpastes were introduced in the 1950s but sales rose dramatically in the United Kingdom during the 1970s (the proportion of toothpaste sales containing fluoride rose from less than 10% in 1970 to more than 90% in 1976).

Although dentists were involved in the development, evaluation, and promotion of fluoridated toothpastes, it was the commercial interests of home-care product manufacturers that drove this development. There are also numerous examples of manufacturers and dental organizations working together to promote oral health that continue to this day.

The commercial marketing of dental care products has other advantages. It has raised the prominence of oral health and reinforces the idea of frequent oral cleaning as a social norm. In addition, the positive ways in which oral healthcare products are advertised has made a healthy mouth seem attainable.

Conclusions

In summary, we have argued that poor oral health is a major public health problem with a huge population burden of oral disease because of the large numbers of people affected. While the incidence of dental caries has fallen in the developed world, the prevalence of the condition would be staggering when compared to any other disease. Periodontal diseases, cancer, and trauma all pose specific problems. All these diseases have significant impacts and create a considerable burden to both individuals and the community. The links between oral and general health indicate that strategies to improve both sets of problems and reduce inequalities should be integrated within the framework advocated by the CSDH. Several factors, including tighter financial pressures, trends in oral health, the deprofessionalization of dentistry, the role of consumerism in oral health, and the need for a better evidence base will modify this endeavour, as will the growth of skill mix, technology, and cosmetic dentistry.

REFERENCES

Alexandridi, F., Tsantila, S., and Pepelassi, E. (2018). Smoking cessation and response to periodontal treatment. *Australian Dental Journal*, 63(2), 140–9.

Andersen, R.M. (1995). Revisiting the behavioral model and access to medical care: does it matter? *Journal of Health and Social Behavior*, 36(1), 1–10.

Andersson, L. (2013). Epidemiology of traumatic dental injuries. *Pediatric Dentistry*, **35**, 102–5.

Baelum, V. (2011). Dentistry and population approaches for preventing dental diseases. *Journal of Dentistry*, 39(Suppl. 2), S9–19.

Bernabé, E., and Marcenes, W. (2011). Income inequality and tooth loss in the United States. *Journal of Dental Research*, 90(6), 724–9.

Blair, Y., Macpherson, L.M., McCall, D.R., McMahon, A.D., and Stephen, K.W. (2004). Glasgow nursery-based caries experience, before and after a community development-based oral health programme's implementation. *Community Dental Health*, 21(4), 291–8.

Centers for Disease Control and Prevention (2010). *Community Water Fluoridation*. [Online] Available at: http://www.cdc.gov/fluoridation/

Commission on Social Determinants of Health (CSDH) (2008). *Closing the Gap in a Generation: Health Equity Through Action on the Social Determinants of Health. Final Report of the Commission on Social Determinants of Health*. Geneva: WHO.

Conway, D.I., Petticrew, M., Marlborough, H., Berthiller, J., Hashibe, M., and Macpherson, L.M. (2008). Socioeconomic inequalities and oral cancer risk: a systematic review and meta-analysis of case-control studies. *International Journal of Cancer*, 122(12), 2811–19.

Conway, D.I., et al. (2015). Estimating and explaining the effect of education and income on head and neck cancer risk: INHANCE consortium pooled analysis of 31 case-control studies from 27 countries. *International Journal of Cancer*, 136(5), 1125–39.

Crall, J.J., and Forrest, C.B. (2018). A life course health development perspective on oral health. In: Halfon, N., Forrest, C.B., Lerner, R.M., and Faustman, E.M. (eds.) *Handbook of Life Course Health Development*. New York: Springer.

Dean, H.T., Arnold, F.A. Jr., Jay, P., and Knutson, J. W. (1950). Studies on mass control of dental caries through fluoridation of the public water supply. *Public Health Reports*, 65(43), 1403–8.

Drisko, C.L. (2014). Periodontal debridement: still the treatment of choice. *Journal of Evidence Based Dental Practice*, **14**, 33–41.

Duangthip, D., Chen, K.J., Gao, S.S., Lo, E.C.M., and Chu, C.H. (2017). Managing early childhood caries with atraumatic restorative treatment and topical silver and fluoride agents. *International Journal of Environmental Research and Public Health*, 14(10), 1204.

Dye, B.A., et al. (2007). Trends in oral health status: United States, 1988–1994 and 1999–2004. Vital and health statistics. *Series 11, Data from the National Health Survey*, 11(248), 1–92.

Dyer, T., Owens, J., and Robinson, P.G. (2013). What matters to patients when their care is delegated to dental therapists? *British Dental Journal*, 214(6), E17.

Eke, P.I., Dye, B.A., Wei, L., Thornton-Evans, G.O., and Genco, R.J. (2012). Prevalence of periodontitis in adults in the United States: 2009 and 2010. *Journal of Dental Research*, 91(10), 914–20.

Elani, H.W., Harper, S., Allison, P.J., Bedos, C., and Kaufman, J.S. (2012). Socio-economic inequalities and oral health in Canada and the United States. *Journal of Dental Research*, 91(9), 865–70.

Frencken, J.E., Sharma, P., Stenhouse, L., Green, D., Laverty, D., and Dietrich, T. (2017). Global epidemiology of dental caries and severe periodontitis–a comprehensive review. *Journal of Clinical Periodontology*, 44(Suppl. 18), S94–S105

Gandini, S., et al. (2008). Tobacco smoking and cancer: a meta-analysis. *International Journal of Cancer*, 122(1), 155–64.

Gigerenzer, G., and Gray, J.A.M. (2011). Launching the century of the patient. In: Gigerenzer, G. and Gray, J.A.M. (eds.) *Better Doctors, Better Patients, Better Decisions*. Cambridge, MA: The MIT Press, pp. 3–29.

Global Burden of Disease Cancer Collaboration (2017). Global, regional, and national cancer incidence, mortality, years of life lost, years lived with disability, and disability-adjusted life-years for 32 cancer groups, 1990 to 2015: a systematic analysis for the global burden of disease study. *JAMA Oncology*, 3(4), 524–48.

Globocan (2010). *Estimated Cancer Incidence, Mortality, Prevalence and Disability-adjusted Life years (DALYs) Worldwide in 2008.* Paris: International Agency for Research on Cancer/WHO. [Online] Available at: http://globocan.iarc.fr/

Gupta, N., et al. (2016). Changing Trends in oral cancer-a global scenario. *Nepal Journal of Epidemiology*, 6(4), 613–19.

Iheozor-Ejiofor, Z., et al. (2015). Water fluoridation for the prevention of dental caries. *Cochrane Database of Systematic Reviews*, 6(6), CD010856.

Johnson, N.W., et al. (2011). Global oral health inequalities in incidence and outcomes for oral cancer: causes and solutions. *Advances in Dental Research*, 23(2), 237–46.

Jones, S. and Lennon, M.A. (eds.). (1997). *Fluoridation. Community Oral Health*. Bath: Wright.

Jordao, L.M.R., Vasconcelos, D.V., da Silveira Moreira, R., and Freire, M.C.M. (2012). Individual and contextual determinants of periodontal health in 12 year old schoolchildren in a Brazilian capital city. *International Journal of Dentistry*, **2012**, 325475.

Joury, E., Bernabé, E., Gallagher, J.E., and Marcenes, W. (2018). Burden of orofacial pain in a socially deprived and culturally diverse area of the United Kingdom. *Pain*, 159(7), 1235–43.

Kandelman, D., Arpin, S., Baez, R.J., Baehni, P.C., and Petersen, P.E. (2012). Oral health care systems in developing and developed countries. *Periodontology 2000*, 60(1), 98–109.

Kankaanpää, R., Seppänen, S., Hiiri, A., Manninen, M., Puska, P., and Lahti, S. (2012). Effect of national recommendations on the sale of sweet products in the upper level of Finnish

comprehensive schools. *Community Dental Health*, 29(2), 149–53.
Kassebaum, N.J., et al. (2017). Global, regional, and national prevalence, incidence, and disability-adjusted life years for oral conditions for 195 countries, 1990–2015: a systematic analysis for the global burden of diseases, injuries, and risk factors. *Journal of Dental Research*, 96(4), 380–7.
Kay, E.J. and Locker, D. (1997). *Effectiveness of Oral Health Promotion: A Review*. London: Health Education Authority.
Krieger, N. (2008). Ladders, pyramids and champagne: the iconography of health inequities. *Journal of Epidemiology and Community Health*, 62(12), 1098–104.
Krieger, N., Dorling, D., and McCartney, G. (2012). Mapping injustice, visualising equity: why theory, metaphors and images matter in tackling inequalities. *Public Health*, 126(3), 256–8.
Kwan, S. and Petersen, P.E. (2010). Oral health: equity and social determinants. In: Blas, E., and Sivasankara Kurup, A.S. (eds.) *Equity, Social Determinants and Public Health Programmes*. Geneva: WHO, pp. 159–76.
Lorenc, T., Petticrew, M., Welch, V., and Tugwell, P. (2013). What types of interventions generate inequalities? Evidence from systematic reviews. *Journal of Epidemiology and Community Health*, 67(2), 190–3.
Lustig, R.H., Schmidt, L.A., and Brindis, C.D. (2012). The toxic truth about sugar. *Nature*, 482(7383), 27–9.
Mackenbach, J.P. (2012). The persistence of health inequalities in modern welfare states: the explanation of a paradox. *Social Science & Medicine*, 75(4), 761–9.
Marino, R., Villa A., and Weitz A. (2006). *Dental Caries Prevention using Milk as the Vehicle for Fluorides: The Chilean Experiences*. Melbourne: School of Dental Science, University of Melbourne.
Marmot, M. (2010). *Fair Society, Healthy Lives: Strategic Review of Health Inequalities in England Post-2010*. London: Marmot Review.
Marmot, M., and Bell, R. (2011). Social determinants and dental health. *Advances in Dental Research*, 23(2), 201–6.
Marinho, V.C. (2014). Cochrane fluoride reviews: an overview of the evidence on caries prevention with fluoride treatments. *Faculty Dental Journal*, 5(2), 78–83.
Marinho, V.C., Worthington, H.V., Walsh, T., and Chong, L.Y. (2015). Fluoride gels for preventing dental caries in children and adolescents. *Cochrane Database of Systematic Reviews*, (6), CD002280.
Moynihan, P., Makino, Y., Petersen, P.E., and Ogawa, H. (2018). Implications of WHO guideline on sugars for dental health professionals. *Community Dentistry and Oral Epidemiology*, 46(1), 1–7.
Murray, J.J., Vernazza, C.R., and Holmes, R.D. (2015). Forty years of national surveys: an overview of children's dental health from 1973–2013. *British Dental Journal*, 219(6), 281–5.
Needleman, I., Nibali, L., and Di Iorio, A. (2015). Professional mechanical plaque removal for prevention of periodontal diseases in adults—systematic review update. *Journal of Clinical Periodontology*, 42(Suppl. 16), S12–35.
Needleman, I., Tucker, R., Giedry-Leeper, and Worthington, H. (2005). Guided tissue regeneration for periodontal intrabony defects—a Cochrane Systematic Review. *Periodontology 2000*, **37**, 106–23
Penchansky, R., and Thomas, J.W. (1981). The concept of access. Definition and relationship to consumer satisfaction. *Medical Care*, 19(2), 127–40.
Petti, S., Masood, M., Messano, G.A., and Scully, C. (2013). Alcohol is not a risk factor for oral cancer in nonsmoking, betel quid non-chewing individuals. A meta-analysis update. *Ann Ig*, 25(1), 3–14.
Purcell, K.R., O'Rourke, K., and Rivis, M. (2015). Tobacco control approaches and inequity—how far have we come and where are we going?. *Health Promotion International*, 30(suppl. 2), ii89–ii101.

Qadri, G., Alkilzy, M., France, M., Hoffmann, W., and Splieth, C. (2018). School-based oral health education increases caries inequalities. *Community Dental Health*, 35(3), 153–9.

Richardson, B. and Richardson, J. (2011). *End the Decay: The Cost of Poor Dental Health and What Should be Done About It*. Melbourne: Brotherhood of St Lawrence.

Righolt, A. J., Jevdjevic, M., Marcenes, W., and Listl, S. (2018). Global-, regional-, and country-level economic impacts of dental diseases in 2015. *Journal of Dental Research*, 97(5), 501–7.

Rouxel, P., and Chandola, T. (2018). Socioeconomic and ethnic inequalities in oral health among children and adolescents living in England, Wales and Northern Ireland. *Community Dentistry and Oral Epidemiology*, 46(5), 426–34.

Sabbah, W., Mortensen, L. H., Sheiham, A., and Batty, D. (2013). Oral health as a risk factor for mortality in middle-aged men: the role of socioeconomic position and health behaviours. *Journal of Epidemiology & Community Health*, 67(5), 392–7.

Sankaranarayanan, R., et al. (2013). Long term effect of visual screening on oral cancer incidence and mortality in a randomized trial in Kerala, India. *Oral Oncology*, 49(4), 314–21.

Satur, J.G., Gussy, M., Morgan, M.V., Calache, H., and Wright, C. (2010). Review of the evidence for oral health promotion effectiveness. *Health Education Journal*, 69(3), 257–66.

Sgan-Cohen, H.D., et al. (2013). Global Oral Health Inequalities Research Agenda (IADR-GOHIRA(R)): a call to action. *Journal of Dental Research*, 92(3), 209–11.

Shield, K.D., et al. (2017). The global incidence of lip, oral cavity, and pharyngeal cancers by subsite in 2012. *CA: A Cancer Journal for Clinicians*, 67(1), 51–64.

Spencer, A.J., and Do, L.G. (2016). Caution needed in altering the ‘optimum’ fluoride concentration in drinking water. *Community Dentistry and Oral Epidemiology*, 44(2), 101–8.

Spencer, A.J., Do, L.G., and Ha, D.H. (2018). Contemporary evidence on the effectiveness of water fluoridation in the prevention of childhood caries. *Community Dentistry and Oral Epidemiology*, 46(4), 407–15

Thomson, W.M., Poulton, R., Milne, B.J., Caspi, A., Broughton, J.R., and Ayers, K.M. (2004). Socioeconomic inequalities in oral health in childhood and adulthood in a birth cohort. *Community Dentistry Oral Epidemiology*, 32(5), 345–53.

Tsakos, G., Demakakos, P., Breeze, E., and Watt, R.G. (2011). Social gradients in oral health in older adults: findings from the English longitudinal survey of aging. *American Journal of Public Health*, 101(10), 1892–9.

White, D., Pitts, N., Steele, J., Sadler, K., and Chadwick, B. (2011). *Disease and Related Disorders —A Report from the Adult Dental Health Survey 2009*. London: The Health and Social Care Information Centre.

Wong, F.S. and Kolokotsa, K. (2004). The cost of treating children and adolescents with injuries to their permanent incisors at a dental hospital in the United Kingdom. *Dental Traumatology*, 20(6), 327–33.

Yaacob, M., et al. (2014). Powered versus manual toothbrushing for oral health. *Cochrane Database of Systematic Reviews*, 6, CD002281.

Yeung, C.A., Chong, L.Y., and Glenny, A.M. (2015). Fluoridated milk for preventing dental caries. *Cochrane Database of Systematic Reviews*, 8, CD003876.

Zaleckiene, V., Peciuliene, V., Brukiene, V., and Drukteinis, S. (2014). Traumatic dental injuries: etiology, prevalence and possible outcomes. *Stomatologija*, 16(1), 7–14.

8.9

Musculoskeletal disorders

Lope H. Barrero and Alberto J. Caban-Martinez

Introduction

What will be included in this chapter?

Musculoskeletal disorders (MSDs) refer to a large variety of health conditions that affect muscles, tendons, ligaments, nerves, and bones. Conditions such as osteoarthritis, osteoporosis, low back pain, rheumatoid arthritis, and carpal tunnel syndrome fall into this large category of health conditions.

Multiple factors related to the intrinsic characteristics of individuals, their environment, and their daily activities can contribute to the occurrence and burden of MSDs. The relative contribution of these factors can vary across populations in different parts of the world. As we learn more about the mechanisms through which these conditions occur, we have been able to improve our understanding of how multiple complex pathways can take a person from a healthy status to a disease state. This new knowledge about the diagnosis and pathogenesis of MSDs has emphasized the fact that in reality, MSDs can include several very different conditions, which in essence implies that no single pathway for diagnosis, treatment, or rehabilitation can be effective, not even in very well defined specific populations.

The importance and major public health burden of MSDs is rooted in the fact that these conditions occur more frequently than many other conditions that can affect human health; and when they do occur, they tend to last longer than other conditions, coexist with other, sometimes deadly conditions that are caused by the same factors or that are complications resulting from them. For example, in the United States, approximately 50 million adults have a doctor diagnose of arthritis, yet 1.2 million, or nearly one in four adults with arthritis (24%), also have heart disease (Cisternas et al. 2009). These characteristics translate directly into a large burden to society that can appear in the form of direct health services costs, reduction in opportunities for those affected and suffering for those affected and their families. Furthermore, because of the differences in access to healthcare, the burden of these conditions may vary substantially between different latitudes, generating questions about equality and social justice.

In this chapter we first present a brief description of some of the most important MSDs of public health importance. We then briefly describe what we know about their causes and what is already known about how to prevent them or reduce their impact to society. We will also describe the scope of the problem from a public health perspective. We also present a brief account on the epidemiology of these conditions; the estimated costs to society; the most important barriers that prevent persons from obtaining some relief. Lastly, we summarize recommendations for public health policy action and for future research.

MSDs definition and classification

There are a broad scope of health conditions that comprise MSDs including inflammatory conditions such as gout and rheumatoid arthritis, age-related diseases such as osteoarthritis and osteoporosis, conditions with ambiguous aetiologies such as fibromyalgia and low back pain, conditions due to injuries from occupational exposures, sports injuries, as well as injuries resulting from slips, trips, and falls. While some of these conditions present with acute onset and short duration, a larger proportion of these disorders reoccur or become lifelong health conditions (Woolf et al. 2010). These conditions can substantially impair both mental (Carder et al. 2013; Holmgren et al. 2013) and physical function and therefore are a major determinant of the health of the world population.

Different classifications of MSDs can be found in the scientific literature based, for example, on body part affected, whether it has a systemic origin or based on the external causal factor (e.g. whether it is work-related or not).

The International Classification of Disease (ch. XIII) provides a useful framework with which to classify groups of diagnosis of MSDs (ICD-10) (http://apps.who.int/classifications/icd10/browse/2010/en#/XIII) and they include:

Arthropathies: Disorders affecting predominantly limb joints including those due to microbiological agents (e.g. rheumatoid arthritis, juvenile arthritis, gout, psoriatic spondylitis).
Systemic connective tissue disorders: Including autoimmune disease and collagen (vascular) disease (e.g. systemic lupus erythematous and Sicca syndrome (Sjören)).
Dorsopathies: Disorders of the spine including deforming dorsophathies, spondylopathies and other dorsopathies (e.g. cervicalgia, sciatica, and low back pain).
Soft tissue disorders: Disorders of muscles (e.g. muscle strain).
Osteopathies and chondropathies: Including disorders of bone density and structure (e.g. osteoporosis); and chondropathies (e.g. juvenile osteochondrosis).
Other disorders: Including conditions nowhere else classified including post-procedural musculoskeletal disorders and biomechanical lesions not elsewhere specified.
Additional MSDs may be found among other groups of diagnosis: MSDs involving nerve damage (e.g. carpal tunnel syndrome) can be found among diseases of the nervous system. Also, MSDs involving sprains and strains for joints and ligaments can be found in the injury category.

Major MSDs: case definitions, natural history, pathophysiology, risk factors, and treatment

While there are many types of MSDs, musculoskeletal tissue degeneration in general, can be initiated by systemic disease, trauma, and repeated biomechanical strain—all of which can lead to disability and pain. For example, autoimmune diseases can initiate and sustain damaging processes to body tissues including musculoskeletal tissue (Smolen and Aletaha 2008). Similarly, hormonal imbalances can generate loss of structure in bones and result in osteoporosis. On the other hand, trauma resulting, for example from a fall at an early age, can initiate abnormal processes of gene expression or result in incomplete repair processes that end up affecting musculoskeletal tissue structures. Also, repeated loads can result in tissue microtrauma that affects its function and eventually accumulates to major damage if recovery is not allowed.

Environmental factors (e.g. eating and physical activity habits) and individual characteristics (e.g. genetic factors and age) are determinants of MSDs (Gabriel and Michaud 2009) influencing the initiation and development of the disease process through complex relations. For example, poor eating habits can result in overweight individuals, which can affect the biomechanical loading on joints and hence lead to osteoarthritis conditions (Guilak 2011; Vincent et al. 2012). Similarly, genetic factors can be involved in the onset of systemic inflammatory disease and can also influence body response at different stages of the disease process. On the other hand, ageing activates several different biochemical processes of tissue degeneration through insufficient expression of tissue components or the occurrence of other age-related diseases with the capability to influence the process of tissue degeneration.

Tissue degeneration is thought to accumulate throughout the life course, although determining the beginning of the disease process is difficult. The accumulated damage can be seen clinically, based on radiographic or serologic evidence. For example, tissue tears and tissue deformation can be seen based on imaging techniques. An important body of research has been developed on new biomarkers to track the onset and development of disease (Naylor and Eastell 2012; Pisetsky et al. 2012; Rousseau and Garnero 2012). Nevertheless, such imaging findings are poor to moderate predictors of disability. For the specific case of the upper extremity, a recent review concluded that there is still limited evidence of an association between quantitative medical imaging biomarkers and neck and shoulder MSDs.

The key feature determining disability and rehabilitation from MSDs among individuals is the presence of pain (Laisné et al. 2012). Once pain occurs as a result of MSDs, most likely it will occur again intermittently throughout the lifetime. If pain lasts enough and/or is perceived as strong enough it will result in disability. How pain is managed is therefore crucial as it ultimately determines the burden of MSDs to individuals, their families, and society. Although the ways in which pain exerts the effect on how persons feel are complex and remains undetermined, it appears that emotional processes (e.g. depression) may be involved in how pain is perceived; and therefore, on how large is the consequence of pain. Cognitive (e.g. increased attention to pain); behavioural (e.g. low levels of physical activity, limitation of adaptive pain coping behaviour, and impairment of social support); and neurophysiological factors (e.g. enhanced CNS pain processing, impaired neuroendocrine function) seem to mediate the pathway from emotional disturbances to final outcomes (Edwards et al. 2011).

Low back pain, neck pain, osteoarthritis, osteoporosis, and rheumatoid arthritis are some of the most common conditions among MSDs causing the largest public health burden to society worldwide. A brief account of the natural history, diagnosis, and treatment of a few major conditions is presented next.

Low back pain

Definition

Low back pain (LBP) is characterized as pain manifesting from the anatomic area between the twelfth rib and the superior aspect of the thigh either with or without radiating leg pain (Krismer and Van Tulder 2007). LBP can result from several different causes. Non-specific LBP has received the most attention from the healthcare arena because it represents more than 80% of the low back pain cases seen in the healthcare setting. These cases of LBP can be initially classified based on whether or not it presents with pain in one or two legs; and based on pain duration.

Pathophysiology

The pathophysiology of LBP can generally be classified as either mechanical or non-mechanical in origin (Biyani and Andersson 2004). Mechanical LBP can be due to disk herniation, where the annulus fibrosis around the disk tears, allowing the disk nucleus to push against nearby nerve roots and cause pain. Another possible mechanical pathway for LBP that is commonly seen in ageing populations are arthritic spurs that impinge about adjacent nerve roots in the spinal canal igniting pain. As we age, bone remodelling can cause spurs to develop at the vertebral bodies and these spurs can directly compress these nerve roots and vascular structures leading to pain, inflammation, and secondary oedema. Non-mechanical LBP can be caused by neoplastic, infectious, inflammatory arthritis or other diseases (e.g. Paget's disease, Scheuermann's disease, Basstrup's disease, etc.) (Jarvik and Deyo 2002).

Natural history

Most people will feel LBP at least once in their lives independently of the root causes; although only a proportion of them are likely to seek healthcare services (Balagué et al. 2012). Among individuals with non-specific LPB, the frequency and duration of pain can vary considerably, with some individuals reporting multiple periods of pain with interspersed period of less severe or no pain. This period of pain could be brief, lasting a few days, or it could be longer, lasting a few weeks or even months that eventually leads to disability. If the pain lasts for a significantly extended period of time beyond the normal healing period of approximately 2 months, then a chronic pain syndrome could arise (Krismer and Van Tulder 2007). Studies have also shown that when pain radiates down one or both legs there is a strong association with more significant loss of function (Hoy et al. 2010a).

While a great majority of adults will experience one episode of back pain during their life, many adults will also experience new episodes of back pain (Jeffries and Grimmer-Somers 2007). In fact, recent studies have classified individuals who seek healthcare for back pain into four categories, including: those persons with 'persistent mild' pain, those who start with mild pain, progressing quickly to 'recovering' with no pain; those with 'severe chronic' pain who have permanently high pain; and those with 'fluctuating' pain whose pain varied between mild and high levels (Dunn et al. 2006).

Risk factors

Risk factors for a new episode of LBP may include: poor general health, female gender, and body weight in women (Croft et al. 1999); genetics and age (incidence is higher in the third decade (Hoy et al. 2010a), occupational exposures and education (Hoy et al. 2010a)). Although by definition non-specific LBP has unknown causes, there are known associations between LBP and degeneration of the lumbar discs seen with clinical imaging. The scientific literature suggests that mechanical factors are associated with self-reported LBP; however, several meta-analyses and systematic reviews have suggested that mechanical factors such as awkward posture, manual material handling, pushing, or pulling, and prolonged occupational sitting are not causal of LBP. Obesity and smoking may also play a role (Balagué et al. 2012). Genetic factors may also be implicated via disc narrowing effects on pain or via two genes (i.e. catechol-O-methyltransferase [COMT] and the μ-opioid receptor [OPRM1])

implicated in pain perception, signalling, psychological processing, and immunity. Interleukin-1 gene cluster polymorphisms are associated with modic changes and might have a pathogenic role. Genotype has also been reported to be associated with the outcome of surgery for degenerative disc disease (Balagué et al. 2012).

Treatment

In general, the recovery from acute LBP either with or without radiating pain down the leg has a favourable recovery profile. While certain interventions such as stress management, shoe inserts, or insoles, back supports, ergonomics or back education, and reduced lifting programmes are not proven to be effective; exercise interventions seem to be the most effective in reducing LBP. For example, intervention studies have shown that when compared to individuals not randomized to the intervention group, those who receive specific training and advice on proper manual material handling either with or without the use of assistive devices, report less sick leave (Balagué et al. 2012).

Neck pain

Definition

Neck pain (NP) encompasses a broad range of different clinical presentations and has been defined as pain in the neck that lasts for at least one day either with or without radiating upper limb pain that is associated with activity limitations (Hoy et al. 2010b). NP can result from different specific causes (e.g. trauma) or serious pathologies (e.g. inflammatory disorder, radiculopathy, or myelopathy); but in the majority of cases, as it is for LBP, NP is non-specific (Hoving et al. 2001).

The Bone and Joint Decade's Task Force on Neck Pain and Its Associated Disorders (Carroll et al. 2008) recommended to classify NP through the following scheme:

Grade I neck pain: No signs or symptoms suggestive of major structural pathology and no or minor interference with activities of daily living; will likely respond to minimal intervention such as reassurance and pain control; does not require intensive investigations or ongoing treatment.
Grade II neck pain: No signs or symptoms of major structural pathology, but major interference with activities of daily living; requires pain relief and early activation/intervention aimed at preventing long-term disability.
Grade III neck pain: No signs or symptoms of major structural pathology, but presence of neurologic signs such as decreased deep tendon reflexes, weakness, and/or sensory deficits; might require investigation and, occasionally more invasive treatments.
Grade IV neck pain: Signs or symptoms of major structural pathology, such as fracture, myelopathy, neoplasm, or systemic disease; this grade requires prompt investigation and treatment.

Pathophysiology

In a large number of individuals, neck pain can be a result of either muscular or ligamentous factors related to posture, poor ergonomics, stress, and/or chronic muscle fatigue. Neck muscle pain can also develop secondarily as a consequence of postural adaptations to a primary source of pain in the shoulder or other upper extremity joints (Bogduk 2011). Individuals with degenerative arthritis in the upper cervical joints can present with severe neck pain that transmits down into the neck or to the back of the ear.

Natural history

Neck pain is first experienced in childhood or adolescence and can manifest several times over the life course. In fact, a large proportion of individuals with neck pain never really experience complete resolution of their pain. Recent research studies have shown that between 50% and 85% of those who experience neck pain at some point in their life, they will report neck pain again 1 to 5 years later (Haldeman et al. 2010). This neck pain experience appears to be similar in the general population, workers and among individuals who experience motor vehicle crashes (Haldeman et al. 2010).

The presentation of neck pain among individuals who seek healthcare can vary and includes concomitant pains such as headache, shoulder, upper limb, upper and/or lower back pain (Leaver et al. 2012). However, among

individuals with non-specific neck pain, the origin of the pain and its progression can also vary. In addition, pain arising from anatomic regions other than the neck can be explained as referred pain from innervated cervical spine structures, or as other regional conditions coexisting with neck pain, or as the presence of a generalized, widespread pain syndrome involving neck pain. Nonetheless, neck pain can be a complex health condition that is difficult to treat and contributes significantly to the burden of MSDs (Leaver et al. 2012).

Several factors can affect the course of neck pain. For example, younger age is associated with a better prognosis, whereas poor health and prior neck pain episodes were associated with a poorer prognosis. Poorer prognosis was also associated with poor psychological health, worrying, and becoming angry or frustrated in response to neck pain. Greater optimism, a coping style that involved self-assurance, and having less need to socialize, were all associated with better prognosis. Specific workplace or physical job demands were not linked with recovery from neck pain (Haldeman et al. 2008).

Risk factors

Poor general health, smoking, higher pain, and lower Short-Form Health Survey (SF-12) mental component scores have been reported to be independently associated with higher level disability resulting from neck pain (Leaver et al. 2012). Several epidemiologic studies suggest that risk factors for neck pain have a multifactorial aetiology, including for example, age, gender, and genetics. However there is no evidence that common degenerative changes in the cervical spine are risk factors for neck pain. Research studies have also documented that modifiable factors for neck pain include smoking, exposure to environmental tobacco, and physical activity participation. Occupational health studies have also documented that high quantitative job demands, low workplace social support, sedentary work position, repetitive work, and precision work have been associated with increased risk of neck pain (Haldeman et al. 2008).

Treatment

Generalized neck pain can be treated conservatively with the use of over-the-counter pain medications, ice, heat, and massage, and strengthening and/or stretching exercises at home. Current strategies suggest that if the pain persists after a few weeks of conservative treatment, then further evaluation is warranted (Haldeman et al. 2010). There is a dearth of high-quality evidence elucidating the best specific treatment for neck pain in general. For instance, there are many reviews as there are controlled trials of therapies for neck pain; yet for many interventions, there is no concordance in the conclusions drawn by reviewers. Few studies have been conducted on the long-term effects of treatments for neck pain. A study conducted by Haldeman et al. (2008), suggested that compared to quality of life years with the use of standard NSAIDs, Cox-2 NSAIDs, exercise, manipulation, and mobilization among adults with neck pain; active treatments are not significantly superior to any other treatment option in the short or long term even after controlling for adverse event risks, treatment effectiveness, and patient-preferences for health outcomes in their models (Linton and van Tulder 2001; Haldeman et al. 2008).

Osteoarthritis

Definition

Osteoarthritis (OA) is a degenerative disorder of the joints that can also be termed degenerative joint disease degenerative arthritis or osteoarthrosis. Together they comprise a group of joint conditions that involve degradation of joint tissues, including the articular cartilage and subchondral bone. Joints of the hip and knee are most frequently affected, but also the joints of the hand, spine, and feet are typically affected. Individuals with these conditions report feeling specific pain, tenderness, stiffness, and locking around the joint. Among persons with chronic joint degeneration one may see muscles shrinkage and ligaments loosening around the joint. Current classification schemes for OA include documentation of loss of cartilage, subchondral sclerosis, subchondral cyst formation, and osteophytes on radiological imaging, and self-reported joint pain (Fransen et al. 2011).

Pathophysiology

OA is said to be primary if originated without a known cause (Punzi et al. 2010). However, OA can also be secondary to other diseases such as congenital joint disorders and diabetes. Traditionally, osteoarthritis was

thought to affect primarily the articular cartilage of synovial joints; however, pathophysiologic changes are also known to occur in the synovial fluid, as well as in the underlying bone structures, the overlying joint capsule, and other joint tissues (Punzi et al. 2010). During the early phases of osteoarthritis, swelling of the cartilage usually occurs, because of the increased production of proteoglycans that reflects an effort by the cartilage's cellular structures to repair cartilage damage. This phase may last for several years or decades and is characterized by intense repair of the articular cartilage. As osteoarthritis progresses further, however, the cartilage eventually begins to soften and lose elasticity and thereby further compromising joint surface integrity. In the long run, this loss of cartilage results in loss of joint space (Punzi et al. 2010).

Natural history

OA is a joint condition that involves a slow, progressive degeneration of the joint and is the most common type of all arthritis conditions. We know that OA develops as a result of articular cartilage damage that can be triggered by a complex interplay of genetic, metabolic, biochemical, and biomechanical factors with secondary involvement of the inflammation pathway. The overall natural process involves interactive degradation and repair processes of cartilage, bone, and synovium. Of all the cells involved in the aetiology and pathogenesis of OA, chondrocytes play a large role. Studies on animals and humans suggest that these cells exhibit numerous metabolic dysfunctions leading to OA. This pathology more than any other disease pathology can result in difficulties in daily activities such as walking, stair climbing, and other lower extremity tasks (Gabriel and Michaud 2009).

Risk factors

OA is said to be a primary disorder if it originates without a known cause; although evidence suggests that risk factors such as genetic predisposition, age, obesity, female sex, greater bone density, joint laxity, and excessive mechanical loading are associated with its aetiology (Punzi et al. 2010) Recent evidence from North American and European cohorts suggest that obesity or heavy occupational physical activity are risk factors for symptomatic knee and hip OA (Gabriel and Michaud 2009; Fransen et al. 2011) In addition, other studies have documented that history of joint injury, frequent stair climbing or frequent lifting of heavy weights are significantly associated with hip OA where the strongest association was among individuals with a history of joint injury. In a case–control study in Japan, there was an association with occupational lifting but no association with obesity (Fransen et al. 2011). In other study, OA has been associated with high obesity levels and cardiovascular disease which can cause synergistic effects (Punzi et al. 2010).

Treatment

While there is no known cure for osteoarthritis, certain treatments can help to reduce pain and maintain joint movement. Treatment generally includes exercise, lifestyle modification (e.g. diet) and pain relievers. If the condition is severe, joint replacement surgery is typically recommended. While specific medications designed to reduce pain (e.g. acetaminophen and opioid analgesics are helpful, they do not have an effect on joint inflammation as do non-steroidal anti-inflammatory drugs). In more severe cases of OA, there are two types on injects that have shown moderate evidence for pain improvement: steroid (i.e. glucocorticoid) injections and injections of a liquid known as hyaluronate designed to mimic the natural fluid found in joint space (Iagnocco and Naredo 2012). Platelet-rich plasma (PRP) is an autologous blood product that contains a high concentration of platelets and has emerged as a viable treatment method for individuals suffering from OA (Bennell et al. 2017). Growing epidemiologic data suggests that PRP for the treatment of OA is efficacious and safe, particularly during early states of OA (Meheux et al. 2016). Once these conservative and minimally invasive approaches have been exhausted, a surgical approach to rehabilitating the osteoarthritis joint may be needed (Iagnocco and Naredo 2012).

Osteoporosis

Definition

Osteoporosis is classified as either primary or secondary depending on what is causing the disease. The primary form of osteoporosis is a metabolic bone disease characterized by low bone mass and microarchitectural deterioration of bone tissue, leading to enhanced bone fragility and increased fracture risk (Riis 1993). Secondary

osteoporosis can result from a variety of the chronic conditions that significantly contribute to bone mineral loss, or it can result from the effects of medications and nutritional deficiencies. The World Health Organization (WHO) defines osteoporosis using bone density such that a bone density score that is 2.5 standard deviations (SDs) or more below the young adult mean value. However, individuals with bone density between scores 1 and 2.5 SDs below average are said to have osteopenia (Kanis et al. 1994). In the case of the spine, decreased bone density imparts increased risk for bone fracture; for every 1 SD decrease in bone density of the spine there is an increased risk for new vertebral fractures by factor of 2.0 to 2.4 (Wasnich 1993).

Pathophysiology

Challenges in bone mass acquisition during human growth and development and accelerated bone loss in the period after peak bone mass are both leading mechanisms for the development of osteoporosis. Studies have shown that both of these mechanisms are influenced by environmental and genetic factors (Iagnocco and Naredo 2012). In postmenopausal women, studies have attributed about two-thirds of the risk of fracture to be related to the premenopausal peak bone mass (Farmer et al. 1984). In twin and mother-daughter dyad studies, about 40–80% of the variability observed in bone mass can be attributed to genetic factors. In fact, biological studies have suggested that genes regulating oestrogen receptors, transforming growth factor-β, and apolipoprotein E and collagen, influence the development of osteoporosis. Conversely, bone loss, is more strongly influenced by individual level factors such as nutritional, behavioural, and pharmacologic factors yet still associated to a lesser degree with oestrogen genetic factors (Ronis et al. 2011).

Natural history

Osteoporosis is a complex and dynamic disease of bone where there is asymptomatic reduction in the quantity of bone mass per unit volume. When the bone mass reaches a very low threshold, the structural integrity and mechanical support provided by the bone fails leading to fractures. The overall natural process of this disease, involves a dysregulation of the homeostatic processes between osteoclastic (bone destruction) and osteoblastic (bone growth) activity. If sufficient remodelling in one direction occurs (i.e. more bone destruction than growth), the bone becomes weakened and could lead to fractures (Hoy et al. 2010b; Hoy et al. 2012).

Risk factors

There are several well-known modifiable and non-modifiable risk factors for osteoporosis. Non-modifiable risk factors include age, body size, race, and family history (Drake et al. 2012; Wright and Saag 2012). Recent epidemiologic studies in the US suggest that men and non-white women at risk add approximately 30 to 54 million affected persons. Racial and ethnic differences in the prevalence of osteoporosis and osteoporotic fractures have been well documented (e.g. Ibrahim et al. 2012). Some studies have also noted that women of Asian descent have lower bone mineral density that white women (ILO 1999). In terms of age, studies have shown that decreased bone mineral density and osteoporotic fracture rates increase throughout the life course. Similarly, men also suffer from osteoporosis yet not at the same rate as women (ILO 2000; Jarvik and Deyo 2002). Modifiable risk factors, on the other hand, include sex hormone deficiencies, diet, certain medical conditions (e.g. anorexia nervosa), and medications, lack of physical activity, heavy alcohol consumption, and smoking (Wright and Saag 2012).

Treatment

The treatment plan for osteoporosis is often multifaceted including pharmacologic and behavioural interventions to reduce risk of fracture and improve bone mineralization. For example, a treatment plan could include activities in exercise, improving nutrition, and educational training on the prevention of falls. Pharmacologic interventions often complement these approaches and include drugs that modify bone reorganization, increase bone density, and subsequently reduce fracture risk. Mineral and vitamin supplementation is also paramount and includes the combination use of calcium and vitamin D (Drake et al. 2012).

Rheumatoid arthritis

Definition

Rheumatoid arthritis (RA) is a disease characterized by chronic inflammation of the joints; however, adjacent tissues and other body organs (i.e. such as the skin, heart, lungs, and eyes) can also be affected. In this condition, the body's immune system has turned on itself and attacks the tissues and structures within the joints. While any joint can be affected by RA, the hands and feet are most commonly affected and have the greatest potential for leading to joint destruction and functional disability (Huizinga and Pincus 2010).

Pathophysiology

While the direct aetiology of RA is not well understood, several risk factors have been associated with the onset of an immune system reaction, including, a joint infection, trauma, and tobacco smoke. These immune system triggers cause the tissues within the joint to enlarge and cause inflammation; particularly among individuals who have a genetic disposition for immune sensitivity. Ultimately, the chronic inflammation in the joint leads to the destruction of various joint tissues, including cartilage.

Natural history

Epidemiologic studies have documented geographic variations in the prevalence of RA where it is found highest among Pima (5.3%) and Chippewa Indians (6.8%), and lowest among individuals from Japan and China (0.2–0.3%) (Helmick et al. 2008) These geographic variations in the prevalence of the disease further suggest that either a genetic or environmental component may be present (Silman and Pearson 2002). Population-based estimates suggest that approximately one-third of individuals with RA become work-disabled within 2 years of disease onset, and approximately 50% are work-disabled after 10 years (Combe 2009; JD 2010). Part of this disability and decrease in health-related quality of life is due to the physical, emotional, and social impact of the disease. While physician rating of patient pain is generally lower than pain self-reported by the patient, disease severity has been associated with the degree of pain experienced by the patient as well as their physical functioning.

Risk factors

Research studies have identified both modifiable and non-modifiable risk factors that contribute to the development of RA. In fact, a large proportion of RA cases are believed to arise from the interaction between genetic (i.e. non-modifiable) and environmental (i.e. modifiable) exposures. For example, the number of new RA cases is typically two to three times higher in women than men and the onset is highest among people in their sixties for both men and woman (Silman and Pearson 2002). Among modifiable risk factors include smoking, nutritional/dietary factors, infectious agents, and reproductive hormonal exposures (Pincus and Callahan 1993). Of these factors, tobacco exposure has been shown to be the strongest and most consistent in terms of being associated with RA, particularly among people who have specific immune markers in the blood (e.g. test positive for anticitrullinated protein/peptide antibodies) (Scott et al. 2010). Other studies have documented mixed evidence of an association between the use of oral contraceptives and a risk of RA (Brennan et al. 1997).

Treatment

Treatment options for RA can vary from the newest RA drugs to inclusion of exercise and NSAIDs as well as individualized care and management tips. There are two types of drugs used in treating RA, classified as first- and second-line medications. The first-line drugs include agents such as aspirin and steroids that are commonly used to reduce the person's pain and inflammation. Second-line drugs are slow acting and include the disease-modifying antirheumatic drugs such as methotrexate (i.e. Rheumatrex and Trexall), and hydroxychloroquine (i.e. Plaquenil) that thwart progressive joint disease and support remission. RA usually requires lifelong treatment, including medications, physical therapy, exercise, education, and surgical intervention depending on the disease severity (Ibrahim et al. 2012).

The scope of the problem

Thousands of millions of people suffer from MSD worldwide. The independence and autonomy of persons to do daily normal activities is affected by MSDs. The value in life that was going to be created by these sufferers can now be considered lost to society. Immeasurable suffering accompanies persons with MSDs and their families.

Families and society (e.g. government, industry) often contribute to social welfare and healthcare programmes that subsidize health services (Fransen et al. 2011).

Understanding and quantifying the magnitude of problem that MSDs bring to society is a difficult task (Caban-Martinez et al. 2011). Most cost estimations include only health expenditures and are available only for selected diseases and for some countries. On the other hand, measures based on disease occurrence (prevalence rates and proportion incidence) can offer a comprehensive overview of the problem in different geographical areas of the world, not without methodological difficulties (e.g. as MSDs have a slow onset, it is frequently hard to identify the moment when the disease starts).

A few important messages can be highlighted from the information available: (1) MSDs occur very frequently; (2) MSDs are expensive; (3) the extent of the problem of MSDs is increasing over time; and (4) There are important geographical differences in the distribution of MSDs.

MSDs occur very frequently

MSDs are among the most frequently occurring conditions in the population worldwide. The Global Burden of Disease (GBD) study estimates that 1,266 million people worldwide suffered from MSDs in 2016 (GBD 2016a, 2016b) Some of the conditions mainly affecting people are LBP, neck pain, and osteoarthritis (Table 8.9.1). Prevalence of musculoskeletal disorders has grown consistently from 15.1% in 1990 to 17.8% in 2016 (GBD 2016a, 2016b). It is important to consider, however, that not all MSDs were included in the study that produced these global estimations (Hoy et al. 2012).

Table 8.9.1 Point prevalence (in thousands) and prevalence rate of major musculoskeletal diseases in 2016 (GHDx, 2016)

MSD Type	Total		Males		Females	
	P	%	P	%	P	%
Low back pain†	511,048.2	7.18	229,687.0	6.47	281,361.2	7.88
Neck pain	290,524.0	4.08	123,233.5	3.47	167,290.5	4.69
Osteoarthritis	301,567.2	4.24	118,331.9	3.33	183,235.3	5.13
Other MSDs	337,859.3	4.74	136,523.4	3.85	201,335.9	5.64

† = Activity-limiting low back pain (+/– pain referred into 1 or both lower limbs) that lasts for at least 1 day.

Source: data from Hoy D., Bain C., Williams G., March L., Brooks P., Blyth F., Woolf A., Vos T. and Buchbinder R. (2012). A Systematic Review of the Global Prevalence of Low Back Pain. *Arthritis and Rheumatism*, 64, 2028–37.

MSDs are expensive

In 2011, the world spent a total of US$ 6.9 trillion in health-related expenses (WHO 2014). This figure includes government expenditures and out-of-pocket expenditures in health. An important part of this cost is attributable to MSDs where among developed nations they are responsible for approximately 25% of total expenses due to illness (Walsh et al. 2008). In addition, approximately 20% of regular primary care visits in the vast majority of countries is due to MSDs (Walsh et al. 2008). The pain experienced by individuals with MSDs is one the most common reasons for self-medications and for visits to a primary care setting. Among countries with ready access to medical care, MSDs account for a considerable number of healthcare visits and expenditures (Murray et al. 2012).

Most of data available in terms of monetary expenditures of healthcare is representative of specific geographical areas and specific MSD; most frequently in developed countries. In the State of Victoria, Canada for example, MSDs were the most frequent reason for outpatient visits to the medical centre (11.6% of the total) between 1997 and 2000. Also, among 26 medical specialties, MSD had the sixth highest number of inpatient episodes (6.2% in years 2000/01). In terms of costs, MSDs rank fifth in the number of bed-days away from work during the 1997 to 2001 period (Osborne et al. 2007).

These expenditures are not unique to Canada. An economic study on low back pain in the United Kingdom estimated that the direct costs of LBP were £1.6 billion in 1998. When factoring the indirect costs associated with LBP, countries such as Sweden (costs per capita: 24 US$ direct costs versus 266 US$ indirect costs, 8% versus

92%) and the Netherlands (24 US$ direct versus 299 US$ indirect, 7% versus 93%) had direct costs, lower than indirect expenditures spent in the United Kingdom alone. In the United States, data from the 1988 National Health Interview Survey documented that general LBP accounted for 149 million lost work days annually while work-related LBP accounted for 102 million work days lost. In total, all cases of LBP were estimated to cost about US$1,230 on average for men and US $773 on average for women with an annual productivity loss of $28 billion (Krismer and Van Tulder 2007).

The burden of MSDs is growing

The burden of MSDs is estimated using the disability-adjusted life years (DALY) metric of the GBD study. This metric aims to quantify how many years are lost due to disability (YLD) and due to mortality (YLL) for a list of hundreds of conditions, which are updated yearly. In simple terms, for example, if a person loses part of her capacity to perform daily activities due to a particular condition (disability weight of the condition), that person loses proportionally part of the years that she was going to live until her expected age of death. The sum of these two loses are the DALYs. For each condition, thousands of sequelae were identified. For example, for the condition 'LBP', four sequelae resulting from combining acute or subacute versus chronic pain, and pain in one or two legs versus without this pain were identified (Hoy et al. 2010a). YLDs are computed as the prevalence of a sequela times the disability weight for that sequel. The YLDs arising from a disease or injury are the sum of the YLDs for each of the sequelae associated with that disease. The value of these estimations roots on the fact that an attempt was made to achieve comparable figures for different geographical areas for two moments in time, years 1990 and 2016.

A few methodological aspects need to be highlighted in order to understand the DALYs: the years of life lost due to mortality are generally low due to MSDs. Part of it is because the chronic nature of these conditions, which causes that their effect is exerted through the loss of quality of life rather than to the loss of years of life. However, there are exceptions. For example, the death rate is increased for all osteoporotic fractures; however, while the death will be not be attributed to the underlying osteoporosis, it will be attributed to the injury. We also see this misclassification in other MSDs such as systemic vasculitis and untreated RA, where the death certificate will likely be labelled as death from end-organ failure or other comorbidity. As a consequence the metric does not fully capture the premature loss of life due to the individual living with the MSD. For conditions with low mortality rates such as osteoarthritis, there might actually be a greater burden of death due to the condition (Woolf et al. 2010).

The general distribution in the burden of disease has shifted from previously greater burden of communicable, maternal, and neonatal causes to non-communicable diseases including MSDs. In 1990, approximately 45.5% of DALYs were caused by communicable, maternal, and neonatal causes, 43.9% from non-communicable diseases, and 10.6% from injuries as compared to estimates in 2016 that had shifted to 27.9%, 61.4%, and 10.7%, respectively (GBD 2016a, 2016b). In absolute terms, worldwide the DALYs due to MSDs grew from 86.6 million ((63.1–112.7) in 1990 to 140 million (102.3–181.7) in 2016 (45% increase); or in relative terms from 1,645 (1,199.3–2,140.7) to 1,894.3(1,384.3–2,457.8) per 100,000 thousand persons (12% increase). These conditions would be responsible in year 2016 for 5.8% of the DALYs, which is a much larger share compared to the 3.6% estimated in year 1990 (GBD 2016a, 2016b).

From 1990 to 2016, the proportion of specific disease conditions in the MSDs has not changed significantly; however, LBP (41.2%), neck pain (20.7%) and osteoarthritis (11.6%) continue currently sharing the largest proportion of the pie among all MSDs in year 2016 (Table 8.9.1) (GBD 2016a, 2016b). Furthermore, LBP is the single MSD condition, which continues being the leading cause of all disability from 1990 to 2016; and has the highest age-standardized rates of YLDs in 133 of 195 countries and territories measured in the GBD study (Vos et al. 2012, 2017). Among DALYs, low back and neck pain combined, are 'the 4th leading cause worldwide in 2016, affecting primarily countries with middle, High-middle and High SDI, where it is the 4th, 3rd and 2nd leading cause', respectively (GBD 2016a, 2016b).

Differences across geographical areas for conditions such as LBP and neck pain are large, although concerns have been raised by the authors of the study regarding potentially methodological issues assigning for the case of neck pain, measurement error as spatial variation (Vos et al. 2017).

Musculoskeletal disease in absolute term is a larger burden for people between 40 and 49 years old (Table 8.9.2) (GBD 2016a, 2016b). However, it is clear that for males start being an important issue (considering all DALYs) as early as 10 years old in relation to other conditions at that age range, which is more or less maintained until the age of death; being one of the most important types of conditions in the range from 35 to 44. For women, MSDs start being important also at age of 10. In relative terms per 100,000 people, the burden remains stable since 1990 to

2016, but it increases monotonically by age until the age of 70–74 years, after which the burden stabilizes or slightly reduces (GBD 2016a, 2016b).

Table 8.9.2 DALYs (in thousands) and 95% confidence intervals related to major MSDs by selected age groups (both sexes) in year 2016

Cause	All age groups	35–39 years	40–44 years	45–49 years	50–54 years	55–59 years	60–64 years	65–69 years	70–74 years	75–79 years	80+ Year
Musculoskeletal disorders	140,030.6 (102,331.5–181,687.7)	11,201.4 (7,651.5–15,440.7)	13,343.0 (9,093.8–18,259.8)	14,993.3 (10,491.7–20,275.4)	14,979.3 (10,534.4–19,901.1)	13,514.7 (9,653.0–18,191.9)	12,452.8 (8,814.4–16,989.2)	10,107.2 (7,196.3–13,272.6)	7,354.7 (5,316.1–9,667.6)	5,471.2 (3,979.8–7,146.4)	5,95 (4,4 7,63
Rheumatoid Arthritis	5,563.4 (3,985.9–7,171.2)	307.7 (209.3–421.2)	398.4 (271.5–537.1)	507.1 (356.2–687.2)	597.1 (420.9–804.5)	631.3 (445.4.5–834.7)	651.2 (461.8–841.0)	564.7 (415.6–717.6)	428.9 (320.5–536.7)	330.0 (249.2–409.8)	371. (282 454.
Osteoarthritis	16,282.9 (11,486.0–22,047.3)	667.1 (465.3–929.2)	1,073.0 (741.7–1,474.1)	1,570.0 (1,094.9–2,170.5)	1,985.4 (1,392.7–2,712.8)	2,173.7 (1,525.6–2,961.3)	2,310.6 (1,616.8–3,131.4)	1,979.2 (1,383.7–2,668.7)	1,503.9 (1,049.6–2,038.45)	1,193.3 (835.5–1,626.6)	1,38 (976 1,88
Low Back Pain	57,648.2 (40,820.5–75,877.0)	4,755.1 (2,974.9–7,209.6)	5,499.4 (3,462.5–8,236.7)	5,755.7 (3,677.7–8,544.8)	5,679.8 (3,627.9–8,424.0)	5,202.6 (3,251.4.1–7,749.2)	4,771.1 (3,065.1–7,144.5)	3,791.2 (2,414.6–5,551.4)	2,806.5 (1,834.7–4,141.0)	2,289.4 (1,493.3–3,243.4)	2,51 (1,7 3,45
Neck Pain	28,.936.4 (19,578.5–40,543.1))	2,567.3 (1,546.8–3,913.5)	3,391.0 (2,049.2–5,342.1)	4,031.4 (2,390.7–6,125.7)	3,779.3 (2,323.3–5,765.7)	2,881.4 (1,745.8–4,299,7)	2,352.0 (1,475.9–3,565.8)	1,905.8 (1,208.5–2,859.3)	1,368.6 (884.5–2,102.4)	986.3 (618.3–1,493.9)	1,02 (677 1,46
Gout	1,071.2 (742.4–1,455.0)	61.1 (39.1–86.7)	83.9 (66.7–149.6)	103.1 (66.7–149.6)	115.2 (73.3–167.3)	125.3 (81.9–180.1)	135.6 (87.9–193.4)	119.3 (80.4–167.3)	92.1 (59.9–131.2)	72.6 (48.7–101.0)	82.4 (54.5 113.
Other musculoskeletal disorders	30,528.6 (21,196.7–42,458.6)	2,843.0 (1,742.3–4,397.7)	2,897.4 (1,761.8–4,546.9)	2,966.0 (1,834.2–4,505.7)	2,822.3 (1,620.0–4,437.5)	2,500.3 (1,549.5–3,854.1)	2,232.2 (1,346.0–3,518.9)	1,747.0 (1,137.8–2,569.2)	1,154.7 (687.8–1,765.8)	599.6 (380.6–872.1)	390. (576 828.

Determinants of MSDs

Four risk factors can be considered major drivers of MSD occurrence worldwide: Obesity (Lim et al. 2012), ageing, the increase in trauma (particularly via road traffic collisions) (Walsh et al. 2008) and repeated mechanical loading (particularly via working conditions) (Lim et al. 2012). At another level of causation, behavioural changes related to diet and physical activity may be driving observed changes in obesity. Major determinants of these behavioural changes include technological development, income increase, globalization, growth of mass media and urbanization (Fig. 8.9.1).

Paradoxically, technological improvements have resulted in different consequences. On the one hand, medical advances and prevention of infant mortality (Anderson and Arias 2003) have made possible that the population ages have contributed to increase human life expectancy (Olshansky et al. 2005) contributing also in this way to ageing of the world population and therefore, to the occurrence of MSDs. Also, technological advances have made possible the reduction of physical demands (Popkin et al. 1989) at least for a part of the population, which may have also contributed to obesity and therefore to the occurrence of MSDs. Lastly, the lack of permeation of such advances in many industries contributes to the existence of large sudden physical demands or permanent low demands that are capable of promoting the occurrence or MSDs.

Other related factors include: (1) worldwide shifts in the trade of technology innovations that affect energy expenditures during leisure, transportation, and work; (2) globalization of modern food processing, marketing, and distribution techniques (most frequently linked with westernization of the world's diet); (3) vast expansion of the global mass media; and (4) other changes that constitute the rubric of the effects resulting from an increased opening of our world economy (Popkin et al. 1989).

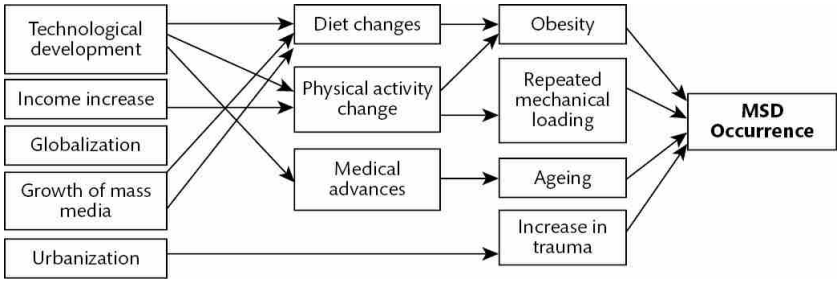


Fig. 8.9.1 Conceptual model relating the major determinants of musculoskeletal disorders.

Obesity

A major worldwide public health concern is the increasing growth of overweight and obese individuals in both developed and now developing countries. The DALYs of musculoskeletal disorders attributable to a high body mass index has been estimated by the Global Burden of Disease study. The most recent available report of this study indicated that for nearly all regions of the world, the burden of MSDs attributable to body mass index (BMI) has grown over the past few years. For the super region Central Europe, Eastern Europe, and Central Asia in 2016, the study attributed 225 DALYs of MSDs per 100,000 people to BMI, which is about 8.6% of all the DALYs per 100,000 people caused by MSDs in the same year. Between 1980 and 2008, the mean BMI (a proxy measurement for overweight and obese levels) worldwide increased by 0.4 kg/m² per decade for men and 0.5 kg/m² per decade for women. Gender and geographic differences in BMI status has been changing over the years. For example, male and female BMIs in 2008 were highest in specific Oceania countries, reaching 33.9 kg/m² for men and 35.0 kg/m² for women in Nauru. Female BMI was lowest in Bangladesh (20.5 kg/m²) and male BMI in Democratic Republic of the Congo 19.9 kg/m², with BMI less than 21.5 kg/m² for both sexes in a few countries in sub-Saharan Africa, and east, south, and Southeast Asia. The United States had the highest BMI of high-income countries. In 2008, an estimated 1.46 billion adults worldwide had BMI of 25 kg/m² or greater; of these 205 million men and 297 million women were obese (Finucane et al. 2011). From 1992 to 2000, using nationally representative data for women aged 20–49 years in 36 countries the proportion of overweight plus obese individuals exceeded underweight in most countries (Mendez et al. 2005). Even more concerning is the fact that the obesity problem in children matches the adult obesity, with the concomitant problems related to diabetes (Popkin 2006).

Multiple factors are contributing to this trend of increasing overweight and obesity worldwide, including most importantly physical activity and diet. Although there is significant variability between countries, it is true that around the world, diet and nutritional food consumption has become increasingly energy-dense and sweeter; and that processed foods are replacing foods that are high in fibre (Popkin 2006). On the other hand, the desire to reduce physical demands has always being an objective. In the advent of the information age, where more modern technologies are used in the workplace, in the shopping arena, and at home, there is likely to be decreased levels of physical activity. The picture of overall shifts in activity, their causes, and their consequences are still incomplete. Since the beginning of time, humans have made efforts to reduce physical effort in the home and at work. In addition, the growing demand for goods that are tasteful including portions higher in fat and sugars has been growing (Popkin et al. 1989). For example, the total calories ingested by US children age 2 to 19 increased on average from 2,065 calories in years 2003–2004 to 2,952 in years 2005_2008; while in adults there was a slight reduction from 2,170 to 2,106 for the same periods (Economic Research Service 2012).

The Economic Research Service (ERS) has conducted a series of studies that give interesting conclusions about the factors that influence diet decisions. Demographic factors (i.e. lower income, younger age, and lower formal education), household characteristics (single head of the household working parents) and acculturation (e.g. acculturated Hispanics), and misperceptions about quality of food at home, are related to lower quality diets or bad habits regarding their diets (skipping meals, eating away from home) (Economic Research Service 2012).

Ageing

The world population is ageing. The median of the world population has gone from 23.9 years in 1950 to 29.6 years in 2015 (Fig. 8.9.2). The population 60 or above worldwide has gone from roughly 202 million (8.0% of the total population) in 1950 to 906 million (12.3% of the total population) in 2015; and the population above 80 years

old went from roughly 14 million (0.6% of the total population) to roughly 126 million in 2015 (1.7% of the total population) (United Nations 2017).

The current median age of the population in Europe, Northern America, and Oceania are much above the world average. The growing trend in Europe is especially concerning as it is the current trend in Asia and Latin American and the Caribbean. In Asia, people over 65 have doubled from 6.8% in 2008 to 16.2% in 2004. It is expected that from 2008 to 2040, the proportion of people over 65 will grow 316% in Singapore, 274% in India, 269% in Malaysia, 261% in Bangladesh and 256% in Philippines. In 2008, Japan has the oldest population, and in absolute terms, China and India have the largest number of people over 65 (106 and 60 million) (Fransen et al. 2011).

This is also important considering that ageing is expected to increase the probability of comorbidity. One report from Scotland indicated that 64.9% of people between 65 and 84 years old suffer from comorbidity (Barnett et al. 2012). A recent review highlights the importance of musculoskeletal disorders as a pervasive component of comorbidity, capable of reducing disproportionally quality of life compared to other chronic conditions (Duffield et al. 2017). Naturally, the presence of comorbidity implies new challenges in terms of treatment accompanied with a significant increase in the cost of health services.

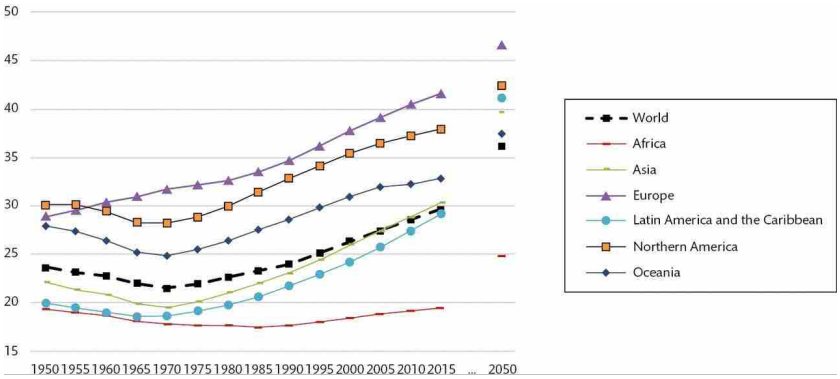


Fig. 8.9.2 Median age of the world population in the past decades and projection until year 2050.
Source: data from United Nations (2017). World Population Prospects 2017 [online]. Available at: https://esa.un.org/unpd/wpp/DataQuery/ [Accessed 20 June 2018].

Occupational factors

Although the importance of workplace factors as determinants of major MSDs such as LBP has been challenged (e.g. Coggon et al. 2018), the paradigm is still today that ergonomic conditions at the workplace can affect significantly the occurrence of this and other conditions, such as neck pain, and other upper-extremity conditions. The Global Burden of Disease study estimated that the DALYs per 100,000 people attributable to ergonomic factors ranged roughly in 2016 between 150 (for regions such as North Africa and Middle East, and Latin America and the Caribbean) and 275 (for the regions Central Europe, Eastern Europe, and Central Asia). It is, however, the regions North Africa and Middle East, and Latin America and the Caribbean are the ones growing in the past years since 1990 to 2016, from roughly 140 DALYs per 100,000 people to over 150 DALYs per 100,000 people.

Future directions

By now it should be clear that there are societal determinants of MSDs which act through several different complex pathways that give way to a significant increase in MSD occurrence, with important effects to human development and quality of life. To tackle such complex problem, different strategies are required at different levels (Fig. 8.9.3).

Policy level

Two actions that are non-specific to MSDs but are the platform for the health of the population are required: Redressing general social inequalities, including education, employment, housing, and food; and improving the

health financing system. Such actions are typically in the range of action of governments and the platform over which other actions are built. No health system alone will be able to address health inequalities if these factors are not addressed. According to the International Labour Organization (ILO), social security benefits that includes cover for lost wages in the event of illness is available to only one in five adults worldwide, and more than half of the world population lacks any form of social protection (ILO 1999; ILO 2000). For example, approximately 5–10% of adults in sub-Saharan Africa and Southern Asia as compared to 20–60% of middle-income countries have social security coverage. Recommendations in this regard have been given by the WHO. Country governments need to make sure that if people pay at the entrance of health services, those payments corresponds to their capacity of payment on a sliding scale. They also need to improve access and utilization of primary preventive services for the poor, despite the challenges of implementing such services in low-resource and economically strapped countries (WHO 2010).

Efforts to finance these types of systems should include especial attention to the need to better distribute available resources. Among Organization for Economic Cooperation and Development (OECD) countries there is a known 20/80 syndrome in which they spend over 80% of the world's resources on health yet make up less than 20% of the world’s total population. While only 6.5% of the African and 3.7% of South East Asian regions Gross Domestic Product (GDP) is spent on healthcare, OECS spend a much larger proportion on health (12.4%). Low-income countries with annual per capita income of less than US $1,005 relied most heavily on out-of-pocket payments to finance healthcare. In these countries the share of out-of-pocket payment in total health expenditure measured in US$ terms was 50% compared to only 13% in countries with incomes higher than US$12,276 per capita (WHO 2010).

Other important actions that are specific to MSD at the policy level are the creation and awareness of the scope of the MSD problem and investing in MSD research. Creating awareness is required so that action is started. Awareness moves political willingness. For example, the recent GBD study has placed LBP as the single leading cause of disability in year 2016 worldwide among 328 selected conditions assessed in this study (Vos et al. 2016). However, the health-related Sustainable Development Goals stayed focused on reducing mortality rather than the main causes of disability; and it has been postulated that tackling disability would be perceived as a more difficult task than tacking mortality (Vos et al. 2017). Nevertheless, the task persists; the information needs to reach and be understood by policy makers to prompt action. Efforts by the Bone and Joint Decade, a Global Alliance for Musculoskeletal Health have focused on raising this awareness of the burden of musculoskeletal conditions by developing sustainable networks and improving access to cost-effective prevention and treatment (Global Alliance for Musculoskeletal Health).

Healthcare systems are now confronted with the growing costs and increased prevalence of MSDs, that will require a coherent policy for dealing with MSDs. Identifying and prioritizing the research efforts on the most effective and affordable strategies for low-resource countries to deal with these conditions is paramount (Murray et al. 2012). The revealed outmost importance of MSDs should have consequences regarding the health system’s capital investments, especially to the light of an ageing population that will suffer comorbidity and considering too current trend in the increase of population BMI. As the expected future burden of MSDs grow, investment efforts into these cost-effective strategies will need to spread over a long-term outlook. In addition, training and educational efforts for the existing and entering healthcare workforce is urgently needed to adequately address the large group of individuals that will be afflicted with MSDs. Innovation in health professional education should include MSDs on a region by region basis, so that trends in demographic and epidemiologic change is fast enough that a forward looking assessment of the burden should be incorporate (Murray et al. 2012).

Important questions are still to be answered to carry out action, which is accomplished through research. The question remains about the research that should be put in place as the needs are great. Research in this area should follow the large framework that is being proposed here. There are needs at the policy level and there are needs at the practice level. At the policy level, there are important unsolved questions about what policies could make the job done regarding prevention of MSDs, access to healthcare and improvement of severity of disability (Vos et al. 2017). At the practice level, there are needs specifically regarding best clinical practices and new treatments. An important area of research priority includes assessing the impact of modifiable risk factors through innovative treatment approaches. Public policy that addresses these risk factors may help significantly reduce the public health burden and cost of MSDs (Haldeman et al. 2008).

Research on the specific risk factors that should be targeted in specific populations is still pending. For example, there is a dearth of epidemiologic research from low and middle-income countries on musculoskeletal disorders (Hoy et al. 2014). Due to significant demographic and environmental differences influencing the onset and progression of various musculoskeletal disorders in these regions, it becomes challenging to reasonably extrapolate some of the risk factor finding from high-income countries to low and middle-income countries. In developing

countries, trends in obesity, the burden of heavy physical labour, diminished access to healthcare variations in the approach to pain treatment and linguistic variation all contribute to differences in reported and treated MSDs (Fransen et al. 2011). Nonetheless, these observations could be lower or higher estimates of the true value, given the general lack of epidemiologic data in these developing countries that could support the development of better policy (Gabriel and Michaud 2009).

Practice level

While research progresses, much is already known on what to do and it should be put in practice. We already know what the major modifiable risk factors are; and ways to improve access to healthcare. While they may require adjustment to be applied in specific contexts, the basic understanding of the problem and potential ways to tackle it have already been proposed.

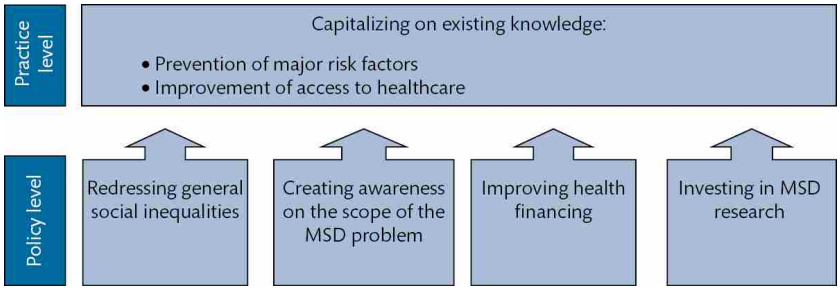


Fig. 8.9.3 Conceptual model for practice and policy levels approaching to addressing musculoskeletal disorders.

For example, in the case of LBP treatment, there are several new scientific and clinical investigations underway to understand innovative technologies, new drugs, and strategic management approaches for individuals with LBP. The healthcare community should be encouraged by the fact that we already know so much about this the challenges of LBP that is affected by a host of genetic, physical, psychological, environmental, cultural, and societal factors (Balagué et al. 2012). Unfortunately individuals with musculoskeletal disorders are too often ignored, and their complaints often misunderstood by healthcare providers. Large population studies have shown that successful management of musculoskeletal pain must occur during the acute phase in order to hinder progression to a more chronic state. Adoption of broader standards of care for these conditions, particularly in low-resources areas is needed (Walsh et al. 2008). This is especially the case when MSDs are comorbid with other health conditions. For example, research has shown that MSDs are closely associated with mental health and physical disability. Among older with MSDs, depression is common, and is associated with functional disability (Punzi et al. 2010). Studies have shown that the older adults depression and pain are better predictors of disability than findings from radiographic imaging. These findings can inform the development of practice guidelines to screening for other comorbid conditions that might improve the overall health of the person with MSDs (Punzi et al. 2010).

Regarding the prevention of known modifiable risk factors for MSDs, practice efforts should focus on new ways to prevent and control obesity and exposures to mechanical demands in the workplace. However, little has been done globally, nationally, or regionally to address systematically obesity (Popkin 2006). From the 1980s to 2008, trends in the mean population BMI have varied substantially between nations. Developing interventions and drafting policies that modify this observed increase in BMI by targeting these modifiable risk factors are highly needed in most countries (Finucane et al. 2011).

Practice efforts to address MSDs should include education on diet. Two noteworthy country efforts can be seen in Finland and South Korea (Holveck et al. 2007). Finland has been reported as one of the countries that have made efforts in the direction of using intersectoral collaboration (Pietinen et al. 2010). They have developed an integrated national price and food-labelling policy coupled with educational programmes on diet and nutrition as well as advocacy efforts from select organizations. A second example is South Korea that has long promoted the use of a traditional diet rich in vegetables and low in fat—very distinct from observed trends in other middle-income countries that have experienced a westernized shift in dietary consumption (Lee et al. 2012). Over the years, the South Korean government has provided educational community workshops to new brides on how to

prepare traditional dishes and avoid incorporation of western style dishes. These efforts have resulted in lower obesity levels and lower intake of energy-dense foods that would be expected from high-income level countries.

Regarding treatment options and practice, there is a need to translate what is already known scientifically to guidelines for clinicians. In the advent of the information age, healthcare professionals can easily become bogged down with multiple sources of information from professional organizations, drug developers, and scientific journals. Data from all these sources could be combined, digested, and presented in a more systematic fashion so that the end user (i.e. healthcare professionals) can benefit from a uniform and up-to-date source of information on best practices to treat MSDs. Translating our basic and clinical science knowledge based into real world bedside practices should be warranted (Punzi et al. 2010). Clinical recommendations that consider presence of comorbidity and treatment burden in such circumstances should also be prioritized (Duffield et al. 2017). However, it appears clear that the formulation of clinical recommendations is not enough to introduce changes in clinical practice; therefore, research on how to permeate effectively these practices is warranted (Buchbinder et al. 2015).

Regarding access to healthcare, efforts need to be made so that available treatments permeate all people in need, especially in more vulnerable populations. This is of course, related to the financing aspect mentioned earlier. For example, among rural regions of the developing world, access to joint replacement surgery is scarce and likely leading to a large number of older adults living with high-grade degenerative joint disease and poor quality of life. The new possibilities of the digital era should be explored to improve health access, considering ethical issues and new barriers related to the access to technology (Lupton 2015; Howard and Kaufman 2018).

Despite access to new information that is made readily available to healthcare providers in developing countries other barriers may be present (Turner 2009; Doughty et al. 2010; Vargas et al. 2010). For example, in order to minimize the sequelae from a bout of LBP, providers in developing countries could incorporate the use of evidence-based guidelines to standardize their management approach and improve MSD treatment outcomes in their community setting. However, despite the progress made in the past two decades in developing and updating guidelines, and adapting them on a national basis, uptake by healthcare providers is not optimum (Global Alliance for Musculoskeletal Health). There are challenges with implementing guidelines in low-resource setting, including the use of specific drugs or technologies that might not be readily available in those areas, as well as pressure from industry or healthcare providers who challenge changes in medical and public health practice (Balagué et al. 2012).

REFERENCES

Anderson, R. and Arias, E. (2003). The effect of revised populations on mortality statistics for the United States, 2000. *National Vital Statistics Reports*, **51**, 1–24.

Balagué, F., Mannion, A., Pellisé, F., and Cedraschi, C. (2012). Non specific low back pain. *The Lancet*, **379**, 482–491.

Barnett, K., Mercer, S., Norbury, M., Watt, G., Wyke, S., and Guthrie, B. (2012). Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study. *The Lancet*, 380(9836), 37–43.

Bennell, K., Hunter, D., and Paterson, K. (2017). Platelet-rich plasma for the management of hip and knee osteoarthritis. *Current Rheumatology Reports*, **19**, 24.

Biyani, A., and Andersson, G. (2004). Low back pain: pathophysiology and management. *Journal of the American Academy of Orthopedic Surgery*, **12**, 106–15.

Bogduk, N. (2011). The anatomy and pathophysiology of neck pain. *Physical Medicine & Rehabilitation Clinics of North America*, **22**, 367–82.

Brennan, P., Bankhead, C., Silman, A., and Symmons, D. (1997). Oral contraceptives and rheumatoid arthritis: results from a primary care-based incident case–control study. *Seminars in Arthritis & Rheumatism*, **26**, 817–23.

Buchbinder, R., Maher, C., and Harris A. (2015). Setting the research agenda for improving health care in musculoskeletal disorders. *Nature Reviews Rheumatology*, 11(10), 597–605.

Caban-Martinez, A., et al. (2011). Arthritis, occupational class, and the aging US workforce. *American Journal of Public Health*, **101**, 1729–34.

Carder, M., Mcnamee, R., Turner S., Hodgson J., Holland F., and Agius, R. (2013). Time trends in the incidence of work-related mental ill-health and musculoskeletal disorders in the UK. *Occupational & Environmental Medicine*, 70(5), 317–24.

Carroll, L., et al. (2008). Bone and joint decade 2000–2010 task force on neck pain and its associated disorders 2008. course and prognostic factors for neck pain in workers: results of the bone and joint decade 2000–2010 task force on neck pain and its associated disorders. *Spine*, 33, S93–100.

Cisternas, M., et al. (2009). Trends in medical care expenditures of us adults with arthritis and other rheumatic conditions 1997 to 2005. *Journal of Rheumatology*, **36**, 2531–8.

Coggon, D., et al. (2018). Drivers of international variation in prevalence of disabling low back pain: findings from the Cultural and Psychosocial Influences on Disability study. *European Journal of Pain*, 23(1), 35–45.

Combe, B. (2009). Progression in early rheumatoid arthritis. *Best Practice & Research in Clinical Rheumatology*, **23**, 59–69.

Croft, P., Papageorgiou, A., Thomas, E., Macfarlane, G., and Silman, A. (1999). Short-term physical risk factors for new episodes of low back pain. Prospective evidence from the South Manchester Back Pain Study. *Spine*, **24**, 1556–61.

Doughty, K., Rothman, L., Johnston, L., Le, K., Wu, J., and Howard, A. (2010). Low-income countries' orthopaedic information needs: challenges and opportunities. *Clinical Orthopaedics & Related Research*, **468**, 2598–603.

Drake, M., et al. (2012). Clinical review. Risk factors for low bone mass-related fractures in men: a systematic review and meta-analysis. *Journal of Clinical Endocrinology & Metabolism*, **97**, 1861–70.

Duffield, S., et al. (2017). The contribution of musculoskeletal disorders in multimorbidity: implications for practice and policy. *Best Practice & Research Clinical Rheumatology*, **31**, 129–44.

Dunn, K., Jordan, K., and Croft, P. (2006). Characterizing the course of low back pain: a latent class analysis. *American Journal of Epidemiology*, **163**, 754–61.
Economic Research Service (2012). *Food Consumption and Nutrition Intakes. United States Department of Agriculture*. [Online] Available at: http://www.ers.usda.gov/data-products/food-consumption-and-nutrient-intakes.aspx
Edwards, R., Cahalan, C., Mensing, G., Smith, M., and Ja, H. (2011). Pain, catastrophizing, and depression in the rheumatic diseases. *Nature Reviews Rheumatology*, **7**, 216–24.
Farmer, M., White, L., Brody, J., and Bailey, K. (1984). Race and sex differences in hip fracture incidence. *American Journal of Public Health*, **74**, 1374–80.
Finucane, M., et al. (2011). National, regional, and global trends in body-mass index since 1980: systematic analysis of health examination surveys and epidemiological studies with 960 country-years and 9.1 million participants. *The Lancet*, **377**, 557–67.
Fransen, M., Bridgett, L., March, L., Hoy, D., Penserga, E., and Brooks, P. (2011). The epidemiology of osteoarthritis in Asia. *International Journal of Rheumatic Diseases*, **14**, 113–21.
Gabriel, S., and Michaud, K. (2009). Epidemiological studies in incidence, prevalence, mortality, and comorbidity of the rheumatic diseases. *Arthritis Research & Therapy*, **11**, 229–44.
Global Burden of Disease (GBD) (2016a). Deaths and DALYs for 2016 with trends since 1990. [Online] Available at: http://ghdx.healthdata.org/gbd-results-tool
Global Burden of Disease (GBD) (2016b). *GBD Compare*. [Online] Available at: https://vizhub.healthdata.org/gbd-compare/
Global Alliance for Musculoskeletal Health. *The Bone and Joint Decade*. [Online] Available at: http://bjdonline.org/
Guilak, F. (2011). Biomechanical factors in osteoarthritis. *Best Practice & Research in Clinical Rheumatology*, **25**, 815–23.
Haldeman, S., Carroll, L., and Cassidy, J. (2010). Findings from the bone and joint decade 2000 to 2010 task force on neck pain and its associated disorders. *Journal of Occupational & Environmental Medicine*, **52**, 424–7.
Haldeman, S., Carroll, L., Cassidy, J., Schubert, J., and Nygren, Å. (2008). The bone and joint decade 2000–2010 task force on neck pain and its associated disorders. *Spine*, 33, S5–S7.
Helmick, C., Felson, D., Lawrence, R., and Workgroup, N.A.D. (2008). Estimates of the prevalence of arthritis and other rheumatic conditions in the United States. Part I. *Arthritis & Rheumatology*, **58**, 15–25.
Holmgren, K., Fjällström-Lundgren, M., and Hensing, G. (2013). Early identification of work-related stress predicted sickness absence in employed women with musculoskeletal or mental disorders: a prospective, longitudinal study in a primary health care setting. *Disability & Rehabilitation*, **35**, 418–26.
Holveck, J., et al. (2007). Prevention, control, and elimination of neglected diseases in the Americas: pathways to integrated, inter-programmatic, inter-sectoral action for health and development *BMC Public Health*, **7**, 6.
Hoving, J., et al. (2001). A critical appraisal of review articles on the effectiveness of conservative treatment for neck pain. *Spine*, **26**, 196–205.
Howard, I. and Kaufman, M. (2018). Telehealth applications for outpatients with neuromuscular or musculoskeletal disorders. *Muscle Nerve*, 58(4), 475–85.
Hoy, D., et al. (2012). A systematic review of the global prevalence of low back pain. *Arthritis & Rheumatism*, **64**, 2028–37.
Hoy, D., Brooks, P., Blythc, F., and Buchbinder, R. (2010a). The epidemiology of low back pain. *Best Practice & Research Clinical Rheumatology*, **24**, 769–81.
Hoy, D., Protani, M., Dea, R., and Buchbinder, R. (2010b). The epidemiology of neck pain. *Best Practice & Research Clinical Rheumatology*, **24**, 783–92.
Hoy, D., Geere, J., Davatchi, F., Meggitt, B., and Barrero, L. (2014). A time for action: opportunities for preventing the growing burden and disability from musculoskeletal conditions in low- and middle income countries. *Best Practice & Research Clinical Rheumatology*, **28**, 377–93.
Huizinga, T., and Pincus, T. (2010). In the clinic. Rheumatoid arthritis. *Annals of Internal Medicine*, 153, ITC1–1–ITC1–15.
Iagnocco, A. and Naredo, E. (2012). Osteoarthritis: research update and clinical applications. *Rheumatology (Oxford)*, 51, vii 2–5.
Ibrahim, I., Owen, S., and Barton, A. (2012). Genetics and the impact on treatment protocols in patients with rheumatoid arthritis. *Expert Reviews in Clinical Immunology*, **8**, 509–11.
ILO (1999). *Unemployment, Social Protection and Crises: Trends and Issues*. Geneva: ILO.
ILO (2000). *Income Security and Social Protection in a Changing World: World Labour Report 2000*. Geneva: ILO.
Jarvik J. and Deyo R. (2002). Diagnostic evaluation of low back pain with emphasis on imaging. *Annals of Internal Medicine*, **137**, 586–97.
Jd., I. (2010). The changing face of rheumatoid arthritis: sustained remission for all? *Nature Reviews Immunology*, **10**, 605–11.
Jeffries, L. and Grimmer-Somers, K. (2007). Epidemiology of adolescent spinal pain a systematic overview of the research literature. *Spine*, **32**, 2630–7.
Kanis, J., Melton, L., Christiansen, C., Johnston, C., and Khaltaev, N. (1994). The diagnosis of osteoporosis. *J Bone Miner Res*, **9**, 1137–1141.
Krismer, M. and Van Tulder, M. (2007). The low back pain group of the bone and joint health strategies for Europe project 2007. Low back pain (non-specific). *Best Practice & Research Clinical Rheumatology*, **21**, 77–91.
Laisné, F., Lecomte, C., and Corbière, M. (2012). Biopsychosocial predictors of prognosis in musculoskeletal disorders: a systematic review of the literature. *Disability Rehabilitation*, **34**, 355–82.
Leaver, A., Maher, C., Mcauley, J., Jull, G., and Refshauge, K. (2012). Characteristics of a new episode of neck pain. *Manual Therapy*, 18(3), 254–7.
Lee, H., Duffey, K., and Popkin, B. (2012). South Korea's entry to the global food economy: shifts in consumption of food between 1998 and 2009. *Asia Pacific Journal of Clinical Nutrition*, **21**, 618–29.
Lim, S., Theo Vos, T., Flaxman, A., Danaei, G., Shibuya, K., and Adair-Rohani, H. (2012). A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010. *The Lancet*, **380**, 2224–60
Linton, S. and Van Tulder, M. (2001). Preventive interventions for back and neck pain problems: what is the evidence? *Spine*, **26**, 778–87.
Lupton, D. (2015). Health promotion in the digital era: a critical commentary. *Health Promotion International*, 30(1), 174–83.
Meheux, C., McCulloch, P., Lintner, D., Varner, K., and Harris J. (2016). Efficacy of intra-articular platelet-rich plasma injections in knee osteoarthritis: a systematic review. Arthroscopy, 32, 495–505.
Mendez, M., Monteiro, C., and Popkin, B. (2005). Overweight exceeds underweight among women in most developing countries. *American Journal of Clinical Nutrition*, 81:714–21.
Murray, C., et al. (2012). Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010. *The Lancet*, **380**, 2197–23.
Naylor, K. and Eastell, R. (2012). Bone turnover markers: use in osteoporosis. *Nature Reviews Rheumatology*, **8**, 379–89.
Olshansky, S., et al. (2005). A potential decline in life expectancy in the United States in the 21st century. *New England Journal of Medicine*, **352**, 1138–45.
Osborne R., Nikpour M., Busija J., Sundararajan V., and Wicks I. (2007). Prevalence and cost of musculoskeletal disorders: a population-based, public hospital system healthcare consumption approach. *The Journal of Rheumatology*, **34**, 2466–75.
Pietinen, P., Männistö, S., Valsta, L., and Sarlio-Lähteenkorva, S. (2010). Nutrition policy in Finland. *Public Health Nutrition*, **13**, 901–6.
Pincus, T. and Callahan, L. (1993). What is the natural history of rheumatoid arthritis?. *Rheumatic Diseases Clinics of North America*, **19**, 123–51.

Pisetsky, D., Ullal, A., Gauley, J., and Tc, N. (2012). Microparticles as mediators and biomarkers of rheumatic disease. *Rheumatology (Oxford)*, **51**, 1737–46.
Popkin, B. (2006). Global nutrition dynamics: the world is shifting rapidly toward a diet linked with noncommunicable diseases. *American Journal of Clinical Nutrition*, **84**, 289–98.
Popkin B., Haines P., and Reidy K. (1989). Food consumption trends of US women: patterns and determinants between 1977 and 1985. *American Journal of Clinical Nutrition*, **49**, 1307–19.
Punzi, L., Oliviero, F., and Ramonda, R. (2010). New horizons in osteoarthritis. *Swiss Medical Weekly*, 140, w13098.
Riis, B. (1993). Biochemical markers of bone turnover II: diagnosis, prophylaxis, and treatment of osteoporosis. *The American Journal of Medicine*, 95, S17–S21.
Ronis, M., Mercer, K., and Chen, J. (2011). Effects of nutrition and alcohol consumption on bone loss. *Current Osteoporosis Reports*, **9**, 53–59.
Rousseau, J. and Garnero, P. (2012). Biological markers in osteoarthritis. *Bone*, **51**, 265–77.
Scott, D., Wolfe, F., and Huizinga, T. (2010). Rheumatoid arthritis. *The Lancet*, **376**, 1094–108.
Silman, A. and Pearson, J. (2002). Epidemiology and genetics of rheumatoid arthritis. *Arthritis Research & Therapy*, 4, S265–72.
Smolen, J. and Aletaha, D. (2008). The burden of rheumatoid arthritis and access to treatment: a medical overview. *European Journal of Health Economics*, 8, S39–47.
Turner, T. (2009). Developing evidence-based clinical practice guidelines in hospitals in Australia, Indonesia, Malaysia, the Philippines and Thailand: values, requirements and barriers. *BMC Health Services Research*, **9**, 235.
United Nations (2017). World population prospects 2017. [Online]. Available at: https://esa.un.org/unpd/wpp/DataQuery/
Vargas, I., Vázquez, M., Mogollón-Pérez, A., and Unger, J. (2010). Barriers of access to care in a managed competition model: lessons from Colombia. *BMC Health Services Research*, **10**, 297.
Vincent H., Raiser S., and Vincent K. (2012). The aging musculoskeletal system and obesity-related considerations with exercise. *Ageing Research Reviews*, **11**, 361–73.
Vos T., et al. (2017). Global, regional, and national incidence, prevalence, and years lived with disability for 328 diseases and injuries for 195 countries, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet*, **390**, 1211–59.
Vos T., et al. (2012). Years lived with disability (YLDs) for 1160 sequelae of 289 diseases and injuries 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010. *The Lancet*, **380**, 2163–96.
Walsh N., et al. (2008). Bone and joint decade task force for standards of care for acute and chronic musculoskeletal pain 2008. Standards of care for acute and chronic musculoskeletal pain: the bone and joint decade (2000–2010). *Archives of Physical Medicine & Rehabilitation*, **89**, 1830–45.
Wasnich R. (1993). Bone mass measurement: prediction of risk. *American Journal of Medicine*, **95**, 6S–10S.
World Health Organization (WHO) (2010). *The World Health Report: Health Systems Financing: The Path to Universal Coverage*. Geneva: WHO.
World Health Organization (WHO) (2014). *WHO Global Health Expenditure Atlas*. Geneva: WHO.
Woolf A., Vos T., and March L. (2010). How to measure the impact of musculoskeletal conditions. *Best Practice & Research Clinical Rheumatology*, **24**, 723–32.
Wright N. and Saag K. (2012). From fracture risk prediction to evaluating fracture patterns: recent advances in the epidemiology of osteoporosis. *Current Rheumatology Reports*, **14**, 205–11.

8.10

Neurological diseases, epidemiology, and public health

Walter A. Kukull, Kumeren Govender, and James Bowen

Introduction to neurological diseases, epidemiology, and public health

Included in this chapter are brief descriptions of headache, traumatic brain injury (TBI), epilepsy, dementia, peripheral neuropathy, parkinsonism, and multiple sclerosis (MS). Diagnosis is critical to epidemiological study but is often difficult because of long, preclinical undetectable periods. Clinical criteria rather than pathological evidence can lead to misclassification of disease and spurious inferences. To some extent biomarkers are becoming accepted as a preclinical indicator of disease or a means to add diagnostic certainty among the symptomatic.

Population-based studies are the best method to avoid spurious, biased, or non-generalizable results, but for most neurological diseases these are seldom performed. Selecting ‘control’ subjects from the same base that gave rise to the cases is methodologically essential but often a luxury ignored by investigators. Cohort studies must start with people free of disease but if preclinical detection of disease process is not possible, risk factor associations may be severely biased.

Determination of risk factor exposure, and estimating whether it may have occurred during a time window which had potential to result in the disease in question (and not to have resulted from the disease) is critical. In acute TBI this becomes obvious, while for Alzheimer’s disease, which may have a 30-year preclinical period, it is cryptic. Researchers seldom have records, biological specimens, or even accurate self-reporting sufficient to establish previous exposure patterns. Biological markers of exposure (except for genotype) are occasionally employed in cancer studies (e.g. DNA adducts), are seldom available for neurological diseases, but may

provide some hope for the future. Biopsy is seldom feasible and autopsy results reflect cross-sectional cumulative effects of disease, leaving the picture additionally confusing.

In less developed countries, the problems of diagnosis and exposure measurement grow in magnitude. Differences in available facilities and local practices are likely the easiest to overcome. Addressing political concerns and suspicions to gain cooperation necessary to begin a study may take additional time and preparation.

Headache

Clinical overview of headaches

Headache is among the most common medical complaints however the pathogenesis of most headache is poorly understood. The nosology of headaches is based on the cause when known and the clinical picture when the cause is unknown. As many as 90% of all primary headaches fall under a few categories, including migraine, tension-type, and cluster headache. While episodic tension-type headache is the most frequent headache type in population-based studies, migraine is the most common diagnosis in patients presenting to primary care physicians with headache. The International Classification of Headache Disorders (ICHD) is the most commonly used system of classifying headaches, now in its third edition based on greater scientific evidence (Headache Classification Committee of the International Headache Society) (Box 8.10.1). This system classifies individual headaches, as most patients may have more than one type of headache. Headaches associated with neurological or systemic diseases are comparatively rare. Idiopathic conditions are far more common and include migraine (with or without aura), tension-type headaches, and cluster headaches. Because these idiopathic headaches are the overwhelming majority, they have the greatest impact on epidemiological studies.

Migraine without aura, previously named common migraine, is a recurrent headache disorder that typically is unilateral. Some people have throbbing pain while others have constant non-throbbing pain. Nausea, photophobia, and phonophobia may occur. The pain often builds over a few hours. The typical length of the headache is 4–72 hours.

Migraine with aura, previously named classic migraine, has an aura consisting of a fully reversible alteration of neurological function that usually precedes the headache. This commonly consists of a central area of visual loss surrounded by a rim of shimmering light (scintillating scotoma). Non-visual auras may include paraesthesias, numbness, weakness, aphasia, or vertigo. Auras usually precede the headache by about 20 minutes, though timing may vary. The headache is most commonly, though not always, unilateral, throbbing, and often associated with nausea and vomiting. It typically lasts a few hours. It may have premonitory symptoms preceding the headache and aura.

Tension-type headaches, the most common type of headache, are usually bilateral with a sensation of pressure or a tight band around the head. They are less likely to have premonitory symptoms or nausea or vomiting. They do not have auras. They usually last longer than migraine headaches, typically an entire or several days. They build up more slowly than migraine headaches.

Box 8.10.1 International Headache Society, abbreviated classification of headache

1. Migraine
    - 1.1 Migraine without aura
    - 1.2 Migraine with aura
2. Tension-type headache
3. Trigeminal autonomic cephalagias (TACs)
4. Other primary headache disorders
5. Headache attributed to trauma or injury to the head and/or neck
6. Headache attributed to cranial and/or cervical vascular disorder
7. Headache attributed to non-vascular intracranial disorder
8. Headache attributed to a substance or its withdrawal
9. Headache attributed to infection
10. Headache attributed to disorder of homoeostasis
11. Headache or facial pain attributed to disorder of cranium, neck, eyes, ears, nose, sinuses, teeth, mouth, or other facial or cranial structure
12. Headache attributed to psychiatric disorder
13. Painful lesions of the cranial nerves and other facial pain
14. Other headache disorders

Source: data from International Headache Society, The International Classification of Headache Disorders: 3rd edition, *Cephalalgia*, Volume 38, Supplement 1, pp. 1–211. Copyright © International Headache Society 2018.

Cluster headaches are named after the tendency to occur in clusters lasting weeks to months. However, other types of headaches may also occur in clusters and the diagnosis is based on the characteristics of the headache rather than the clustering. The headache develops abruptly. During a cluster, it usually occurs between one and eight times a day, often at the same time of day. The pain is more short-lived than that of other idiopathic headaches and generally subsides within 3 hours. The pain is often more severe than that seen with migraine and patients are often agitated during the attack.

In a single patient, less severe headaches tend to be tension-type while more severe headaches are migraine. Over time a patient's headaches often change from classic migraine in youth to tension-type headaches in middle age. Headache frequency may increase with age, becoming chronic daily headaches. Some term these headaches as 'transitioned migraines'. This term is not included in the ICHD and many of these cases are actually due to medication overuse.

Prevalence of headaches

Case definition, sample characteristics, methodology, and analytic designs have differed, sometimes substantially, raising questions of comparability. One-year prevalences vary from 20–90% (all headaches), 3–24.6% (migraine), 9.8–72.3% (tension-type), and 0.5–7.3% (chronic headache) (Stovner et al. 2007).

Several studies have evaluated the prevalence of migraine, reflecting the availability of validated survey instruments for this disease (Natoli et al. 2010). For migraine, age-standardized prevalence from the Global Burden of Disease Study 2016 was 14.4% (95% confidence interval (CI) 13.8–15) (Stovner et al. 2018). Prevalence peaks at 30–39 years of age, and is higher in women, White people, and those with lower incomes (Lipton et al. 2007). Less than half report their migraines to a physician and approximately 40% of these did not receive a correct diagnosis (Lipton et al. 1998). Susceptibility to migraine may be affected by ethnicity since migraine appears to be more common in Caucasians compared to Africans,

African Americans, or Asian Americans (Osuntokun et al. 1992; Stewart et al. 1996).

There are far fewer studies on tension-type headaches, though it is the most common type of headache. Up to 89% of people suffer tension-type headaches at some time, but for most this is infrequent (Kaniecki 2012). About 25% have weekly tension-type headaches and 2–3% chronic tension-type headaches (Lyngberg et al. 2005a). The prevalence appears to be increasing, though the factors leading to this are not entirely clear. About half of people with tension-type headaches improve within a decade (Lyngberg et al. 2005b). Chronic daily headache from medication overuse is increasingly recognized (Tepper 2012).

Cluster headaches are the least common of the idiopathic headache disorders. Pooled data showed a lifetime prevalence of 124 per 100,000 [CI 101–151] and a 1-year prevalence of 53 per 100,000 (CI 26-95) (Fischera et al. 2008).

Familial and genetic risks of headaches

Genetic influences on migraine have investigated principally familial aggregation and twins. Studies of single genes responsible for familial hemiplegic migraine provide insight into migraine pathophysiology (Ward 2012). The overall lifetime prevalence of migraine with aura for monozygotic (MZ) and dizygotic (DZ) twins is 7%, similar to population surveys (Gervil et al. 1999; Ulrich et al. 1999). Concordance in MZ twins is 34% compared to 12% for DZ twins. For migraine without aura the concordance for MZ twins is 28% and 18% for DZ. This indicates a potential genetic contribution but modifying environmental influences must also be important. Concordance is higher for MZ than DZ twins whether raised together or apart, and about 50% of the variance is explained by genetic factors with the remaining half ‘non-shared environmental factors’, and measurement error (Ziegler et al. 1998; Russell et al. 2007).

The relative risk (RR) of migraine in first-degree relatives of migraineurs is 1.88 compared to a control group (Stewart et al. 2006). The RR was higher in relatives of those whose headaches began before 16 years of age (RR = 2.50) than those whose headaches began later (RR = 1.44). The risk in relatives of those with severe headaches had a relative risk of 2.38 (CI 1.56–3.62) while relatives of those with less severe pain had a relative risk

of 1.52 (0.99–2.34). A few candidate genes have been associated with migraine, including the KCNK18 gene that encodes for TRESK, a two-pore domain potassium channel (Lafrenière et al. 2010) and the CSNK1D gene, which encodes casein kinase I isoform delta (Brennan et al. 2013). However, in a systematic re-evaluation of 27 promising candidate migraine genes in a large data set compiled from genome-wide association studies, with over 5,175 migraineurs and 13,972 control subjects, found that none of the candidate genes reached the threshold for statistical significance (de Vries et al. 2016). Thus, which candidate loci and genes are truly implicated in the pathogenesis of migraine is still uncertain.

Risk factors for headaches

Stress or psychological factors are often cited as contributing to tension-type headaches. However, the frequency of these factors is similar in patients with migraine and tension-type headaches. With chronicity, psychological changes increase in frequency, suggesting that psychological issues are secondary to chronic headaches rather than causative. There is an increased risk of depression in migraineurs and also an increased risk for migraine among those with prior depression, suggesting the two disorders may have mechanisms in common. The issues of migraine and depression have been reviewed (Frediani and Villani 2007).

Chronic daily headaches are more common in women, white people, and those with less education. An improvement in headache frequency after 1 year was more common in those with higher education, non-white people, and those who were married (Scher et al. 2003). However, in adolescents with a family history of migraine, migraine prevalence is the same in low- versus high-income groups suggesting that genetic factors overwhelm environmental risk factors in those genetically susceptible (Hagen et al. 2002). One of the strongest risk factors for headaches is prior headaches (RR = 4.15) (Boardman et al. 2006). Pain in other areas of the body is associated with headache (RR = 1.43).

Stroke and migraine

Migraine has been associated with ischaemic stroke and to a lesser degree with haemorrhagic stroke in a number of studies (Alhazzani and Goddeau

2013). Additional study will be needed to adequately describe the true relationship between migraine and stroke.

Costs and public health impact of headaches

The estimated 23 million people with migraines in the United States may miss 150 million workdays each year with an associated cost of up to US$17 billion (Cady 1999; Hu et al. 1999). Many more people suffer with decreased effectiveness at work than actually miss work days (Schwartz et al. 1997). In Europe, costs of headaches are substantial with mean per-person annual costs of €3,561 for medication overuse, €1,222 for migraine, and €303 for tension-type headaches (Linde et al. 2012). Health-related quality of life is lower in those with migraine, and this is related to the frequency of attacks (Wang et al. 2001). There are many challenges in making cost determinations due to limits on available epidemiological data (Leonardi et al. 2005). However, the best estimate suggests that migraine accounts for 1.4% of all years of healthy life lost worldwide. It is the 19th leading cause of years of healthy life lost in men and 12th in women. Although new advances such as neurostimulators and a recent new class of compounds called gepants to treat migraine attacks have been developed, this major public health issue deserves continued attention.

Traumatic brain injury

Clinical overview of traumatic brain injuries

TBI is commonly divided into two categories by mechanism of action: (1) penetrating or (2) closed head injuries. These also may be characterized as resulting from direct contact injuries or from concussive ‘acceleration deceleration’ types of injuries (Werner and Engelhard 2007). Among penetrating injuries, greater tissue damage is caused by high-velocity penetrating objects than by lower-velocity ones. Both penetrating and closed injuries may perturb or impair cerebral blood flow (CBF) leading to ischaemia and altered cerebral metabolism. Impaired CBF resulting from cerebral vasospasm predicts poorer outcome and may occur in approximately 30% of TBI cases (Werner and Engelhard 2007). Brain

contusions may lead to both blood vessel and cell membrane injury resulting in vasogenic or cytotoxic oedema. Hydrocephalus may develop due to blockage of the routes of normal cerebrospinal fluid (CSF) flow. Finally, closed head injuries may lead to diffuse axonal injury. Diffuse axonal injury results in balls of axonal material occurring at axon transection sites or sites of altered axonal flow. Intracranial pressure monitoring is a concern; however, a recent trial does not support it as superior to other forms of neurological monitoring (Chesnut et al. 2012). Neurological damage due to TBI may grow after the injury because of impaired cerebral perfusion and other factors.

For mild head injuries, including skull fracture, concussion, and unspecified intracranial or head injury (Bazarian et al. 2005), the severity of the injury is usually measured by the degree of post-traumatic amnesia. More recently, concerns for chronic traumatic encephalopathy (CTE) due to repetitive mild head injuries have become a substantial concern as a cause of dementia and dysfunction years later, especially for those engaged in contact sports (Baugh et al. 2012; McKee et al. 2013). Head injury severity is usually classified by the Glasgow Coma Scale (GCS): scores of 13–15 are minor, 9–12 are moderate, 5–8 are severe, and 4 or less are very severe; and by length of coma (Jennett et al. 1976; Sherer et al. 2007). The GCS is universally accepted as a tool for TBI classification because of its reproducibility, simplicity, and predictive value for overall prognosis. Though it is limited by confounding factors such as endotracheal intubation, medical sedation, and paralysis and intoxication. An alternative to this scoring system is the Full Outline of UnResponsiveness (FOUR) Score which includes a brainstem examination however may be somewhat complicated to perform by non-neurologists and has a shorter track-record of prognostic prediction.

Incidence and prevalence of traumatic brain injuries

The US Centers for Disease Control and Prevention (CDC) estimates there were 2.5 million emergency department visits for TBI in 2013 in the United States, an increase from 534 per 100,000 in 2007 to 787 in 2013 per 100,000, with at least part of the increase attributable to falls in older adults (CDC 2003, 2007). Approximately 52,844 of those resulted in death (TBI mortality 17.1 per 100,000 population) based on all reported emergency

department visits, hospitalizations, and deaths. Cassidy et al. (2004) compiled a review of more than 160 published studies and found that approximately 70–90% of all treated TBIs are likely to be mild which results in an expected rate for *treated* mild TBI of 100–300 per 100,000 population. Almost paradoxically, Rivara et al. found that among children with TBI the burden of disability caused by TBI was primarily due to mild TBI (Rivara et al. 2012). TBI incidence in 2013 among children aged 0–4 was the highest of all age groups at 1,591.5 per 100,000, while hospitalization and death rates were highest among those aged 65 and older. In 2013, falls accounted for 47.2% of TBIs, struck by/against events for 15.4% and motor vehicle traffic for 13.7%. Overall, males had higher rates of TBIs compared with females (959 versus 810.8 per 100,000). Males had a more than fivefold higher rate of intentional self-harm-related TBIs (11.7 versus 2.3 per 100,000). The leading cause of TBI-related deaths varied by age. While falls were the most common cause of mortality for persons 65 years and older, intentional self-harm was the leading cause of death for persons between 45–64 years. Motor vehicle collisions were the leading cause of death for persons 15–24, while homicide was the leading cause of death for children ages 0–4 years (Taylor et al. 2017).

Risk factors of traumatic brain injuries

Falls

Falls are perhaps the most common cause of TBI and incidence increases dramatically with advanced age. Those aged ≥75 years had the highest rate per 100,000 population of fall-related TBIs (1,859), followed by 0–4 years (1,119.3), and 65–74 years (539.8) (Taylor et al. 2017). Use of four or more medications may significantly increase risk of TBI from falling.

Falls are also an important cause of TBI among the very young. The 15-State CDC surveillance system allowed calculation of TBI hospitalization rates due to unintentional falls for children aged 0–11 months compared to those aged 12–23 months. Fall-related TBI hospitalizations were 71.5 per 100,000 person-years in the younger group as compared to 36.5 per 100,000 person-years in the older group. Although TBI in toddlers and young children is due primarily to household or playground falls, motor vehicle accidents are the predominant cause in older children (Keenan and Bratton 2006).

Intentional self-harm

The most common cause of TBI-related death changed, from motor vehicle crashes in 2007 to intentional self-harm in 2013. Fatalities due to intentional self-harm were found predominately among males, while firearms were the predominant principal mechanisms of injury. This increase in fatalities is consistent with an overall increase in the United States suicide rates and represents the importance of providing comprehensive and coordinated prevention efforts including social support enhancement, reducing stigma for help seeking and providing help for those greatest at risk (Curtin et al. 2016; Taylor et al. 2017).

Vehicle accidents

The highest rates of TBIs attributable to motor vehicle crashes per 100,000 population were among those aged 15–24 years (258.3), 25–34 years (182.9), and 35–44 years (126.5). Injuries resulting from motor vehicle accidents, whether occurring to occupants or as pedestrian or cyclist versus motor vehicle, result in many of the more severe TBIs. Mandatory seat belt and child seat laws could have an effect on reducing occupant injuries, while bicycle and motorcycle helmet use may reduce the occurrence of TBI in those constituent categories.

Violence

Assaults are estimated to have accounted for approximately 10% of the TBIs occurring in the United States in 2003 or roughly 160,000 TBIs (Rutland-Brown et al. 2006).

As much as 20% of TBI may be the result of violence, roughly half of these are due to firearms. The age group at highest risk is 15–24 years. While males appear to be more likely to sustain an injury due to violence, women may be more likely to die as a result. The incidence of firearm-related injury or death was 42 per 100,000. The homicide rate varies dramatically by age and ethnic group. Overall it was about 14 per 100,000, but rose to 73 per 100,000 in African American males and further to 164.2 per 100,000 among 15–34-year-old African American males.

Abuse and domestic violence are important causes of TBI among women and among children (Monahan and O'Leary 1999; Valera and Berenbaum 2003; Banks 2007). Monahan estimates that about 35% of the 2–3 million

women battered each year by their domestic partner sustain TBI as a result. Keenan et al. (2003) found that physical abuse may be the most frequent cause of TBI in children aged less than 2 years with an estimated incidence rate of approximately 17 per 100,000 person-years; infants had a higher incidence than older children and boys were at somewhat greater risk than girls.

War and sociopolitically directed violence

By the end of 2012 approximately two million US and UK men and women will have served in the wars in Iraq and Afghanistan resulting in thousands of deaths and tens of thousands of severe injuries. TBI is a major problem for the United States military; the Department of Defense reports that between 2000 and 2017, 375,230 military personnel suffered TBI. Injuries related to explosive blasts may account for many or most of these (Warden 2006) and effects of these types of injuries years later could have important individual and public health consequences for many years to come. Not only do blasts account for much of the observable and immediate injuries, but they may also lead to more mild or unreported TBI that can be coupled with post-traumatic stress disorder (PTSD) (Hoge et al. 2004; Holbrook et al. 2005; Okie 2005). People who may have experienced TBI or blast-related TBI could ultimately be at higher risk for dementia as they enter ages 60–80.

Implications for public health

The potential years of productive life lost due to TBI varies with its cause, and causes are differentially age-related. The risk of seizures following TBI is increased up to 17-fold in patients with severe injuries (Annegers et al. 1998). While injuries related to falls and motor vehicles are the most common, those who are most severely affected, have the poorest outcomes, and require the most continuing care, tend to be the very young and the elderly. Sports injuries tend to be mild and resolving, but in the case of repetitive injury may result in serious sequelae years later. War-related TBIs could be the source of an unexpected public health burden because of the large numbers of soldiers serving and the high incidence of blast and other injuries coupled with the development of PTSD, psychiatric symptoms,

cognitive deficits, and degenerative dementias or other neurological disorders.

Epilepsy

Clinical overview of epilepsy

The International League Against Epilepsy (ILAE) Commission on Classification and Terminology is currently the most commonly used means of organizing epileptic syndromes and has proposed substantial changes to the 1989 classification system in 2010 and made further revisions in 2017 (Table 8.10.1) (Fisher et al. 2017). The current framework allows for diagnosis at four levels, depending on information and resources that are available, and also addresses the broad concepts of aetiology and associated comorbidities at all levels.

Seizures are classified into generalized onset, focal onset, and unknown onset. Characteristics of individual seizures are divided into focal onset seizures (formerly named partial or local seizures) that begin in a localized part of the brain, and generalized onset seizures that begin diffusely in the brain. The areas of brain initially involved determine the symptoms of location-related seizures. A focal aware seizure corresponds to the previous term ‘simple partial seizure’. A focal impaired awareness seizure corresponds to the prior term ‘complex partial seizure’. The seizure type ‘focal to bilateral tonic-clonic’ is a special seizure type, corresponding to the 1981 phrase ‘partial onset with secondary generalization’.

Generalized seizures (formerly *grand mal*) begin in widespread areas of the brain. The most common type of generalized seizure is noted for muscle stiffening followed by jerking (tonic-clonic). Generalized seizures are divided into motor and non-motor (absence) seizures. Further subdivisions are similar to those of the 1981 classification. Absence seizures (formerly *petit mal*) consist of brief episodes of staring and lack of responsiveness. Myoclonic seizures involve brief jerks of muscles rather than repetitive clonic movements. Tonic seizures have generalized muscle stiffening. Atonic seizures involve sudden loss of muscle tone.

Table 8.10.1 The International League Against Epilepsy 2017 classification of epileptic seizures

I. Generalized onset seizures	
Motor	*Non-motor (absence)*
Tonic-clonic Clonic Tonic Myoclonic Myoclonic-tonic-clonic Myoclonic-atonic Atonic Epileptic spasms	• Typical • Atypical • Myoclonic • Eyelid myoclonia
II. Focal onset seizures	
Motor onset	*Nonmotor onset*
Aware Impaired awareness	• Aware • Impaired awareness
Automatisms Atonic Clonic Epileptic spasms Hyperkinetic Myoclonic Tonic	• Autonomic • Behaviour arrest • Cognitive • Emotional • Sensory
Focal to bilateral tonic-clonic	• Focal to bilateral tonic-clonic
III. Unknown onset seizures	
Motor	*Nonmotor*
Tonic-clonic Epileptic spasms	• Behaviour arrest
IV. Unclassified seizures	

Source: data from Fisher, R.S. Position Paper of the ILAE Commission for Classification and Terminology, *Epilepsia*, Volume 58, Issue 4, pp. 1522–30, Copyright © 2017.

Incidence and prevalence of epilepsy

Aspects of epilepsy epidemiology have been reviewed (Fiest et al. 2017). Epidemiological studies of epilepsy are challenged by the difficulty in making a correct diagnosis and the sophisticated technology and expertise required for proper diagnosis and classification. A substantial number of patients diagnosed with epilepsy do not have seizures on careful testing (Uldall et al. 2006). There is also difficulty with standardizing case ascertainment with markedly different prevalence rates using different definitions of epilepsy (Svendsen et al. 2007).

In a meta-analysis the overall incidence of epilepsy was 61.4/100,000, with 139 in low/middle-income countries and 48.9 in high-income countries (Fiest et al. 2017). Population-based and prospective studies had higher incidence rates than clinic-based and retrospective studies. Incidence is higher in children but both incidence and prevalence decrease with calendar time among children and increase among the elderly (Hauser et al. 1996; Hirtz et al. 2007). Others found an increased incidence in men, older people, black people, and those with pre-existing disability or comorbid conditions (Kaiboriboon et al. 2013). A study of children in the United Kingdom found a cumulative incidence of epilepsy by age 5 to be 1% in those born between 1994 and 1996 and 0.53% in those born between 2003 and 2005 (Meeraus et al. 2013). The annual incidence declined by 4%/year from 2001 to 2008. However, a longitudinal study from Rochester, Minnesota, United States, found a slightly higher estimate of epilepsy incidence over time (Zarrelli et al. 1999).

The prevalence of epilepsy is estimated to be 7.6/1000 (Fiest et al. 2017). However, prevalence varies in different regions. In an African population, a greater than twofold variability was found in prevalence across regions (Ngugi et al. 2013). Younger patients more commonly had seizures attributable to trauma and older patients to infectious causes. There were similar findings between regions of Taiwan (Chen et al. 2012). In general, poor countries have more than twice the burden of epilepsy compared to higher-income regions with much of the difference explained by preventable infections, perinatal care, and trauma (Newton and Garcia 2012). Differences in study methodology, survey methods, case definition,

the availability of sophisticated medical technology in a community, and cultural differences make epilepsy prevalence studies particularly challenging.

Mortality of epilepsy

People with epilepsy may experience two to three times the risk of death as their unaffected counterparts (Thurman et al. 2017). The standardized mortality ratio for people with recurrent seizures was approximately fourfold higher than expected. A longitudinal study found a threefold excess in all-cause mortality, and a sevenfold increase among those under age 20 (Shackleton et al. 1999). Some of the risk of death in epilepsy reflects underlying diseases like brain neoplasms and cerebrovascular disease. However, there is also an increased mortality compared to the non-epileptic population due to accidents (often drowning and burns) and suicide (Lhatoo and Sander 2005). Sudden unexplained death in people with epilepsy (SUDEP) is increasingly recognized as an important cause of death in epilepsy. This increases with the duration and severity of epilepsy (Devinsky 2011). It is more common in adults. The incidence of SUDEP varies from 0.09 to 2.65 in community samples, 1.2 to 5.9 in epilepsy centres, and 6.0 to 9.3 among patients referred for surgery for epilepsy. In a population-based cohort of children with epilepsy who were followed for 40 years, SUDEP occurred in 9%, accounting for 38% of all deaths.

Infectious causes of epilepsy

In developing countries, infections are a much more important cause of epilepsy than in the United States and Europe. Parasitic, bacterial, and viral infections combine with hereditary factors, perinatal damage, head trauma, and toxic exposures to contribute to high rates of epilepsy in these countries (Newton and Garcia 2012). An example of an important infectious risk factor is advanced human immunodeficiency virus (HIV) disease whereby HIV itself or associated opportunistic infections such as toxoplasmosis or tuberculomas may lead to epilepsy. Another important infectious cause is *Taenia solium* cysticercosis (pork tapeworm) which causes neurocysticercosis. This is a major cause of epilepsy worldwide (Del Brutto

2012). Other infectious causes of epilepsy are various febrile illnesses, particularly encephalitis, and malaria.

Genetics of epilepsy

A genetic aetiology is commonly based on family aggregation and twin studies. However, only a small proportion of patients have a known genetic mutation, but this is changing rapidly with advances in molecular technologies. Over 200 genes are now associated with epilepsy and include monogenic, complex, mitochondrial, chromosomal, and imprinting. Screening methods have been proposed for evaluating patients with epilepsy (Lemke et al. 2012). Epigenetic factors may also influence circuit excitability and neurogenesis in epilepsy (Roopra et al. 2012). Many types of epilepsy are now recognized as genetic disorders. However, these are rare forms of epilepsy and the genetic contribution to the majority of cases remains unexplained. First-degree relatives of people with epilepsy have a two- to fourfold increase in epilepsy risk (Annegers et al. 1982). A number of epilepsies have now been shown to be due to channelopathies, mutations in ion channels (Avanzini et al. 2007). It is expected that many more genes associated with epilepsy will be identified.

Costs and public health burden of epilepsy

Only 42% of those over age 18 with epilepsy are employed compared to 70% of those without epilepsy (Libby et al. 2012). People with epilepsy have higher work absences. Holding other variables constant, people with epilepsy had a loss of productivity of $9,504 compared to people without epilepsy. Indirect costs may account for 85% of the total, and the largest share of direct costs is attributable to patients with intractable epilepsy (Begley et al. 2000). Seizure frequency is inversely associated with health-related quality of life (Leidy et al. 1999). Seizure-free individuals report a quality of life similar to the general population; however, more seizures lead to a poorer quality of life, regardless of additional comorbidity, and irrespective of gender. Effective seizure control appears to be important in reducing costs as well as increasing patient quality of life.

Dementia

Clinical overview of dementia

Dementia presents with a progressive loss of a person's usual and customary cognitive function from any of several domains. The deficits must represent a decline from previous level of function and be severe enough to interfere with daily function and independence. This often begins with memory problems in Alzheimer's disease (AD) but could also begin with language deficits or disinhibition in frontotemporal lobar degeneration (FTLD), or with deficits in executive function in vascular cognitive impairment, or with sleep disorders or hallucinations in Lewy body disease. Regardless of the beginning domain, cognitive impairment associated with most dementia tends to progress and affect other cognitive domains with time. Dementia often includes behavioural changes, such as agitation, wandering, personality change, or depression, as well as sleep disturbances and psychiatric symptoms. As knowledge of specific subtypes of dementia has grown, the research community has come to rely more on specific clinical criteria for those subtypes. While the Diagnostic and Statistical Manual (DSM) (American Psychiatric Association 1994) criteria were originally constructed to provide for mutually exclusive diagnostic subgroups, it is now widely acknowledged that more than one underlying 'brain disease' may coexist and lead to the expression of dementia.

Dementia has many causes including neurodegenerative, vascular, infectious, and traumatic. Among the neurodegenerative dementias that predominate with ageing are AD, Lewy body disease (aka dementia with Lewy bodies (DLB) or PD dementia) and FTLD dementias. Cerebrovascular disease may coexist with and contribute to any of these or may be a primary cause on its own as vascular dementia (VaD). Prion disease, and associated spongiform encephalopathies (e.g. Creutzfeldt–Jakob disease (CJD) and variant CJD) are rare, but potentially transmissible from ingestion of animal tissue or from contact or transplantation of human tissue. Other disorders and insults which can result in dementia include drug-induced conditions, alcoholism, Huntington's disease, subdural hematoma, brain tumours, hydrocephalus, vitamin B_{12} deficiency, multiple medical conditions, hypothyroidism, and neurosyphilis.

Criteria for the clinical diagnosis of AD dementia (McKhann et al. 2011), mild cognitive impairment due to AD (Albert et al. 2011), and preclinical

AD (Sperling et al. 2011) emphasize the addition of biomarkers to characterize the underlying disease pathology. Amyloid imaging and CSF assays for amyloid-beta protein, tau, and phospho-tau proteins are of greatest importance in preclinical AD because no cognitive symptoms are present. Neuropathological criteria formerly were the clinical-pathological NIA-Reagan consensus criteria (Anonymous 1997) ranking the likelihood that pathological features (amyloid plaques and neurofibrillary tangles) existed in sufficient frequency to have caused the clinically observable dementia. New neuropathological criteria (Hyman et al. 2012) form a diagnosis based only on pathological features present, regardless of clinical dementia. The research community thus has taken a major step to define AD by pathology that is present (in both clinical and neuropathological criteria) and has recognized formally that asymptomatic AD occurs. This opens the door to conceptualization of secondary prevention.

Mild cognitive impairment

Clinically observable cognitive decline, not yet at the level meeting criteria for dementia, has been called mild cognitive impairment (MCI) (Petersen 2004). As a prodrome to dementia or simply an impaired state, it may have many underlying causes; MCI itself is not an aetiological diagnosis, but a characterization of affected cognition. The portion of MCI which is due to AD has now been provided with formal clinical criteria and biomarker indications (Albert et al. 2011). Recognition that the disease process in AD may have been underway for as many as 10–25 years and is potentially detectable by CSF biomarkers and/or amyloid imaging (Sperling et al. 2011; Bateman et al. 2012) changes the way epidemiologists should look at risk factor studies and the timing of exposures sufficient to cause disease onset. Pathological onset also explains in part why potential treatments given to mild or moderately impaired subjects would have little chance of success. What would a rehabilitative treatment for dementia due to AD potentially look like? Possibly in addition to slowing or stopping the formation of pathology, it would need to facilitate neuronal regeneration, branching, and connections, if recovery might happen.

Tracking and describing prognostic factors associated with cognitive decline has been the subject of many investigations. However, in the case of AD and probably other underlying causes, the timing and course of decline

may be more likely to be tied to the frequency and location of damage due to pathology than it might be to education, for example. Determining and describing potentially modifiable factors which might change the course of decline in those with symptomatic disease, is one of the major challenges of AD and other dementias. The original criteria for MCI (Petersen 2007) allowed MCI to represent a variety of aetiologies in addition to AD. On the non-amnestic side, it could easily include vascular cognitive impairment as well as cognitive impairment due to FTLD or Lewy body disease. However, the amnestic side of MCI is regarded by many to represent primarily early AD (Morris et al. 2001; Morris and Cummings 2005; Morris 2006). People diagnosed with MCI also are known to recover and not go on to develop dementia. The potential progression of cognitive decline is described well by Sperling et al. (2011) as a continuum, rather than reclassification into separate categories of dementia or MCI.

Dementia and Alzheimer's disease

Prevalence and incidence of dementia and AD

Almost two decades ago, Evans and colleagues reported prevalence estimates for dementia and AD based on a community study in East Boston, Massachusetts, United States (Evans et al. 1989), placing an heuristic upper bound on estimated prevalence of dementia in United States' communities: rising from 3% among those 65–74 years of age to 47% in those over age 85, with over 80% of the observed dementia classified as AD. Evans later projected that 10.3 million people would have AD in the year 2050. A recent meta-analysis of prevalence studies worldwide (Prince et al. 2013) shows remarkably similar age-specific prevalence and age-adjusted summary prevalence across all regions of the world, with the lowest prevalence being in sub-Saharan Africa and possibly the highest in Latin America and the Caribbean. The global estimate of people living with dementia in 2016 is reported to be 43.8 million by the Global Burden of Disease Study (Nichols et al. 2019). By 2050, it is estimated that 115.38 million people will be living with dementia. An estimated $420 billion was spent on dementia care in the United States and Western Europe in 2010; worldwide $604 billion were spent (Wimo et al. 2013). An incidence rate meta-analysis (Jorm and Jolley 1998) gathered data from 23 studies.

Incidence was estimated for Europe, the United States, and East Asia; dementia, AD, and VaD rates were computed. Incidence rates for clinically diagnosed dementia rose from 3.6 per 1,000 person-years (aged 65–69) to 37.7 per 1,000 person-years (aged 85–89) in Europe and from 2.4 to 27.5 per 1,000 person-years for the same age groups in the United States.

Similarly Rocca et al. (1998) re-analysed dementia and AD incidence data for 1975 through 1984 and observed that annual age-specific incidence rates have appeared to stay quite stable with time during that interval. Launer et al. (1999), combining cohorts enrolled in Denmark, France, the Netherlands, and the United Kingdom, determined that the incidence of dementia rose from 2.5 per 1,000 person-years at age 65–69 to 85.6 per 1,000 person-years in those aged 90 and older; while AD rose from 1.2 per 1,000 person-years to 63.5 per 1,000 person-years across the same age groups.

Suppose that the incidence of AD were estimated by the development of positive biomarkers including asymptomatic people. How would such a change shift the age distribution and the incidence rates?

While the tendency has been to attempt to count primary cause or pathologically ‘pure’ cases, dementia causing aetiologies (e.g. Alzheimer, vascular, Lewy body, FTLD) appear to commingle frequently in community dwelling cases. This commingling is consistent with the view that these diseases may exist as multiple amyloidoses resulting from aggregation of misfolded proteins (Morimoto 2006).

Distinguishing between normal people, those with MCI and those asymptomatic people with occult dementia-related pathology, is now possible to some degree as noted in Sperling et al. (2011) based on newly developed neuroimaging techniques (Klunk et al. 2003; Mathis et al. 2003; Small 2004) and CSF biomarkers. These ‘biomarkers’ may provide important clues about pathological onset of disease as well as inform the study of risk factors and critical periods of exposure prior to disease onset (Mintun et al. 2006; Bacskai et al. 2007; Fagan et al. 2007; Rowe et al. 2007).

Risk (and protective) factors for AD

The attempt to identify environmental risk factors for AD has been disappointing and conflicting, for methodological as well as biological

reasons. Because the pathological disease may begin from 10 to 25 years prior to any symptoms, the time window for exposures which could be relevant for onset is shifted even further back in time. Few observational studies were able to place any exposures except for genetic factors in this time range. Thus case–control studies were likely to be methodologically compromised by diagnostic misclassification, as well as ill-timed or missing exposure data. Cohort study designs fare only slightly better because the initial cohort would include asymptomatic people; for example, the base cohort would not effectively exclude prevalent disease, causing diagnostic misclassification. Thus, the risk or protective factors reported by epidemiological case–control or cohort studies to date must be viewed with considerable caution; the notable exception is the genetic risk factor, apolipoprotein E (APOE) e4 allele. APOE may exert independent effects or it may modify or confound the effect of environmental risk factors (Corder et al. 1993; Launer et al. 1999; Mehta et al. 1999).

Higher educational level has been proposed as decreasing risk of AD, but the relationship is complex (e.g. Hendrie 1998; Geerlings et al. 1999; Koepsell et al. 2007). The idea that higher education confers greater ‘cognitive reserve’ to be accessed when disease strikes is tantalizing, though biologically unsubstantiated. Koepsell et al. in a clinico-pathological study found that there was ‘no evidence of larger education-related differences in cognitive function when AD neuropathology was more advanced’, concluding further that ‘Higher Mini-Mental State Examination scores among more educated people with mild or no AD may reflect better test-taking skills or cognitive reserve, but these advantages may ultimately be overwhelmed by AD neuropathology’ (Koepsell et al. 2007). Quality of early life environment, as measured by number of siblings and area of residence, was associated with developing AD in late life, but educational level was not significantly associated with risk of AD (Moceri et al. 2000).

Anti-inflammatory medications have been studied for their potential protective effect on AD (Breitner 1996; In t’Veld et al. 2001; Szekely et al. 2007). The Alzheimer’s disease anti-inflammatory prevention trial (ADAPT) was launched in 2001. The trial enrolled 2,625 subjects, aged 70 years or older, and randomized them to placebo, celecoxib, or naproxen sodium treatment arms. The trial itself was stopped early because of an excess of cardiovascular events in the treatment arms. A preliminary analysis based on data available following trial closure showed no

significant effect of either celecoxib or naproxen as compared to placebo for the prevention of AD.

Oestrogen replacement therapy as a protective factor has shown a remarkably similar chequered research results history as anti-inflammatory medications, with the bulk of epidemiological evidence favouring protection and a fair amount of biological plausibility also marshalled in its support. Despite this large accumulation of epidemiological and laboratory evidence supporting the potential protective effect of oestrogen against the onset of AD, the Women's Health Initiative Memory Study (a randomized controlled trial of oestrogen and oestrogen plus progestin vs. placebo) showed a statistically significant *increased* risk of dementia among those given the active treatments (Shumaker et al. 2003, 2004). As a potential treatment for AD dementia, results also have shown no effect on AD (Henderson et al. 2000; Mulnard et al. 2000).

Factors related to cardiovascular disease and metabolic syndrome (including diabetes) may have an influence on AD and other dementias even though the exact mechanisms of interaction may not yet have been elucidated (Launer et al. 2000; Yaffe et al. 2004; Kivipelto and Solomon 2006; Laitinen et al. 2006; Li et al. 2007; Muller et al. 2007). The rationale for this approach appears to be that it is sound medical practice to treat hypertension, high cholesterol, diabetes, and obesity, and if more complete and effective treatment can be accomplished it may also, secondarily, serve to reduce the occurrence of dementia through vascular or other pathways which may or may not directly impact AD pathology. Cholinesterase inhibitors and memantine are currently the only available symptomatic medications for global functioning and cognition in these patients. Cholinesterase inhibitors target the acetylcholine deficit arising from the loss of neurons in the nucleus basalis of Meynert and its projections in patients with dementia. These may be considered symptomatic therapies and are not to be neuroprotective or to alter the underlying disease trajectory.

Genetics and AD

Great progress was made regarding the genetics of AD during the late 1980s and 1990s with the discovery of presenilin-1 and -2 and amyloid precursor protein mutations as familial, autosomal dominant causes and the

APOE e4 allele as a ‘risk factor’. Several reviews of AD genetics describing this period of growth in greater detail include: Hardy et al. (1998), Levy-Lahad et al. (1998), St George-Hyslop (2000), and Haass (2004).

The strongest and most consistent genetic risk factor for late-onset, non-autosomal dominant AD today is the APOE genotype. The association was first described from Dr Allen Roses’ laboratory (Corder et al. 1993; Saunders et al. 1993; Strittmatter et al. 1993). APOE naturally occurs as three different alleles (epsilon 2, epsilon 3, and epsilon 4) which pair to form one of six genotypes for each individual. Genotypes containing the epsilon 4 allele are associated with increased risk of AD; homozygous epsilon 4 greatly increases risk (e.g. greater than eightfold). Since the initial description of increased risk associated with the epsilon 4 allele, many investigators have observed the association. Discussion of APOE genotype is now included in nearly all genetic and risk factor studies of AD, either as a focus or as a potential confounder/effect modifier of an association. Despite the huge volume of studies including the APOE genotype, relatively little is known concerning how the e2, e3, and e4 alleles actually work to influence the risk of AD.

There have been a number of GWAS of AD, using large numbers of subjects (e.g. Hollingworth et al. 2011; Naj et al. 2011), which have identified and replicated a range of new common variants. Recently two important studies have shown an increased AD risk due to TREM2 (Guerreiro et al. 2013) and in African Americans an increased risk associated with ABCA7 (Reitz et al. 2013). Because these and other current GWAS have included strong replication they are regarded as superior to most candidate gene studies that occurred between the mid-1990s and 2010. More such studies have been performed; in addition, data sharing has allowed other investigators to examine associations between these identified loci and clinical characteristics (endophenotypes) such as cognitive decline.

Vascular dementia

VaD is difficult to describe clinically and neuropathologically. The search for ‘pure’ cases of VaD and AD may have obstructed progress of understanding how these diseases act and interact. Vascular features

commonly coexist with AD in the elderly: newer sophisticated imaging and other measures, as well as neuropathology, bear this out (Au et al. 2006). Several different criteria have been developed to diagnose VaD (Chui et al. 1992; Roman et al. 1993; American Psychiatric Association 1994). Much of the pioneering work in the clinical definition and recognition of VaD can be attributed to Hachinski and colleagues. The Hachinski Ischaemic Score continues to be used to indicate the extent of cerebrovascular involvement in cognitive decline (Hachinski 1994).

Pathologically, VaD is an enigma, because of the lack of clear definitions for white matter damage, infarcts, and lesion location in strategic areas of the brain—VaD, in effect, lacks a pathological gold standard against which to compare the clinical diagnosis or clinical criteria and measure their effectiveness (Murray et al. 2007). Thus, the clinical diagnostic criteria have been shown to be difficult and potentially unreliable in practice, even when applied by well-experienced research investigators (Chui et al. 2000).

Lewy body disease and frontotemporal lobar degeneration

Two additional types of dementia, Lewy body disease (McKeith 2006) and FTLD (Cairns et al. 2007a), have been separated from AD based on their clinical presentations but more specifically on their pathology. Lewy body disease or DLB may be diagnosable in up to 30% of dementia cases. PD is also a ‘Lewy body’ disease and if Parkinson’s cases survive long enough they will become demented as a result (Zaccai et al. 2005; Tsuang et al. 2006). Lewy body disease or DLB often presents with fluctuating cognitive performance, visual hallucinations, REM sleep behaviour disorder, and/or parkinsonism. Memory impairment may not be as prominent in the early stages of the disease as deficits in attention, frontal subcortical skills, and visuospatial ability (Knopman et al. 2001; Korczyn and Reichmann 2006).

FTLD has undergone a revolution in classification since initial work on progranulin (Gass et al. 2006; Rademakers et al. 2007) followed by the careful study of the involvement of TDP-43 (Neumann et al. 2006; Van Deerlin et al. 2010) and the discovery of association with C9ORF72 (DeJesus-Hernandez et al. 2011) includes specific disease subtypes and may be the most frequently occurring dementia in people under the age of 65 (Mackenzie and Rademakers 2007). Behavioural variant frontotemporal dementia (bvFTD) represents a behavioural subtype with changes such as

loss of personal awareness, loss of social graces, disinhibition, overactivity, restlessness, impulsivity, distractibility, hyperorality, withdrawal from social contact, apathy, or inertia, and stereotyped or perseverative behaviours. The memory loss is variable and often appears to be due to lack of concern or effort. Frontal lobe impairments are notable including abstraction, planning, and self-regulation of behaviour. FTLD may also be associated with parkinsonism or motor neuron disease. Pathologically, approximately 40% of FTLDs are related to mutations in the microtubule-associated protein tau (*MAPT*) gene on chromosome 17 and are therefore 'tauopathies'; these include Pick's disease, corticobasal degeneration, and progressive supranuclear palsy (Rademakers and Hutton 2007; Tolnay and Frank 2007). About 50% of FTLDs are not tauopathies but are reactive to ubiquitin (FTLD-U) rather than mutations in *MAPT* (Mackenzie and Rademakers 2007). Recently, it was discovered that mutations in the progranulin gene, also on chromosome 17, were an important cause of FTLD-U disorders (Baker et al. 2006; Cruts et al. 2006; Gass et al. 2006) including primary progressive aphasia (Mesulam et al. 2007). The pathological protein involved in these disorders as well as in motor neuron disease is TDP-43 (Mackenzie et al. 2006; Neumann et al. 2006; Cairns et al. 2007b; Kwong et al. 2007). These are arguably the most exciting and important findings for dementia and FTLD in recent times. FTLD is rare compared to AD and even so is thought to be frequently misdiagnosed as a psychiatric disorder because of its early onset and unusual clinical symptoms.

Peripheral neuropathy

Clinical overview of peripheral neuropathy

Though 'peripheral neuropathy' may refer to any disease of peripheral nerves, it usually describes systemic diseases that affect peripheral nerves rather than focal diseases affecting an isolated nerve. Most of these diseases affect longer nerves first with symptoms developing first in the feet and progressing up the legs. Few peripheral neuropathies affect shorter proximal nerves first. By the time symptoms reach the knees, the hands become symptomatic followed by the anterior trunk and crown of the head. Symptoms depend on the type of nerve fibre involved. Involvement of motor fibres leads to weakness, muscle wasting, and hyporeflexia. If

longstanding, motor neuropathies may lead to high arches (pes cavus) or hammer toes. Sensory nerve involvement leads to loss of sensation, distorted sensation (dysaesthesias), or spontaneous unpleasant sensations (paraesthesias). Autonomic neuropathies lead to postural hypotension, sexual dysfunction, bowel dysfunction, bladder dysfunction, and gastroparesis. The size of the affected nerve fibre can often be suggested by the history with large fibre disease causing reflex loss, vibration loss, and joint position loss. Small fibre disease often leads to autonomic dysfunction, dysaesthesias, loss of pain sensation, and loss of temperature sensation.

Electrodiagnostic testing is used to diagnose and classify peripheral neuropathies. Nerve conduction velocities can classify peripheral neuropathies into demyelinating and axonal types. Demyelinating neuropathies have disproportionate slowing and increased latencies of nerve conduction speeds. Axonal diseases cause disproportional loss of amplitude with relative preservation of conduction speed. Nerve conduction studies measure only the fast-conducting large diameter fibres. Electromyography (EMG) measures the electrical activity of muscle fibres. A loss of muscle fibre innervation by large myelinated neurons leads to increased insertional activity, positive waves, fibrillation potentials, polyphasic motor unit potentials, and decreased recruitment patterns on EMG. Occasionally, nerve or muscle biopsies are used for diagnosis of polyneuropathies.

Generally, polyneuropathies involve many peripheral nerves and result in autonomic neuropathies, sensory loss, or weakness. Mononeuropathies, as the name implies, involve a single nerve injury or entrapment, for example, carpal tunnel syndrome (CTS) or Bell’s palsy. Peripheral nerve disorders are often classified as hereditary or acquired. Charcot–Marie–Tooth syndrome is a well-known hereditary form. Acquired nerve disorders are associated with trauma/compression, diabetes, alcoholism, and other nutritional and metabolic problems. They may also be related to infectious causes such as Guillian–Barré syndrome, leprosy, Lyme disease, or HIV infection; or they may be caused by toxic exposures to metals (e.g. lead, mercury), industrial chemicals, or therapeutic drugs (e.g. antineoplastic agents) (Rowland and Merritt 1995). Little is known about the epidemiology of peripheral neuropathies, though they are common. The few studies that have been conducted find prevalence rates ranging from 2.4 to

8% (Martyn and Hughes 1997). Prevalence increases considerably with age to 54% of those aged 85 and older (Mold et al. 2004).

Carpal tunnel syndrome

CTS is probably the most common neuropathy. Carpal tunnel release surgery is one of the most common hand surgeries. Occupational CTS from repetitive, higher-impact actions differs from non-occupational CTS by occurring nearly equally among men and women and at a lower mean age (37 vs. 51 years) than non-occupational CTS (Franklin et al. 1991). Based on US workmen's compensation records, an incidence of 1.74/1,000 full-time equivalent jobs was observed.

In contrast to the occupational incidence, the general population CTS incidence was 3.46/1000 person-years (Nordstrom et al. 1998). The apparent increase in incidence may reflect a true change or may be affected by popular knowledge of the condition and diagnostic suspicion. Prevalence of symptoms in relation to true disease prevalence is also a consideration (Atroshi et al. 1999). Reported CTS symptoms of tingling, pain, and numbness have a prevalence of about 14%, whereas CTS was confirmed in less than 3%.

CTS patients with inflammatory arthritis, diabetes, and hypothyroidism were more likely to receive CTS release surgery (Solomon et al. 1999). Heredity is the single strongest predictor of CTS with half of the risk of CTS due to genetic factors (Hakim et al. 2002). Using MZ and DZ twins, the heritability was estimated to be 0.46 (95% CI 0.34–0.58). There is a bimodal age distribution with peaks at 50–54 and 75–84 years of age (Bland and Rudolfer 2003). In a case–control study, elevated body mass index was more common among cases than among control subjects (Stallings et al. 1997).

Diabetes mellitus

Diabetes is a common, yet complex cause of both mono- and polyneuropathies. Diabetic peripheral neuropathy may affect 26–47% of those with diabetes (Barrett et al. 2007) though others have found lower rates of 7.8–17.4% (Dyck et al. 2012). More effective glucose control could reduce the risk to some extent. The cost of diabetic peripheral neuropathy,

particularly in those with pain, can be substantial (Currie et al. 1998; Barrett et al. 2007).

Microvascular disease, chronic hyperglycaemia, and type of diabetes are the most important predictors of polyneuropathy (Dyck et al. 1999). Autonomic neuropathy is strongly influenced by hyperglycaemia and is associated with increased mortality (Orchard et al. 1996). Elevated diastolic blood pressure, ketoacidosis, elevated fasting triglyceride level, and microalbuminuria are also risk factors for diabetic polyneuropathy (Tesfaye et al. 1996).

Nutritional neuropathies

Vitamin B_{12} deficiency is the most common of a number of nutritional deficiencies that can lead to peripheral neuropathy. A Cuban epidemic of peripheral neuropathy from 1992 to 1993 affected over 50,000 people with an incidence of 461 per 100,000 people caused by a lack of availability of meat and dairy products (Roman 1994). Optic and peripheral forms of the disease were observed. A nutritional deficiency is suspected and treatment with multivitamins, in particular B vitamins, stopped the outbreak.

Peripheral neuropathy due to infection

HIV infection is an important cause of peripheral neuropathies including diffuse peripheral neuropathy, Guillain–Barré syndrome, mononeuropathies, neuropathies due to secondary infections, and neuropathies due to medications (Brew 2003). The incidence and prevalence vary too much between studies to allow accurate estimates.

Guillain–Barré syndrome is due to an autoimmune attack on the peripheral nerve myelin. Previously considered an idiopathic disease, it is now recognized that a significant proportion of cases are associated with prior infections including *Campylobacter jejuni* and cytomegalovirus (Hardy et al. 2011). Though the prevalence of leprosy is declining, it remains one of the most important causes of peripheral neuropathy in the developing world (Suzuki et al. 2012). Fortunately, early treatment can prevent many of the disabling and disfiguring effects of the disease.

Parkinsonism and Parkinson's disease

The Parkinson's research community is somewhat fragmented with regard to how idiopathic PD should be described and characterized: a) should full clinical symptomatology (e.g. cardinal signs) be required; b) should evidence of PD pathophysiology (e.g. Lewy body pathology) be sufficient to say PD exists; c) should PD be viewed as a systemic neurodegenerative disorder in contrast to focus on movement disorder (Langston 2006; Litvan et al. 2007)? The argument for expanding the view of PD continues that limiting the clinical and pathological view of the disease itself limits the potential for research progress (Langston 2006).

Clinical overview of PD

There are four cardinal features of clinical PD: tremor, rigidity, bradykinesia, and in the later stages postural gait changes. The clinical diagnosis of PD usually requires two or more of these symptoms. In severe cases, patients may be unable to move (freezing) when they encounter minor obstacles such as doorways or cracks. While movement disorder specialists make the diagnosis of parkinsonism with some degree of confidence, PD usually requires histopathological confirmation. In an attempt to increase the accuracy and validity of clinical diagnosis, improvements in clinical diagnostic criteria have been proposed (Gelb et al. 1999; Jankovic et al. 2000).

Parkinsonism includes several major subclasses: idiopathic parkinsonism (PD), symptomatic parkinsonism (e.g. drug induced, toxin induced, and other specific causes), 'parkinson-plus' syndromes (e.g. multiple system atrophy (MSA), progressive supranuclear palsy), and hereditary degenerative diseases (e.g. Hallervorden–Spatz disease, Huntington disease). PD or idiopathic parkinsonism comprises approximately 80% of parkinsonism.

MSA is sometimes misdiagnosed as PD; it is a relatively rare and very debilitating condition usually involving progressive autonomic failure plus poor responsiveness to levodopa or cerebellar ataxia. There is new evidence that MSA like PD and DLB is a synucleinopathy (Goedert and Spillantini 1998; Dickson et al. 1999; Armstrong et al. 2006).

Incidence and prevalence of PD

The frequency of PD varies according to the diagnostic criteria, study population, and epidemiologic methods used. The prevalence of PD in industrialized countries is generally estimated at 0.3% of the entire population and about 1% in people over 60 years of age (Nussbaum and Ellis 2003). PD prevalence has been reported with dramatic inconsistency; typically prevalence has been reported in the range of about 50 per 100,000 to 200 per 100,000 of the population, with a maximum of about 350 per 100,000 people. Combined results of five European studies (de Rijk et al. 1997) included 14,636 people age 65 or older; after age adjusting to the European 1991 standard population the prevalence of PD was reported as 1.6 per 100 population (presumably age 65 or older), this translates to about 1,600 per 100,000. In addition, the age-specific prevalence of PD was reported to increase per 100,000 population as follows: 41 in 40 to 49 years; 107 in 50 to 59 years; 173 in 55 to 64 years; 428 in 60 to 69 years; 425 in 65 to 74 years; 1,087 in 70 to 79 years; and 1903 in older than age 80 (Pringsheim et al. 2014).

Incidence rates for PD and parkinsonism carry many of the same caveats as for prevalence. A review estimates standardized incidence rates of PD are 8–18 per 100,000 person-year (De Lau and Breteler 2006). Onset of PD is rarely before age 50 years and a sharp increase of the incidence is observed after age 60 years. Age-specific incidence rates are, of course, able to be compared directly whereas summary population rates may be dependent on age structure of the population. How ageing contributes to the clinical expression or degenerative processes of PD, as well as how ageing may act with genetic and environmental risk factors, to initiate pathophysiology or onset of clinical symptoms remain key questions for the epidemiology of parkinsonism and PD.

Risk factors of PD

Researcher hopes to identify consistent environmental causes of PD were increased by the observation that 1-methyl-4-phenyl-1, 2,3,6-tetrahydropyridine (MPTP) caused acute PD shortly after ingestion (Langston and Ballard 1983; Langston et al. 1983). The structure of MPTP and its metabolism products are somewhat similar to common pesticides

raising interest in such exposures as risk factors or causes of PD. The past predominance of environmental causes has at least partially given way to studies of genetic and pathophysiological aetiologies (Litvan et al. 2007; Do et al. 2011).

An excellent review of pesticide/herbicide exposure as risk factors for PD was recently published by Tanner et al. (2011). Determination of critical exposure time periods and duration of exposures is always an obstacle to overcome when studying PD risk factors because the initiation of pathophysiology may occur many years prior to any symptoms of the diagnosis of PD. Further, the time from exposure to onset of pathophysiology or clinical disease may vary dramatically. Consider the rather immediate onset of clinical symptoms following MPTP exposure compared with what may be the case with agricultural pesticides (Gorell et al. 1998; Petrovitch et al. 2002; Firestone et al. 2005). Interest in pesticide exposure as a potential cause led to a biotransformation gene approach for evaluating the occurrence of susceptible people. Specifically, some people may be more, or less, able to metabolize environmental toxins because of polymorphic genes involved in metabolism. One of the family of cytochrome P-450 biotransformation genes, *CYP2D6*, is involved in metabolism of debrisoquine (structurally similar to pesticides); some polymorphic forms are 'poor' metabolizers and others are normal or rapid metabolizers of debrisoquine. Initial studies appeared to show that poor metabolizers were at increased risk of PD but later studies and meta-analyses fail to support this conclusion (Maraganore et al. 2000; Scordo et al. 2006). A similar approach has been taken to identify susceptibles focusing on polymorphic forms of glutathione transferases (GST) which are involved in the metabolism of pesticides and other xenobiotics (Menegon et al. 1998). Continued effort to identify gene–environment interaction in this way may eventually prove fruitful, but success is rather limited to date.

An association between exposure to metals and PD has been described by Gorell et al. (1999). And more recently an association with welding has been reported by Racette et al. (2012). Heavy metals such as iron, copper, manganese, and lead have been raised as potential risk factors in the past and Racette et al. may provide some measure of replication to those findings.

Smoking has been rather consistently associated with decreased risk of PD in reported reviews and in individual studies. Although the negative

association with smoking is one of the more consistent findings, it is not completely without alternative explanations. Reasons for the plausibility of such an association revolve around the neuroprotective potential of nicotine. Another possible explanation is that patients who develop PD are more likely to quit smoking or less likely to smoke in the first place than those who don't develop the disease. This is explained as dopamine is an important component of the brain's reward system, and those who will later develop signs of PD do not engage in reward-seeking behaviours, including smoking, because dopamine is significantly reduced in the basal ganglia years before symptoms of PD appear. Most epidemiological studies of PD use 'prevalent' or existing cases in their studies. The low incidence of PD effectively precludes concentrating on only newly diagnosed cases in all but the largest of studies (or in very large cohort studies). When attempting to identify a cross-sectional sample of cases for enrolment into a case–control study, it can be shown that those patients who have had the disease the longest are the most likely to be included. The most severe, short-duration, or rapidly declining cases tend to be missed. If PD cases who had a history of smoking were much more likely to die sooner than 'controls' who were smokers, then smoking would appear to be protective because of the excess of smoking 'controls'. Despite such an argument, however, the PD research community appears to be relatively confident in the potential protective effect of smoking (Ritz et al. 2007).

Dietary-related factors have been shown in some studies to reduce risk of PD: fish and vegetables (Gao et al. 2007), coffee and caffeine (Ross et al. 2000), dietary folate, and vitamin B_{12} among others (de Lau et al. 2006). Body mass index has been reported to increase the risk of PD (Hu et al. 2006), while gout and hyperuricaemia has been reported to protect against it (de Lau et al. 2005; Alonso et al. 2007).

As recently as two decades ago it was commonly taught that PD was caused entirely by environmental factors. Now with the explosion of genetic knowledge and technology and an understanding of some of the proteins potentially involved in pathogenesis, Hardy states that nearly half of the disease risk has been explained by genes so far discovered (Hardy 2010). Do et al. (2011) discuss the recent GWAS and genes involved in PD risk. Glucocerebrocidase has been found to be deficient in PD brains and similarly the *GBA* gene is associated with Lewy body disease causing PD and dementia (Tsuang et al. 2012). While the list of genes involved in PD

continues to grow, the acknowledged recessive gene causes appear to be *parkin* (*PARK2*), *PINK1* (*PARK6*), and *DJ-1* (*PARK7*). Autosomal dominant and genetic risk factors include *SNCA*, *UCHL1*, *LRRK2*, *GBA*, and *MAPT*. As discussed by Do and others, the list continues to grow as do the parallels with other neurodegenerative diseases such as AD and FTLD.

Public health impact of PD

PD is progressive and debilitating. While initial treatments with levodopa and similar medications effectively quell most motor symptoms, their effectiveness begins to subside in about 50% of patients after 3–5 years. To date, there have been no true disease-modifying treatments available and we are only beginning to examine genetic pathways leading to pathogenesis or to drug targets that may modify the disease process prior to onset of symptoms (Siderowf et al. 2012). This aspiration is still a long way off, it appears. With increasing motor problems comes increased healthcare cost and decreased quality of life (de Boer et al. 1999); dementia is now regarded as an almost inevitable consequence of PD, assuming the patient survives long enough, leading to the need for long-term care in many instances. The overlap between PD dementia and DLB is of great interest because of the potential underlying synuclein pathology and therefore the potential genetic relationships as well (Rocca et al. 2007; Tsuang et al. 2012).

Multiple sclerosis

Clinical overview of MS

Although MS is relatively uncommon, it has a great impact on society. The impact of this disease is disproportionately large because it strikes young adults, is expensive to treat, and can cause disability. Only 21% of MS patients have no work limitations and 29% remain in the workforce (Minden et al. 2004). MS has substantial medical costs (Minden et al. 2004; Adelman et al. 2013). Because of lost earnings and increased healthcare costs, MS is the third leading cause of significant disability in the 20–50-year age range (LaRocca et al. 1984).

Clinically, MS is characterized by demyelinating lesions of central nervous system white matter causing motor, sensory, cerebellar, visual, brainstem, autonomic, spinal cord pathways, and other symptoms. These often occur in attacks with complete or partial recovery, with symptoms remaining stable between exacerbations (relapsing/remitting disease) (Lublin and Reingold 1996). Alternatively, symptoms may slowly progress in the absence of exacerbations (primary progressive disease). Relapsing/remitting cases may evolve into secondary progressive disease with a slow worsening of baseline function. When the disease results in death, the immediate cause is usually infectious, secondary to urinary tract involvement or pneumonia.

The length of exacerbations can be shortened by corticosteroids. There are currently nine US Food and Drug Administration-approved medications to slow the disease, with others under review. In addition to disease-modifying therapies, symptomatic treatments are often required. A multitude of new treatments are being tested, giving hope for more effective treatments in the future.

Definition of MS

The current diagnostic criteria are the revised McDonald criteria (Polman et al. 2011). These require two or more episodes of neurological deficit at different times and different locations within the nervous system. MS attacks used to fulfil the criteria must have symptoms typical of those seen with MS and have objective findings on examination or paraclinical tests. The new criteria allow diagnosis by clinical presentation alone, but also allow magnetic resonance imaging (MRI) findings to be used to demonstrate dissemination in time or location. The revised McDonald criteria allow the disease to be diagnosed at an earlier stage than previous criteria.

Case ascertainment of MS

Because of the necessity for neurological expertise and special studies to make reliable diagnoses, reported worldwide prevalence may not be completely comparable, especially where differences in the availability and quality of healthcare exist. This is especially true in light of the highly

variable clinical presentation of the disease, and the variable course. The requirement for repeated attacks before a diagnosis is made and the often vague nature of the initial clinical symptoms leads to difficulties in determining exact incidence figures in a timely manner.

Prevalence of MS

MS prevalence is easier to determine than incidence, particularly considering the difficulty in determining the time of disease onset. With the new McDonald criteria, MRI availability, and increasing awareness of the disease, the time from symptom onset until diagnosis is rapidly shortening (Marrie et al. 2005). The reported prevalence of MS varies widely with latitude, from one per 100,000 or less near the equator, to over 150 per 100,000 in some high-latitude areas (Koch-Henriksen and Sorensen 2011). Migrants carry the risk of their geographic residence before their mid-teens (Alter et al. 1966a, 1966b). The universal association between latitude and risk of MS has been challenged by findings from a 2010 systematic review and meta-analysis of epidemiologic studies of MS (Koch-Henriksen and Sorensen 2011). The results showed that, while the prevalence of MS increased with geographic latitude in Western Europe, North America, and Australia/New Zealand, the incidence of MS increased with latitude only in Australia/New Zealand, and not in Western Europe or North America. Hence, there was no latitudinal gradient for MS incidence in the northern hemisphere. In the absence of association with incidence, the observed latitudinal gradient of MS prevalence could be explained by other factors, such as survival time, diagnostic accuracy, and ascertainment probability.

Risk factors of MS

Gender

Approximately 75% of people with MS are female. However, in the first half of the twentieth century, males predominated (Kurtzke 2005). Gender rates changed to a female preponderance about mid-century and the female/male ratio continues to increase by about 0.85%/year (Orton et al. 2006; Wallin et al. 2012). The disease is less active during pregnancy and more active during the 3 months immediately postpartum (Confavreux et al. 1998; Langer-Gould and Beaber 2013). The changes in MS activity with

hormones and pregnancy may be due to alterations in immune system function, but the temporal changes in gender ratios remain unexplained.

Ethnic background of MS

In general, MS occurs with greater frequency in white people, particularly those of northern European ancestry. Other ethnic groups have a lower incidence and prevalence of MS, but this appears to be changing. The relative risk of MS in black compared with white male veterans increased from 0.44 in the Second World War/Korean conflict to 0.67 in Vietnam to 1.16 in the Gulf War (Wallin et al. 2012). During the same time periods the relative risk among black females compared to white males increased from 1.28 to 2.86 to 3.62, catching up to white females with 1.79, 2.99, and 3.54 respectively.

Genetic susceptibility of MS

In Caucasians, some alleles of the HLA complex are associated with MS susceptibility, particularly HLA-DRB1.1501 (Haines et al. 1996). In addition, other groups of genes that influence the immune response or myelin structure have been investigated (Sadovnick et al. 1991). To date, large GWAS have identified over 100 polymorphisms associated with MS, but associations are modest, explaining less than 25% of the heritability of the disease (Hafler et al. 2007; Sawcer et al. 2011; Lin et al. 2012). Nevertheless, these studies confirm the importance of genes related to immune function.

In the Danish twin study, there was concordance for MS in 24% of MZ twins and 3% of dizygotic twins (Hansen et al. 2005). The standardized incidence ratio for MZ twins compared to non-twins was 1.23 and 0.78 for DZ twins. There is also an increased risk of MS in other first-degree relatives of people with MS. In a Danish study, the relative risk of MS in first-degree relatives of people with MS was 7.1 (95% CI 5.8–8.8) (Nielsen et al. 2005). The lifetime risk was calculated to be 2.8% for male and 2.9% for female first-degree relatives. Some have found that the risk of MS is higher in children of MS fathers than MS mothers (odds ratio 1.99, 95% CI 1.05–3.77) (Kantarci et al. 2006). This increased rate of transmission to offspring by a subgroup in which the disease is less common (men) suggests a genetic contribution in which genes are concentrated in men with the disease. However, others have reported that the risk of MS in children whose parents have MS is similar regardless of whether the affected parent

was the father or mother (Herrera et al. 2007). A Canadian study found that the age-adjusted risk of MS in siblings was 3.11% (95% CI 2.39–3.83) (Ebers et al. 2004). This group also found an excess of mother–daughter pairs compared to father–son pairs and suggested that epigenetic factors may contribute (Chao et al. 2010). These studies indicate that there is an important genetic contribution to the disease, but environmental factors also play a major role.

Environmental factors of MS

In the presence of inherent susceptibility, some external factors seem to be associated with MS. People with MS have a later age of exposure to common childhood exanthematous diseases, and lower birth orders though not all studies have supported this finding (Bager et al. 2006). There have been reports of clusters of disease, thought to have been related to environmental exposures, but on investigation these supposed clusters have generally not been beyond expected variability.

The strong relationship between MS and latitude has led some to postulate that areas with less sun exposure, and thus lower levels of vitamin D, may be responsible for the disease (Simon et al. 2012). Military recruits with low 25-hydroxyvitamin D (25(OH) D) levels had a subsequent increased risk of MS. Low vitamin D levels are associated with risk of MS in children. The risk of MS is lower in offspring of mothers with increased vitamin D intake. Patients with MS generally have low 25(OH)D levels. A small trial of 25(OH)D supplementation in MS patients found no difference in MRI activity (Stein et al. 2011). Other randomized treatment trials of 25(OH)D are underway.

The relationship between MS prevalence, latitude, migration, and socioeconomic factors has led some to suggest that MS is related to the degree and timing of exposure to various infectious organisms (Fleming and Cook 2006). This hygiene hypothesis states that exposure to infectious agents later in life (due to high levels of hygiene) leads to changes in the immune system that eventually lead to MS. Epstein–Barr virus (EBV) is associated with MS (Ascherio and Munger 2007; Owens and Bennett 2012). MS is extremely rare in people who have never had EBV. Furthermore, those exposed to EBV later in childhood have a two- to threefold increase in risk relative to those who acquire EBV in early childhood. Antibody titres to EBV are increased in MS, often years before

recognition of the disease, but this may reflect a general upregulation of the immune system. Cytomegalovirus appears to be protective in children (Waubant et al. 2011).

Both ever-smokers and current smokers have an increased risk of MS (Jafari and Hintzen 2011). Children exposed to cigarette smoke have a higher risk of developing MS (Mikaeloff et al. 2007). Smoking is associated with increased disability, more MRI activity, and conversion to secondary progressive MS.

Overview of MS

The risk of MS appears to be due to environmental factors acting on genetically predisposed people. Eventually, these factors lead to an immune attack on the central nervous system, resulting in symptoms. The genetic factors are polygenetic. The environmental factors remain uncertain. Immunomodulating treatments have played an important role in slowing the disease, but their benefits are only partial. Much remains to be clarified about the mechanisms of the disease process, and it is hoped that a better knowledge of these mechanisms will lead to improved treatments.

Conclusion

Presenting current and useful research information on a number of neurological conditions is a difficult task. This chapter has attempted to address that challenge for some selected neurological conditions and referencing heavily to current or classic research papers. Conditions such as headache have substantial public health impact because of the age groups affected, their prevalence, and the lost productivity (or economic loss) related to them. MS, a relatively common neurological disease, can affect individuals in young adulthood, decrease their productivity, and ultimately make them dependent on others. TBI occurring in youth or young adulthood can cause years of extra medical care in addition to lost productivity among those who survive the immediate event. Epilepsy may have onset throughout the life course, it may result from trauma or may be caused by specific genes, among other causes. While there are intractable forms of epilepsy, great strides have been made in seizure control enabling

patients to lead relatively full and normal lives. Neurodegenerative diseases, such as PD and AD, rob productivity, functional ability, and independence from older individuals; they also force huge increases in healthcare costs. Without question neurological diseases have substantial public health effects.

Determining the incidence and prevalence for most of the diseases and conditions in this chapter is quite an inexact science. The conditions are often difficult to define and detect in the population and for the most part they are not regarded as ‘reportable’ conditions. Therefore, we gain insight as to disease occurrence primarily from limited but (hopefully) well-designed and conducted studies. As mentioned in the introduction to this chapter, the epidemiological study of neurological conditions is a complicated matter. Problems with diagnostic inaccuracy and insidious disease onset influence our ability to observe risk factor associations; factors related to survival may be mistaken for risk/protective factors.

The recent work of the Human Genome Project and the HapMap Project have greatly influenced technology and have now made possible genome-wide studies that could not have been imagined a decade ago. The contribution of genes that in and of themselves cause disease may be smaller than that of genes which act together with other genes in complex ways or act to metabolize or potentiate environmental exposures. The epigenetic interaction between gene and environment will be increasingly well studied in the future. Description of gene products and functions may lead to specific drug therapies never before possible. The genetic information presented in this chapter, while relatively current, may become obsolete quickly. The fields of genetics and molecular biology are moving rapidly. It is a challenge also for epidemiologists to apply the knowledge gained by the genetic researchers to the design and analysis of epidemiological studies. The diagnosis of neurological conditions may be made more accurately and earlier with genetic information. Science and the public health will benefit beyond even our current expectations.

Epidemiology must take advantage of these molecular advances. Many scholars have written on pros and cons of reductionism in science. Much of epidemiology lies in its public health context, and the same is likely to be true for genetic influences on neurological diseases. Arrays of genes may identify susceptible individuals; however, those individuals may avoid disease unless met with specific environmental or behavioural exposures.

The tasks of public health and epidemiology will still involve prevention, the non-random occurrence of disease, and its environmental context—in addition to heredity. The tools to address those tasks will continue to be refined.

REFERENCES

Adelman, G., Rane, S.G., and Villa, K.F. (2013). The cost burden of multiple sclerosis in the United States: a systematic review of the literature. *Journal of Medical Economics*, **16**, 639–47.

Albert, M.S., et al. (2011). The diagnosis of mild cognitive impairment due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association work-groups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & Dementia*, **7**, 270–9.

Alhazzani, A. and Goddeau, R.P. (2013). Migraine and stroke: a continuum of association in adults. *Headache*, 53(6), 1023–7.

Alonso, A., Rodriguez, L.A., Logroscino, G., and Hernan, M.A. (2007). Gout and risk of Parkinson disease: a prospective study. *Neurology*, **69**, 1696–700.

Alter, M., Leibowitz, U., and Halpern, L. (1966a). Multiple sclerosis in European & Afro-Asian populations of Israel. A clinical appraisal. *Acta Neurologica Scandinavica*, 42(Suppl. 19), 47–54.

Alter, M., Leibowitz, U., and Speer, J. (1966b). Risk of multiple sclerosis related to age at immigration to Israel. *Archives of Neurology*, **15**, 234–37.

American Psychiatric Association (1994). *Diagnostic and Statistical Manual of Mental Disorders: DSM-IV*. Washington, DC: American Psychiatric Association.

Annegers, J.F., Hauser, W.A., Anderson, V.E., and Kurland, L.T. (1982). The risks of seizure disorders among relatives of patients with childhood onset epilepsy. *Neurology*, **32**, 174–9.

Annegers, J.F., Hauser, W.A., Coan, S.P., and Rocca, W.A. (1998). A population-based study of seizures after traumatic brain injuries. *The New England Journal of Medicine*, **338**, 20–4.

Anonymous (1997). Consensus recommendations for the postmortem diagnosis of Alzheimer's disease. The National Institute on Aging, and Reagan Institute Working Group on Diagnostic Criteria for the Neuropathological Assessment of Alzheimer's Disease. *Neurobiology of Aging*, 18, S1–2.

Armstrong, R.A., Cairns, N.J., and Lantos, P.L. (2006). Multiple system atrophy (MSA): topographic distribution of the alpha-synuclein-associated pathological changes. *Parkinsonism & Related Disorders*, **12**, 356–62.

Ascherio, A. and Munger, K.L. (2007). Environmental risk factors for multiple sclerosis. Part I: the role of infection. *Annals of Neurology*, **61**, 288–99.

Atroshi, I., Gummesson, C., Johnsson, R., Ornstein, E., Ranstam, J., and Rosen, I. (1999). Prevalence of carpal tunnel syndrome in a general population. *Journal of the American Medical Association*, **282**, 153–8.

Au, R., et al. (2006). Association of white matter hyperintensity volume with decreased cognitive functioning: the Framingham Heart Study. *Archives of Neurology*, **63**, 246–50.

Avanzini, G., Franceschetti, S., and Mantegazza, M. (2007). Epileptogenic channelopathies: experimental models of human pathologies. *Epilepsia*, 48(Suppl. 2), 51–64.

Bacskai, B.J., et al. (2007). Molecular imaging with Pittsburgh Compound B confirmed at autopsy: a case report. *Archives of Neurology*, **64**, 431–4.

Bager, P., et al. (2006). Sibship characteristics and risk of multiple sclerosis: a nationwide cohort study in Denmark. *American Journal of Epidemiology*, **163**, 1112–17.

Baker, M., et al. (2006). Mutations in progranulin cause tau-negative frontotemporal dementia linked to chromosome 17. *Nature*, **442**, 916–19.
Banks, M.E. (2007). Overlooked but critical: traumatic brain injury as a consequence of interpersonal violence. *Trauma, Violence & Abuse*, **8**, 290–8.
Barrett, A.M., Lucero, M.A., Le, T., Robinson, R.L., Dworkin, R.H., and Chappell, A.S. (2007). Epidemiology, public health burden, and treatment of diabetic peripheral neuropathic pain: a review. *Pain Medicine*, 8(Suppl. 2), S50–62.
Bateman, R.J., et al. (2012). Clinical and bio-marker changes in dominantly inherited Alzheimer's disease. *The New England Journal of Medicine*, **367**, 795–804.
Baugh, C.M., et al. (2012). Chronic traumatic encephalopathy: neurodegeneration following repetitive concussive and subconcussive brain trauma. *Brain Imaging and Behavior*, **6**, 244–54.
Bazarian, J.J., McClung, J., Shah, M.N., Cheng, Y.T., Flesher, W., and Kraus, J. (2005). Mild traumatic brain injury in the United States, 1998–2000. *Brain Injury*, **19**, 85–91.
Begley, C.E., et al. (2000). The cost of epilepsy in the United States: an estimate from population-based clinical and survey data. *Epilepsia*, **41**, 342–51.
Bland, J.D. and Rudolfer, S.M. (2003). Clinical surveillance of carpal tunnel syndrome in two areas of the United Kingdom, 1991–2001. *Journal of Neurology, Neurosurgery, & Psychiatry*, **74**, 1674–9.
Boardman, H.F., Thomas, E., Millson, D.S., and Croft, P.R. (2006). The natural history of headache: predictors of onset and recovery. *Cephalalgia*, **26**, 1080–8.
Breitner, J.C. (1996). Inflammatory processes and anti-inflammatory drugs in Alzheimer's disease: a current appraisal. *Neurobiology of Aging*, **17**, 789–94.
Brennan, K.C., et al. (2013). Casein kinase iδ mutations in familial migraine and advanced sleep phase. *Science Translational Medicine*, **5**, 56.
Brew, B.J. (2003). The peripheral nerve complications of human immunodeficiency virus (HIV) infection. *Muscle & Nerve*, **28**, 542–52.
Cady, R.K. (1999). Focus on primary care female population with migraine. *Obstetrical & Gynecological Survey*, 54, S7–13.
Cairns, N.J., et al. (2007a). Neuropathologic diagnostic and nosologic criteria for frontotemporal lobar degeneration: consensus of the Consortium for Frontotemporal Lobar Degeneration. *Acta Neuropathologica*, **114**, 5–22.
Cairns, N.J., et al. (2007b). TDP-43 in familial and sporadic frontotemporal lobar degeneration with ubiquitin inclusions. *American Journal of Pathology*, **171**, 227–40.
Cassidy, J.D., et al. (2004). Incidence, risk factors and prevention of mild traumatic brain injury: results of the WHO Collaborating Centre Task Force on Mild Traumatic Brain Injury. *Journal of Rehabilitation Medicine*, **43**, 28–60.
Centers for Disease Control and Prevention (CDC) (2003). Nonfatal fall-related traumatic brain injury among older adults—California, 1996–1999. *Morbidity and Mortality Weekly Report*, **52**, 276–8.
Centers for Disease Control and Prevention (CDC) (2007). Nonfatal traumatic brain injuries from sports and recreation activities—United States, 2001–2005. *Morbidity and Mortality Weekly Report*, **56**, 733–7.
Chao, M.J., et al. (2010). Parent-of-origin effects at the major histocompatibility complex in multiple sclerosis. *Human Molecular Genetics*, **19**, 3679–89.
Chen, C.C., Chen, L.S., Yen, M.F., Chen, H.H., and Liou, H.H. (2012). Geographic variation in the age- and gender-specific prevalence and incidence of epilepsy: analysis of Taiwanese National Health Insurance-based data. *Epilepsia*, **53**, 283–90.
Chesnut, R.M., et al. (2012). A trial of intracranial-pressure monitoring in traumatic brain injury. *The New England Journal of Medicine*, **367**, 2471–81.

Chui, H.C., et al. (2000). Clinical criteria for the diagnosis of vascular dementia: a multicenter study of comparability and interrater reliability [see comments]. *Archives of Neurology*, **57**, 191–6.

Chui, H.C., Victoroff, J.I., Margolin, D., Jagust, W., Shankle, R., and Katzman, R. (1992). Criteria for the diagnosis of ischemic vascular dementia proposed by the State of California Alzheimer's Disease Diagnostic and Treatment Centers [see comments]. *Neurology*, **42**, 473–80.

Confavreux, C., Hutchinson, M., Hours, M.M., Cortinovis-Tourniaire, P., and Moreau, T. (1998). Rate of pregnancy-related relapse in multiple sclerosis. Pregnancy in Multiple Sclerosis Group. *The New England Journal of Medicine*, **339**, 285–91.

Corder, E.H., et al. (1993). Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families [see comments]. *Science*, **261**, 921–3.

Cruts, M., et al. (2006). Null mutations in progranulin cause ubiquitin-positive frontotemporal dementia linked to chromosome 17q21. *Nature*, **442**, 920–4.

Currie, C.J., Morgan, C.L., and Peters, J.R. (1998). The epidemiology and cost of inpatient care for peripheral vascular disease, infection, neuropathy, and ulceration in diabetes. *Diabetes Care*, **21**, 42–8.

Curtin, S.C., Warner, M., and Hedegaard, H. (2016). Increase in suicide in the United States, 1999–2014. *NCHS Data Brief*, **241**, 1–8.

De Boer, A.G., Sprangers, M.A., Speelman, H.D., and De Haes, H.C. (1999). Predictors of health care use in patients with Parkinson's disease: a longitudinal study. *Movement Disorders*, **14**, 772–9.

Dejesus-Hernandez, M., et al. (2011). Expanded GGGGCC hexanucleotide repeat in noncoding region of C9ORF72 causes chromosome 9p-linked FTD and ALS. *Neuron*, **72**, 245–56.

De Lau, L.M., Koudstaal, P.J., Hofman, A., and Breteler, M.M. (2005). Serum uric acid levels and the risk of Parkinson disease. *Annals of Neurology*, **58**, 797–800.

De Lau, L.M., Koudstaal, P.J., Witteman, J.C., Hofman, A., and Breteler, M.M. (2006). Dietary folate, vitamin B_{12}, and vitamin B_6 and the risk of Parkinson disease. *Neurology*, **67**, 315–18.

Del Brutto, O.H. (2012). Neurocysticercosis. *Continuum (Minneapolis Minn)*, **18**, 1392–416.

De Lau, L.M. and Breteler, M.M. (2006). Epidemiology of Parkinson's disease. *The Lancet Neurology*, **5**, 525–35.

De Rijk, M.C., et al. (1997). Prevalence of parkinsonism and Parkinson's disease in Europe: the EUROPARKINSON Collaborative Study. European Community Concerted Action on the Epidemiology of Parkinson's disease. *Journal of Neurology, Neurosurgery, and Psychiatry*, **62**, 10–15.

de Vries, B., et al. (2016). Systematic re-evaluation of genes from candidate gene association studies in migraine using a large genome-wide association data set. *Cephalalgia*, **36**, 604–14.

Devinsky, O. (2011). Sudden, unexpected death in epilepsy. *The New England Journal of Medicine*, **365**, 1801–11.

Dickson, D.W., Lin, W., Liu, W.K., and Yen, S.H. (1999). Multiple system atrophy: a sporadic synucleinopathy. *Brain Pathology*, **9**, 721–32.

Do, C.B., et al. (2011). Web-based genome-wide association study identifies two novel loci and a substantial genetic component for Parkinson's disease. *PLoS Genetics*, 7, e1002141.

Dyck, P.J., et al. (2012). Impaired glycemia and diabetic polyneuropathy: the OC IG Survey. *Diabetes Care*, **35**, 584–91.

Dyck, P.J., Davies, J.L., Wilson, D.M., Service, F.J., Melton, L.J., 3rd, and O'Orien, P.C. (1999). Risk factors for severity of diabetic polyneuropathy: intensive longitudinal assessment of the Rochester Diabetic Neuropathy Study cohort. *Diabetes Care*, **22**, 1479–86.

Ebers, G.C., Sadovnick, A.D., Dyment, D.A., Yee, I.M., Willer, C.J., and Risch, N. (2004). Parent-of-origin effect in multiple sclerosis: observations in half-siblings. *The Lancet*, **363**,

1773–4.
Evans, D.A., et al. (1989). Prevalence of Alzheimer's disease in a community population of older persons. Higher than previously reported. *Journal of the American Medical Association*, **262**, 2551–6.
Fagan, A.M., Roe, C.M., Xiong, C., Mintun, M.A., Morris, J.C., and Holtzman, D.M. (2007). Cerebrospinal fluid tau/beta-amyloid(42) ratio as a prediction of cognitive decline in nondemented older adults. *Archives of Neurology*, **64**, 343–9.
Fiest, K.M., et al. (2017). Prevalence and incidence of epilepsy: a systematic review and meta-analysis of international studies. *Neurology*, **88**, 296–303.
Firestone, J.A., Smith-Weller, T., Franklin, G., Swanson, P., Longstreth, W.T., Jr., and Checkoway, H. (2005). Pesticides and risk of Parkinson disease: a population-based case-control study. *Archives of Neurology*, **62**, 91–5.
Fischera, M., Marziniak, M., Gralow, I., and Evers, S. (2008). The incidence and prevalence of cluster headache: a meta-analysis of population-based studies. *Cephalalgia*, **28**, 614–18.
Fisher, R.S., et al. (2017). Operational classification of seizure types by the International League Against Epilepsy: Position Paper of the ILAE Commission for Classification and Terminology. *Epilepsia*, **58**, 522–30.
Fleming, J.O. and Cook, T.D. (2006). Multiple sclerosis and the hygiene hypothesis. *Neurology*, **67**, 2085–6.
Franklin, G.M., Haug, J., Heyer, N., Checkoway, H., and Peck, N. (1991). Occupational carpal tunnel syndrome in Washington State, 1984–1988. *American Journal of Public Health*, **81**, 741–6.
Frediani, F. and Villani, V. (2007). Migraine and depression. *Neurological Sciences*, 28(Suppl. 2), S161–5.
Gao, X., et al. (2007). Prospective study of dietary pattern and risk of Parkinson disease. *American Journal of Clinical Nutrition*, **86**, 1486–94.
Gass, J., et al. (2006). Mutations in progranulin are a major cause of ubiquitin-positive frontotemporal lobar degeneration. *Human Molecular Genetics*, **15**, 2988–3001.
Geerlings, M.I., et al. (1999). Cognitive reserve and mortality in dementia: the role of cognition, functional ability and depression. *Psychological Medicine*, **29**, 1219–26.
Gelb, D.J., Oliver, E., and Gilman, S. (1999). Diagnostic criteria for Parkinson disease. *Archives of Neurology*, **56**, 33–9.
Gervil, M., Ulrich, V., Kyvik, K.O., Olesen, J., and Russell, M.B. (1999). Migraine without aura: a population-based twin study. *Annals of Neurology*, **46**, 606–11.
Goedert, M., and Spillantini, M.G. (1998). Lewy body diseases and multiple system atrophy as alpha-synucleinopathies. *Molecular Psychiatry*, **3**, 462–5.
Gorell, J.M., Johnson, C.C., Rybicki, B.A., Peterson, E.L., and Richardson, R.J. (1998). The risk of Parkinson's disease with exposure to pesticides, farming, well water, and rural living. *Neurology*, **50**, 1346–50.
Gorell, J.M., Rybicki, B.A., Cole Johnson, C., and Peterson, E.L. (1999). Occupational metal exposures and the risk of Parkinson's disease. *Neuroepidemiology*, **18**, 303–8.
Guerreiro, R., et al. (2013). TREM2 variants in Alzheimer's disease. *The New England Journal of Medicine*, **368**, 117–27.
Haass, C. (2004). Take five—BACE and the gamma-secretase quartet conduct Alzheimer's amyloid beta-peptide generation. *EMBO Journal*, **23**, 483–8.
Hachinski, V. (1994). Vascular dementia: a radical redefinition. *Dementia*, **5**, 130–2.
Hafler, D.A., et al. (2007). Risk alleles for multiple sclerosis identified by a genomewide study. *The New England Journal of Medicine*, **357**, 851–62.
Hagen, K., Vatten, L., Stovner, L.J., Zwart, J.A., Krokstad, S., and Bovim, G. (2002). Low socio-economic status is associated with increased risk of frequent headache: a prospective study of

22718 adults in Norway. *Cephalalgia*, **22**, 672–9.
Haines, J.L., et al. (1996). A complete genomic screen for multiple sclerosis underscores a role for the major histocompatibility complex. The Multiple Sclerosis Genetics Group. *Nature Genetics*, **13**, 469–71.
Hakim, A.J., Cherkas, L., El Zayat, S., MacGregor, A.J., and Spector, T.D. (2002). The genetic contribution to carpal tunnel syndrome in women: a twin study. *Arthritis & Rheumatism*, **47**, 275–9.
Hansen, T., Skytthe, A., Stenager, E., Petersen, H.C., Bronnum-Hansen, H., and Kyvik, K.O. (2005). Concordance for multiple sclerosis in Danish twins: an update of a nationwide study. *Multiple Sclerosis*, **11**, 504–10.
Hardy, J. (2010). Genetic analysis of pathways to Parkinson disease. *Neuron*, **68**, 201–6.
Hardy, J., Duff, K., Hardy, K.G., Perez-Tur, J., and Hutton, M. (1998). Genetic dissection of Alzheimer's disease and related dementias: amyloid and its relationship to tau [published erratum appears in Nat Neurosci 1998 Dec;1(8):743]. *Nature Neuroscience*, **1**, 355–8.
Hardy, T.A., Blum, S., McCombe, P.A., and Reddel, S.W. (2011). Guillain– Barré syndrome: modern theories of etiology. *Current Allergy and Asthma Reports*, **11**, 197–204.
Hauser, W.A., Annegers, J.F., and Rocca, W.A. (1996). Descriptive epidemiology of epilepsy: contributions of population-based studies from Rochester, Minnesota. *Mayo Clinic Proceedings*, **71**, 576–86.
Headache Classification Committee of the International Headache Society (2018). The International Classification of Headache Disorders: 3rd edition. *Cephalalgia*, **38**, 1–211.
Henderson, V.W., et al. (2000). Estrogen for Alzheimer's disease in women: randomized, double-blind, placebo-controlled trial. *Neurology*, **54**, 295–301.
Hendrie, H.C. (1998). Epidemiology of dementia and Alzheimer's disease. *American Journal of Geriatric Psychiatry*, 6, S3–18.
Herrera, B.M., et al. (2007). Parental transmission of MS in a population-based Canadian cohort. *Neurology*, **69**, 1208–12.
Hirtz, D., Thurman, D.J., Gwinn-Hardy, K., Mohamed, M., Chaudhuri, A.R., and Zalutsky, R. (2007). How common are the 'common' neurologic disorders? *Neurology*, **68**, 326–37.
Hoge, C.W., Castro, C.A., Messer, S.C., Mcgurk, D., Cotting, D.I., and Koffman, R.L. (2004). Combat duty in Iraq and Afghanistan, mental health problems, and barriers to care. *The New England Journal of Medicine*, **351**, 13–22.
Holbrook, T.L., Hoyt, D.B., Coimbra, R., Potenza, B., Sise, M., and Anderson, J.P. (2005). Long-term posttraumatic stress disorder persists after major trauma in adolescents: new data on risk factors and functional outcome. *Journal of Trauma*, **58**, 764–9; discussion 769–71.
Hollingworth, P., et al. (2011). Common variants at ABCA7, MS4A6A/MS4A4E, EPHA1, CD33 and CD2AP are associated with Alzheimer's disease. *Nature Genetics*, **43**, 429–35.
Hu, G., Jousilahti, P., Nissinen, A., Antikainen, R., Kivipelto, M., and Tuomilehto, J. (2006). Body mass index and the risk of Parkinson disease. *Neurology*, **67**, 1955–9.
Hu, X.H., Markson, L.E., Lipton, R.B., Stewart, W.F., and Berger, M.L. (1999). Burden of migraine in the United States: disability and economic costs. *Archives of Internal Medicine*, **159**, 813–18.
Hyman, B.T., et al. (2012). National Institute on Aging-Alzheimer's Association guidelines for the neuropathologic assessment of Alzheimer's disease. *Alzheimer's & Dementia*, **8**, 1–13.
In t'Veld, B.A., et al. (2001). Nonsteroidal antiinflammatory drugs and the risk of Alzheimer's disease. *The New England Journal of Medicine*, **345**, 1515–21.
Jafari, N. and Hintzen, R.Q. (2011). The association between cigarette smoking and multiple sclerosis. *Journal of the Neurological Sciences*, **311**, 78–85.
Jankovic, J., Rajput, A.H., Mcdermott, M.P., and Perl, D.P. (2000). The evolution of diagnosis in early Parkinson disease. Parkinson Study Group. *Archives of Neurology*, **57**, 369–72.

Jennett, B., Teasdale, G., Braakman, R., Minderhoud, J., and Knill-Jones, R. (1976). Predicting outcome in individual patients after severe head injury. *The Lancet*, **1**, 1031–4.
Jorm, A.F. and Jolley, D. (1998). The incidence of dementia: a meta-analysis. *Neurology*, **51**, 728–33.
Kaiboriboon, K., Bakaki, P.M., Lhatoo, S.D., and Koroukian, S. (2013). Incidence and prevalence of treated epilepsy among poor health and low-income Americans. *Neurology*, **80**, 1942–9.
Kaniecki, R.G. (2012). Tension-type headache. *Continuum (Minneapolis Minn)*, **18**, 823–34.
Kantarci, O.H., et al. (2006). Men transmit MS more often to their children vs. women: the Carter effect. *Neurology*, **67**, 305–10.
Keenan, H.T. and Bratton, S.L. (2006). Epidemiology and outcomes of pediatric traumatic brain injury. *Developmental Neuroscience*, **28**, 256–63.
Keenan, H.T., Runyan, D.K., Marshall, S.W., Nocera, M.A., Merten, D.F., and Sinal, S.H. (2003). A population-based study of inflicted traumatic brain injury in young children. *Journal of the American Medical Association*, **290**, 621–6.
Kivipelto, M., and Solomon, A. (2006). Cholesterol as a risk factor for Alzheimer's disease—epidemiological evidence. *Acta Neurologica Scandinavica Supplement*, **185**, 50–7.
Klunk, W.E., et al. (2003). Imaging the pathology of Alzheimer's disease: amyloid-imaging with positron emission tomography. *Neuroimaging Clinics of North America*, **13**, 781–9, ix.
Knopman, D.S., et al. (2001). Practice parameter: diagnosis of dementia (an evidence-based review). Report of the Quality Standards Subcommittee of the American Academy of Neurology. *Neurology*, **56**, 1143–53.
Koch-Henriksen, N. and Sorensen, P.S. (2011). Why does the north–south gradient of incidence of multiple sclerosis seem to have disappeared on the northern hemisphere? *Journal of the Neurological Sciences*, **311**, 58–63.
Koepsell, T.D., Kurland, B.F., Harel, O., Johnson, E.A., Zhou, X.H., and Kukull, W.A. (2007). Education, cognitive function, and severity of neuropathology in Alzheimer disease. *Neurology*, 68, A169–70.
Korczyn, A.D., and Reichmann, H. (2006). Dementia with Lewy bodies. *Journal of the Neurological Sciences*, **248**, 3–8.
Kurtzke, J.F. (2005). Epidemiology and etiology of multiple sclerosis. *Physical Medicine & Rehabilitation Clinics of North America*, **16**, 327–49.
Kwong, L.K., Neumann, M., Sampathu, D.M., Lee, V.M., and Trojanowski, J.Q. (2007). TDP-43 proteinopathy: the neuropathology underlying major forms of sporadic and familial frontotemporal lobar degeneration and motor neuron disease. *Acta Neuropathologica*, **114**, 63–70.
Lafrenière, R.G., et al. (2010). A dominant-negative mutation in the TRESK potassium channel is linked to familial migraine with aura. *Nature Medicine*, **16**, 1157.
Laitinen, M.H., et al. (2006). Fat intake at midlife and risk of dementia and Alzheimer's disease: a population-based study. *Dementia and Geriatric Cognitive Disorders*, **22**, 99–107.
Langer-Gould, A. and Beaber, B.E. (2013). Effects of pregnancy and breastfeeding on the multiple sclerosis disease course. *Clinical Immunology*, 149(2), 244–50.
Langston, J.W. (2006). The Parkinson's complex: parkinsonism is just the tip of the iceberg. *Annals of Neurology*, **59**, 591–6.
Langston, J.W., and Ballard, P.A., Jr. (1983). Parkinson's disease in a chemist working with 1-methyl-4-phenyl-1,2,5,6-tetrahydropyridine. *The New England Journal of Medicine*, **309**, 310.
Langston, J.W., Ballard, P., Tetrud, J.W., and Irwin, I. (1983). Chronic Parkinsonism in humans due to a product of meperidine-analog synthesis. *Science*, **219**, 979–80.
Larocca, N.G., et al. (1984). Field testing of a minimal record of disability in multiple sclerosis: the United States and Canada. *Acta Neurologica Scandinavica Supplement*, **101**, 126–38.

Launer, L.J., et al. (1999). Rates and risk factors for dementia and Alzheimer's disease: results from EURODEM pooled analyses. EURODEM Incidence Research Group and Work Groups. *European Studies of Dementia. Neurology*, **52**, 78–84.

Launer, L.J., et al. (2000). CASCADE: a European collaborative study on vascular determinants of brain lesions. Study design and objectives. *Neuroepidemiology*, **19**, 113–20.

Leidy, N.K., Elixhauser, A., Vickrey, B., Means, E., and Willian, M.K. (1999). Seizure frequency and the health-related quality of life of adults with epilepsy. *Neurology*, **53**, 162–6.

Lemke, J.R., et al. (2012). Targeted next generation sequencing as a diagnostic tool in epileptic disorders. *Epilepsia*, **53**, 1387–98.

Leonardi, M., Steiner, T.J., Scher, A.T., and Lipton, R.B. (2005). The global burden of migraine: measuring disability in headache disorders with WHO's Classification of Functioning, Disability and Health (ICF). *Journal of Headache Pain*, **6**, 429–40.

Levy-Lahad, E., Tsuang, D., and Bird, T.D. (1998). Recent advances in the genetics of Alzheimer's disease. *Journal of Geriatric Psychiatry and Neurology*, **11**, 42–54.

Lhatoo, S.D. and Sander, J.W. (2005). Cause-specific mortality in epilepsy. *Epilepsia*, *46*(Suppl. 11), 36–9.

Li, G., et al. (2007). Age-varying association between blood pressure and risk of dementia in those aged 65 and older: a community-based prospective cohort study. *Journal of the American Geriatrics Society*, **55**, 1161–7.

Libby, A.M., Ghushchyan, V., McQueen, R.B., Slejko, J.F., Bainbridge, J.L., and Campbell, J.D. (2012). Economic differences in direct and indirect costs between people with epilepsy and without epilepsy. *Medical Care*, **50**, 928–33.

Lin, R., Charlesworth, J., Van Der Mei, I., and Taylor, B.V. (2012). The genetics of multiple sclerosis. *Practical Neurology*, **12**, 279–88.

Linde, M., et al. (2012). The cost of headache disorders in Europe: the Eurolight project. *European Journal of Neurology*, **19**, 703–11.

Lipton, R.B., Bigal, M.E., Diamond, M., Freitag, F., Reed, M.L., and Stewart, W.F. (2007). Migraine prevalence, disease burden, and the need for preventive therapy. *Neurology*, **68**, 343–9.

Lipton, R.B., Stewart, W.F., and Simon, D. (1998). Medical consultation for migraine: results from the American Migraine Study. *Headache*, **38**, 87–96.

Litvan, I., et al. (2007). The etiopathogenesis of Parkinson disease and suggestions for future research. Part I. *Journal of Neuropathology & Experimental Neurology*, **66**, 251–7.

Lublin, F.D. and Reingold, S.C. (1996). Defining the clinical course of multiple sclerosis: results of an international survey. National Multiple Sclerosis Society (USA) Advisory Committee on Clinical Trials of New Agents in Multiple Sclerosis. *Neurology*, **46**, 907–11.

Lyngberg, A.C., Rasmussen, B.K., Jorgensen, T., and Jensen, R. (2005a). Has the prevalence of migraine and tension-type headache changed over a 12-year period? A Danish population survey. *European Journal of Epidemiology*, **20**, 243–9.

Lyngberg, A.C., Rasmussen, B.K., Jorgensen, T., and Jensen, R. (2005b). Prognosis of migraine and tension-type headache: a population-based follow-up study. *Neurology*, **65**, 580–5.

Mackenzie, I.R., et al. (2006). The neuropathology of frontotemporal lobar degeneration caused by mutations in the progranulin gene. *Brain*, **129**, 3081–90.

Mackenzie, I.R. and Rademakers, R. (2007). The molecular genetics and neuropathology of frontotemporal lobar degeneration: recent developments. *Neurogenetics*, **8**, 237–48.

Maraganore, D.M., Farrer, M.J., Hardy, J.A., Mcdonnell, S.K., Schaid, D.J., and Rocca, W.A. (2000). Case-control study of debrisoquine 4-hydroxylase, N-acetyltransferase 2, and apolipoprotein E gene polymorphisms in Parkinson's disease. *Movement Disorders*, **15**, 714–19.

Marrie, R.A., Cutter, G., Tyry, T., Hadjimichael, O., Campagnolo, D., and Vollmer, T. (2005). Changes in the ascertainment of multiple sclerosis. *Neurology*, **65**, 1066–70.

Martyn, C.N. and Hughes, R.A. (1997). Epidemiology of peripheral neuropathy. *Journal of Neurology, Neurosurgery, & Psychiatry*, **62**, 310–18.

Mathis, C.A., Wang, Y., Holt, D.P., Huang, G.F., Debnath, M.L., and Klunk, W.E. (2003). Synthesis and evaluation of 11C-labeled 6-substituted 2-arylbenzothiazoles as amyloid imaging agents. *Journal of Medicinal Chemistry*, **46**, 2740–54.

McKee, A.C., et al. (2013). The spectrum of disease in chronic traumatic encephalopathy. *Brain*, **136**, 43–64.

McKeith, I.G. (2006). Consensus guidelines for the clinical and pathologic diagnosis of dementia with Lewy bodies (DLB): report of the Consortium on DLB International Workshop. *Journal of Alzheimer's Disease*, **9**, 417–23.

Mckhann, G.M., et al. (2011). The diagnosis of dementia due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & Dementia*, **7**, 263–9.

Meeraus, W.H., Petersen, I., Chin, R.F., Knott, F., and Gilbert, R. (2013). Childhood epilepsy recorded in primary care in the UK. *Archives of Disease in Childhood*, **98**, 195–202.

Mehta, K.M., et al. (1999). Head trauma and risk of dementia and Alzheimer's disease: the Rotterdam study. *Neurology*, **53**, 1959–62.

Menegon, A., Board, P.G., Blackburn, A.C., Mellick, G.D., and Le Couteur, D.G. (1998). Parkinson's disease, pesticides, and glutathione transferase polymorphisms. *The Lancet*, **352**, 1344–6.

Mesulam, M., et al. (2007). Progranulin mutations in primary progressive aphasia: the PPA1 and PPA3 families. *Archives of Neurology*, **64**, 43–7.

Mikaeloff, Y., Caridade, G., Tardieu, M., and Suissa, S. (2007). Parental smoking at home and the risk of childhood-onset multiple sclerosis in children. *Brain*, **130**, 2589–95.

Minden, K., Niewerth, M., Listing, J., Biedermann, T., Schontube, M., and Zink, A. (2004). Burden and cost of illness in patients with juvenile idiopathic arthritis. *Annals of the Rheumatic Diseases*, **63**, 836–42.

Mintun, M.A., et al. (2006). [11C]PIB in a non-demented population: potential antecedent marker of Alzheimer disease. *Neurology*, **67**, 446–52.

Moceri, V.M., Kukull, W.A., Emanuel, I., Van Belle, G., and Larson, E.B. (2000). Early-life risk factors and the development of Alzheimer's disease. *Neurology*, **54**, 415–20.

Mold, J.W., Vesely, S.K., Keyl, B.A., Schenk, J.B., and Roberts, M. (2004). The prevalence, predictors, and consequences of peripheral sensory neuropathy in older patients. *Journal of the American Board of Family Medicine*, **17**, 309–18.

Monahan, K., and O'Leary, K.D. (1999). Head injury and battered women: an initial inquiry. *Health & Social Work*, **24**, 269–78.

Morimoto, R.I. (2006). Stress, aging, and neurodegenerative disease. *The New England Journal of Medicine*, **355**, 2254–5.

Morris, J.C. (2006). Mild cognitive impairment is early-stage Alzheimer disease: time to revise diagnostic criteria. *Archives of Neurology*, **63**, 15–16.

Morris, J.C., and Cummings, J. (2005). Mild cognitive impairment (MCI) represents early-stage Alzheimer's disease. *Journal of Alzheimer's Disease*, **7**, 235–9; discussion 255–62.

Morris, J.C., et al. (2001). Mild cognitive impairment represents early-stage Alzheimer disease. *Archives of Neurology*, **58**, 397–405.

Muller, M., Tang, M.X., Schupf, N., Manly, J.J., Mayeux, R., and Luchsinger, J.A. (2007). Metabolic syndrome and dementia risk in a multiethnic elderly cohort. *Dementia and Geriatric Cognitive Disorders*, **24**, 185–92.

Mulnard, R.A., et al. (2000). Estrogen replacement therapy for treatment of mild to moderate Alzheimer disease: a randomized controlled trial. Alzheimer's Disease Cooperative Study. *Journal of the American Medical Association*, **283**, 1007–15.
Murray, M.E., Knopman, D.S., and Dickson, D.W. (2007). Vascular dementia: clinical, neuroradiologic and neuropathologic aspects. *Panminerva Medica*, **49**, 197–207.
Naj, A.C., et al. (2011). Common variants at MS4A4/MS4A6E, CD2AP, CD33 and EPHA1 are associated with late-onset Alzheimer's disease. *Nature Genetics*, **43**, 436–41.
Natoli, J.L., et al. (2010). Global prevalence of chronic migraine: a systematic review. *Cephalalgia*, **30**, 599–609.
Neumann, M., et al. (2006). Ubiquitinated TDP-43 in frontotemporal lobar degeneration and amyotrophic lateral sclerosis. *Science*, **314**, 130–3.
Newton, C.R. and Garcia, H.H. (2012). Epilepsy in poor regions of the world. *The Lancet*, **380**, 1193–201.
Ngugi, A.K., et al. (2013). Prevalence of active convulsive epilepsy in sub-Saharan Africa and associated risk factors: cross-sectional and case-control studies. *The Lancet Neurology*, **12**, 253–63.
Nichols, E., et al. (2019). Global, regional, and national burden of Alzheimer's disease and other dementias, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *The Lancet Neurology*, **18**, 88–106.
Nielsen, N.M., et al. (2005). Familial risk of multiple sclerosis: a nationwide cohort study. *American Journal of Epidemiology*, **162**, 774–8.
Nordstrom, D.L., Destefano, F., Vierkant, R.A., and Layde, P.M. (1998). Incidence of diagnosed carpal tunnel syndrome in a general population. *Epidemiology*, **9**, 342–5.
Nussbaum, R.L. and Ellis, C.E. (2003). Alzheimer's disease and Parkinson's disease. *New England Journal of Medicine*, **348**, 1356–64.
Okie, S. (2005). Traumatic brain injury in the war zone. *The New England Journal of Medicine*, **352**, 2043–7.
Orchard, T.J., Ce, L.L., Maser, R.E., and Kuller, L.H. (1996). Why does diabetic autonomic neuropathy predict IDDM mortality? An analysis from the Pittsburgh Epidemiology of Diabetes Complications Study. *Diabetes Research and Clinical Practice*, 34(Suppl.), S165–71.
Orton, S.M., et al. (2006). Sex ratio of multiple sclerosis in Canada: a longitudinal study. *The Lancet Neurology*, **5**, 932–6.
Osuntokun, B.O., et al. (1992). Prevalence of headache and migrainous headache in Nigerian Africans: a community-based study. *East African Medical Journal*, **69**, 196–9.
Owens, G.P. and Bennett, J.L. (2012). Trigger, pathogen, or bystander: the complex nexus linking Epstein-Barr virus and multiple sclerosis. *Multiple Sclerosis*, **18**, 1204–8.
Petersen, R.C. (2004). Mild cognitive impairment as a diagnostic entity. *Journal of Internal Medicine*, **256**, 183–94.
Petersen, R.C. (2007). Mild cognitive impairment: current research and clinical implications. *Seminars in Neurology*, **27**, 22–31.
Petrovitch, H., et al. (2002). Plantation work and risk of Parkinson disease in a population-based longitudinal study. *Archives of Neurology*, **59**, 1787–92.
Polman, C.H., et al. (2011). Diagnostic criteria for multiple sclerosis: 2010 revisions to the McDonald criteria. *Annals of Neurology*, **69**, 292–302.
Prince, M., Bryce, R., Albanese, E., Wimo, A., Ribeiro, W., and Ferri, C.P. (2013). The global prevalence of dementia: a systematic review and metaanalysis. *Alzheimer's & Dementia*, **9**, 63–75.e2.
Pringsheim, T., Jette, N., Frolkis, A., and Steeves, T.D. (2014). The prevalence of Parkinson's disease: a systematic review and meta-analysis. *Movement Disorders*, **29**, 1583–90.

Racette, B.A., et al. (2012). Increased risk of parkinsonism associated with welding exposure. *Neurotoxicology*, **33**, 1356–61.
Rademakers, R., et al. (2007). Phenotypic variability associated with progranulin haploinsufficiency in patients with the common 1477C—>T (Arg493X) mutation: an international initiative. *The Lancet Neurology*, **6**, 857–68.
Rademakers, R., and Hutton, M. (2007). The genetics of frontotemporal lobar degeneration. *Current Neurology and Neuroscience Reports*, **7**, 434–42.
Reitz, C., et al. (2013). Variants in the ATP-binding cassette transporter (ABCA7), apolipoprotein E 4, and the risk of late-onset Alzheimer disease in African Americans. *Journal of the American Medical Association*, **309**, 1483–92.
Ritz, B., et al. (2007). Pooled analysis of tobacco use and risk of Parkinson disease. *Archives of Neurology*, **64**, 990–7.
Rivara, F.P., et al. (2012). Incidence of disability among children 12 months after traumatic brain injury. *American Journal of Public Health*, **102**, 2074–9.
Rocca, W.A., et al. (2007). Risk of cognitive impairment or dementia in relatives of patients with Parkinson disease. *Archives of Neurology*, **64**, 1458–64.
Rocca, W.A., Cha, R.H., Waring, S.C., and Kokmen, E. (1998). Incidence of dementia and Alzheimer's disease: a reanalysis of data from Rochester, Minnesota, 1975–1984. *American Journal of Epidemiology*, **148**, 51–62.
Roman, G.C. (1994). An epidemic in Cuba of optic neuropathy, sensorineural deafness, peripheral sensory neuropathy and dorsolateral myeloneuropathy. *Journal of the Neurological Sciences*, **127**, 11–28.
Roman, G.C., et al. (1993). Vascular dementia: diagnostic criteria for research studies. Report of the NINDS-AIREN International Workshop [see comments]. *Neurology*, **43**, 250–60.
Roopra, A., Dingledine, R., and Hsieh, J. (2012). Epigenetics and epilepsy. *Epilepsia*, 53(Suppl. 9), 2–10.
Ross, G.W., et al. (2000). Association of coffee and caffeine intake with the risk of Parkinson disease. *Journal of the American Medical Association*, **283**, 2674–9.
Rowe, C.C., et al. (2007). Imaging beta-amyloid burden in aging and dementia. *Neurology*, **68**, 1718–25.
Rowland, L.P. and Merritt, H.H. (1995). *Merritt's Textbook of Neurology*. Baltimore, MD: Williams & Wilkins.
Russell, M.B., Levi, N., and Kaprio, J. (2007). Genetics of tension-type headache: a population based twin study. *American Journal of Medical Genetics Part B*, **144**, 982–6.
Rutland-Brown, W., Langlois, J.A., Thomas, K.E., and Xi, Y.L. (2006). Incidence of traumatic brain injury in the United States, 2003. *Journal of Head Trauma Rehabilitation*, **21**, 544–8.
Sadovnick, A.D., Bulman, D., and Ebers, G.C. (1991). Parent-child concordance in multiple sclerosis. *Annals of Neurology*, **29**, 252–5.
Saunders, A.M., et al. (1993). Apolipoprotein E epsilon 4 allele distributions in late-onset Alzheimer's disease and in other amyloid-forming diseases [see comments]. *The Lancet*, **342**, 710–11.
Sawcer, S., et al. (2011). Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature*, **476**, 214–19.
Scher, A.I., Stewart, W.F., Ricci, J.A., and Lipton, R.B. (2003). Factors associated with the onset and remission of chronic daily headache in a population-based study. *Pain*, **106**, 81–9.
Schwartz, B.S., Stewart, W.F., and Lipton, R.B. (1997). Lost workdays and decreased work effectiveness associated with headache in the workplace. *Journal of Occupational & Environmental Medicine*, **39**, 320–7.
Scordo, M.G., Dahl, M.L., Spina, E., Cordici, F., and Arena, M.G. (2006). No association between CYP2D6 polymorphism and Alzheimer's disease in an Italian population.

Pharmacological Research, **53**, 162–5.
Shackleton, D.P., Westendorp, R.G., Trenite, D.G., and Vandenbroucke, J.P. (1999). Mortality in patients with epilepsy: 40 years of follow up in a Dutch cohort study [see comments]. *Journal of Neurology, Neurosurgery, & Psychiatry*, **66**, 636–40.
Sherer, M., Struchen, M.A., Yablon, S.A., Wang, Y., and Nick, T.G. (2007). Comparison of indices of TBI severity: Glasgow coma scale, length of coma, post-traumatic amnesia. *Journal of Neurology, Neurosurgery, & Psychiatry*, 79(6), 678–85.
Shumaker, S.A., et al. (2004). Conjugated equine estrogens and incidence of probable dementia and mild cognitive impairment in postmenopausal women: Women's Health Initiative Memory Study. *Journal of the American Medical Association*, **291**, 2947–58.
Shumaker, S.A., et al. (2003). Estrogen plus progestin and the incidence of dementia and mild cognitive impairment in postmenopausal women: the Women's Health Initiative Memory Study: a randomized controlled trial. *Journal of the American Medical Association*, **289**, 2651–62.
Siderowf, A., et al. (2012). Impaired olfaction and other prodromal features in the Parkinson At-Risk Syndrome Study. *Movement Disorders*, **27**, 406–12.
Simon, K.C., Munger, K.L., and Ascherio, A. (2012). Vitamin D and multiple sclerosis: epidemiology, immunology, and genetics. *Current Opinion in Neurology*, **25**, 246–51.
Small, G.W. (2004). Neuroimaging as a diagnostic tool in dementia with Lewy bodies. *Dementia and Geriatric Cognitive Disorders*, 17(Suppl. 1), 25–31.
Solomon, D.H., Katz, J.N., Bohn, R., Mogun, H., and Avorn, J. (1999). Nonoccupational risk factors for carpal tunnel syndrome. *Journal of General Internal Medicine*, **14**, 310–14.
Sperling, R.A., et al. (2011). Toward defining the preclinical stages of Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & Dementia*, **7**, 280–92.
Stallings, S.P., Kasdan, M.L., Soergel, T.M., and Corwin, H.M. (1997). A case-control study of obesity as a risk factor for carpal tunnel syndrome in a population of 600 patients presenting for independent medical examination. *Journal of Hand Surgery*, **22**, 211–15.
Stein, M.S., et al. (2011). A randomized trial of high-dose vitamin D2 in relapsing-remitting multiple sclerosis. *Neurology*, **77**, 1611–18.
Stewart, W.F., Bigal, M.E., Kolodner, K., Dowson, A., Liberman, J.N., and Lipton, R.B. (2006). Familial risk of migraine: variation by proband age at onset and headache severity. *Neurology*, **66**, 344–8.
Stewart, W.F., Lipton, R.B., and Liberman, J. (1996). Variation in migraine prevalence by race. *Neurology*, **47**, 52–9.
St George-Hyslop, P.H. (2000). Molecular genetics of Alzheimer's disease. *Biological Psychiatry*, **47**, 183–99.
Stovner, L., et al. (2007). The global burden of headache: a documentation of headache prevalence and disability worldwide. *Cephalalgia*, **27**, 193–210.
Stovner, L.J., et al. (2018). Global, regional, and national burden of migraine and tension-type headache, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *The Lancet Neurology*, **17**, 954–76.
Strittmatter, W.J., et al. (1993). Apolipoprotein E: high-avidity binding to beta-amyloid and increased frequency of type 4 allele in late-onset familial Alzheimer disease. *Proceedings of the National Academy of Sciences of the United States of America*, **90**, 1977–81.
Suzuki, K., Akama, T., Kawashima, A., Yoshihara, A., Yotsu, R.R., and Ishii, N. (2012). Current status of leprosy: epidemiology, basic science and clinical perspectives. *Journal of Dermatology*, **39**, 121–9.
Svendsen, T., Lossius, M., and Nakken, K.O. (2007). Age-specific prevalence of epilepsy in Oppland County, Norway. *Acta Neurologica Scandinavica*, **116**, 307–11.

Szekely, C.A., et al. (2007). NSAID use and dementia risk in the Cardiovascular Health Study. Role of APOE and NSAID type. *Neurology*, **68**, 1800–8.
Tanner, C.M., et al. (2011). Rotenone, paraquat, and Parkinson's disease. *Environmental Health Perspectives*, **119**, 866–72.
Taylor, C.A., Bell, J.M., Breiding, M.J., and Xu, L. (2017). Traumatic brain injury-related emergency department visits, hospitalizations, and deaths—United States, 2007 and 2013. Morbidity and mortality weekly report. *Surveillance Summaries (Washington, D.C. : 2002)*, **66**, 1–16.
Tepper, S.J. (2012). Medication-overuse headache. *Continuum (Minneapolis Minn)*, **18**, 807–22.
Tesfaye, S., et al. (1996). Prevalence of diabetic peripheral neuropathy and its relation to glycaemic control and potential risk factors: the EURODIAB IDDM Complications Study. *Diabetologia*, **39**, 1377–84.
Thurman, D.J., et al. (2017). Epidemiology Commission of the International League Against Epilepsy. The burden of premature mortality of epilepsy in high-income countries: a systematic review from the Mortality Task Force of the International League Against Epilepsy. *Epilepsia*, **58**, 17–26.
Tolnay, M., and Frank, S. (2007). Pathology and genetics of frontotemporal lobar degeneration: an update. *Clinical Neuropathology*, **26**, 143–56.
Tsuang, D., al. (2006). Predicting Lewy body pathology in a community-based sample with clinical diagnosis of Alzheimer's disease. *Journal of Geriatric Psychiatry and Neurology*, **19**, 195–201.
Tsuang, D., et al. (2012). GBA mutations increase risk for Lewy body disease with and without Alzheimer disease pathology. *Neurology*, **79**, 1944–50.
Uldall, P., Alving, J., Hansen, L.K., Kibaek, M., and Buchholt, J. (2006). The misdiagnosis of epilepsy in children admitted to a tertiary epilepsy centre with paroxysmal events. *Archives of Disease in Childhood*, **91**, 219–21.
Ulrich, V., Gervil, M., Kyvik, K.O., Olesen, J., and Russell, M.B. (1999). Evidence of a genetic factor in migraine with aura: a population-based Danish twin study. *Annals of Neurology*, **45**, 242–6.
Valera, E.M. and Berenbaum, H. (2003). Brain injury in battered women. *Journal of Consultant Clinical Psychology*, **71**, 797–804.
Van Deerlin, V.M., et al. (2010). Common variants at 7p21 are associated with frontotemporal lobar degeneration with TDP-43 inclusions. *Nature Genetics*, **42**, 234–9.
Wallin, M.T., et al. (2012). The Gulf War era multiple sclerosis cohort: age and incidence rates by race, sex and service. *Brain*, **135**, 1778–85.
Wang, S.J., Fuh, J.L., Lu, S.R., and Juang, K.D. (2001). Quality of life differs among headache diagnoses: analysis of SF-36 survey in 901 headache patients. *Pain*, **89**, 285–92.
Ward, T.N. (2012). Migraine diagnosis and pathophysiology. *Continuum (Minneapolis Minn)*, **18**, 753–63.
Warden, D. (2006). Military TBI during the Iraq and Afghanistan wars. *Journal of Head Trauma Rehabilitation*, **21**, 398–402.
Waubant, E., et al. (2011). Common viruses associated with lower pediatric multiple sclerosis risk. *Neurology*, **76**, 1989–95.
Werner, C., and Engelhard, K. (2007). Pathophysiology of traumatic brain injury. *British Journal of Anaesthesia*, **99**, 4–9.
Wimo, A., Jonsson, L., Bond, J., Prince, M., Winblad, B., and Alzheimer's Disease International (2013). The worldwide economic impact of dementia 2010. *Alzheimer's & Dementia*, **9**, 1–11.e3.
Yaffe, K., et al. (2004). The metabolic syndrome, inflammation, and risk of cognitive decline. *Journal of the American Medical Association*, **292**, 2237–42.

Zaccai, J., Mccracken, C., and Brayne, C. (2005). A systematic review of prevalence and incidence studies of dementia with Lewy bodies. *Age & Ageing*, **34**, 561–6.

Zarrelli, M.M., Beghi, E., Rocca, W.A., and Hauser, W.A. (1999). Incidence of epileptic syndromes in Rochester, Minnesota: 1980–1984. *Epilepsia*, **40**, 1708–14.

Ziegler, D.K., Hur, Y.M., Bouchard, T.J., Jr., Hassanein, R.S., and Barter, R. (1998). Migraine in twins raised together and apart. *Headache*, **38**, 417–22.

8.11

Infectious diseases and prions

Davidson H. Hamer, Amira Khan, and Zulfiqar A. Bhutta

Introduction

Despite considerable progress in attaining reductions in global morbidity and mortality, infectious diseases continue to account for a substantial burden of global morbidity, disability, and mortality. While non-communicable diseases (NCDs) continue to be the leading cause of global mortality (Liu et al. 2016), diseases such as lower respiratory infections (LRIs), diarrhoea, and tuberculosis (TB) are still among the top ten killers for all ages and sexes (World Health Organization [WHO] 2018b). Moreover, the emergence of many new viral, bacterial, fungal, and parasitic pathogens as well as rising antimicrobial resistance are current challenges.

As a result of changes in a variety of different environmental, social, economic, and public health factors, morbidity and mortality due to infectious diseases have declined in industrialized countries during the last 150 years with the result being a gradual transition to chronic diseases including cardiovascular disease, diabetes mellitus, and cancer as major causes of mortality in these countries today. However, in contrast, in less-developed countries, infectious diseases continue to contribute substantially to the overall burden of disease.

Detailed information on the definitions of infectious diseases, modes of transmission, and their control are provided in Chapter 11.3. An overview of issues related to emerging and re-emerging infections is provided in Chapter 8.17. Similarly, detailed information on diseases caused by sexually transmitted infections, HIV/acquired immunodeficiency syndrome (AIDS), tuberculosis, and malaria can be found in Chapters 8.12–8.15. This chapter will review the global burden of common infectious diseases in children and adults, determinants of the high infectious disease burden in resource-poor countries, and important aspects of the clinical manifestations, diagnosis, and

treatment of the handful of infectious diseases that account for the major share of morbidity and mortality in children and adults worldwide.

Burden of infectious diseases

At the beginning of the twentieth century, infectious diseases were the leading cause of death throughout the world. At that time, three diseases—pneumonia, diarrhoea, and tuberculosis—were responsible for about 30% of deaths in the United States. During the last century, there has been a decline in infectious diseases mortality in the United States from 797 deaths per 100,000 in 1900 to 36 per 100,000 in 1980.

In the late twentieth century, substantial reduction in child mortality occurred in low- and middle-income countries (LMICs). The fall in the number of child deaths from 1960 to 1990 averaged 2.5% per year and the risk of dying in the first 5 years of life halved—a major achievement in child survival. In the period from 1990 to 2001, mortality rates dropped an average of 1.1% annually, mostly after the neonatal period. Deaths among children under 5 years of age dropped from 12.7 million in 1990 to about 5.9 million in 2015 and further decreased to 5.6 million in 2016 (UN IGME 2017). The annual rate of under-5 mortality reduction showed a steady increase from 1.8% in 1990–2000 to 3.9% in 2000–2015 (UN IGME 2015). Much of this decline in under-5 deaths is attributable to the reduction in mortality secondary to infectious diseases. The WHO estimates that between 2000 and 2012, mortality in children due to measles, diarrhoea and pneumonia reduced by 80%, 50% and 40%, respectively (WHO 2018c).

However, despite this promising decline, Millennium Development Goal (MDG) 4, which called for a two-thirds reduction in under-5 mortality rate from 1990 base figures, was not met. It has been recognized that this was largely due to the relatively slower progress in reducing neonatal mortality. Neonatal mortality has decreased by 49% as compared to a 62% reduction in mortality in children 1–59 months (UN IGME 2017).

As deaths in children under the age of 5 have decreased, the proportion of these deaths that occur during the newborn period has increased. Out of the 5.6 million under-5 deaths in 2016, 2.6 million were neonates, which accounts for nearly half of all under-5 mortality. Preterm birth complications (16%), pneumonia (16%), intrapartum-related events (11%), and diarrhoea (8%) account for more than half of the child deaths globally (UN IGME 2017).

Thus, despite many public health successes, infectious diseases remain one of the main killers of children under-5. Moreover, neonatal sepsis, malaria, and AIDS account for 7%, 5% and 1% of all under-5 mortality, respectively (UN IGME 2017). The distribution of these major causes of mortality in various WHO regions is presented in Figure 8.11.1.

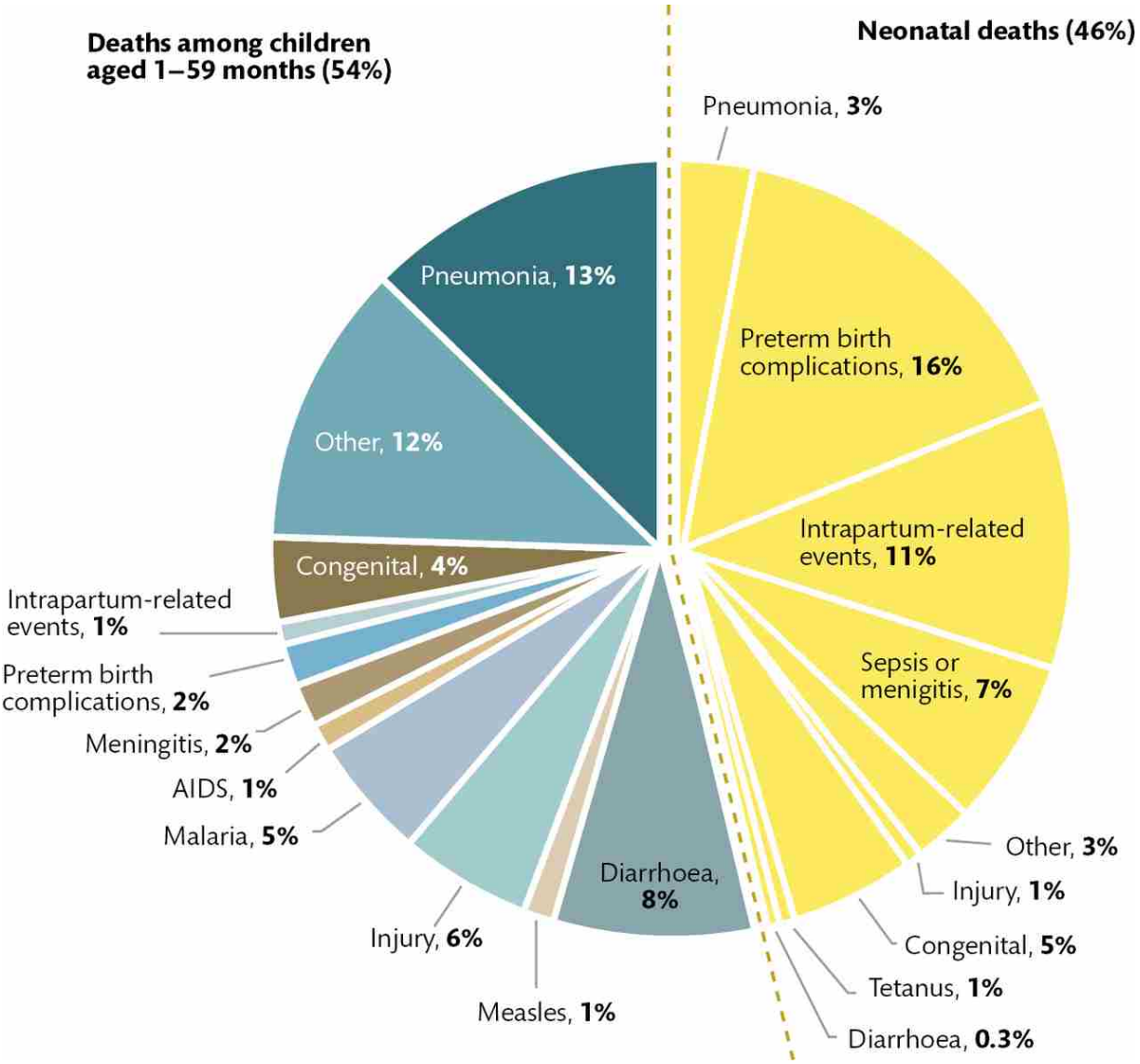


Fig. 8.11.1 Levels and trends in child mortality, 2017.

Reprinted from United Nations Inter-agency Group for Child Mortality Estimation (UN IGME), 'Levels & Trends in Child Mortality: Report 2017, Estimates Developed by the UN Inter-agency Group for Child Mortality Estimation', United Nations Children's Fund, New York, 2017, with permission from UNICEF.

Nearly 80% of all under-5 deaths occur in two regions of the world; sub-Saharan Africa and Southern Asia (UN IGME 2017). With an under-5 mortality rate of 79 deaths per 1,000 live births, sub-Saharan Africa remains the deadliest region for young children. Only six countries (India, Nigeria, Pakistan, the Democratic Republic of Congo, Ethiopia, and China) account for half of global deaths in children less than 5 years with the predominant causes being neonatal disorders and preventable infections such as pneumonia and diarrhoea. (Fig. 8.11.2; Table 8.11.1)

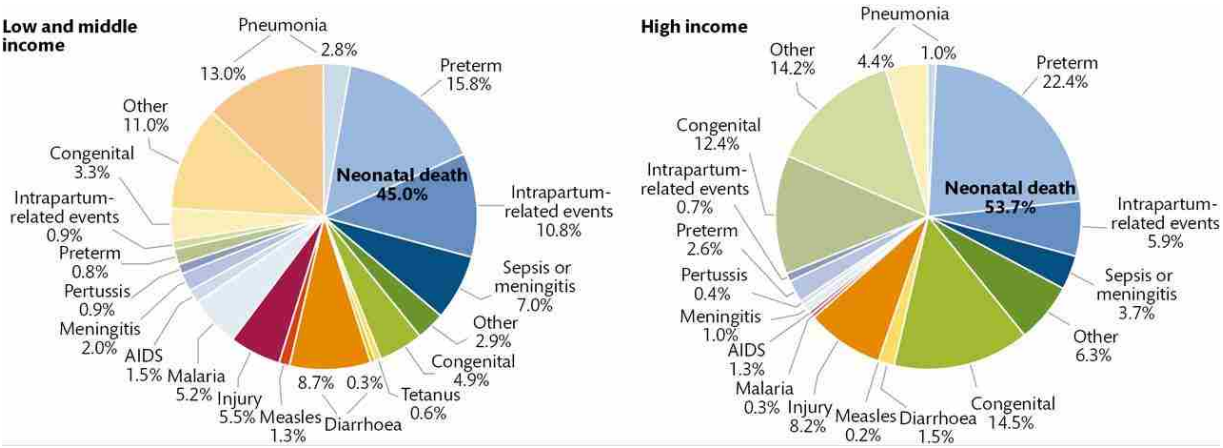


Fig. 8.11.2 Causes of childhood deaths among live-born children less than five, by income (2015). Reprinted from: Liu, L., Hill, K., Oza, S., Hogan, D., Cousens, S., Mathers, C., Stanton, C., Lawn, J. and Black, R. (2016). Levels and Causes of Mortality Under Age Five. In: R. Black, R. Laxminarayan, M. Temmerman and N. Walker, ed., *Reproductive, Maternal, Newborn and Child Health—Disease Control Priorities*, 3rd ed. [online] Washington DC: International Bank for Reconstruction and Development/The World Bank. Available at: http://dcp-3.org/chapter/1989/levels-and-causes-mortality-under-age-five [Accessed 14 May 2018].

Table 8.11.1 Global deaths, age-standardized death rates per 100,000, and per cent change between 2006 and 2016 for both sexes combined for all GBD causes

	All age deaths (thousands)		Age-standardized death rate (per 100,000)	
	2016	%Δ 2006–2016	2016	%Δ 2006–2016
All causes	54,698.6 (54,028.7–55,514.9)	4.1 (2.8–5.6)*	832.7 (822.7–845.0)	–16.8 (–17.9 to –15.7)*
Communicable, maternal, neonatal, and nutritional disorders	10,558.0 (10,097.7–11,143.4)	–23.9 (–26.1 to –21.6)	154.1 (147.1–163.1)	–32.3 (–34.2 to –30.3)*
HIV/AIDS and tuberculosis	2246.8 (2172.8–2314.5)	–34.7* (–36.5 to –32.9)	31.0 (30.0–31.9)	–44.8 (–46.3 to –43.3)
Tuberculosis	1213.1 (1161.5–1265.4)	–20.9 (–24.5 to –17.9)*	17.3 (16.5–18.1)	–36.1 (–39.2 to –33.8)
HIV/AIDS	1033.8 (987.4–1081.6)	–45.8 (–47.7 to –43.7)	13.7 (13.1–14.3)	–52.8 (–54.4 to –51.0)*
Disease syndromes, bacterial and viral diseases				
Neonatal sepsis and other neonatal infections	243.0 (205.0–317.7)	–11.8 (–21.9 to 1.5)	3.5 (3.0–4.6)	–11.5 (–21.7 to 1.8)
Diarrhoeal diseases	1655.9 (1244.1–2366.6)	–24.2 (–32.2 to –14.2)*	25.1 (18.8–36.0)	–35.9 (–42.4 to –27.2)
Lower respiratory infections	2377.7 (2145.6–2512.8)	–8.2 (–12.4 to –3.9)	36.8 (33.2–38.9)	–22.1 (–25.3 to –18.9)
Upper respiratory infections	2.3 (2.0–2.7)	–26.2 (–37.4 to –12.1)	0.0 (0.0–0.0)	–34.9 (–44.7 to – 22.9)*
Otitis media	1.1 (0.8–1.5)	–41.1 (–53.8 to –22.1)	0.0 (0.0–0.0)	–49.2 (–59.4 to –33.5)*
Meningitis	318.4 (262.5–408.7)	–8.4 (–19.3 to 9.5)	4.5 (3.8–5.8)	–16.3 (–26.0 to 0.2)

	All age deaths (thousands)		Age-standardized death rate (per 100,000)	
	2016	%Δ 2006–2016	2016	%Δ 2006–2016
Encephalitis	102.9 (83.9–138.4)	–1.9 (–19.1 to 20.9)	1.5 (1.2–2.0)	–14.1 (–28.3 to 5.0)
Diphtheria	1.1 (0.8–1.5)	–66.4 (–77.7 to –48.3)	0.0 (0.0–0.0)	–68.2 (–78.9 to –50.8)
Whooping cough	73.0 (38.9–126.1)	–36.3 (–63.5 to 17.6)	1.0 (0.6–1.8)	–38.3 (–64.7 to 13.7)
Tetanus	36.7 (22.2–47.2)	–59.5 (–65.5 to –52.9)*	0.5 (0.3–0.7)	–62.4 (–67.8 to –56.3)*
Measles	68.1 (25.5–146.1)	–72.5 (–77.0 to –67.8)	1.0 (0.4–2.1)	–73.3 (–77.9 to –69.1)*
Varicella and herpes zoster	12.5 (11.4–13.9)	–15.2 (–22.2 to –7.9)	0.2 (0.2–0.2)	–28.9 (–34.4 to –23.2)
Neglected tropical diseases and malaria				
Malaria	719.6 (594.6–863.0)	–25.9 (–41.4 to –6.1)*	10.2 (8.4–12.3)	–30.5 (–45.3 to –12.0)*
Leishmaniasis	13.7 (7.7–23.0)	–54.1 (–57.9 to –49.8)*	0.2 (0.1–0.3)	–58.9 (–62.1 to –55.1)*
Dengue	37.8 (10.9–52.7)	81.8 (42.3 to 132.6)*	0.5 (0.2–0.7)	60.9 (26.1 to 105.8)*
Rabies	13.3 (7.2–19.1)	–47.4 (–56.2 to –35.3)*	0.2 (0.1–0.3)	–53.4 (–61.2 to –42.7)
Ascariasis	4.9 (4.0–6.1)	–39.2 (–50.0 to –24.6)*	0.1 (0.1–0.1)	–42.9 (–53.0 to –29.3)*
Other neglected tropic diseases	27.1 (19.2–34.0)	3.1 (–23.2 to 34.3)	0.4 (0.3–0.5)	–4.8 (–28.3 to 23.4)

95% UIs are in parentheses. Asterisks denote statistically significant increases or decreases ($p<0.05$).

Source: data from *The Lancet*, Volume 390, Issue 10100, GBD 2016 Causes of Death Collaborators, Global, regional, and national age–sex specific mortality for 264 causes of death, 1980–2016: a systematic analysis for the Global Burden of Disease Study 2016, pp. 1151–210, Copyright © 2017.

MDG 6 focused on arresting the spread of infectious diseases, with a specific mention of HIV and malaria. Although the goal was not achieved in

full, some of the targets were met, with the incidence of HIV and malaria declining. Of the 36.7 million people living with HIV in 2016, more than 20 million had access to antiretroviral therapy (ART). With increasing access to ART, the incidence of HIV decreased by 40% between 1990 and 2013 (WHO 2015). Moreover, since 2010 new HIV infections have gone down by nearly 11%; 1.7 million in 2016 as compared to 1.9 million in 2010 (UNAIDS 2017).

The Joint United Nations Programme on HIV/AIDS (UNAIDS) estimates that 43% of all new HIV infections occur in Eastern and Southern Africa, and the region carries the greatest burden of the disease. However, the proportion of AIDS-related deaths in the area has declined by an estimated 42% between 2010 and 2016. Similarly, most other regions such as Asia and the Pacific, Western and Central Africa, and Middle East, and North Africa saw a decline in new HIV cases during this period.

LRIs are the third leading cause of disability-adjusted life years (DALYs) worldwide, with 60% of the LRI DALYs coming from children under-5. Diarrhoeal diseases are the fifth leading cause of global DALYs. HIV/AIDS is tenth on the list as compared to fifth in 2006, while malaria and tuberculosis rank thirteenth and eighteenth, respectively. No communicable disease is among the top ten leading causes of DALYs in high-income countries. In contrast, pneumonia, HIV/AIDS, diarrhoea, and malaria rank among the top ten causes of death and DALYs in LMICs (Murray et al. 2012; GBD 2016 DALYs and HALE Collaborators 2017). Table 8.11.2 shows the estimated number of deaths in under-5 children by cause while Table 8.11.3 shows a comparison of infectious disease mortality in children in 2000 and 2015.

Table 8.11.2 Estimated numbers of deaths by cause in 2015 in children under 5

	Estimated number (millions)
Neonates aged 0–27 days	
Preterm birth complications	0.943 (0.832–1.066)
Intrapartum-related complications	0.631 (0.550–0.723)
Sepsis or meningitis	0.402 (0.280–0.522)
Congenital abnormalities	0.305 (0.260–0.382)
Pneumonia[a]	0.161 (0.111–0.239)
Tetanus	0.034 (0.018–0.084)
Diarrhoea[b]	0.018 (0.010–0.070)
Other conditions	0.187 (0.142–0.240)
Children aged 1–59 months	
Pneumonia[a]	0.760 (0.651–0.943)
Diarrhoea[b]	0.509 (0.401–0.661)
Injuries	0.326 (0.272–0.410)
Malaria	0.306 (0.225–0.452)
Congenital abnormalities	0.207 (0.165–0.259)
Meningitis	0.115 (0.091–0.162)
Preterm birth complications	0.112 (0.061–0.168)
AIDS	0.086 (0.076–0.101)
Measles	0.074 (0.038–0.268)
Intrapartum-related complications	0.058 (0.028–0.092)
Pertussis	0.054 (0.053–0.060)

Uncertainty range (UR) is defined as the 2.5–97.5 centile. Other conditions among children aged 1–59 months included causes originated during the perinatal period, cancer, severe malnutrition, and other specified causes. Intrapartum-related events were formerly referred to as ‘birth asphyxia’.

Source: data from Liu L., et al. (2016) ‘Global, regional, and national causes of under-5 mortality in 2000–15: an updated systematic analysis with implications for the Sustainable Development Goals’ *The*

Lancet 388(10063) pp. 3027–35.

Table 8.11.3 Estimated numbers of deaths caused by infections in neonates and children aged 1–59 months, at the global level in 2000–2015

	2000	2015*
Neonates		
Pneumonia	0.296 million	0.161 million
Sepsis/meningitis	0.532 million	0.402 million
Tetanus	0.161 million	0.034 million
Diarrhoea	0.04 million	0.018 million
Children (1–59 months)		
Pneumonia	1.4 million	0.76 million
Meningitis	0.284 million	0.115 million
Diarrhoea	1.2 million	0.509 million
Measles	0.485 million	0.074 million
Malaria	0.725 million	0.306 million
AIDS	0.22 million	0.086 million
Pertussis	0.056 million	0.054 million

* Neonatal sepsis/meningitis, neonatal pneumonia, neonatal tetanus, childhood pneumonia, childhood diarrhoea, meningitis, diarrhoea, measles, malaria, and AIDS all showed >30% decline in mortality rates from 2000 to 2015

Source: data from Liu L., et al. (2016) 'Global, regional, and national causes of under-5 mortality in 2000–15: an updated systematic analysis with implications for the Sustainable Development Goals' *The Lancet* 388(10063) pp. 3027–35.

Factors responsible for the decline in infectious diseases in industrialized nations during the last century include improved nutrition, safer food, and water supplies, improved hygiene and sanitation, the introduction of antimicrobial agents, and immunizations, all of which resulted in decreased host susceptibility and reductions in disease transmission (Nakatani 2016). While there have been substantial reductions in morbidity and mortality due

to communicable diseases in the last century, there remain significant gaps in child and adult mortality between rich and poor countries.

Apart from the immediate causes of infections in childhood, a number of determinants contribute to the high burden of infectious diseases in LMICs. These include several distal determinants such as income, social status, and education, which work through an intermediate level of environmental and behavioural risk factors. These risk factors, in turn, lead to the proximal causes of death (nearer in time to the terminal event), such as undernutrition, infectious diseases, and injury (WHO 2008). The major social determinants affecting the mortality and morbidity of young children include poverty, indoor air pollution, malnutrition, inequity, lack of education, inadequate health-related social behaviours and practices, and other social and cultural determinants of health.

Specific disease categories

As described in the previous section, a limited number of infectious diseases are responsible for a substantial proportion of the total global burden of morbidity and mortality, especially in resource-limited areas of the world. This section will provide an overview of the major types of infectious diseases responsible for the bulk of acute and chronic morbidity and mortality of children and adults worldwide. Detailed information on approaches for the prevention and control of these diseases is available in Chapter 11.3.

Neonatal sepsis

Sepsis and meningitis are significant causes of morbidity and mortality in newborns, particularly in preterm and low birth weight (LBW) infants, and are estimated to cause 420,000 deaths annually; nearly 16% of neonatal mortality (Laxminarayan and Bhutta 2016). Neonatal sepsis may be defined using clinical criteria (Table 8.11.4) and/or microbiological testing, by positive blood and/or cerebrospinal fluid (CSF) cultures. It may also be classified according to the time of disease onset: early onset sepsis (EOS) and late onset sepsis (LOS). Meningitis can occur as a part of sepsis in both the EOS and LOS time periods or as focal infection with LOS disease. The distinction has clinical relevance, as EOS disease is mainly due to bacteria acquired before and during delivery, and LOS disease to bacteria acquired

after delivery (nosocomial or community sources). In the literature, however, there is little consensus as to what age limits apply, with EOS ranging from 48 hours to 6 days after delivery. Notwithstanding this limitation, the most commonly used cut off is <7 days (EOS) versus ≥7 days (LOS).

Table 8.11.4 Clinical criteria for diagnosis of neonatal sepsis and meningitis

Sepsis	Meningitis
Symptoms [a] Convulsions Inability to feed Unconsciousness Lethargy Fever (>37.7°C or feels hot) Hypothermia (<35.5°C or feels cold)	**General signs** • Drowsiness • Reduced feeding • Unconsciousness • Lethargy • High-pitched cry • Apnoea
Signs Severe chest indrawing Reduced movement Crepitations Cyanosis	**Specific signs** • Convulsions • Bulging fontanelle

[a] The more symptoms a neonate has, the higher the probability of the disease.

Global and regional epidemiology

An estimated 2.6 million neonates died in 2016, 46% of the total under-5 deaths that year. Among newborn deaths, major causes included preterm birth complications (35%), intrapartum-related complications, previously labelled as birth asphyxia (24%), whereas sepsis or meningitis accounted for 15% of neonatal deaths (UN IGME 2017). Despite major progress in neonatal care, overall case fatality rates from sepsis range from 2 to as high as 50%.

Unfortunately, hospitals in developing countries are also hot beds of infection transmission, especially multidrug-resistant nosocomial infections. Reported rates of neonatal sepsis vary from 6.5 to 38 per 1,000 live hospital-

born babies and the rates of bloodstream infection range from 1.7 to 33 per 1,000 live births, with rates in Africa clustering around 20 and in South Asia around 15 per 1,000 live births. According to estimates, nearly 40% of neonatal mortality in LMICs is associated with hospital-acquired infections (Zaidi et al. 2005). Factors responsible for hospital-acquired neonatal sepsis include lack of aseptic technique for procedures, inadequate hand hygiene and glove use, deficient sterilization and disinfection practices, overuse of invasive devices, re-use of disposable supplies without adequate sterilization, re-use of single-use vials, overcrowded and understaffed labour and delivery rooms, unhygienic bathing and skin care, contaminated bottle feedings, inappropriate and prolonged use of antibiotics, and lack of effective infection control practices.

Aetiology

In general, Gram-negative pathogens are responsible for a substantial proportion of EOS. In contrast to industrialized countries where group B streptococci (GBS) are common causes of neonatal sepsis, *Klebsiella pneumoniae* is an important aetiology in developing countries. LOS is most commonly due to *E. coli*, *S. aureus*, *S. pyogenes*, *S. pneumoniae*, and *Salmonella* spp. The organisms causing neonatal sepsis and meningitis in LMICs are listed in Table 8.11.5. More recently the results of the Aetiology of Neonatal Infection in South Asia (ANISA) study were published; these showed a similar pattern as previously for LMICs (Table 8.11.6) (Saha et al. 2018).

Table 8.11.5 Distribution of all isolates by age of onset

Organism isolated	N	<7 days of life % (95% CI)	N	8–59 days of life % (95% CI)	N	60–90 days of life % (95% CI)
Staphylococcus aureus	47	14.6 (10.8–18.5)	276	14.3 (12.7–15.9)	1	2.3 (0.0–6.7)
Group A streptococci/*Streptococcus pyogenes*	4	1.2 (0.0–2.4)	55	2.8 (2.1–3.5)	8	18.2 (6.8–29.6)
Group B streptococci (GBS)	22	6.8 (4.1–9.6)	42	2.2 (1.6–2.9)	0	0.0 (–)
Group D streptococci/enterococci	6	1.9 (0.4–3.4)	15	0.8 (0.4–1.2)	0	0.0 (–)
Streptococcus pneumoniae	14	4.3 (2.1–6.5)	111	5.7 (4.7–6.7)	14	31.8 (18.0–45.6)
Other/unspecified *Streptococcus* spp.	25	7.7 (4.8–10.6)	51	2.6 (1.9–3.3)	0	0.0 (–)
Other/unspecified Gram–positives[a]	3	0.9 (0.0–1.9)	117	6.0 (4.9–7.1)	0	0.0 (–)
All Gram–positives	**121**	37.5 **(32.2–42.8)**	**667**	34.5 **(32.4–36.6)**	**23**	52.3 **(37.5–67.1)**
Klebsiella pneumoniae	24	7.4 (4.6–10.3)	235	12.2 (10.7–13.7)	2	4.5 (0.0–10.6)
Other/unspecified *Klebsiella* spp.	10	3.1 (1.2–5.0)	41	2.1 (1.5–2.7)	0	0.0 (–)
Escherichia coli	50	15.5 (11.6–19.5)	337	17.4 (15.7–19.1)	1	2.3 (0.0–6.7)
Pseudomonas spp.	23	7.1 (4.3–9.9)	180	9.3 (8.0–10.6)	1	2.3 (0.0–6.7)
Enterobacter spp.	13	4.0 (1.9–6.1)	71	3.7 (2.7–4.5)	0	0.0 (–)
Serratia spp.	0	0.0 (–)	40	2.1 (1.5–2.7)	1	2.3 (0.0–6.7)
Proteus spp.	6	1.9 (0.4–3.4)	7	0.4 (0.1–0.7)	0	0.0 (–)
Salmonella spp.	1	0.3 (0.0–0.9)	33	1.7 (1.1–2.3)	6	13.6 (3.5–23.7)
Haemophilus influenzae	3	0.9 (0.0–1.9)	37	1.9 (1.3–2.5)	4	9.1 (0.6–17.6)

Organism isolated	N	<7 days of life % (95% CI)	N	8–59 days of life % (95% CI)	N	60–90 days of life % (95% CI)
Neisseria meningitidis	0	0.0 (–)	14	0.7 (0.3–1.1)	0	0.0 (–)
Acinetobacter spp.	20	6.2 (3.6–8.8)	116	6.0 (4.9–7.1)	3	6.8 (0.0–14.2)
Other/unspecified Gram-negatives[b]	47	14.6 (10.8–18.5)	69	3.6 (2.8–4.4)	1	2.3 (0.0–6.7)
All Gram-negatives	**197**	61.0 **(55.7–66.3)**	**1,180**	**61.0 (58.8–63.2)**	**19**	43.2 **(28.6–57.8)**
Non-stated/undetermined	5	1.5 (0.2–2.8)	87	4.5 (3.6–5.4)	2	4.5 (0.0–10.6)
Total	**323**	100.0 **(n/a)**	**1934**	100.0 **(n/a)**	**44**	100.0 **(n/a)**

[a] Includes data for *Aerococcus* spp., *Bacillus* spp., and others.

[b] Includes data for *Citrobacter* spp., *Moraxella* spp., *Shigella* spp., *Aeromonas* spp. and others. Data were also extracted for coagulase-negative staphylococci: ≤ 7 days of life—3 isolates; 7–59 days of life—880 isolates; 60–90 days of life—0 isolates.

Adapted from Waters, D. et al., Aetiology of community-acquired neonatal sepsis in low- and middle-income countries, *Journal of Global Health*, Volume 1, Issue 2, pp. 154–70, Copyright © 2011, reproduced under the Creative Commons Attribution License 3.0.

Table 8.11.6 Estimated aetiology-specific incidence, based on results of a partial latent class model (Aetiology of Neonatal Infections in South Asia [ANISA] Study)

Incidence per 1,000 live births	
Pathogen class	**Overall per 1,000 live births (95% CI)**[4]
Bacteria[1]	
All bacteria tested	13.24 (11.18, 15.63)
Bordetella spp.	0.66 (0.34, 1.32)
Chlamydia pneumoniae	0.07 (0.04, 0.15)
Chlamydia trachomatis	0.21 (0.12, 0.41)
Escherichia coli	1.43 (0.88, 2.18)
Group A *streptococcus*	0.25 (0.23, 0.29)
Group B *streptococcus*	0.93 (0.54, 1.43)
pan–*Haemophilus influenza*	0.38 (0.21, 0.79)
Klebsiella pneumoniae	1.49 (0.97, 2.06)
Mycoplasma pneumoniae	0 (0, 0)
Neisseria meningitidis	0.15 (0.1, 0.26)
Pseudomonas aeruginosa	0.24 (0.11, 0.52)
Salmonella spp.	1.07 (0.44, 2.11)
Staphylococcus aureus	0.88 (0.53, 1.4)
Streptococcus pneumoniae	0.96 (0.58, 1.64)
Ureaplasma spp.	2.38 (1.62, 3.17)
Other blood culture[2]	2.14 (1.71, 2.6)
Viruses[1]	
All viruses tested	10.13 (9.04, 11.55)
Adenovirus	0.42 (0.22, 0.77)

Incidence per 1,000 live births	
Pathogen class	**Overall per 1,000 live births (95% CI)[4]**
Cytomegalovirus	0.69 (0.3, 1.27)
Enterovirus/rhinovirus	1.12 (0.69, 1.96)
Human metapneumovirus	0.34 (0.22, 0.61)
Human parechovirus	0.14 (0.07, 0.27)
Influenza type A	0.42 (0.2, 0.78)
Influenza type B	0.44 (0.31, 0.76)
Parainfluenza 1	0.41 (0.26, 0.72)
Parainfluenza 2	0.06 (0.03, 0.13)
Parainfluenza 3	0.58 (0.37, 1.01)
Respiratory syncytial virus	5.39 (4.84, 6.31)
Rubella	0.12 (0.07, 0.23)
Other/None[3]	60.42 (57.69, 62.87)

[1] Rates are based on the estimated pathogen prevalence among pSBI episodes, derived from partial latent class model (see text for methods) that evaluated 5,253 pSBI episodes identified from a population of 63,114 registered live births.

[2] The pathogen class 'Other Blood Culture' includes all bacteria that grew on blood culture but did not have an associated assay on the ANISA molecular diagnostic panel; this class was not estimated directly from the model but indirectly by calculating the product of the number of blood culture isolates in the 'Other Blood Culture' class and the average estimated etiologic proportion attributed to pathogens with multiple tests that included blood culture.

[3] The pathogen class 'Other/None' includes any pSBI episode that was not attributed by the partial latent class model (see text for details) to one of the evaluated ANISA pathogen classes.

[4] Aggregated incidence was calculated by weighting site-specific incidence estimates by average registered live births per month.

Source: data from: Saha, S., et al (2018). Causes and incidence of community-acquired serious infections among young children in south Asia (ANISA): an observational cohort study. *Lancet*. [online] Available at: https://www.thelancet.com/pdfs/journals/lancet/PIIS0140-6736(18)31127-9.pdf [Accessed 7 Jul. 2018].

Evidence-based interventions to address neonatal infections

The lack of success in achieving the fourth MDG is largely attributable to the relatively slow reduction in neonatal mortality. Newborn survival, now a focus of global and regional agendas, is very visible in the Sustainable

Development Goals (SDGs). The SDGs call for a reduction of neonatal mortality to at least 12 per 1,000 live births, by 2030. Real progress in saving newborns will depend upon provision of a good mix of preventive and therapeutic services. Scaling up proven interventions addressing maternal risk factors and underlying causes could reduce neonatal infections by 84% (Bhutta et al. 2014).

Medical treatment

Reaching and treating sick newborn infants promptly is critical to survival. Normally, health facilities use a combination of ampicillin and gentamicin, with or without vancomycin, for cases of neonatal sepsis (Table 8.11.7). If there is a high risk of *S. aureus*, cloxacillin with gentamicin is a recommended option. However, increasing antibiotic resistance among common organisms causing neonatal sepsis in both community and hospital settings presents a challenge to the selection of appropriate antibiotics. Case management of neonatal infections is mainly provided through child health services, both in facilities and through family-community care. Similarly, integrated management of childhood illnesses (IMCI) has been widely implemented as the main approach for addressing child health in health systems. However, IMCI management guidelines rely on the sick child being brought to a health facility. The modification of IMCI to include the neonatal period (IMNCI) and recent expansion to community settings has now been included as a public health strategy in many countries including India.

Following the demonstration of significant reduction in neonatal mortality with the use of oral co-trimoxazole and injectable gentamicin by community health workers (CHWs), this strategy could be employed in circumstances where referral is difficult. Moreover, the African Neonatal Sepsis Trial (AFRINEST) showed that simplified antibiotic regimens, of injectable gentamicin, oral amoxicillin and, in certain cases, injectable procaine benzylpenicillin, given at home or a primary healthcare facility, for possible serious bacterial infections (PSBI), were comparable in effectiveness to parenteral WHO-recommended antibiotic regimens (Tshefu et al. 2015). The current WHO recommendation for treating PSBIs is 7-day treatment with parenteral benzylpenicillin with gentamicin or ampicillin with gentamicin at a hospital or treatment by frontline health worker with intramuscular gentamicin and oral amoxicillin (WHO 2015). Currently in some health systems, outreach health workers, community nutrition, and child

development workers are being trained to visit all mothers and neonates at home two to three times within the first 10 days, starting soon after birth, to provide home-based preventive care/health promotion and to detect neonates with sickness requiring referral.

Table 8.11.7 Antimicrobial therapy of neonatal meningitis and sepsis

Patient group	Likely aetiology	Antimicrobial choice	
		Developed countries	**Developing countries**
Sepsis			
Immunocompetent neonates	***Developed countries:***	Ampicillin or penicillin plus an aminoglycoside	Ampicillin or penicillin plus gentamicin
	Group B streptococci, *E. coli*		
	Developing countries:		Or co-trimoxazole plus gentamicin
	Klebsiella spp.		
	Pseudomonas spp.		
	Salmonella spp.		
Meningitis			
Immunocompetent neonates (age <3 months)	***Developed countries:***	Ampicillin plus ceftriaxone or cefotaxime	Ampicillin plus gentamicin
	Group B streptococci		
	E.*coli*		
	L.*monocytogenes*		
	Developing countries:		
	S.*pneumoniae*		
	E.*coli*		
Immunodeficient	Gram-negative bacilli	Ampicillin plus ceftazidime	
	L.*monocytogenes*		

Preventive measures

Preventive interventions need to bridge the continuum of care from pregnancy, through childbirth and the neonatal period, and beyond. Lack of positive health-related behaviour, education, and poverty are underlying causes of many neonatal deaths, either through increasing the prevalence of risk factors such as maternal infection or by reducing access to effective care.

Attempts to reduce the proportion of LBW births at the population level have had limited success. Many deaths in preterm babies and in those born at term with LBW can be prevented with extra attention to warmth, feeding, and prevention or prompt treatment of infections. In LMICs, 90% of mothers deliver babies at home without a skilled health professional present; globally 50 million births occur at home (Lawn et al. 2014). Simple low-cost interventions, notably tetanus toxoid vaccination, provision of clean delivery kits, exclusive breastfeeding, counselling for birth preparedness, and breastfeeding promotion through peer counsellors and women's groups, have been shown to reduce newborn morbidity and mortality (Bhutta et al. 2014). Postnatally, kangaroo mother care for LBW infants, topical emollient therapy, hand washing, attention to environmental hygiene and sterilization, and antibiotics for neonatal infections, are additional health system measures. Alcohol-based antiseptics for hand hygiene are an appealing innovation because of their efficacy in reducing hand contamination and their ease of use, especially when sinks and supplies for hand washing are limited.

Early initiation of breastfeeding affects neonatal health outcomes through several mechanisms. Mothers who suckle their offspring shortly after birth have a greater chance of successfully establishing and sustaining breastfeeding throughout infancy and provide a variety of immune and non-immune components that accelerate intestinal maturation, resistance to infection, and epithelial recovery from infection. Prelacteal feeding with non-human milk antigens may disrupt normal physiological gut priming. WHO currently recommends dry cord care for newborns, and application of chlorhexidine for those born at home in areas with high mortality risk (WHO 2013). Chlorhexidine application to the cord stump in such settings can reduce omphalitis by 27% and provide a 23% decrease in risk of neonatal mortality (Imdad et al. 2013). A closely related issue is the need for general skin care. Topical emollient therapy has been shown to decrease risk of nosocomial infections by 50% and neonatal mortality by 27% (Salam et al. 2013).

At least two doses of tetanus toxoid should be given during pregnancy so that protective antibodies can be transferred to the foetus, to protect it from

neonatal tetanus. Women with a history of prolonged rupture of membranes, especially if preterm, should be given prophylactic antibiotics. Maternal antibiotic therapy in this situation is effective in reducing maternal and neonatal infection-related morbidities. A detailed review of antenatal, intrapartum, and postnatal interventions noted that antibiotics for preterm premature rupture of membranes (PPROM) could reduce infection-related neonatal mortality and preterm-related complications by 2% each (Bhutta et al. 2014). A domiciliary cadre of trained birth attendants potentially can be trained to recognize PPROM and provide referral and, possibly, initial antimicrobial therapy.

An important aspect of prevention of neonatal infection in developed countries relates to GBS disease. The joint guidelines developed and implemented in the United States have led to a significant reduction in the burden of disease. Most newborns born to mothers with risk of GBS colonization undergo a full diagnostic evaluation and empiric therapy.

In recent years, the importance of hospital-acquired infections in newborn infants, frequently with multidrug-resistant organisms, has been recognized. It is imperative that preventive strategies such as hand washing, preventing unnecessary hospital stay and environmental control are strictly enforced.

Gastrointestinal tract infections

Diarrhoea, the most common manifestation of intestinal tract infections, is a leading cause of preventable death in most developing countries where its greatest impact is seen in infants and children. Infectious diarrhoea may be accompanied by numerous complications (Table 8.11.8).

Table 8.11.8 Complications of gastrointestinal infections

Complication	Causative pathogens
Dehydration	*Vibrio cholerae*, *Cryptosporidium parvum* (especially in immunocompromised hosts), enterotoxigenic *Escherichia coli* (ETEC), rotavirus
Severe vomiting	Staphylococcal food poisoning, norovirus, rotavirus
Haemorrhagic colitis	*Campylobacter jejuni*, enterohaemorrhagic *E. coli* (EHEC), *Salmonella*, *Shigella*, *V. parahaemolyticus*
Toxic megacolon, intestinal perforation	EHEC, *Shigella*, *C. jejuni* (rare), *Clostridium difficile* (rare), *Salmonella* (rare), *Yersinia* (rare)
Haemolytic uraemic syndrome (HUS), thrombotic thrombocytopenic purpura (TTP)	EHEC, *Shigella*, *C. jejuni* (rare)
Reactive arthritis	C.*jejuni*, *Shigella*, *Salmonella*, *Yersinia*
Malabsorption/malnutrition	*Cyclospora cayetanensis*, *Giardia lamblia*, *C. parvum* or *C. hominis* (especially immunocompromised hosts)
Distant metastatic infection	*Salmonella*, *C. jejuni* (rare), *Yersinia* (rare)
Guillain–Barré syndrome	C.*jejuni* (rare)

Invasive diarrhoea refers to diarrhoea caused by bacterial pathogens that invade the bowel mucosa, causing inflammation and tissue damage and may cause blood in stools (bloody diarrhoea). Invasive diarrhoea accounts for approximately 10% of diarrhoeal episodes in children under 5 years of age and approximately 15% of diarrhoea-associated deaths in this age group worldwide (GBD Diarrhoeal Diseases Collaborators 2017). Although less frequent, bloody diarrhoea generally lasts longer, is associated with higher risk of complications and case fatality rates, and is more likely to adversely affect a child's growth. The burden of disease is mainly in younger age groups; 72% of deaths from diarrhoea and 81% of deaths from pneumonia happen in children younger than 2 years.

Global and regional epidemiology

The aetiology and severity of gastrointestinal infections are determined by several epidemiological factors. Young children and the elderly are at greatest

risk for more severe disease and complications. Childhood malnutrition and unsafe water, sanitation, and hygiene are considered the two leading risk factors for diarrhoea. The presence of underlying medical conditions, especially those that compromise immunity, greatly enhances the risk of acquiring an infection and its ultimate severity. Poor sanitation, inadequate water supplies, and increasing globalization of food transport systems all predispose to the development of large epidemics of food- and water-borne outbreaks of gastrointestinal disease. Seasonal or cyclic weather variations also influence the epidemiology of diarrhoeal disease and food poisoning (see Hamer and Gorbach 2010).

Several reviews have evaluated diarrhoea burden and mortality rates. A review carried out more than two decades ago estimated that 4.6 million children died annually from diarrhoea. Kosek et al. updated these estimates by reviewing 60 studies of diarrhoea morbidity and mortality published between 1990 and 2000 (Kosek et al. 2003). They concluded that diarrhoea accounted for 21% of all deaths at less than 5 years of age and caused 2.5 million deaths per year, although morbidity rates remained relatively unchanged. More recently, the GBD study estimates that diarrhoea incidence and mortality fell by nearly 6% and 21%, respectively, between 2005 and 2015. However, there were 2.39 billion episodes of diarrhoea in 2015 and it killed more than 1.3 million people in 2015, with nearly 0.5 million of these deaths occurring in children under-5 (GBD Diarrhoeal Diseases Collaborators 2017). Accounting for 8% of under-5 deaths, persistent high rates of diarrhoea morbidity may have significant long-term effects on linear growth and physical and cognitive function in children. Fortunately, the leading risk factors for diarrhoea are on the decline, with diarrhoea DALYs attributable to unsafe water and poor sanitation having fallen by 13.4% and those secondary to childhood undernutrition by 10% between 2005 and 2015. Figure 8.11.3 shows the 2015 country-specific global trends for diarrhoea mortality in children under-5.

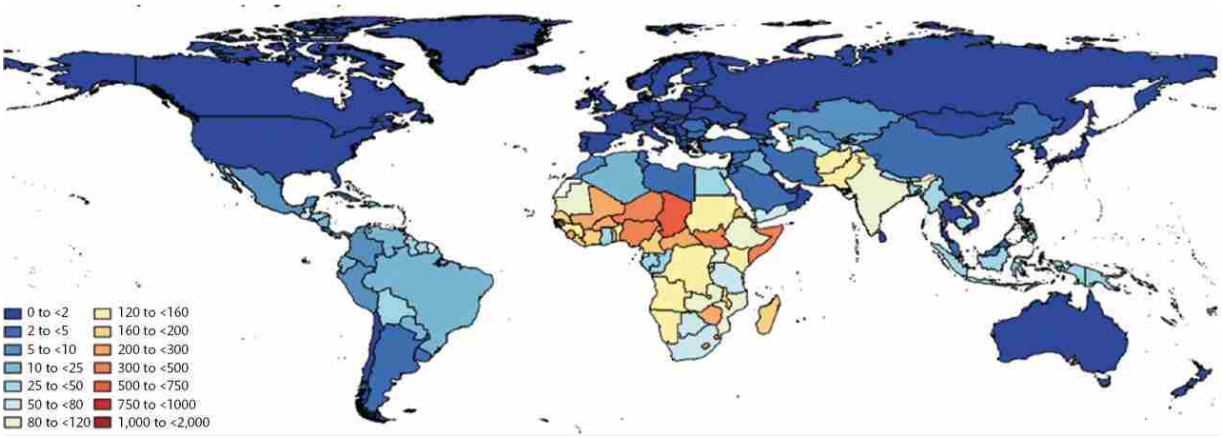


Fig. 8.11.3 Diarrhoea mortality rate in children younger than 5 years in 2015.
Reprinted from GBD Diarrhoeal Diseases Collaborators (2017). Estimates of global, regional, and national morbidity, mortality, and aetiologies of diarrhoeal diseases: a systematic analysis for the Global Burden of Disease Study 2015. The Lancet Infectious Diseases, 17(9), pp.909–48 under the CC BY 4.0 license.

Aetiology

Common aetiologies of non-inflammatory diarrhoea include enterotoxigenic *E. coli* (ETEC) and other strains such as enteroaggregative (EAEC), diffusely adhering, and enteropathogenic *E. coli*, *Vibrio cholerae*; non-01 choleras such as *V. vulnificus*; parasites including *Giardia lamblia*, *Cryptosporidium parvum*, and microsporidia; and several different virus species including rotavirus, noroviruses, and astroviruses. Acute inflammatory diarrhoea is the result of infection with bacterial enteropathogens such as *Shigella*, *Campylobacter*, *Salmonella* spp., enterohaemorrhagic *E. coli* (EHEC), *V. parahaemolyticus*, and *Clostridium difficile*. Among the parasites, *Entamoeba histolytica* is the most common cause of dysenteric illness although *Balantidium coli*, *Schistosoma mansoni*, *S. japonicum*, *Trichuris trichiura*, hookworms, and *Trichinella spiralis* can all cause bloody, mucoid diarrhoea. In 2015, rotavirus, *Shigella*, and *Salmonella* were responsible for the most diarrhoeal deaths. While aetiology of diarrhoeal mortality was similar in the under-5 age group, the top three causes of diarrhoeal deaths were rotavirus, *Cryptosporidium* and *Shigella* (GBD Diarrhoeal Diseases Collaborators 2017). The MAL-ED study conducted in eight LMICs reported norovirus, rotavirus, and *Campylobacter* spp. among the commonest diarrhoea-causing pathogens in children under 2 years (Platts-Mills et al. 2015). Estimates of

rotavirus and cholera as major diarrhoea aetiologies associated with increased risk of mortality are shown in Table 8.11.9.

Table 8.11.9 Diarrhoea episodes and deaths by WHO region

Several diarrhoea episodes	African region	American region	Eastern Mediterranean region	European region	Southeast Asian region	Western Pacific region	World
Rotavirus							
Episodes (×103)	2,557 (1,901–3,026)	1,129 (833–1,330)	1,344 (990–1,576)	660 (488–781)	2,382 (1,753–790)	1,781 (1,312–2,090)	9,853 (7,277–11,593)
Proportion of all severe episodes (%)	26.8%	23.4%	31.3%	25.9%	25.5%	32.6%	27.3%
Vibrio cholerae							
Episodes (×103)	38 (28–45)	–	–	–	416 (306–488)	2 (2–3)	456 (336–536)
Proportion of all severe episodes (%)	0.4%	–	–	–	4.5%	0.04%	1.3%
Diarrhoea deaths							
Rotavirus							
Deaths (×103)	95 (52.5–151.3)	2.6 (1.9–5.6)	30.2 (20.1–48.0)	1.6 (1.1–2.8)	58.1 (47.7–74.7)	5.5 (2.0–7.7)	192.7 (133.1–284.4)
Proportion of diarrhoea deaths (%)	26.8%	23.4%	31.3%	25.9%	25.5%	32.6%	27.1%
V.*cholera*							
Episodes (×103)	1.4 (0.8–2.3)	–	–	–	10.2 (8.4–13.2)	0.1 (0.02–0.1)	11.7 (7.9–16.8)

Several diarrhoea episodes	African region	American region	Eastern Mediterranean region	European region	Southeast Asian region	Western Pacific region	World
Proportion of diarrhoea deaths (%)	0.4%	–	–	–	4.5%	0.04%	1.6%

Source: data from *The Lancet*, Volume 381, Issue 9875, Fischer Walker, C.L. et al., Global burden of childhood pneumonia and diarrhoea, pp. 1405–16, Copyright © 2013.

Infectious microorganisms in contaminated food and drink are the main source of travellers' diarrhoea. High-risk foods include uncooked vegetables, salsa, meat, and seafood. Tap water, ice, unpasteurized milk and dairy products, salads, and unpeeled fruits are also associated with an increased risk. Although many different pathogens may be responsible, the leading culprits are various forms of *E. coli*, particularly ETEC and EAEC. Food poisoning is most commonly caused by the consumption of food contaminated with bacteria or bacterial toxins. Food poisoning can also be due to parasites (e.g. trichinosis), viruses (e.g. hepatitis A), and other toxins (e.g. *Amanita* mushrooms). The most well-recognized causes of bacterial food poisoning are the following: *Clostridium perfringens*, *S. aureus*, *Vibrio* spp. (including *V. cholerae* and *V. parahaemolyticus*), *Bacillus cereus*, *Salmonella* spp., *C. botulinum*, *Shigella* spp., toxigenic *E. coli* (ETEC and EHEC), and certain species of *Campylobacter*, *Yersinia*, *Listeria*, and *Aeromonas*.

Issues in presentation and diagnosis

Gastrointestinal infections usually result in three principal syndromes: non-inflammatory diarrhoea, inflammatory diarrhoea, and systemic disease. Non-inflammatory diarrhoea primarily involves the small intestine, whereas inflammatory diarrhoea predominantly affects the colon. The location of infection influences the clinical characteristics and certain diagnostic features of the diarrhoeal disease (Table 8.11.10). Thus, organisms that target the small intestine tend to produce watery, potentially dehydrating diarrhoea, while those infecting the large intestine cause bloody mucoid diarrhoea associated with tenesmus.

Table 8.11.10 Clinical features and aetiologies of diarrhoeal diseases

Feature	Site of infection	
	Small intestine	**Large intestine**
Pathogens	• *Escherichia coli* (enteropathogenic *E. coli*, enterotoxigenic *E. coli*) • *Cryptosporidium parvum* • *Giardia lamblia* • Norovirus • Rotavirus • *Vibrio cholerae*	• *E. coli* (EIEC, EHEC) • *Entamoeba histolytica* • *Shigella* spp.
Location of pain	Mid-abdomen	Lower abdomen, rectum
Volume of stool	Large	Small
Blood in stool	Rare	Common
Faecal leucocytes	Rare	Common (except in amoebiasis)
Sigmoidoscopy	Normal	Mucosal ulcers, haemorrhagic foci, friable mucosa

Acute watery diarrhoea can be rapidly dehydrating, with stool losses of 250 mL/kg/day or more, a quantity that quickly exceeds total plasma and interstitial fluid volumes, and is incompatible with life unless aggressive fluid therapy can be provided. Such dramatic dehydration is usually due to rotavirus, ETEC, or *V. cholerae*, and it is most dangerous in the very young.

Persistent diarrhoea is defined as diarrhoea lasting 14 days or longer, is manifested by malabsorption, nutrient losses, and wasting, and is typically associated with malnutrition. Although persistent diarrhoea accounts for 8–20% of diarrhoea episodes, it is associated with a disproportionately increased risk of death. Persistent diarrhoea more commonly follows an episode of bloody diarrhoea and is associated with a tenfold higher risk of mortality.

Inflammatory diarrhoea is a manifestation of invasive intestinal infection that is associated with intestinal damage and nutritional deterioration, often with systemic manifestations including fever. Although clinicians often use the term bloody diarrhoea interchangeably with dysentery, the latter is a

syndrome consisting of the frequent passage of characteristic, small-volume, bloody mucoid stools, abdominal cramps, and tenesmus. Bloody diarrhoea is usually caused by *Shigella* or some strains of *E. coli*. Agents that cause bloody diarrhoea or dysentery can also provoke a form of diarrhoea that does not present clinically with visible blood in the stool, although mucosal damage and inflammation are present and faecal blood and white blood cells are usually detectable by microscopy.

Because of the significant morbidity and costs associated with infectious diarrhoea, making a specific laboratory diagnosis can be useful epidemiologically, diagnostically, and therapeutically. A definitive diagnosis is achieved mainly through study of faecal specimens, using bacteriological culture, viral culture, or direct electron microscopy for viral particles, and identification of microbial antigens (viruses, bacteria, parasites, or toxins). DNA probes, polymerase chain reaction (PCR), and immunodiagnostic tests can now be used to identify several pathogens in stool specimens.

Evidence-based interventions

Increased use of oral rehydration therapy, improved nutrition, increased breastfeeding, better supplemental feeding, and improvement in hygiene and sanitation are believed to have contributed to the decline in morbidity and mortality of diarrhoea. Syndromic diagnosis provides important clues to optimal management and is both programmatically and epidemiologically relevant. The correct treatment of diarrhoea requires mothers to recognize the problem and seek medical care promptly, and health workers to give oral rehydration solution (ORS) or other fluids to prevent or treat dehydration, dispense an appropriate antibiotic when needed, provide advice on appropriate feeding, and provide follow-up, especially for children at increased risk of serious morbidity or death. Low-osmolarity ORS and zinc supplementation (10–20 mg/day) have led to significantly improved diarrhoea outcomes. A recent series in *The Lancet* (Bhutta et al. 2013b) highlighted the role of 15 core interventions to address childhood pneumonia and diarrhoea, of which several are also common nutrition interventions underscored recently (Table 8.11.11) (Bhutta et al. 2013a).

Medical treatment

Since the most devastating consequences of acute infectious diarrhoea result from fluid losses, the major goal of treatment is the replacement of fluid and electrolytes. While the intravenous route of administration has been traditionally used, ORS has been shown to be effective and logistically more practical and less costly to administer, especially in developing countries. ORS is the treatment of choice for mild-to-moderate diarrhoea in both children and adults, as long as vomiting is not a major feature of the gastrointestinal infection. ORS can also be used in severely dehydrated patients after initial parenteral rehydration.

Table 8.11.11 Interventions for prevention and treatment of diarrhoea, pneumonia, and undernutrition

Diarrhoea	Pneumonia	Nutrition
• Breastfeeding promotion • Improved water source, sanitation, and hygiene • Preventive vitamin A supplementation • Preventive zinc supplementation • Rotavirus vaccine • ORS • Zinc—for treatment of diarrhoea • Antibiotics for dysentery	• Breastfeeding promotion • Improved water source, sanitation, and hygiene • Preventive vitamin A supplementation • Preventive zinc supplementation • Hib vaccine • Pneumococcal vaccine • Case management of neonatal infections • Oral antibiotics: case management of pneumonia in children	• Breastfeeding promotion • Improved water source, sanitation, and hygiene • Preventive vitamin A supplementation • Preventive zinc supplementation • Periconceptional folic acid supplementation or fortification • Appropriate complementary feeding • Maternal balanced energy protein supplementation • Maternal calcium supplementation • Multiple micronutrient supplementation in pregnancy • Management of moderate and severe acute malnutrition

Source: data from *The Lancet*, Volume 382, Issue 9890, Bhutta, Z.A. et al., Evidence based interventions for improvement of maternal and child nutrition: what can be done and at what cost?, pp. 452–77, Copyright © 2013 Elsevier Ltd All rights reserved.

Although there is no doubt about the value of ORS in treating dehydrating diarrhoea, the optimal sodium concentration of the solution remains in dispute, particularly in regard to the treatment of mild-to-moderate diarrhoea.

The high concentration of sodium (90 mmol) in the standard WHO ORS formulation may cause hypernatraemia and even seizures in children with non-cholera watery diarrhoea. Consequently, lower concentrations of sodium and a reduced osmolarity solution have been found to be effective for rehydration and not to be associated with any serious adverse clinical events. Thus, United Nations Children's Fund (UNICEF) and WHO recommend the use of low-osmolarity ORS to prevent and treat dehydration. ORS has major advantages over home fluids in the management of children with significant dehydration and coupled with zinc therapy (20 mg zinc sulphate daily for 10 days) is the mainstay of treatment of childhood diarrhoea. At current global coverage, it is estimated that ORS reduces diarrhoea mortality by 69%, which would increase to 93% if coverage was 100% (Munos et al. 2010). Moreover, zinc supplementation has been shown to reduce the duration of a diarrhoea episode by 25% and a 30% reduction in stool volume.

Dietary abstinence, the traditional approach to an acute diarrhoeal illness, restricts the intake of necessary calories, fluids, and electrolytes. During an acute attack, the patient often finds it more comfortable to avoid spicy, high-fat, and high-fibre foods, all of which can increase stool volume and intestinal motility. Although giving the bowels a rest provides symptomatic relief, continued oral intake of fluids and foods is critical for both rehydration and the prevention of malnutrition. In children, it is particularly important to restart feeding as soon as the child is willing to accept oral intake.

Because certain foods and fluids can increase intestinal motility, it is wise to avoid fluids such as coffee, tea, cocoa, and alcoholic beverages. Ingestion of milk and dairy products can potentiate fluid secretion and increase stool volume. Besides the oral rehydration therapy outlined earlier, acceptable beverages for mildly dehydrated adults include fruit juices and various bottled soft drinks. Carbonated drinks should be allowed to 'de-fizz' by letting them stand in a glass before ingestion. Soft, easily digestible foods are generally acceptable to the patient with acute diarrhoea.

Since most patients with infectious diarrhoea, even those with a recognized pathogen, have a mild, self-limited course, neither a stool culture nor specific treatment is required for such cases. For more severe cases, however, empirical antimicrobial therapy should be instituted, pending the results of stool and blood cultures. Gastrointestinal infections likely to respond to antibiotic treatment include cholera, giardiasis, cyclosporiasis, shigellosis, *E. coli* diarrhoea in infants, symptomatic travellers' diarrhoea, and *C. difficile* diarrhoea. The choice of antimicrobial drug should be based on *in vitro*

sensitivity patterns, which vary from region to region. A fluoroquinolone antibiotic is a good choice for empirical therapy, since these agents have broad-spectrum activity against virtually all bacterial pathogens responsible for acute infectious diarrhoea (except *C. difficile*). Resistance to fluoroquinolones in South and Southeast Asia is an increasing problem.

In patients with severe community-acquired diarrhoea—characterized by more than four stools per day lasting for at least 3 days or more with at least one associated symptom such as fever, abdominal pain, or vomiting—there is a high likelihood of isolating a bacterial pathogen. In this setting, a short course of a fluoroquinolone, namely 1–3 days' duration, will generally provide prompt relief with a low risk of adverse effects. Fluoroquinolones will not be effective for parasitic infections—specific antiparasitic drugs should be prescribed after identification of the offending pathogen in stool smears.

The duration of antimicrobial therapy has not been clearly defined. While courses of anywhere from 3 to 10 days of treatment have been recommended, there are several studies that included severe forms of diarrhoea which suggested that a single dose is as effective as more prolonged therapy. When treatment is indicated, a number of studies have shown that the combination of an antimicrobial drug and an antimotility drug provides the most rapid relief of diarrhoea.

Preventive measures

Diarrhoeal disease affects rich and poor, old and young, and those in developed and developing countries, yet a strong relationship exists between poverty, an unhygienic environment, and the number and severity of diarrhoeal episodes—especially for children under 5 years. Poverty also restricts the ability to provide age-appropriate, nutritionally balanced diets, or to modify diets when diarrhoea develops so as to mitigate and repair nutrient losses. The impact is exacerbated by the lack of adequate, available, and affordable medical care. Thus, preventive and management strategies for diarrhoea must have an equity focus.

Malnutrition is an independent predictor of the frequency and severity of diarrhoeal illness and can lead to a vicious cycle in which sequential diarrhoeal disease leads to increasing nutritional deterioration, impaired immune function, and greater susceptibility to infection.

Family knowledge about diarrhoea must be reinforced in areas such as prevention, nutrition, hand washing, and hygiene, measles vaccination, preventive zinc supplements, and when and where to seek care. It is estimated that, in the 1990s, more than 1 million deaths related to diarrhoea may have been prevented each year, largely attributable to the promotion and use of these therapies.

A 2016 meta-analysis estimates that nearly half of all diarrhoea episodes can be prevented by breastfeeding (Victora et al. 2016). Not breastfeeding is associated with a 165% increase in diarrhoea incidence in 0–5-month-old infants (Bhutta et al. 2014). Continued breastfeeding during the diarrhoea episode provides nutrients to the child, prevents weight loss, and improves recovery from diarrhoea. Contaminated and poor-quality complementary foods are associated with increased diarrhoea burden and stunting. Appropriate, safe, and aptly initiated complementary feeding has been shown to reduce mortality by 6% in young children. Diarrhoea frequently causes fever, altering host metabolism and leading to the depletion of body stores of nutrients. These losses must be replenished during convalescence, which takes much longer than the illness does to develop. For these reasons, appropriate feeding strategies during diarrhoea episodes are a cornerstone of treatment.

Various studies suggest that zinc-deficient populations are at increased risk of developing diarrhoeal diseases, respiratory tract infections, and growth retardation.). A recent systematic review concluded that zinc therapy reduces diarrhoea-related admissions by 23%, duration of acute diarrhoea by 0.5 days and persistent diarrhoea by 0.68 days (Lazzerini and Wanzira 2016). Poor sanitation, lack of access to clean water, and inadequate personal hygiene are responsible for an estimated 90% of childhood diarrhoea. A recent Cochrane review estimates that promotion of hand washing in high-income and low-middle-income settings can reduce diarrhoea incidence by 30% (Ejemot-Nwadiaro et al. 2015) while some other reviews have shown hand washing with soap to lead to a diarrhoea risk reduction of nearly 48% (Black et al. 2003, 2008). Hand washing promotion strategies have also been shown to reduce diarrhoea burden with ancillary benefits in community settings.

Immunization represents an ideal way to prevent certain bacterial and viral diseases, but has not yet proved successful for combating many gastrointestinal pathogens. The cholera vaccine that has been available for decades suffers from low efficacy, a moderate risk of side effects, and a short duration of action. Newer oral cholera vaccines, such as the inactivated B

subunit vaccine, are highly effective for prevention of severe cholera. The rotavirus vaccine is estimated to reduce the incidence of severe rotavirus infection by nearly 74% and in 2009, the WHO recommended the vaccine be incorporated into national immunization programmes, particularly in LMICs (Khan and Bhutta 2017). According to GAVI, since programme inception the rotavirus vaccine has reached more than 54 million children (GAVI 2016). Measles is known to predispose to diarrhoeal disease secondary to measles-induced immunodeficiency, and it is estimated that measles vaccine at varying levels of coverage (45–90%) could prevent 44–64% of measles cases, 0.6–3.8% of diarrhoeal episodes, and 6–26% of diarrhoeal deaths among children under 5 years. A review of the Demographic Health Survey data from four LMICs concluded that measles vaccination was associated with a decrease in diarrhoea as well as acute respiratory infection (ARI) in children (Bawankule et al. 2017).

Acute respiratory infections

ARIs are classified as upper or lower respiratory tract infections. Upper respiratory tract infections include the common cold, otitis media, sinusitis, and pharyngitis while lower respiratory tract infections include laryngitis, tracheitis, bronchitis, bronchiolitis, pneumonia, and any combination thereof. ARIs are not only confined to the respiratory tract but may also have systemic effects due to extension of infection into the bloodstream, the production of microbial toxins, inflammation, and reduced lung function.

LRIs, especially bronchiolitis and pneumonia, are the most common causes of both illness and mortality in children under 5 years. In adults, pneumonia and influenza are major causes of morbidity and mortality in developed as well as less-developed nations.

Global and regional epidemiology

The Global Burden of Disease (GBD) study estimates that 2.75 million deaths in 2015 were a consequence of LRIs, which include pneumonia and bronchiolitis. The major proportion of pneumonia deaths, nearly 704,000 according to GBD (920,000 as per UN IGME estimates), occurred in children under-5; however, owing to widely available antibiotic therapy and vaccines such *Haemophilus influenzae* type B (Hib) and pneumococcal conjugate, the number of LRI deaths in this age group have fallen by nearly 37% between

2005 and 2015 (Institute for Health Metrics and Evaluation n.d.). However, LRI mortality is significant in even older age groups, especially in the elderly. The GBD study notes that LRIs caused an estimated 1.27 million deaths in adults 70 years or older (GBD 2015 LRI Collaborators 2017).

Pneumonia causes most ARI deaths, and is the second largest killer of children under-5, causing almost 16% of all under-5 deaths globally. Among under-5 children pneumonia continues to be the leading killer, with 12.8% of deaths attributable to it (Liu et al. 2016). However, in sub-Saharan Africa pneumonia remains the leading cause of under-5 deaths with nearly 490,000 deaths, 16.6% of the total under-5 mortality. South Asia, a high-burden region, saw a decrease in under-5 deaths attributable to pneumonia, where in 2015 nearly 15.1% of all under-5 deaths were due to pneumonia as compared to 19.2% in 2000. Notably, 65% of total episodes of childhood pneumonia occur only in 15 high-burden countries, namely, Afghanistan, Angola, Burkina Faso, China, Democratic Republic of the Congo, Ethiopia, India, Indonesia, Kenya, Mali, Niger, Nigeria, Pakistan, Tanzania, and Uganda. The global incidence of pneumonia in the 0–4 age group is estimated at 0.19 episodes per child per year. Regional incidence was estimated to be highest in Southeast Asia (0.26) and Africa (0.27) (Walker et al. 2013). Serious neonatal infections account for nearly 25% of neonatal mortality globally, and it is difficult to disentangle sepsis and deaths from pneumonia. With the inclusion of neonatal pneumonia, recent estimates indicate that pneumonia is the single largest contributor to child mortality, accounting for about 16% of all under-5 deaths globally. It is also important to note that in contrast to diarrhoeal deaths where mortality rates have declined dramatically, despite the introduction of a global programme for the control of ARIs decades ago, there has been only a small reduction in the overall burden of pneumonia deaths. A comprehensive review of the global burden of pneumonia estimated that 81% of the deaths from childhood pneumonia occur in the first 2 years. Twelve per cent (12%) of childhood pneumonia episodes progress to severe disease, with a case fatality ratio of 8.9% for severe pneumonia. Predictably, the highest proportion of severe pneumonia cases were in Southeast Asian (39%) and African (43%) regions. Most severe pneumonia cases in children are attributable to *Streptococcus pneumoniae* (18.3%), which also accounts for 33% of childhood pneumonia mortality worldwide.

Despite gains in the availability and quality of healthcare in industrialized countries, ARIs, especially lower respiratory tract infections, remain a major cause of morbidity and mortality for adults and children. In fact, even today

pneumonia and influenza together are the eighth most common cause for death among adults in the United States. Recovery from pneumonia in the elderly takes longer and complications and mortality are also more frequent than in younger populations. Pneumonia is one of the most common causes of hospitalization and decreased activities of daily living among the elderly. Risk factors for death from pneumonia in adults include advanced age, alcohol consumption, leukopenia, bacteraemia, hypoxaemia, comorbid conditions such as diabetes mellitus, congestive heart failure, active malignancies, and immunosuppression, and certain signs and symptoms including hypothermia, hyperthermia, tachypnoea, hypotension, and altered mental status. Many factors such as the presence of certain comorbid medical conditions, use of certain drugs, changes in physiochemical characteristics of the non-specific host defence system such as cilia and mucus of the respiratory tract, malnutrition, and mechanical devices contribute to an increased incidence of pneumonia among the elderly. However, an important predisposing factor to the increased incidence of infections is the age-associated decline in immune responsiveness. Changes in immune response not only decrease resistance to pathogens but also contribute to increased morbidity and mortality due to respiratory infections.

Aetiology

A wide range of different bacterial and viral pathogens is responsible for community-acquired pneumonia in children and adults. Foremost among them is *S. pneumoniae*, which accounts for up to half of all cases (Table 8.11.12). During recent years, the role of viral pathogens in the aetiology of acute lower respiratory infections (ALRIs) has been increasingly described. While influenza is well recognized as a cause of viral pneumonia, several studies have demonstrated the importance of parainfluenza virus, respiratory syncytial virus (RSV), adenovirus, and human metapneumovirus.

S. pneumoniae is implicated in 55.4% of LRI mortality in all ages, and nearly the same (55.8%) for LRI deaths in the under-5 age group. Following *S. pneumoniae*, Hib, RSV, and influenza were the leading aetiological agents for pneumonia mortality in children less than 5 (GBD LRI 2015). However, with a changing disease burden, more urbanization, and widespread vaccination, there is an increasing interest to study the changing aetiology of childhood pneumonia, to inform newer treatment options and vaccines (Levine et al. 2012).

Issues in presentation and diagnosis

The WHO has defined pneumonia solely on the basis of clinical findings obtained by visual inspection and setting respiratory rate cut-offs. Currently, the standard WHO algorithm for ARIs in children 2–59 months defines non-severe pneumonia as cough or difficult and fast breathing (respiratory rate of 50 breaths per minute or more for children aged 2–11 months; or respiratory rate of 40 breaths per minute or more for children aged 12–59 months) and/or chest indrawing. Severe pneumonia is defined as having cough or difficult breathing, with at least one or more of the following signs: central cyanosis, severe respiratory distress, inability to drink, lethargy, or convulsions. Additionally, severe pneumonia could also have fast breasting and chest indrawing. This current algorithm is based on the updated classification WHO introduced in 2013 (WHO 2014). The previous classification categorized ‘chest indrawing pneumonia’ as severe pneumonia requiring referral and treatment with parenteral antibiotics. However, with updated research and evidence, WHO now recommends classifying and treating ‘chest indrawing pneumonia’, the same as ‘pneumonia’, i.e. therapy at home with oral antibiotics. This has major implications for saving lives of children with pneumonia in difficult circumstances where referral may not be possible. It is recognized that mortality in children due to ARIs could be reduced by one-half if early detection and appropriate treatment could be provided.

Table 8.11.12 Pathogen-specific causes of childhood and adult pneumonia

Age range	Most common causative organism
Neonates (from birth to 30 days after birth)	*Streptococcus pyogenes*, *Staphylococcus aureus*, and *Escherichia coli*
Infants (from 3 weeks to 4 months)	S.*pneumoniae*
Infants older than 4 months and preschool-aged children	Respiratory viruses and *S. pneumoniae*
Children in developing countries	S.*aureus* and *Haemophilus influenzae* including non-typable strains
Adults—outpatient	S.*pneumoniae*, *Mycoplasma pneumoniae*, *H. influenzae*, *Chlamydophila pneumoniae*, and respiratory viruses
Adults—inpatient	S.*pneumoniae*, *M. pneumoniae*, *C. pneumoniae*, *H. influenzae*, *Legionella pneumophila*, respiratory viruses, and aspiration
Adults—intensive care unit	S.*pneumoniae*, *S. aureus*, *L. pneumophila*, Gram-negative bacilli, and *H. influenzae*

In contrast to the simple, clinical definition of pneumonia recommended by the WHO for use in developing countries, pneumonia in resource-rich countries is usually based on the presence of characteristic signs and symptoms (e.g. dry or productive cough, tachypnoea, fever, focal findings on respiratory examination), hypoxaemia, and the presence of infiltrates on chest radiograph. In general, the elderly tend to present with fewer or atypical symptoms of pneumonia than younger patients and therefore non-specific features such as fever or mental status change may be indicators of an underlying lower respiratory tract infection.

Microbiological studies can also be pursued to support the diagnosis of pneumonia due to specific infectious agents and to facilitate decision-making for antibiotic management. While broad-spectrum empirical antimicrobial coverage is recommended in various guidelines, there is a potential risk for clinical failure and increased mortality if inappropriate antibiotic therapy is initiated. If available and adequate quality specimens can be obtained, blood cultures and sputum Gram stain and culture should be performed, especially with increasing severity of illness. Rapid diagnostic tests may be useful when specific diagnoses are being considered such as RSV, influenza, or *L. pneumophila*.

Clinical approaches for the management of childhood pneumonia are significantly hampered by the lack of a gold standard, as classic microbiological methods have poor sensitivity and current algorithms lack sufficient specificity. It is therefore likely that community strategies for the recognition and management of pneumonia by ancillary health workers that rely on simple clinical criteria, other than auscultation, will overdiagnose bacterial pneumonia. There are legitimate concerns that widespread use of first-line antibiotics for all ARIs will lead to loss of effectiveness.

Evidence-based interventions

Only in the early 1980s, long after immunization and diarrhoea-control programmes were launched, did the international community become aware of the epidemiological magnitude of pneumonia in children. This led the WHO and the UNICEF to decide that reduction of mortality from pneumonia should be the main objective of the initial ARI programme. Since early microbiological studies of lung aspirates taken from hospitalized, untreated children with pneumonia in developing countries showed that bacteria were present in more than 50% of cases and it was recognized that bacterial pathogens were responsible for the most severe cases, it became apparent that prompt treatment with a full course of effective antibiotics could be life-saving. However, healthcare providers see only 43% of children with pneumonia in LMICs, wand only 29% receive antibiotics (Bhutta et al. 2013b).

Antibiotic treatment of pneumonia

Although recommendations for antibiotic therapy for pneumonia are based on aetiological diagnosis, identification of the causative organism in routine clinical care is very difficult and empirical antibiotic therapy is often instituted. Guidelines for the treatment of pneumonia in immunocompetent adults in industrialized nations generally recommend a macrolide or doxycycline for outpatients; an advanced macrolide such as azithromycin or clarithromycin plus a beta-lactam or a respiratory fluoroquinolone alone for less acutely ill inpatients; and a beta-lactam plus either an advanced macrolide or respiratory fluoroquinolone for adults requiring intensive care (Mandell et al. 2007).

The various modalities for antibiotic treatment in children according to disease severity, as recommended by the current WHO guidelines, are shown in Table 8.11.13.

The revised WHO treatment guidelines for ARIs also provide empiric treatment recommendations for childhood pneumonia treatment in HIV-affected areas. Ampicillin plus gentamicin or ceftriaxone are recommended as first-line treatment for HIV-infected and HIV-exposed children under 5 years, who develop pneumonia or severe pneumonia, while ceftriaxone alone is recommended as second line if the ampicillin/gentamicin combination is not effective. Moreover, the guidelines recommend cotrimoxazole therapy, as an additional treatment, for suspected *Pneumocystis jirovecii* (formerly *P. carinii*) pneumonia in HIV-infected and HIV-exposed infants (2–12 months) (WHO 2014). There are calls for further research into WHO's severe pneumonia treatment protocols for children with severe acute malnutrition (SAM) to reduce treatment failure and mortality in this group (Chisti et al. 2015a).

Table 8.11.13 Treatment of paediatric pneumonia according to disease severity

Signs/symptoms		
Cough or difficulty in breathing with:	**Classification**	**Treatment**
Fast breathing: ≥ 50 breaths/min in child aged 2–11 months ≥ 40 breaths/min in child aged 1–5 years Chest indrawing	Pneumonia	• Home care • Give appropriate antibiotics for 5 days • Advise the caregiver to return if danger signs appear • Soothe the throat and relieve the cough with a safe remedy • Follow-up after 3 days
Central cyanosis, severe respiratory distress, and/or danger signs (inability to drink, lethargy, convulsions)	Severe pneumonia	• Admit to hospital • Give the recommended antibiotics • Give oxygen • Manage airway • Antipyretics

Oxygen therapy for severe pneumonia can be life-saving. Recent evidence indicates that, where available, oxygen delivery by continuous positive airway pressure (CPAP) improves outcomes in paediatric severe pneumonia cases (Chisti et al. 2015b).

Integrated management of childhood infections (IMCI)

In the mid-1980s, the WHO initiated a control programme for ARIs focusing on management of cases by health workers. The current case management of ARIs has been incorporated into the global IMCI which train health workers to recognize fast breathing, lower chest wall indrawing, or danger signs in children with respiratory symptoms (such as cyanosis or inability to drink).

Evidence from Asia and Africa has shown that well-trained CHWs can use the WHO algorithm to classify fast breathing and/or chest indrawing pneumonia and provide home-treatment with amoxicillin and refer severe pneumonia cases. Community case management (CCM) of pneumonia can play a key role in providing prompt care in low-resource settings and

reducing pneumonia morbidity and mortality. However, there is concern that this approach may lead to an overdiagnosis of pneumonia and inappropriate antibiotic therapy, thereby contributing to antimicrobial resistance.

Recognizing the key role CHWs can play when adequately trained and equipped, the UNICEF and WHO are supporting the integrated community case management (iCCM) strategy in many LMICs, where CHWs identify and treat children for pneumonia, diarrhoea and malaria (in malaria-endemic areas) using appropriate treatment methods such as ORS and zinc, oral antibiotics or artemisinin-based combination therapy (ACT).

Preventive measures

Poverty, overcrowding, air pollution, malnutrition, harmful traditional practices, and delayed and inappropriate case management are important underlying determinants for high ARI case fatality rates. Preventive strategies for pneumonia include immunizing children, hand washing, promoting adequate nutrition (including exclusive breastfeeding and zinc intake), and reducing indoor air pollution (Bhutta 2007).

Three vaccines have the potential to substantially reduce deaths from pneumonia in children under-5. Two kinds of vaccines are currently available against pneumococci: a 23-valent polysaccharide vaccine (PPSV23), which is more appropriate for adults than children, and different formulations of protein-conjugated polysaccharide vaccines (PCV), most important of which contain 7, and 13 serotypes of *S. pneumoniae* (PCV7 and PCV 13, respectively). Rates of invasive pneumococcal disease (IPD) have decreased since the introduction of the PCV7 for use in infants in the United States in 2000 and PCV13 in 2010. While the Hib vaccine is now being used in many resource-poor countries, scaling up of the PCV lags behind. Evidence indicates that Hib vaccination can reduce pneumonia by 18% while the pneumococcal conjugate vaccine can reduce it by 23–35% (WHO/UNCF 2013).

Controlled trials of hand washing promotion in child-care centres in developed countries have reported significant reduction (12–32%) in rates of upper respiratory tract infections. A community-based cluster randomized trial of hand washing promotion from Pakistan also reported that frequent hand washing (with or without soap) led to a 50% reduction in pneumonia incidence (Luby et al. 2005).

About 3 billion people still rely on solid fuels, 2.4 billion on biomass, and the rest on coal, mostly in China. More than half of all the deaths and 83% of DALYs lost attributable to solid fuel use occur as a result of lower respiratory tract infections (LRTIs) in children under-5. Reduction in household air pollution by 50% by appropriate use of clean fuels and lower emission stoves, could reduce severe pneumonia by 33% (WHO/UNICEF 2013). A recent Cochrane review concluded that zinc supplementation is associated with a reduction in the incidence and prevalence of pneumonia in children 2–59 months (Lassi et al. 2016). However, the evidence is still not sufficient for WHO to recommend zinc supplementation in LRIs.

Public health implications

Despite a long-standing global programme for the control of ARIs there have been limited reductions in the overall burden of deaths from pneumonia. The bulk of deaths from childhood pneumonia disproportionately affect the poor who have higher exposure rates to risk factors for developing ARIs, such as overcrowding, poor environmental conditions, malnutrition, and also limited access to health services. The importance of reaching the poor with pneumonia in community settings must be underscored. Such strategies involve recognizing and ambulatory management of pneumonia in community settings through CHWs, assuring transportation and access to facilities for severe pneumonia and availability of antibiotics.

Meningitis in neonates, children, and adults

Acute meningitis is a potentially fatal infection caused by several microorganisms including bacteria, viruses, parasites, and fungi. In addition, meningitis is associated with a risk of chronic morbidity and developmental disability. Although the exact incidence of meningitis in developing countries is uncertain, case fatality rates range from 10 to 30%. The recent GBD 2016 study estimated that there were 318,400 deaths from meningitis in 2016, a decrease of 8.4% from 2006 (GBD 2016 Causes of Death Collaborators 2017). Even if effective treatment is provided, between 10 and 50% of survivors still develop neurological sequelae. Successful outcome of meningitis relates to several factors including age, time, and clinical stability before effective antibiotic treatment, species of microorganism, number of bacteria or quantity of active bacterial products in the CSF at the time of

diagnosis, intensity of the host's inflammatory response, and time elapsed to sterilize CSF cultures.

Aetiology

The three most common bacterial pathogens, *S. pneumoniae*, *H. influenzae*, and *N. meningitidis*, account for more than 80% of cases of bacterial meningitis in the United States (US) overall (Brouwer et al. 2010), although *Listeria monocytogenes* is a greater problem for the elderly, immunocompromised, and pregnant women (Table 8.11.14). Although *S. pneumoniae* remains the leading cause of bacterial meningitis in the US, its incidence fell from 0.8/100,000 people (1997) to 0.3/100,000 people, a decrease attributable to vaccination (Castelblanco et al. 2014). There is a relative paucity of microbiological information from developing countries, but beyond the neonatal period, the main agents of meningitis include Hib, *S. pneumoniae*, and *Neisseria meningitidis* with reported case fatality rates of 7.7, 10, and 3.5%, respectively.

Table 8.11.14 Aetiology and treatment of bacterial meningitis

Patient group	Common organisms	Antimicrobial therapy
Immunocompetent children (age ≥ 3 months–18 years)	H.*influenzae* S.*pneumoniae* N.*meningitidis*	Developing countries: ampicillin plus chloramphenicol Developed countries: vancomycin plus a third-generation cephalosporin (cefotaxime or ceftriaxone[a])
Immunodeficient, pregnant, and elderly (> 50 years)	L.*monocytogenes* S.*pneumoniae* N.*meningitides* Aerobic Gram-negative bacilli	Vancomycin plus ampicillin plus either cefepime or meropenem
Neurosurgical problems and head trauma	S.*aureus* S.*pneumoniae*	Vancomycin plus a third-generation cephalosporin or meropenem

[a] For resistant *S. pneumoniae*, the American Academy of Pediatrics recommends vancomycin plus cefotaxime or ceftriaxone as empiric therapy.

Source: data from Brouwer, M.C. et al., Epidemiology, diagnosis, and antimicrobial treatment of acute bacterial meningitis, *Clinical Microbiology Reviews*, Volume 23, Number 3, pp. 467–92, Copyright © 2010, American Society for Microbiology. All Rights Reserved.

Issues in presentation and diagnosis

The clinical features that may help in diagnosing meningitis are summarized in Table 8.11.15. In general, clinicians should have a low threshold for investigating and excluding meningitis in children, as features may be non-specific. The clinical diagnosis can be confirmed by lumbar puncture and the examination of CSF. The CSF will have a cloudy appearance, elevated protein, increased leucocyte counts with a predominance of neutrophils, and the presence of pathogens on Gram stain and/or culture provide a definitive diagnosis of bacterial meningitis. A bacterial meningitis score has been shown to be a reliable clinical scoring system to help in decision-making with regards to meningitis (Nigrovic et al. 2012). If patients do not have one of the following factors—positive CSF Gram stain, CSF absolute neutrophil count ≥1,000 cells/microlitre, CSF protein ≥80 mg/dl, peripheral blood absolute neutrophil count ≥10,000 cells/microlitre, or history of seizure before or at the time of presentation—they are very unlikely to have bacterial meningitis (negative predictive value = 99.9%) (Nigrovic et al. 2007). In recent years, PCR techniques for diagnosis of bacterial meningitis have become available with sensitivities ranging from 80 to 100% and specificities of 95 to 100% (Brouwer et al. 2012). Moreover, PCR can be used to detect enteroviruses and help in distinguishing bacterial vs. viral meningitis.

Table 8.11.15 Signs and symptoms of paediatric meningitis

Symptoms or presenting history	Signs
Vomiting, inability to feed and drink Headache or pain in back of neck Convulsions Irritability Signs of raised intracranial pressure History of recent head trauma Unequal pupils Rigid posture or posturing Focal paralysis in any limbs or trunk Irregular breathing	• Stiff neck, repeated convulsions • Fontanelle bulging • Petechiae or purpura • Irritability • Lethargy, evidence of head trauma

Medical treatment

The mainstay of treatment is prompt antibiotic therapy for suspected bacterial meningitis. Antibiotics need to be started before the results of CSF culture and sensitivity are available. This requires selection of an appropriate antibiotic, known to be effective against the common bacterial pathogens prevalent locally. An increasing number of beta-lactamase-producing strains of Hib are resistant to ampicillin, and a smaller number of chloramphenicol acetyltransferase-producing strains are resistant to chloramphenicol. Additionally, the proportion of CSF isolates of *S. pneumoniae* that is non-susceptible to penicillin, ceftriaxone, and cefotaxime has also increased. Currently, the drugs for suspected or confirmed bacterial meningitis include vancomycin plus an extended spectrum cephalosporin with the addition of ampicillin if *L. monocytogenes* is suspected (Table 8.11.14). If these options are not available, then ampicillin plus either gentamicin or chloramphenicol may be used. If sepsis is suspected, then cases should be treated with ampicillin or penicillin plus an aminoglycoside, until meningitis is confirmed.

Very early parenteral administration of corticosteroids (before or with initiation of antibiotics) significantly reduces severe adverse outcomes and case fatality rates. Similarly, there is evidence to suggest that restriction of fluids in the first 48 hours may improve outcomes. A recent Cochrane systematic review found that corticosteroids protect against severe hearing loss and neurological sequelae, however, did not reduce overall mortality among individuals with community-acquired bacterial meningitis in high-income countries. The beneficial effect was not seen in low-income settings (Brouwer et al. 2015).

Preventive measures

Although poverty, malnutrition, overcrowding are important risk factors for disease, delayed and inappropriate case management is a common determinant of adverse outcomes. The development of effective vaccines has been a major factor in the reduction of the burden of meningitis in the developed world. These include the Hib, pneumococcal conjugate, and meningococcal vaccines.

Haemophilus influenzae type b vaccine

Currently three Hib conjugate vaccines are available for use in infants and young children with comparable efficacy. Hib vaccines also protect against other infections caused by the pathogen such as pneumonia, joint infections, and bacteraemia. All industrialized countries now include Hib vaccine in their national immunization programmes, resulting in the virtual elimination of invasive Hib disease. There is comparable impressive evidence of benefit from several developing countries following introduction of Hib vaccine. With support from GAVI, Hib coverage, as part of the pentavalent vaccine, in the 73 GAVI-supported countries is nearly 80% (pentavalent third dose coverage). WHO recommends the vaccine be given as a three-dose primary series at the same time as diphtheria, pertussis, and tetanus (DTP) vaccine, starting at 6 weeks of age, with the second and third dose at 4–8 week intervals.

Pneumococcal vaccine

The older 23-valent pneumococcal polysaccharide vaccine is unsuitable for use in young children. The development of the 7-valent, 10-valent, and 13-valent PCVs is a major advance in the control of IPD. In the US, the 7-PCV was included in routine vaccinations of infants and children under 2 years in 2000 and by 2001 the incidence of all IPD in this age group had declined by 69%. With GAVI support, 57 countries, mostly from Latin America, Asia, and sub-Saharan Africa, have introduced pneumococcal conjugate vaccine as part of their routine immunization programmes.

Meningococcal vaccine

Meningococcal polysaccharide vaccine is available for serogroups A, C, W-135, and Y. This quadrivalent vaccine has been introduced in several developed countries as part of routine vaccine schedules, especially for adolescents who will be rooming in crowded dormitories while attending university. In many LMICs, this vaccine has been replaced by the quadrivalent meningococcal conjugate vaccine, which is more immunogenic and, because it induces memory cells, is likely to lead to a longer lasting protective immune response. Recent evidence of the efficacy and safety of a conjugate meningococcal B vaccine suggests that in the near future we will have highly effective vaccines for all major meningococcal serogroups (Vesikari et al. 2013). The success of MenAfriVac (meningococcal A

conjugate vaccine) mass vaccination campaigns and trials in the African meningitis belt is notable and has led to revised WHO recommendations for these countries to introduce the vaccine into routine immunization subsequently (WHO 2015).

Typhoid fever

Typhoid fever, a systemic disease caused by *Salmonella enterica* serovar Typhi, is an acute illness characterized by protean and non-specific symptoms, including fever and gastrointestinal infection. The systemic disease caused by *S. Paratyphi* (A, B, or C) is also clinically similar; both typhoid and paratyphoid are collectively labelled as enteric fevers. The emergence of drug-resistant strains, especially multidrug-resistant (MDR) and extensively drug-resistant (XDR) *S. Typhi*, resistant to ampicillin, chloramphenicol, trimethoprim–sulphamethoxazole, fluoroquinolones and more recently third-generation cephalosporins is a growing problem (Klemm et al. 2018).

Global and regional epidemiology

The global incidence of typhoid fever in 2000 was estimated to be 21.6 million cases, with more than 200,000 deaths (Crump et al. 2004). More recent literature shows varied estimates of 11–21 million cases and nearly 145,000–161,000 deaths per year, with the highest incidence in sub-Saharan Africa, followed by North Africa/Middle East and South Asia (Mogasale et al. 2014; Antillón et al. 2017). In South Asia, community-based studies indicate that a large proportion of cases occur in children under-5, with significant morbidity and mortality. Table 8.11.16 shows the regional distribution of typhoid in LMICs.

Table 8.11.16 Total cases and incidence of typhoid for the Global Burden of Disease regions and subregions made up of low- and middle-income countries

Area/region	Cases (millions)	Incidence (per 100,000 person-years)
All LMICs	17.8	293
Central Europe, Eastern Europe, and Central Asia	0.1	28
Latin America and Caribbean	1.0	169
North Africa and Middle East	2.6	557
Sub-Saharan Africa	7.2	762
Southeast Asia, East Asia, and Oceania	2.21	108
South Asia	3.6	204

Source: The burden of typhoid fever in low- and middle-income countries: A meta-regression approach Marina Antillón1, Joshua L. Warren, Forrest W. Crawford, Daniel M. Weinberger, Esra Ku¨ru¨m, Gi Deok Pak, Florian Marks, Virginia E. Pitzer. 2017. PLoS NTD.

There have been dramatic point-source outbreaks of typhoid related to contamination of food sources or water supply. The use of contaminated ground water, consumption of street foods, and poor personal hygiene are common risk factors for infection.

Issues in presentation and diagnosis

After ingestion in contaminated food or water, *S. Typhi* penetrates the small bowel mucosa and makes its way rapidly to the lymphatics and the bloodstream. Following an initial bacteraemia, the organism is sequestered in cells of the reticuloendothelial system where it multiplies and re-emerges several days later in recurrent waves of bacteraemia, an event that initiates the symptomatic phase of infection. The incubation period ranges from 5 to 14 days.

Typhoid fever is a febrile illness of prolonged duration, characterized by delirium, persistent bacteraemia, splenomegaly, and a variety of systemic manifestations. Children commonly present with fever, headache, and abdominal discomfort, diarrhoea, anorexia, dry cough, or myalgia and constipation. In the later phase of illness, more specific physical signs

including hepatomegaly and splenomegaly may be observed. Rose spots may be seen in an early stage of the illness in fair-skinned children and large proportions have a centrally coated tongue. Pulse–temperature dissociation is present in some patients. In approximately 50% of patients, there is no change in bowel habits; in fact, constipation is more common than diarrhoea in children with typhoid fever.

In LMICs, as per WHO recommendations, healthcare providers are advised to consider typhoid (if malaria has been excluded) in children if there is fever (especially fever ≥7 days) and any of the following; abdominal pain, vomiting, constipation, headache, cough or transient rash (Khan and Bhutta 2017). Moreover, in infants, typhoid can present as an acute febrile illness occasionally leading to hypothermia and shock.

As a result of recurrent waves of bacteraemia, patients with typhoid fever can develop pneumonia, pyelonephritis, osteomyelitis, septic arthritis, and meningitis. Intestinal haemorrhage and perforation, the most common complications, often occur in the third week of infection or during convalescence. The most serious complication, intestinal perforation, occurs in 0.5–3% of the patients with typhoid, and because they occur most commonly in areas where optimal medical care is not readily available, it may be associated with case fatality rates ranging from 4.8 to 30.5%. The elucidation of a *S. Typhi*-specific toxin (Song et al. 2013) not only provides exciting new information on the pathogenesis of typhoid fever but also opens exciting possibilities of development of newer vaccines and interventions for this disease.

Diagnosis of typhoid and paratyphoid fever requires culture of blood, bone marrow, stools, or urine to confirm growth of *S. Typhi*. Laboratory findings commonly include leukopenia, thrombocytopenia, proteinuria, and elevated transaminases, but these are relatively non-specific and uncommon. In developing countries, culture facilities are expensive and mostly confined to hospitals, while most typhoid patients are diagnosed clinically and treated in outpatient settings. In other instances, serological diagnosis may be made with the Widal test. The latter, though low-cost, is insufficiently sensitive in endemic areas. Other diagnostic tests, such as the Typhidot and Tubex, which detect IgM antibodies against specific *S. Typhi* antigens have been developed (Bhutta 2006). Since these newer assays have not proven to have adequate sensitivity and specificity for routine use in community settings, there is a need for further refinement in serological or molecular diagnosis of the

disease. PCR and ELISA tests are modern diagnostic techniques which are rapid and reliable, though at a higher cost.

Medical treatment

In the pre-antibiotic era, typhoid fever case fatality rates approached 20%. In addition to supportive measures such as rest, hydration, soft diet and fever management, antibiotic therapy is central in the treatment of typhoid. Treatment with effective antimicrobial agents—ampicillin, chloramphenicol, co-trimoxazole, and later ciprofloxacin—has progressively reduced case fatality rates to less than 1%, except for MDR isolates. Fluoroquinolones and third-generation cephalosporins are effective in MDR typhoid but over the last few years there have been increasing reports in Asia of *S. Typhi* strains with reduced fluoroquinolone susceptibility. The presence of nalidixic acid resistance *in vitro* is associated with clinical treatment failures with fluoroquinolones. Alternative therapies including azithromycin or third-generation cephalosporins are recommended in these circumstances. However, in recent years a new MDR strain, named 'extensively drug-resistant (XDR) *S. Typhi*' by the researchers, resistant to chloramphenicol, ampicillin, cotrimoxazole, fluoroquinolones, and third-generation cephalosporins has been identified (Klemm et al. 2018). Patients with this strain were successfully treated with azithromycin and meropenem.

Given the considerable morbidity and higher mortality rates reported with MDR typhoid in children, appropriate antibiotic therapy must be instituted promptly and, when appropriate facilities are available, treatment choices should be guided by susceptibility testing.

Preventive measures

Detection of sources of infection related to recent typhoid fever in household contacts, commercial food handlers, or contaminated drinking water sources is essential to design effective preventive measures for disease containment.

The WHO, since 2000, has recommended two licensed vaccines for prevention of disease in children >2 years and adults: Ty21a (an attenuated strain of *S. Typhi* administered orally) and Vi (the purified bacterial polysaccharide vaccine, given parenterally). More recently, WHO has recommended the typhoid conjugate vaccine for routine use in children over 6 months of age and adults up to 45 years of age in typhoid-endemic countries

(WHO 2018a). The WHO recommends the typhoid conjugate vaccine above the previous two vaccines, given its improved immunological properties and long-lasting protection.

Dengue fever

In recent years, dengue fever, has become the most common mosquito-borne viral infectious disease and now represents a major international public health concern. Dengue virus belongs to the genus *Flavivirus* (single-stranded, non-segmented RNA viruses), which includes four serologically distinct serotypes (DEN-1, DEN-2, DEN-3, and DEN-4). Variations in virus strains within and between the four serotypes influence disease severity. There is limited protection across serotypes. Secondary infections (particularly with serotype 2) are more likely to result in severe disease and dengue haemorrhagic fever.

Global and regional epidemiology

Humans and mosquitoes are the principal hosts of dengue virus although some non-human primates can also be infected. Dengue epidemics occur during the warm, humid, rainy seasons, which favour breeding conditions for the mosquito vector, *Aedes aegypti*. More than two-fifths of the world's population (~2.5 billion) lives in areas potentially at risk for dengue, which is endemic in more than 100 countries across the globe, with tropical areas of Asia, the Western Pacific, Latin America, and the Caribbean being the most seriously affected regions.

In some case series, dengue fever has been reported as the second most frequent cause of hospitalization (after malaria) among travellers returning from the tropics. There are an estimated 390 million dengue infections annually, with nearly 100 million manifesting clinically (Bhatt et al. 2013). Around 95% of cases occur in children less than 15 years of age, with infants representing 5% of the cases.

Issues in presentation and diagnosis

The incubation period can vary from 3 to 14 days (typically between 5 and 7 days) and viraemia can persist up to 12 days (typically 4–5 days). Fever usually lasts for 5–7 days; fevers persisting beyond 10–14 days suggest another diagnosis. The clinical features of dengue vary with patient age. Most dengue infections, especially in children, are minimally symptomatic or

asymptomatic. Children may also present with atypical syndromes such as encephalopathy and fulminant liver failure.

Classic dengue fever is characterized by a high fever of abrupt onset, sometimes with two peaks (saddle back fevers), severe myalgias, arthralgia, retro-orbital pain, headaches, and any of the three types of rashes, including a petechial rash, diffuse erythematous rash with isolated patches of normal skin, and a morbilliform rash, haemorrhagic manifestations, and leukopenia. Other manifestations include flushed facies, sore throat, cough, cutaneous hyperaesthesia, and taste aberrations. Convalescence may be prolonged and complicated by profound fatigue and depression.

Warning signs preceding onset of severe disease include abdominal pain or tenderness, persistent vomiting, mucosal bleed, clinical fluid accumulation, liver enlargement, lethargy, and an increase in haematocrit with a fall in platelet count. Appearance of any of these signs indicates the need for strict observation and prompt management to avoid complications.

WHO defines severe dengue (dengue haemorrhagic fever) by the presence of one or more of the following signs; plasma leakage that may cause shock and/or fluid accumulation and/or severe bleeding and/or severe organ impairment. An estimated 20–30% of cases of severe dengue progress to shock (dengue shock syndrome) and 40–50% of those with shock die.

Infection with one serotype is thought to produce lifelong immunity to that serotype but only partial immunity to the others. Previous infection with a specific serotype followed by infection with a new serotype greatly increases the risk of dengue haemorrhagic fever.

Dengue virus serotypes are distinguishable by complement fixation and neutralization test. Other diagnostic tests for patients with dengue include packed cell volume, platelet count, liver function tests, prothrombin time, partial thromboplastin time, electrolytes, and blood gas analysis. Laboratory findings commonly associated with dengue include leukopenia, lymphocytosis, increased concentration of liver enzymes, and thrombocytopenia. Diagnosis can be confirmed with several laboratory tests, especially the haemagglutination inhibition test and IgG or IgM enzyme immunoassays.

Evidence-based interventions

Rapid urbanization has led to an increase in the environmental factors that contribute to the proliferation of *Aedes* mosquitoes. These include

uncontrolled urban development, inadequate management of water and waste, presence of a range of large water stores, and disposable, non-biodegradable containers that become habitats for the larvae. Addressing these issues is key in controlling the dengue vector and spread of disease.

Medical treatment

No specific therapeutic agents exist for dengue fever apart from analgesics and medications to reduce fever. Treatment is supportive; steroids, antivirals, or carbazochrome (which decreases capillary permeability) have no proven role. In contrast, ribavirin, interferon alpha, and 6-azauridine have shown some antiviral activity *in vitro*. Mild or classic dengue is treated with antipyretic agents such as acetaminophen, bed rest, and fluid replacement (usually administered orally and only rarely parenterally); most cases can be managed on an outpatient basis.

The management of dengue haemorrhagic fever and the dengue shock syndrome is purely supportive. Aspirin and other non-steroidal anti-inflammatory drugs should be avoided owing to the increased risk for Reye's syndrome and haemorrhage.

Preventive measures

In the absence of a vaccine, vector control is the only effective preventive measure. At a personal level, the risk of mosquito bites may be reduced using protective clothing and repellents. The most effective preventive measure for travellers in areas where dengue is endemic is to avoid mosquito bites by using insect repellents containing *N*, *N*-diethyl-3-methyl-benzamide (DEET) or picaridin. The insect repellents should be used in the early morning and late afternoon, when *Aedes* mosquitoes are most active.

At a public health level, the risk of dengue fever outbreaks can be reduced by removing neighbourhood sources of stagnant water, especially within homes, using larvicides (especially for containers that cannot be eliminated), use of aerosolized insecticides, and introduction of predatory crustaceans.

The first dengue vaccine (CYD-TDV), Dengvaxia, manufactured by Sanofi, a live recombinant tetravalent vaccine was licensed in 2015, for use in individuals 9 years and older (Sabchareon et al. 2012). However, a 2017 mass vaccination campaign in Philippines was halted after 14 deaths were reported out of the 800,000 children who were vaccinated. Following that, after much deliberation and review of evidence, WHO issued recommendations to only

use the vaccine in dengue-seropositive persons, as vaccinating those not previously infected can lead to a reaction (WHO 2018c). However, in the absence of rapid diagnostic tests for dengue serostatus the utility of this vaccine remains limited. In the meantime, phase 3 trials are in progress for the Takeda Dengue vaccine (TAK-003) and one from the Instituto Butantan in Brazil.

Parasitic infections

A broad range of parasites plagues humans worldwide. While certain parasites, such as the *Plasmodium* species that cause malaria, are well recognized and have received intensive international support for research and programmatic control interventions, others are considered among the world's most neglected diseases. Some of the main neglected tropical parasitic diseases include the protozoan infections such as human African trypanosomiasis, visceral leishmaniasis, and American trypanosomiasis (Chagas disease) and helminthic infections such as the soil-transmitted nematodes (ascariasis, hookworms, trichuriasis), schistosomiasis, lymphatic filariasis, onchocerciasis, and dracunculiasis.

Of the 20 major helminth infections of humans, the commonest are the geohelminths. The most prevalent species are the roundworm (*Ascaris lumbricoides)*, whipworm (*Trichuris trichiura)* and hookworms *(Necator americanus and Ancylostoma duodenale)*. In many low-income countries, it is more common to be infected than not. Indeed, a child growing up in an endemic community can be infected soon after weaning, and continue to be infected and constantly re-infected for life.

Global and regional epidemiology

Recent global estimates indicate that more than a quarter of the world's population is infected with one or more helminths. The geographic distribution of roundworms in many tropical and subtropical regions closely parallels socioeconomic and sanitary conditions. In locales where several species of intestinal parasites are found, co-infection with *A. lumbricoides*, *T. trichiura*, and hookworms is common. In LMICs, about 1.2 billion people are infected with the roundworm, while more than 700 million are infected with hookworm or whipworm (Hotez et al. 2004). However, some recent studies report lower numbers attributing hookworm infection to nearly 440 million,

roundworm to 820 million and whipworm to 465 million people (Pullan et al. 2014). The GBD 2010 study estimated that 152,300 people died from neglected tropical diseases, which were predominantly parasitic infections with the notable exception of rabies (Lozano et al. 2012). While mortality is not common, parasitic infections are responsible for substantial morbidity and indirectly contribute to death from other diseases due to their negative impact on nutritional status. The World Bank ranks geohelminths as the greatest cause of poor health in children 5–15 years of age globally compared to any other infection.

Issues in presentation

Children of school age are at greatest risk from the clinical manifestations of disease. Studies have shown associations between helminth infection and undernutrition, iron deficiency anaemia, stunted growth, poor school attendance, and poor performance in cognition tests. Some 44 million pregnancies are currently complicated by maternal hookworm infection, placing both mothers and children at higher risk of anaemia and death during pregnancy and delivery. Intense whipworm infection in children may result in trichuris dysentery syndrome, the classic signs of which include bloody diarrhoea, anaemia, growth retardation, and occasionally rectal prolapse. Heavy burdens of both roundworm and whipworm are associated with protein energy malnutrition and deficiencies of certain micronutrients such as vitamin A.

Medical treatment

The WHO recommends the use of albendazole, mebendazole, pyrantel pamoate, and levamisole (Table 8.11.17). Albendazole and mebendazole, have high efficacy against roundworm and moderate efficacy against whipworm. Single-dose mebendazole is much less effective against hookworm, with cure rates typically below 60%.

Preventive measures

Better sanitation reduces soil and water transmission as transmission of geohelminths depends on transmission in environments contaminated with egg-carrying faeces. At high coverage, the provision of adequate sanitation is the most effective intervention to eliminate helminthic infections. With high

costs, implementing this strategy is difficult where resources are limited. Both the World Bank and the WHO promote helminth control programmes and consider it as one of the most cost-effective strategies to improve health in developing countries. These programmes emphasize mass drug administration as a major component of control.

Recommended drugs for use in public health settings include albendazole (single dose: 400 mg, reduced to 200 mg for children, 12 and 24 months), mebendazole (single dose: 500 mg), and levamisole or pyrantel palmoate. Medicines are recommended annually where prevalence is >20%, but twice annually where it is >50%.

Home delivery of antihelminthics is problematic for several reasons and thus school-based deworming programmes are preferred. These have been shown to boost school participation and are practical as schools offer a readily available, extensive, and sustained infrastructure with a skilled workforce that can be readily trained. In Kenya, such a programme reduced school absenteeism by a quarter, with the largest gains among the youngest children. Perhaps even more importantly, this study showed that those children who had not been treated benefited from the generally lowered transmission rate in the schools. However, more recently a Cochrane review concluded that although individual treatment of children had nutritional benefits, mass treatment programmes did not improve average nutritional status, haemoglobin levels, cognitive skills, or school performance (Taylor-Robinson et al. 2015). These measures must be coupled with community behaviour change strategies with the aim of reducing contamination of soil and water by promoting the use of latrines and hygienic behaviour. Without a change in defecation habits, periodic deworming cannot attain a stable reduction in transmission.

Table 8.11.17 Diagnosis and treatment of major intestinal nematode infections

Organism	Type of specimen	Specimen preparation	Size of eggs or larvae (µm)	Drug of choice	Alternative therapies
Trichuris trichiura	Stool	Direct smear or concentration	50–54 × 23	Mebendazole, 100 mg orally (PO) twice daily (bid) × 3 days	Albendazole, 400 mg PO once
Ascaris lumbricoides	Stool	Direct smear or concentration	45–70 × 35–50	Mebendazole, 100 mg PO bid × 3 days or albendazole, 400 mg PO once or pyrantel pamoate, 11 mg/kg PO once (max 1 g)	Piperazine citrate, 75 mg/kg twice daily (max. 1 g) by nasogastric tube × 2–3 days until resolution of obstruction
Ancylostoma duodenale Necator americanus	Stool	Direct smear or concentration	55–70 × 35–45	Mebendazole, 100 mg PO twice daily × 3 days	Albendazole, 400 mg PO once or pyrantel pamoate, 11 mg/kg PO × 3 days (max. 1 g)
Enterobius vermicularis	Adhesive tape preparation	Direct microscopy	50–60 × 20–30	Mebendazole, 100 mg PO once or pyrantel pamoate, 11 mg/kg PO once Repeat in 2 weeks	Albendazole, 400 mg PO once. Repeat in 2 weeks
Strongyloides stercoralis	Stool, duodenal aspirate	Concentration or Baermann method	400–500 × 15	Ivermectin, 150–200 µg/kg PO × 1–2 days[a]	Albendazole 400 mg PO once or twice daily for 3 days

[a] Intrarectal ivermectin is an option for treatment of high-grade strongyloidiasis.

Antimicrobial resistance

Overuse and misuse of antimicrobials, especially antibiotics, lead to genetic changes in the organisms, making them resistant to therapeutic agents. In addition to the threat of infectious diseases, antimicrobial resistance (AMR) is a growing crisis in global health. Bacteria, viruses, and parasites have

developed resistance to antimicrobials, rendering standard treatments ineffective. Infections with these ‘super bugs’ tend to persist and spread swiftly. According to an estimate, globally, 70 billion doses of antibiotics are consumed yearly. However, the burden of morbidity and mortality attributable to AMR can only be estimated broadly, especially in LMICs where indiscriminate use of antimicrobials is common, and surveillance is lacking. Realizing the importance of addressing this issue, the WHO now supports a system, the Global Antimicrobial Resistance Surveillance System (GLASS) that supports the collection and study of the regional and global data on AMR.

A recent ground-breaking study conducted in three countries of Africa found that mortality was lower among post-neonatal, preschool children following mass distribution of the antibiotic azithromycin in that group (Keenan et al. 2018). The potential effect of such an intervention in high-childhood mortality regions can be huge, however, implementation has to be with caution considering the danger of antibiotic resistance.

Prions

Prion diseases, also known as transmissible spongiform encephalopathies (TSEs), are responsible for rapidly progressive, fatal neurodegenerative diseases of humans and several other mammalian species. The term ‘prion’ is derived from *pro*teinaceous and *in*fectious. An unconventional transmissible protein that lacks nucleic acid is the probable cause; this prion protein appears to consist of a modified isoform of a normal cellular prion protein (PrP^{c}) that is converted into PrP^{Sc}, which results in major changes in the physicochemical properties of the PrP leading to abnormal folding of normal cellular proteins. Prion diseases are associated with brain pathology characterized by diffuse spongiform degeneration. The transmissibility was established by intracerebral inoculation of various primate species using passaged prions from infected humans.

Epidemiology

Human prion diseases can be sporadic, inherited, or transmissible. The sporadic forms include kuru where infection occurs through ritualistic cannibalism and sporadic Creutzfeldt–Jakob disease (CJD). CJD can also occur in a familial manner due to germ-line mutations in the PrP gene or as a

result of infection from bovine prions (variant CJD (vCJD)). The TSEs include bovine spongiform encephalopathy (BSE) in cattle; scrapie in sheep; chronic wasting diseases in deer, elk, and mules; feline spongiform encephalopathy in cats (possibly secondary to ingestion of prion-contaminated bovine tissues and bone meal); transmissible mink encephalopathy; and exotic ungulate encephalopathy in kudu, nyala, and oryx (Ramasamy et al. 2003). Despite surveillance in the United States and Canada, there is limited evidence of cross-species infection of chronic wasting disease from animal reservoirs to humans (Waddell et al. 2018).

The major human forms of infection are sporadic CJD, vCJD, and an inherited form, the prion disease called Gerstmann–Straussler–Scheinker disease. Iatrogenic CJD can also occur as a result of infected dura mater grafts, improperly sterilized depth electrodes, transplanted corneas, and human growth hormone and gonadotrophin derived from cadaveric pituitaries.

An epidemic of BSE in the United Kingdom that began in 1986 affected nearly 200,000 cattle and was followed by human cases of vCJD starting in 1994 (Brown et al. 2001). The source of the BSE outbreak resulted from a change in the process of rendering carcasses of livestock, which were used as a protein-rich nutritional supplement for feeding ruminants and other animals. Banning the use of meat and bone meal helped to eventually control the epidemic of BSE. In the mid-1990s, the United Kingdom saw a rise in human vCJD cases with 87 definite or probable cases occurring there by November 2000 and a few cases identified as well in France and Ireland. Since the incubation period of vCJD is not known, it was hypothesized to be as short as 5–10 years if the BSE outbreak in the mid to late 1980s was the source. Between 2000 and 2005, the number of annual deaths due to vCJD gradually declined. In total, there were 176 cases between 1995 and 2011 in the United Kingdom and 49 recorded in 11 other countries.

Clinical manifestations

Classic CJD usually presents in the elderly (median age of onset 68 years) and is characterized by memory loss, visual problems, and psychiatric symptoms, followed by progressive cognitive impairment and finally progression to an akinetic mutism state (Iwasaki I. 2017). Death usually occurs within 4–5 months. In contrast, vCJD presents at a median age of 27 years, with an illness duration of 13–14 months, and is clinically

characterized by prominent behavioural and psychiatric symptoms, painful dysaesthesias, and the delayed onset of neurological signs. In addition, vCJD cases all are homozygous for methionine at codon 129 of the PrP gene and have marked accumulation of protease-resistant PrP in brain tissue and florid plaques on neuropathology in large numbers in contrast to classic CJD where these are rare to absent.

Diagnosis and treatment

Neuropathological evaluation by Western blot or immunohistochemistry is the definitive approach for diagnosing human prion diseases. This method can also be used to monitor for vCJD and other forms of CJD. Electroencephalography (EEG) and magnetic resonance imaging (MRI) may be useful adjuncts to diagnosis in patients with CJD. Periodic sharp waves on EEG are often present in classic CJD but absent in vCJD whereas the 'pulvinar sign', an abnormal signal in the posterior thalami on T2- and diffusion-weighted MRI images, is present in the majority of patients with vCJD but is rare or absent in CJD. An important recent advance in diagnostics is the real-time quaking-induced conversion test of CSF, which has high sensitivity and specificity for sporadic and variant CJD (McGuire et al. 2016).

There is no known therapy for any of the prion diseases. Prognosis is poor with progression to death in a median of 4–5 months for CJD and a little over a year for vCJD.

REFERENCES

Antillón, M., et al. (2017). The burden of typhoid fever in low- and middle-income countries: a meta-regression approach. *PLoS Neglected Tropical Diseases*, 11(2), e0005376.

Bhatt, S., et al. (2013). The global distribution and burden of dengue. *Nature*, **496**, 504–7.

Bhutta, Z., et al. (2014). Can available interventions end preventable deaths in mothers, newborn babies, and stillbirths, and at what cost?. *Obstetrical & Gynecological Survey*, 69(11), 641–3.

Bhutta, Z.A. (2006). Current concepts in the diagnosis and treatment of typhoid fever. *BMJ*, **333**, 78–82.

Bhutta, Z.A. (2007). Dealing with childhood pneumonia in developing countries: how can we make a difference? *Archives of Diseases of Childhood*, **92**, 286–8.

Bhutta, Z.A., et al. (2013a). Evidence based interventions for improvement of maternal and child nutrition: what can be done and at what cost? *The Lancet*, 382(9890), 452–77.

Bhutta, Z.A., et al. (2013b). Interventions to address deaths from childhood pneumonia and diarrhoea equitably: what works and at what cost? *The Lancet*, **381**, 1417–29.

Black, R.E., Morris, S.S., and Bryce, J. (2003). Where and why are 10 million children dying every year? *The Lancet*, **361**, 2226–34

Black, R.E., et al. (2008). Maternal and child undernutrition: global and regional exposures and health consequences. *The Lancet*, 371(9608), 243–60.

Brouwer, M., McIntyre, P., Prasad, K., and van de Beek, D. (2015). Corticosteroids for acute bacterial meningitis. *Cochrane Database of Systematic Reviews*, (9), CD004405.

Brouwer, M.C., Thwaites, G.E., Tunkel, A.R., and van de Beek, D. (2012). Dilemmas in the diagnosis of acute community-acquired bacterial meningitis. *The Lancet*, **380**, 1684–92.

Brouwer, M.C., Tunkel, A.R., and van de Beek, D. (2010). Epidemiology, diagnosis, and antimicrobial treatment of acute bacterial meningitis. *Clinical Microbiology Reviews*, **23**, 467–92.

Brown, P., Will, R.G., Bradley, R., Asher, D.M., and Detwiler, L. (2001). Bovine spongiform encephalopathy and variant Creutzfeldt–Jakob disease: background, evolution, and current concerns. *Emerging Infectious Diseases*, **7**, 6–14.

Bawankule, R., Singh, A., Kumar, K., and Shetye, S. (2017). Does measles vaccination reduce the risk of acute respiratory infection (ARI) and diarrhea in children: a multi-country study? *PLoS One*, 2(1), e0169713.

Castelblanco, R., Lee, M., and Hasbun, R. (2014). Epidemiology of bacterial meningitis in the USA from 1997 to 2010: a population-based observational study. *The Lancet Infectious Diseases*, 14(9), 813–19.

Chisti, M., et al. (2015a). Treatment failure and mortality amongst children with severe acute malnutrition presenting with cough or respiratory difficulty and radiological pneumonia. *PLoS One*, 10(10), e0140327.

Chisti, M., et al. (2015b). Bubble continuous positive airway pressure for children with severe pneumonia and hypoxaemia in Bangladesh: an open, randomised controlled trial. *The Lancet*, 386(9998), 1057–65.

Crump, J.A., Luby, S.P., and Mintz, E.D. (2004). The global burden of typhoid fever. *Bulletin of the World Health Organization*, **82**, 346–53.

Ejemot-Nwadiaro, R.I., Ehiri, J.E., Arikpo, D., Meremikwu, M.M., and Critchley, J.A. (2015). Hand washing promotion for preventing diarrhoea. *Cochrane Database of Systematic Reviews*, (9), CD004265.

Fischer Walker, C.L., et al. (2013). Global burden of childhood pneumonia and diarrhoea. *The Lancet*, **381**, 1405–16.

GAVI—The Vaccine Alliance (2016). *GAVI Progress Report*. [Online] Available at: https://www.gavi.org/sites/default/files/publications/progress-reports/Gavi-Progress-Report-2016.pdf

GBD 2015 LRI Collaborators (2017). Estimates of the global, regional, and national morbidity, mortality, and aetiologies of lower respiratory tract infections in 195 countries: a systematic analysis for the Global Burden of Disease Study 2015. *The Lancet Infectious Diseases*, 17(11), 1133–61.

GBD 2016 Causes of Death Collaborators (2017). Global, regional, and national age–sex specific mortality for 264 causes of death, 1980–2016: a systematic analysis for the Global Burden of Disease Study 2016. *The Lancet*, 390(10100), 1151–210.

GBD 2016 DALYs and HALE Collaborators (2017). Global, regional, and national disability-adjusted life-years (DALYs) for 333 diseases and injuries and healthy life expectancy (HALE) for 195 countries and territories, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *The Lancet*, 390(10100), 1260–344.

GBD Diarrhoeal Diseases Collaborators (2017). Estimates of global, regional, and national morbidity, mortality, and aetiologies of diarrhoeal diseases: a systematic analysis for the Global Burden of Disease Study 2015. *The Lancet Infectious Diseases*, 17(9), 909–48.

Hamer, D.H. and Gorbach, S.L. (2010). Gastrointestinal infections. In: Warrell, D.A., Cox, T.M., and Firth, J.D. (eds.) *Oxford Textbook of Medicine* (5th ed.). Oxford: Oxford University Press pp. 2424–34.

Hotez, P.J., et al. (2004). Hookworm infection. *The New England Journal of Medicine*, **351**, 799–807.

Imdad, A., Bautista, R., Senen, K., Uy, M., Mantaring III, J., and Bhutta, Z. (2013). Umbilical cord antiseptics for preventing sepsis and death among newborns. *Cochrane Database of Systematic Reviews*, (5), CD008635.
Institute for Health Metrics and Evaluation (n.d.). *Global Burden of Disease (GBD)*. [Online] Available at: http://www.healthdata.org/gbd
Iwasaki, Y. (2017). Creutzfeldt–Jakob disease. *Neuropathology*, **37**, 174–88.
Keenan, J., et al. (2018). Azithromycin to reduce childhood mortality in sub-Saharan Africa. *New England Journal of Medicine*, 378(17), 1583–92.
Khan, A. and Bhutta, Z. (2017). Childhood infectious diseases: overview. In: Quah, S., (ed.) *The International Encyclopedia of Public Health* (2nd ed.). Oxford: Elsevier, pp. 517–38.
Klemm, E.J., et al. (2018). Emergence of an extensively drug-resistant Salmonella enterica Serovar Typhi clone harboring a promiscuous plasmid encoding resistance to fluoroquinolones and third generation cephalosporins. *mBio* 9(1), e00105–18.
Kosek, M., Bern, C., and Guerrant, R.L. (2003). The magnitude of the global burden of diarrheal disease from studies published 1992–2000. *Bulletin of the World Health Organization*, **81**, 197–204.
Lassi, Z., Moin, A., and Bhutta, Z. (2016). Zinc supplementation for the prevention of pneumonia in children aged 2 months to 59 months. *Cochrane Database of Systematic Reviews*, (12), CD005978.
Lawn, J., et al. (2014). Every Newborn: progress, priorities, and potential beyond survival. *The Lancet*, 384(9938), 189–205.
Laxminarayan, R. and Bhutta, Z. (2016). Antimicrobial resistance—a threat to neonate survival. *The Lancet Global Health*, 4(10), e676–7.
Lazzerini, M. and Wanzira, H. (2016). Oral zinc for treating diarrhoea in children. *Cochrane Database of Systematic Reviews*, 12(12), CD005436.
Levine, O., et al. (2012). The pneumonia etiology research for child health project: a 21st century childhood pneumonia etiology study. *Clinical Infectious Diseases*, 54(suppl 2), S93–S101.
Liu, L., et al. (2016). Global, regional, and national causes of under-5 mortality in 2000–15: an updated systematic analysis with implications for the Sustainable Development Goals. *The Lancet*, 388(10063), 3027–35.
Lozano, R., et al. (2012). Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010. *The Lancet*, **380**, 2095–128.
Luby, S., et al. (2005). Effect of handwashing on child health: a randomised controlled trial. *The Lancet*, 366(9481), 225–33.
Mandell, L.A., et al. (2007). Infectious Diseases Society of America/American Thoracic Society consensus guidelines on the management of community-acquired pneumonia in adults. *Clinical Infectious Diseases*, 44, S27–S72.
McGuire, LI., et al. (2016). Cerebrospinal fluid real-time quaking-induced conversion is a robust and reliable test for sporadic Creutzfeldt–Jakob disease: an international study. *Annals of Neurology*, **80**, 160–5.
Mogasale, V., et al. (2014). Burden of typhoid fever in low-income and middle-income countries: a systematic, literature-based update with risk-factor adjustment. *The Lancet Global Health*, 2(10), e570–80.
Munos, M., Walker, C., and Black, R. (2010). The effect of oral rehydration solution and recommended home fluids on diarrhoea mortality. *International Journal of Epidemiology*, 39(suppl. 1), i75–i87.
Murray, C.J.L., et al. (2012). Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010. *The Lancet*, **380**, 2197–223.
Nakatani, H. (2016). Global strategies for the prevention and control of infectious diseases and non-communicable diseases. *Journal of Epidemiology*, 26(4), 171–8.

Nigrovic, L., Malley, R., and Kuppermann, N. (2012). Meta-analysis of bacterial meningitis score validation studies. *Archives of Disease in Childhood*, 97(9), 799–805.
Nigrovic, L.E., et al. (2007). Clinical prediction rule for identifying children with cerebrospinal fluid pleocytosis at very low risk of bacterial meningitis. *Journal of the American Medical Association*, **297**, 52–60.
Platts-Mills, J., et al. (2015). Pathogen-specific burdens of community diarrhoea in developing countries: a multisite birth cohort study (MAL-ED). *The Lancet Global Health*, 3(9), e564–75.
Pullan, R., Smith, J., Jasrasaria, R., and Brooker, S. (2014). Global numbers of infection and disease burden of soil transmitted helminth infections in 2010. *Parasites & Vectors*, 7(1), 37.
Ramasamy, I., Law, M., Collins, S., and Brooke, F. (2003). Organ distribution of prion proteins in variant Creutzfeldt–Jakob disease. *The Lancet Infectious Diseases*, **3**, 214–22.
Sabchareon, A., et al. (2012). Protective efficacy of the recombinant CYD tetravalent dengue vaccine in Thai schoolchildren: a randomised controlled phase 2b trial. *The Lancet*, **380**, 1559–67
Saha, S.K., et al. (2018). Causes and incidence of community-acquired serious infections among young children in south Asia (ANISA): an observational cohort study. *Lancet*. [Online] Available at: https://www.thelancet.com/pdfs/journals/lancet/PIIS0140-6736(18)31127-9.pdf
Salam, R., Das, J., Darmstadt, G., and Bhutta, Z. (2013). Emollient therapy for preterm newborn infants—evidence from the developing world. *BMC Public Health*, 13(Suppl 3), S31.
Song, J., Gao, J., and Gálan, J. (2013). Structure and function of the Salmonella Typhi chimeric A2B5 typhoid toxin. *Nature*, **499**, 350–4.
Taylor-Robinson, D., Maayan, N., Soares-Weiser, K., Donegan, S., and Garner, P. (2015). Deworming drugs for soil-transmitted intestinal worms in children: effects on nutritional indicators, haemoglobin, and school performance. *Cochrane Database of Systematic Reviews*, (7), CD000371.
Tshefu, A., et al. (2015). Simplified antibiotic regimens compared with injectable procaine benzylpenicillin plus gentamicin for treatment of neonates and young infants with clinical signs of possible serious bacterial infection when referral is not possible: a randomised, open-label, equivalence trial. *The Lancet*, 385(9979), 1767–76.
UNAIDS.org. (2017). Fact sheet - Latest statistics on the status of the AIDS epidemic. [Online] Available at: http://www.unaids.org/en/resources/fact-sheet
United Nations Inter-agency Group for Child Mortality Estimation (UN IGME) (2015). *Levels and Trends in Child Mortality Report 2015*. [Online] Available at: https://www.un.org/en/development/desa/population/publications/mortality/child-mortality-report-2015.asp
United Nations Inter-agency Group for Child Mortality Estimation (UN IGME) (2017). *Levels and Trends in Child Mortality Report 2017*. [Online] Available at: https://www.un.org/en/development/desa/population/publications/mortality/child-mortality-report-2017.asp
Vesikari, T., et al. (2013). Immunogenicity and safety of an investigational multicomponent, recombinant, meningococcal serogroup B vaccine (4CMenB) administered concomitantly with routine infant and child vaccinations: results of two randomised trials. *The Lancet*, **381**, 825–35.
Victora, C.G., et al. for The Lancet Breastfeeding Series Group (2016). Breastfeeding in the 21st century: epidemiology, mechanisms, and lifelong effect. *The Lancet*, **387**, 475–90.
Waddell, L., Greig, J., Mascarenhas, M. Otten, A., Corrin, T., and Hierlihy, K. (2018). Current evidence on the transmissibility of chronic wasting disease prions to humans—a systematic review. *Transboundary & Emerging Diseases*, **65**, 37–49.
Walker, C., et al. (2013). Global burden of childhood pneumonia and diarrhoea. *The Lancet*, 381(9875), 1405–16.
World Health Organization (WHO) (2015). *Summary of WHO Position Paper on Meningococcal A Conjugate Vaccine: Updated Guidance*. Geneva: WHO. [Online] Available at: https://www.who.int/immunization/policy/position_papers/pp_menA_2015_summary.pdf?ua=1

World Health Organization (WHO) (2008). *Closing the Gap in a Generation: Health Equity Through Action on the Social Determinants of Health*. Geneva: WHO. [Online] Available at: https://www.who.int/social_determinants/final_report/csdh_finalreport_2008.pdf

World Health Organization (WHO) (2013). *WHO Recommendations on Postnatal Care of the Mother and Newborn*. Geneva: WHO. [Online] Available at: https://www.who.int/maternal_child_adolescent/documents/postnatal-care-recommendations/en/

World Health Organization (WHO) (2014). *Revised WHO Classification and Treatment of Childhood Pneumonia at Health Facilities*. Geneva: WHO. [Online] Available at: https://www.who.int/maternal_child_adolescent/documents/child-pneumonia-treatment/en/

World Health Organization (WHO) (2015). *Guideline: Managing Possible Serious Bacterial Infection in Young Infants When Referral is not Feasible*. Geneva: WHO. [Online] Available at: https://www.who.int/maternal_child_adolescent/documents/bacterial-infection-infants/en/

World Health Organization (WHO) (2015). *Health in 2015 from MDGs to SDGs*. Geneva: WHO. [Online] Available at: https://www.who.int/gho/publications/mdgs-sdgs/en/

World Health Organization (WHO) (2018a). *WHO Recommends Use of First Typhoid Conjugate Vaccine*. Geneva: WHO. [Online] Available at: http://www.who.int/immunization/newsroom/press/who_recommends_use_first_typhoid_conjugate_vaccine/en/

World Health Organization (WHO) (2018b). *Executive Summary: CYD:TDV Dengue Vaccine*. Geneva: WHO. [Online] Available at: http://www.who.int/immunization/sage/meetings/2018/april/1_WHO_DengueExecSum_SAGE_Apr2018.pdf

World Health Organization (WHO) (2018c). *The Top 10 Causes of Death*. Geneva: WHO. [Online] Available at: http://www.who.int/en/news-room/fact-sheets/detail/the-top-10-causes-of-death

World Health Organization/The United Nations Children's Fund (WHO/UNCF) (2013). *Ending Preventable Child Deaths from Pneumonia and Diarrhoea by 2025–The integrated Global Action Plan for Pneumonia and Diarrhoea (GAPPD)*. Geneva: World Health Organization/The United Nations Children's Fund. [Online] Available at:

Zaidi, A., Huskins, W., Thaver, D., Bhutta, Z., Abbas, Z., and Goldmann, D. (2005). Hospital-acquired neonatal infections in developing countries. *The Lancet*, 365(9465), 1175–88.

8.12

Sexually transmitted infections

Noah Kojima and J.D. Klausner

Introduction to sexually transmitted infections

Sexually transmitted infections (STIs) refer to many pathogens that are commonly spread through sexual contact including vaginal and anal intercourse and oral sex. *Chlamydia trachomatis* (chlamydia), *Neisseria gonorrhoeae* (gonorrhoea), and *Treponema pallidum* (syphilis) cause the majority of worldwide bacterial STIs. Viral STIs include: human papillomavirus (HPV), herpes simplex virus type 1 and 2 (genital herpes; HSV-1 and HSV-2), hepatitis A, B, and C (unrelated viruses that can cause liver disease), as well as human immunodeficiency virus (HIV). Parasites like *Trichomonas vaginalis*, the causative agent of trichomoniasis, are also a major contributor to the global burden of STIs. In total, there are over 30 organisms and hundreds of those organism subtypes that have been recognized as sexually transmissible (Table 8.12.1). The newest agents recognized as sexually transmissible include Zika and Ebola viruses.

STIs are common. The global annual incidence of curable (bacterial and protozoal) STIs estimated by the World Health Organization (WHO) was 357.4 million cases in 2012 (WHO 2015). Furthermore, among countries with public health surveillance systems, STIs are among the most frequently reported infections. However, compared to bacterial or protozoal STIs, viral STIs are far more prevalent with an estimated 708 million cases of Herpes Simplex Virus and HPV infection.

Given the high numbers of STIs that exist in almost every nation, it is not surprising to find that regardless of a nation’s resources, STI-related symptoms can account for a large number of healthcare visits and substantial costs. STIs are also stigmatized and may also result in psychosocial consequences such as anxiety, shame, and stress. Additionally, sequelae like

infertility, adverse pregnancy outcomes, and malignancies, and the morbidities of STIs increase healthcare costs.

Although STIs are commonly associated with genital sores, discharges, or rashes, most STIs are asymptomatic. Many countries around the world continue to use syndromic management (the identification and management of persons with symptoms) as a way to manage STIs clinically, however many non-specific symptoms can lead to overtreatment for suspected STIs and lack of symptoms can lead to under treatment. Therefore, in countries with available infrastructure, screening asymptomatic people at risk for STIs is important for individual prevention and public health. While STI case rates are highest in adolescents and young adults, the most serious health outcomes occur later in life and are disproportionately borne by women and infants. For those reasons, STI prevention and control is an important public health investment for every country.

There are many methods that can be used to reduce the risk of STI acquisition. Vaccinations protect against some viral STIs, like hepatitis A, hepatitis B, and wart or cancer-causing HPV types. Additionally, barrier protection, e.g. male and female condoms, reduce or even eliminate the risk of STI spread or acquisition.

There are many diagnostic tests for STIs. Various fluids, including saliva, urine, blood, vaginal or anal secretions, and self-collected tip of the urethra swabs, can be used for testing. It is important to test any orifice that may have been exposed to an STI during sexual activity because STIs are often localized mucosal infections that are not detected systemically in the blood or urine. Additionally, tests utilize many different biochemical mechanisms that can allow rapid point-of-care testing for some STIs, like HIV and syphilis. Other testing systems, like polymerase chain reaction (PCR), normally require specialized laboratory equipment although new point-of-care PCR-based systems enable detection of STIs in 60–90 minutes.

Bacterial STIs and trichomonal infections are curable. Unfortunately, individuals do not gain immunity to an STI, and are equally susceptible to reinfection. In contrast to bacterial STIs, viral STIs are not medically curable with the exception of hepatitis C infection (rarely sexually transmitted). Other STIs like genital herpes and HIV infection are incurable, although medicines exist to manage chronic infections and reduce the risk of transmission. Some viral STIs and their sequelae are preventable with vaccines, e.g. HPV, hepatitis B virus (HBV), and hepatitis A virus (HAV).

Others, like HPV, HBV, and HCV infection can spontaneously clear without medical intervention.

The field of STIs is constantly changing. New threats, like the Zika virus are emerging, and some cases were sexually transmitted, while known STIs continue to cause challenges, for example, antibiotic-resistant *Neisseria gonorrhoeae*. At the same time, new infection prevention tools are being created. For example, pre-exposure prophylaxis for HIV infection as well as post-exposure prophylaxis for syphilis and chlamydia. Additionally, new methods for diagnosis, like point-of-care HIV and syphilis testing, are being mobilized around the world.

In this chapter we summarize the global epidemiology of STIs and their associated health consequences, and report on factors affecting the spread of STIs in the community. We also discuss STI prevention and control as a public health intervention, relying on many interrelated interventions working together to reduce STI incidence and prevalence in the community. Finally, we consider some of the most likely challenges and opportunities in STI prevention anticipated over the next few decades.

Table 8.12.1 Sexually transmitted pathogens and associated diseases or syndromes

Pathogen	Associated disease or syndrome
Bacteria	
Neisseria gonorrhoeae	*Adult:* cervicitis, urethritis, proctitis, pharyngitis, Bartholinitis, endometritis, pelvic inflammatory disease (PID), infertility, chronic pelvic pain, orchitis, epididymitis, urethral stricture, prostatitis, perihepatitis, disseminated infection, inflammatory arthritis syndrome, enhanced HIV risk, often asymptomatic (up to 2/3 women, 1/3 men)
	Maternal: ectopic pregnancy, maternal death, preterm rupture of membranes *Infant:* neonatal conjunctivitis, corneal scarring, blindness, premature birth, low birth weight (LBW)
Chlamydia trachomatis	*Adult:* cervicitis, urethritis, proctitis, pharyngitis, Bartholinitis, endometritis, PID, infertility, chronic pelvic pain, orchitis, epididymitis, urethral stricture, prostatitis, perihepatitis, disseminated infection, inflammatory arthritis syndrome, lymphogranuloma venereum (LGV)—anogenital ulcer or inguinal swelling, enhanced HIV risk, often asymptomatic (up to 2/3 women, 1/3 men)
	Maternal: ectopic pregnancy, maternal death, preterm rupture of membranes
	Infant: neonatal conjunctivitis, pneumonia, premature birth, LBW
Mycoplasma genitalium	*Adult:* urethritis, cervicitis, PID, enhanced HIV risk
Treponema pallidum (syphilis)	*Adult:* genital ulcer (chancre), local adenopathy, skin rashes, condyloma lata, hepatitis, arthritis, enhanced HIV risk; bone, cardiovascular (e.g. aortic disease) and central nervous system disease (e.g. meningitis, stroke, cranial nerve abnormalities, optic atrophy, tabes dorsalis, general paresis)
	Maternal: fetal loss, stillbirth, preterm delivery *Infant:* neonatal death, congenital syphilis, LBW
Haemophilus ducreyi (chancroid)	*Adult:* genital ulcers, inguinal adenitis, disfiguring lesions, tissue destruction, enhanced HIV risk
Calymmatobacterium granulomatis (Donovanosis)	*Adult:* Nodular swellings and chronic ulcerative lesions of inguinal and anogenital areas (also called granuloma inguinale)
Shigella spp.	*Adult:* shigellosis in homosexual men
Salmonella spp.	*Adult:* enteritis, proctocolitis in homosexual men

Pathogen	Associated disease or syndrome
Campylobacter spp.	*Adult:* enteritis, proctocolitis in homosexual men
Viruses	
Human immunodeficiency virus types 1 and 2	*Adult:* HIV-related disease, opportunistic infections, lymphomas, AIDS *Maternal:* vertical transmission to infants
	Infant: HIV infection
Herpes simplex virus types 1 and 2	*Adult:* anogenital vesicular lesions and ulcerations, recurrent genital ulcers, cold sores, cervicitis, urethritis, pharyngitis, proctitis, chronic pain, arthritis, aseptic meningitis, hepatitis, enhanced HIV risk
	Maternal: vertical transmission to infants
	Infant: ulcerations of skin, eye, mucous membranes; encephalitis, disseminated infection with hepatitis, pneumonitis, encephalitis; neurological abnormalities
Human papilloma virus (> 30 genital genotypes identified)	*Adult:* anogenital and oral warts; intraepithelial neoplasia and carcinoma of the cervix, penis, vulva, vagina, anus; oropharyngeal cancer; recurrent respiratory papillomatosis
	Maternal: vertical transmission to infant *Infant:* recurrent respiratory papillomatosis
Hepatitis B virus	*Adult:* acute hepatitis, liver cirrhosis, end-stage liver disease, hepatocellular cancer
	Maternal: vertical transmission to infant
	Infant: cirrhosis, end-stage liver disease, primary liver cancer
Hepatitis A virus	*Adult:* acute hepatitis A
Hepatitis C virus	*Adult:* acute hepatitis C, liver cirrhosis, end-stage liver disease, hepatocellular cancer
Cytomegalovirus (CMV)	*Adult:* heterophil-negative infectious mononucleosis, hepatitis *Infant:* primary infection of the newborn, hepatitis, sepsis, deafness, mental retardation
Molluscum contagiosum virus	*Adult:* genital molluscum contagiosum
Human T-lymphotrophic retrovirus, type 1	*Adult:* human T-cell leukaemia or lymphoma

Pathogen	Associated disease or syndrome
Human herpesvirus 8 (HHV-8)	*Adult:* Kaposi's sarcoma, primary effusion lymphoma, Castleman's disease
Zika	*Maternal:* vertical transmission to infant
	Infant: microcephaly
Protozoa	
Trichomonas vaginalis	*Adult:* vaginitis, cervicitis, urethritis, endometritis, salpingitis, enhanced HIV risk *Maternal:* chorioamnionitis, preterm delivery *Infant:* LBW, pneumonitis, fever, vaginal discharge in female infants
Entamoeba histolytica	*Adult:* amoebiasis in men who have sex with men
Giardia lamblia	*Adult:* giardiasis in men who have sex with men
Ectoparasites	
Phthirus pubis	*Adult:* pubic lice infestation
Sarcoptes scabiei	*Adult:* scabies, Norwegian (disseminated) scabies
	Infants: Norwegian (disseminated) scabies

Sexually transmitted infections

The global burden of STIs includes the overall number of cases of infection, a host of short- and long-term adverse health outcomes in individuals including death, and the financial consequences (healthcare and other costs) associated with STIs and their associated health, and psychosocial impact for both individuals and economies.

Global STI numbers

The WHO estimated that approximately 357 million new cases of four curable STIs, i.e. *Chlamydia trachomatis*, *Neisseria gonorrhoeae*, *Treponema pallidum*, and *Trichomonas vaginalis*, occurred in 2012 (WHO 2015). There were 130.9 million cases of chlamydia, 78.3 million cases of gonorrhoea, 5.6 million cases of syphilis, and 142.6 million cases of trichomoniasis. The majority of countries, 108 of 198, have an STI

surveillance system in place, however STI surveillance has not been globally standardized, so it is difficult to assess the spread of concerning infections, like untreatable *Neisseria gonorrhoeae*. While new cases of STIs were common, the areas with the highest rates of STIs were Africa, the Americas, and the Western Pacific. Compared to prior STI prevalence estimations, the overall global prevalence of curable STIs seems to be decreasing. However, due to differences in STI surveillance practices, many STI prevalence estimations are poor because they are limited to small studies among accessible populations and case-based surveillance reports.

Compared to the WHO report on curable STIs from 2008, there were fewer overall cases of STIs estimated in 2012 (357 million versus 498.9 million, respectively). While the number of cases of chlamydia increased from 105.7 million cases to 130.9 million cases, the worldwide number of cases of gonorrhoea, syphilis, and trichomoniasis decreased (Fig. 8.12.1). In accordance with the WHO global strategy for STI prevention and control, over 90% of countries worldwide have adopted syndromic management of STIs. The WHO has recommended monitoring two syndromes: urethral discharge among men and genital ulcer disease among men and women. Urethral discharge is most commonly caused by *Neisseria gonorrhoeae* infection and *Chlamydia trachomatis* infection. Genital ulcer disease is classically caused by syphilis and chancroid, however viral infections, like herpes simplex virus, can also cause this syndrome. Genital ulcers facilitate HIV transmission. In 2014, 56 countries reported data on genital ulcer disease (Fig. 8.12.2). From those data, the African Region had the highest burden of genital ulcer disease, followed by the Americas and the Western Pacific.

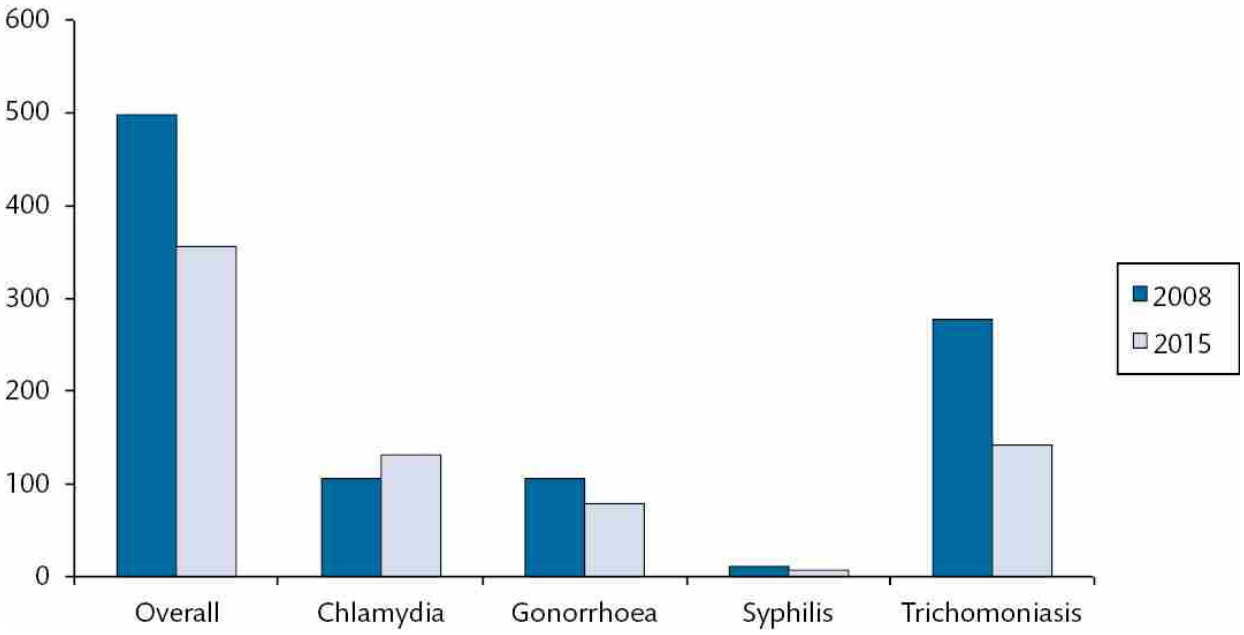


Fig. 8.12.1 New cases (millions) of four curable STIs in 2008 and 2015.

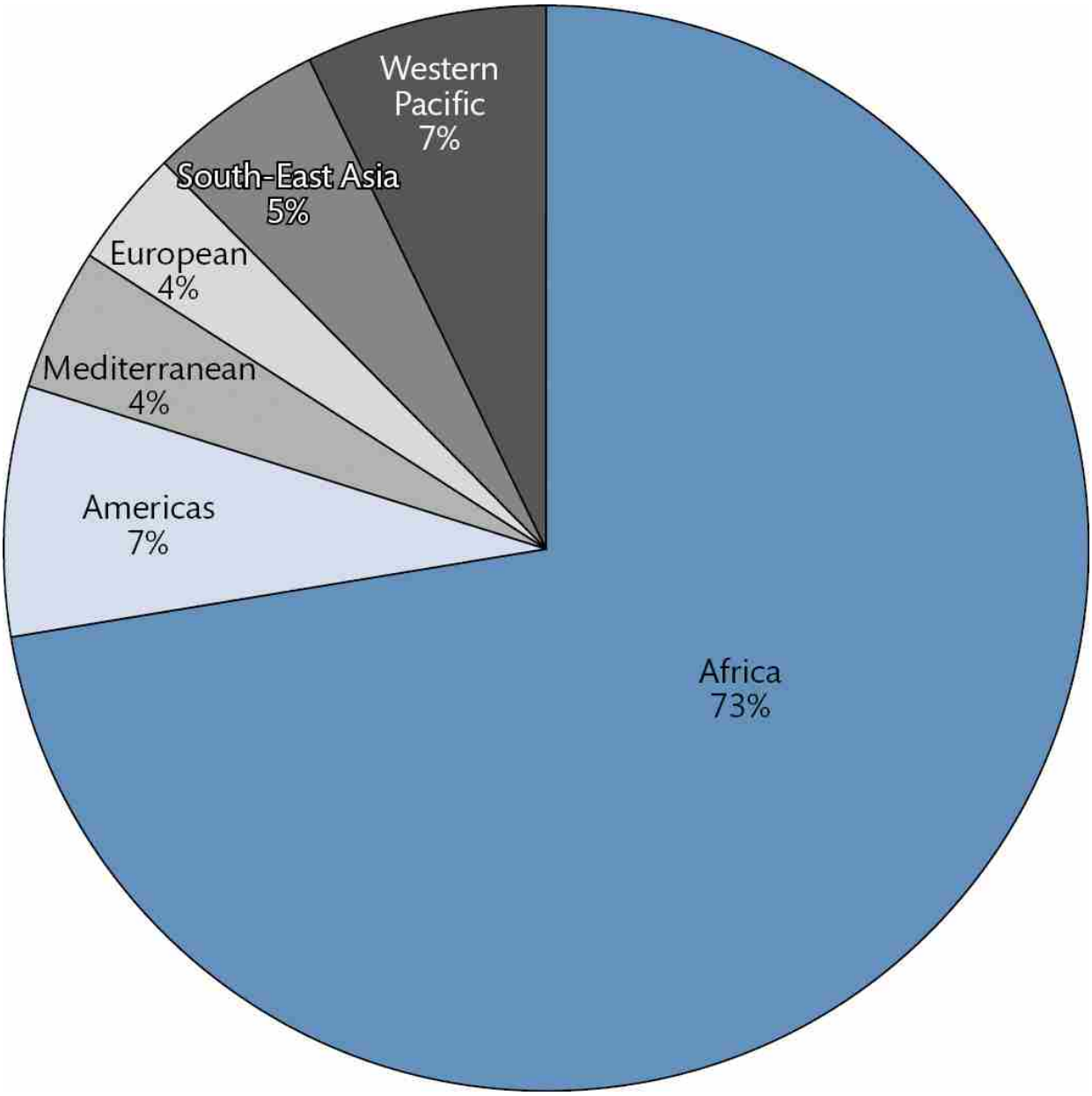


Fig. 8.12.2 Proportion of genital ulcer disease burden by geographic regions, 2014.

While syndromic management is a useful strategy to treat STIs in those presenting to clinics and estimate disease burden, laboratory-based STI testing has superior STI sensitivity and specificity. However, the accessibility and feasibility of laboratory-based STI testing is more limited worldwide due to the lack of political will, absence of infrastructure, and the cost of testing. There are many different laboratory-based methods to test for STIs, mainly *Neisseria gonorrhoeae* infection and syphilis. Microscopy with Gram stain and bacterial culture can detect *Neisseria gonorrhoeae*. Newer

nucleic acid-amplification-based tests have become increasingly useful in the detection of *Neisseria gonorrhoeae* due to their higher sensitivity and specificity. Point-of-care, rapid syphilis tests have become useful in initial syphilis screening, however that test must be used on those without a known history of syphilis or treatment, because the test cannot distinguish between past or current infection. The confirmation of a diagnosis of syphilis necessitates two tests, treponemal and non-treponemal tests, which are not always available worldwide. Therefore, national syphilis estimates can be discordant with the actual burden of syphilis.

Viral STIs

In contrast to bacterial STIs, the only curable viral STI is hepatitis C. Other STIs like genital herpes and HIV infection are incurable; however, they can be managed with antiviral medicines. While HIV infection may be the most well-known viral STI and contributes to a total of 36.9 million people living with a chronic infection, there are an estimated 417 million cases of herpes simplex virus type-2 infection, and roughly 291 million women are infected with human papillomavirus at any given time, worldwide.

Genital herpes infection is a chronic, life-long viral infection caused by HSV-1 or HSV-2. It was estimated that in 2012, there were 417 million people aged 15 to 49 years living with HSV-2 infection. More women had HSV-2 infection when compared to men (267 million persons versus 150 million persons), and this was attributed to women being more susceptible to infection than men. Worldwide, Africa had the highest prevalence of HSV-2 infection, followed by the Western Pacific and Southeast Asia (Fig. 8.12.3) (Looker et al. 2015). Genital herpes disease and transmission can be reduced with medical therapy.

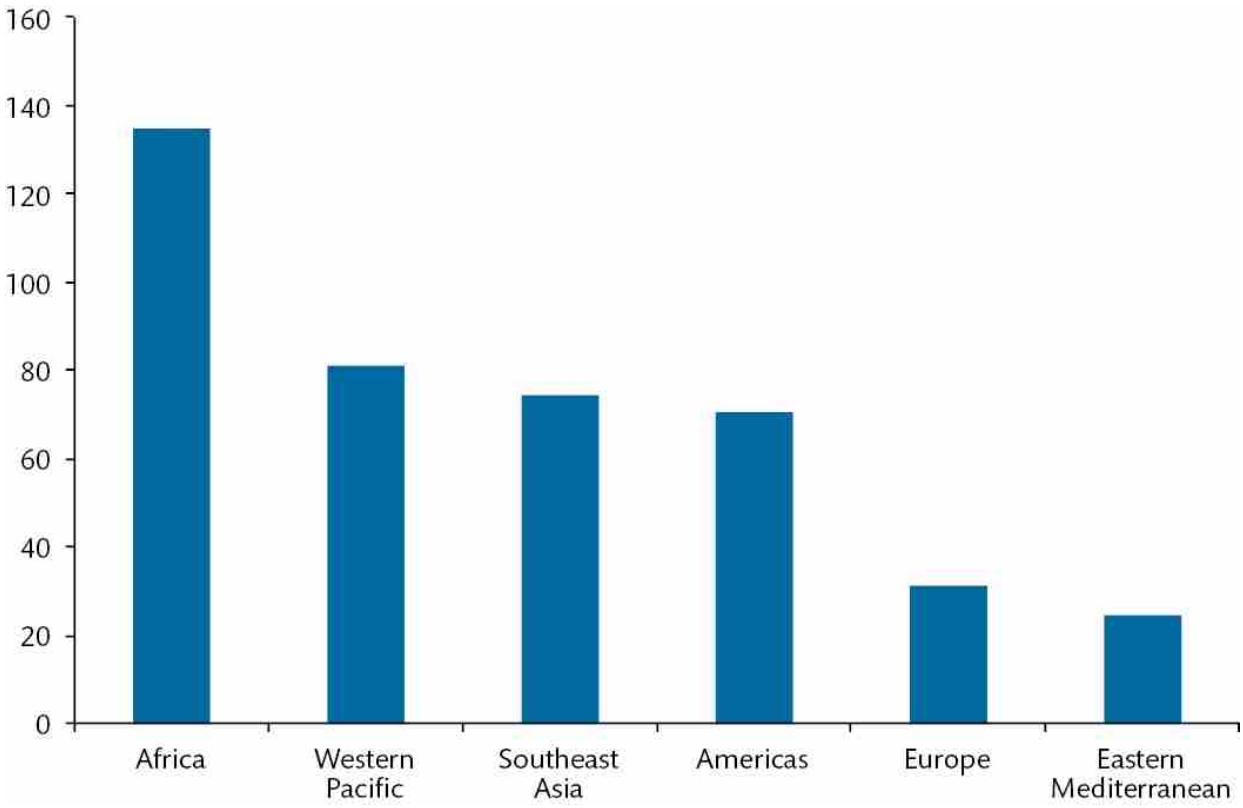


Fig. 8.12.3 Estimated prevalence (millions) of Herpes simplex virus-2 infection by region, 2012. Source: data from Looker K. et al 'Global Estimates of Prevalent and Incident Herpes Simplex Virus Type 2 Infections in 2012'. *PLoS One*, 10, e114989.

Roughly 291 million women worldwide are infected with human papillomaviruses, a group of viruses that consist of over 100 different sub-types of viruses, of which about 30 primarily infect the oropharynx (throat) and genitals. Among those viruses, there are at least 13 cancer-causing varieties, which are classified as 'high-risk' viruses. Cervical cancer is the second most common cancer among women worldwide. In 2012, nearly 270,000 women died from cervical cancer, and nearly 230,000 of those deaths occurred in low- and middle-income countries. In low-income countries, the relevant disability-adjusted life years (DALYs) lost to cervical cancer are second only to oropharyngeal cancer (American Cancer Society 2015). Two human papillomaviruses, type 16 and 18, cause 70% of cervical cancers and precancerous cervical lesions. Worldwide, human papillomaviruses type 16 is the most prevalent, followed by type 18 (Bruni et al. 2010). In the United States, there are 35,000 HPV-associated cancers annually (61% in females and 39% in males), with an annual cost of US $8

billion for screening and treatment of HPV-associated health outcomes (Jemal et al. 2013).

HIV infection continues to cause significant morbidity and mortality. Since access to antiretroviral drugs expanded in the early 2000s, an estimated 20 million have been treated and 5.2 million AIDS-related deaths have been averted in low- and middle-income countries (UNAIDS 2013b). Furthermore, children have seen the largest decrease in HIV incidence, which has been reduced by over half in the last 10 years (UNAIDS 2013a). While great progress has been made in the reduction in the spread of HIV, 36.9 million people are living with chronic HIV infection. Sub-Saharan Africa continues to carry the largest burden of those living with, dying from, and acquiring HIV infection (Fettig et al. 2014).

Viral hepatitis caused 1.34 million deaths in 2015, with an increasing trend (WHO 2017a). Many unrelated viruses that all cause liver disease are grouped together as viral hepatitis. While sexual transmission of hepatitis A, B, C, and D have been observed; hepatitis B is the most commonly sexually transmitted hepatitis virus. Most deaths caused by the viral hepatitis were due to chronic liver disease (720,000 deaths due to cirrhosis), and followed by primary liver cancer (470,000 deaths due to hepatocellular carcinoma). Hepatitis B and C were responsible for 96% of all hepatitis related mortality. Globally, it is estimated that 257 million people are living with chronic hepatitis B infection and 71 million people were living with a chronic hepatitis C infection in 2015. Hepatitis B has the highest prevalence in Africa and Western Pacific, whereas hepatitis C has the highest prevalence in the Europe. In May 2016, the World Health Assembly endorsed a Global Health Sector Strategy on viral hepatitis 2016–2021 to reduce new infections by 90%, and mortality by 65% by 2030. Other viral hepatitides are described in Table 8.12.2.

Table 8.12.2 Viral hepatitis

A	Caused approximately 11,000 deaths in 2015. A vaccine is available. It is primarily transmitted through exposure to contaminated food or water. It can be acquired through sexual activity.
B	Caused over 880,000 deaths in 2015. A vaccine is available. It is primarily transmitted through blood or semen.
C	Caused over 400,000 deaths in 2015. Medicine is available to cure chronic hepatitis C infections. It is primarily transmitted through blood.
D	Older interferon-based medicines can be used to treat this virus with suboptimal results. It is an incomplete virus and can only infect persons with a current hepatitis B infection. It is transmitted mostly through exposure to blood. Persons with hepatitis B infection with a hepatitis D coinfection lead to worse health outcomes. Most experts estimate that 5% of persons infected with hepatitis B are co-infected with hepatitis D.
E	Caused approximately 44,000 deaths and 20 million infections in 2015. A vaccine to prevent hepatitis E infection has been developed and licensed in China; however, it does not have a wide global availability. Hepatitis E is transmitted via the faecal-oral route, primarily by contaminated water. It has a higher case fatality among pregnant women.

While most STI acquisition occurs during sexual activity, mother-to-child transmission (MTCT) of STIs can occur *in utero* or perinatally. In 2018, UNAIDS estimated that 180,000 children acquired HIV infection, predominantly through perinatal transmission or contaminated breast milk (UNAIDS 2018). Antenatal HIV screening and maternal treatment programmes have greatly reduced the frequency of MTCT of HIV infection, when compared to the estimated 420,000 children that acquired HIV infection at the peak of the epidemic in 2003. In eastern and southern Africa, the area with the highest frequency of MTCT of HIV infection in the world, 93% of the 940,000 pregnant women with HIV infection received antiretroviral prophylaxis in 2017, reducing the average frequency of MTCT of HIV to less than 10%.

T. pallidum, the causative agent of syphilis, can readily cross the placenta during pregnancy and can cause morbidity and mortality among mothers and newborns. Between 2008 and 2012, the estimated total number of maternal infections and adverse pregnancy outcomes attributed to congenital syphilis decreased by 38% from 520,000 to 350,000 cases, which consisted of 143,000 early fetal deaths/stillbirths, 62,000 neonatal deaths, 44,000 preterm/low weight births, and 102,000 infants with syphilis (WHO 2015). While declines in congenital syphilis were seen in all regions, the largest

declines were observed in South-East Asia. As of 2018, many countries have been successful in the dual elimination of MTCT of HIV and syphilis, including Cuba, Thailand, Belarus, Anguilla, Antigua, and Barbuda, Bermuda, Cayman Islands, Montserrat, and St. Kitts and Nevis (WHO 2014; WHO 2017b). The WHO acknowledged Armenia for eliminating MTCT of HIV and the Republic of Moldova for eliminating MTCT of syphilis. Strategies to eliminate MTCT of HIV include prevention of unintended pregnancies, increased access to testing and counselling, medical therapy, safe delivery practices, and appropriate infant feeding practices (WHO 2014). Strategies to eliminate of MTCT of syphilis include early antenatal testing and treatment. Other STIs, like *Chlamydia trachomatis*, *Neisseria gonorrhoeae*, and HSV infections, can be transmitted from mothers to newborns as a result of exposure to the pathogen as the neonate passes through the birth canal. Strategies to control those infections are discussed later.

More research has emerged since the 2015 Zika virus outbreak in Brazil. Zika virus is now recognized as sexually transmissible, although its primary route of transmission (>99%) is through mosquitoes (Baud et al. 2017). Among pregnant women, Zika virus infection has been associated with severe adverse infant outcomes including miscarriage, stillbirth, and microcephaly. Additionally, new research supports that MTCT of Zika virus can occur perinatally (Blohm et al. 2018).

For neonatal HSV-1 and HSV-2 infection, a devastating infection of newborns with > 50% mortality, it is estimated that approximately 14,000 cases occurred annually globally between 2010 and 2015 (4,000 cases for HSV-1 and 10,000 cases for HSV-2) (Looker et al. 2017). In the Americas, Europe, and Western Pacific, HSV-1 contributed to more cases than HSV-2, whereas in Africa, HSV-2 contributed to more infections. For congenital cytomegalovirus (CMV) infection, an infection causing deafness and neurodevelopmental problems, it is estimated that the incidence of congenital CMV infection among all live births ranges from 0.6% to 0.7% in developed countries and 1% to 5% in developing counties (Marsico and Kimberlin 2017), far more common than Zika virus.

STI-associated morbidity

The presentation of an STI can range from an asymptomatic infection to a life threatening one. See Table 8.12.1 for a list of STI-related diseases or syndromes. The acquisition of STIs can also lead to stigma, shame, and psychological harm (Cunningham et al. 2009). Additionally, many STIs can lead to longer-term health problems, which can be categorized into four broad categories: female reproductive health disorders, e.g. infertility; adverse pregnancy outcomes, e.g. neonatal death; malignancies, e.g. cervical cancer associated with HPV; and increased risk of HIV infection acquisition.

Chlamydia trachomatis and *Neisseria gonorrhoeae* infections are important causes of infertility (CDC 2018c). While it is possible to test for and treat those infections, 10–15% of women with chlamydial infection who go untreated will develop pelvic inflammatory disease. Infections in the upper genital tract can cause permanent damage from inflammation and scarring of the fallopian tubes, uterus, and surrounding tissues leading to infertility. Scarring of the fallopian tubes also increases the risk of ectopic pregnancy, which can be a serious medical or surgical emergency. *Chlamydia trachomatis* infection has also been associated with HPV-induced cervical cancer (Madeleine et al. 2007). Screening programmes have been shown to be effective in reducing rates of chlamydial and gonorrhoeal female reproductive health disorders (Workowski et al. 2015). Among women with HIV infection, *Trichomonas vaginalis* infection has been associated with increased risk for pelvic inflammatory disease.

Many STIs are associated with adverse pregnancy outcomes, causing morbidity and mortality among newborns. One of the most devastating STIs in pregnancy is syphilis. It is estimated that fetal loss and still birth were 21% more frequent, neonatal deaths were 9.3% more frequent, and prematurity or low birth weight were 5.8% more frequent among pregnant women with syphilis versus pregnant women without syphilis (Gomez et al. 2013). Additionally, 15% of newborns of mothers with untreated syphilis had clinical evidence of congenital syphilis. For *Chlamydia trachomatis* and *Neisseria gonorrhoeae* infections, exposure to the pathogen occurs during delivery. Untreated *Chlamydia trachomatis* infection causes significant perinatal morbidity including blindness and pneumonia. Untreated *Neisseria gonorrhoeae* infection can cause rhinitis, vaginitis, urethritis, blindness, sepsis, arthritis, and meningitis. After delivery, prophylactic antibiotic ointments prevent eye infections from those infections. Prenatal screening and treatment prevent chlamydial and gonococcal infections among neonates

(Workowski et al. 2015). Apart from the well-characterized features of HIV infection, many other viral STIs cause morbidity and mortality among newborns. Neonatal herpes is associated with skin, eye, and mouth disease, lasting neurological impairment, disseminated infection, and, has an estimated 60% fatality rate among newborns without treatment (Looker et al. 2017). Recurrent respiratory papillomatosis can be a serious disease among newborns that acquire perinatal HPV infection (CDC 2017). Among pregnant women with hepatitis B virus, without post-exposure prophylaxis, roughly 40% of infants develop a chronic infection. Among those infants, approximately one-fourth of those will eventually die from sequela of chronic liver disease (CDC 2018d). More recently, it has been found that Zika virus infection during pregnancy is associated with microcephaly, abnormal brain and eye development, hearing loss, and damage to muscles and bones (CDC 2018b).

STIs are also associated with certain malignancies, which can occur during a chronic infection. Major STI-associated malignancies include anogenital cancers associated with HPV, hepatocellular carcinoma associated with chronic hepatitis, some lymphomas are associated with HIV infection, and Kaposi's sarcoma is associated with human herpes virus type 8 infection. Low- and middle-income counties have a higher STI-related cancer burden than high-income counties.

Where there are more than 100 types of HPVs, 13 HPVs have been identified as cancer causing (WHO 2018). Among the cancer-causing HPVs, HPV type 16 and 18 cause 70% of all cervical cancers and precancerous cervical lesions. HPV is also linked with cancers of the anus, vulva, vagina, and penis (Crosbie et al. 2013). Worldwide, cervical cancer is the fourth most frequent malignancy among women. In 2012, there were an estimated 530,000 new cases of cervical cancer among women, with approximately 270,000 deaths (WHO 2018). Over 85% of those deaths occurred in low- and middle-income counties. The progression from chronic HPV infection to cervical cancer is insidious, taking 5 to 10 years to develop among women with weakened immune systems, e.g. infection with HIV, and 15 to 20 years to develop among women with normal immune systems. In high-income counties, routine screening and treatment prevents up to 80% of cervical cancers. As of 2018, cervical cancer can be screened with conventional Pap testing and liquid-based cytology, visual inspection with acetic acid, and molecular testing for cancer-causing HPV types. Highly effective and safe

vaccinations are available for cancer-causing HPV types. Increased use of HPV vaccines has great potential to prevent HPV-associated cancers among women and men worldwide.

In 2015, it was estimated that 470,000 deaths were due to hepatocellular carcinoma secondary to viral hepatitis (WHO 2017a). The majority of cases of hepatocellular carcinoma are associated with chronic hepatitis B and hepatitis C virus infection (El-Serag 2012). Nearly all childhood and approximately half of all cases of hepatocellular carcinoma can be attributed to chronic hepatitis B infection. Additionally, it is estimated that 2.7 million people are living with a hepatitis B and HIV co-infection (WHO 2017a). The prevalence of hepatitis B is highest in the Western Pacific, Africa, and followed by South-East Asia. For prevention, the global coverage of hepatitis B vaccination reached 84% in 2015 reducing rates of transmission in the first five years of life. While challenges are still present, the WHO has created strategies to eliminate viral hepatitis by 2030 by increasing global surveillance, scaling up testing and treatment infrastructure, and the development of new diagnostics and treatments.

Patients with HIV infection are at increased risk of many cancers due to immunosuppression (Yarchoan and Uldrick 2018). AIDS-defining malignancies include non-Hodgkin's lymphoma, Kaposi's sarcoma, and cervical cancer. Patients with HIV infection are also at increased risk of developing lung cancer, anal cancer, Hodgkin's lymphoma, oral cavity cancer, pharyngeal cancer, hepatocellular carcinoma, vulvar cancer, and penile cancer, however those are non-AIDS-defining malignancies.

STIs can also increase risk of transmitting and acquiring HIV infection. STIs that cause genital ulcers, like syphilis or genital herpes, can increase the risk of acquiring or transmitting HIV infection by two to three times (WHO 2016). Other STIs, like *Chlamydia trachomatis* and *Neisseria gonorrhoeae*, and *Trichomonas vaginalis* infections, and STI-related conditions like bacterial vaginosis, seem to increase risk of HIV transmission and acquisition due to mucosal inflammation (Galvin and Cohen 2004; Ward and Ronn 2010). STIs, like syphilis, have been observed to increase HIV viral load and decrease CD4 T-cell counts (Buchacz et al. 2004).

Global STI management costs

The WHO Global Health Sector Strategy to manage global STIs was proposed in 2016, and set goals to end STI epidemics as a major public health concern by 2030. To achieve those targets, 70% of counties need to have a STI surveillance system in place; 70% of countries need to have 95% of pregnant women screened for syphilis and HIV, with 95% of those with syphilis receiving treatment; 70% of key populations need access to STI and HIV services; 70% of countries need to provide or have access to STI services in all HIV, reproductive health, family planning, and ante- and post-natal care services; 70% of countries need to deliver HPV vaccines, and those countries that provide that service need 90% national coverage in at least 80% of districts; and 70% of countries report antimicrobial resistance in *Neisseria gonorrhoeae* infections (WHO 2016). While some targets, like HPV vaccination, will generate significant returns on investment, initial implementation of those goals will be expensive. A recent analysis estimated that the total implementation cost will be US $18.2 billion over 2016 to 2021 in 117 low- and middle-income counties, with an initial $3.26 billion for HPV vaccination, $3.69 billion for HPV screening, $2.54 billion for chlamydial screening, and $1.4 billion for syphilis screening (Korenromp et al. 2017). Clinical STI management will add another $3 billion, with $1.4 billion going to diagnostic testing for *Chlamydia trachomatis* and *Neisseria gonorrhoeae* infection and $0.8 billion going to service delivery.

Transmission of sexually transmitted infections

Many risks factors have been associated with acquiring a STI that include behavioural risks as well as biologic risks. Behavioural risk factors include age of sexual initiation, number(s) and type(s) of sex partner(s), and use of barrier protection. Biological risk factors include sex, age, and developmental physiological changes. In addition to personal risk factors, community risk factors are also present, like the local prevalence of STIs, availability of screening and treatment services as well as social, economic, demographic, and cultural risk factors.

STI transmission dynamics

An important mathematical model to understand the population dynamics of STI transmission was proposed by May et al. in 1987 for HIV infection

(May and Anderson 1987). The model states that the average number of new infections caused by each infected person is determined by three factors: (a) the likelihood that an infected person will transmit the pathogen to an uninfected partner during sexual activity; (b) the average number of new sex partners formed between infected and non-infected people, and (c) the average amount of time an infected person can transmit the pathogen. Increasing any of those factors increases the population prevalence and *vice versa*.

With that model in mind, approaches of STI prevention programmes used to control the spread of STIs can be categorized. Commonly, STI prevention programmes aim to reduce the average amount of time an infected person can transmit the pathogen by testing and treating STIs, however many different interventions have been described and implemented. A list of different STI control strategies is included in Box 8.12.1.

Box 8.12.1 Strategies to control STIs

- Clinical testing and treating
- Community-based screening
- Comprehensive sexual education
- Vaccination
- Barrier protection, e.g. condoms
- Medical male circumcision
- Pre-exposure and post-exposure prophylaxis for HIV and other STIs
- Expedited partner therapy

Medical male circumcision

Medical male circumcision has been proven to reduce risk of male acquisition of HIV and other STIs (CDC 2018a). It is hypothesized that male circumcision reduces the risk of HIV and other STI acquisition because by removing the foreskin, the exposed penile skin thickens creating an effective barrier to infection. Additionally foreskin removal reduces the number of susceptible mucosal immune cells as well as alters the genital microbiota. Three randomized clinical trials showed that adult medical male

circumcision lowered the risk for HIV infection acquisition by 51% to 60% among heterosexual males (Auvert et al. 2005; Bailey et al. 2007; Gray et al. 2007). Additionally, male circumcision reduced the risk of genital ulcer disease prevalence by 47% and cancer-causing HPV prevalence by 23% to 47% (Auvert et al. 2009; Tobian et al. 2009; Serwadda et al. 2010; Mehta et al. 2012). Partners of circumcised men also receive protection from STIs. A large randomized trial performed in Kenyan and Ugandan involving HIV-serodiscordant heterosexual couples, found a reduction in the incidence of syphilis incidence by 42% among men and 59% among women (Pintye et al. 2014). The literature also supports that male circumcision is associated with reduced risk of other STIs among women including oncogenic human papillomavirus (HPV) genotypes and of contracting cervical cancer (Svare et al. 2002; Drain et al. 2006; Wilson et al. 2013; Grabowski et al. 2016). The evidence of the benefits of male circumcision is mixed for some other conditions like herpes simplex virus type 2 (Borkakoty et al. 2010; Mugo et al. 2011; Mujugira et al. 2011; Davis et al. 2017), genital ulcers (Gray et al. 2009), *Chlamydia trachomatis* infection (Castellsague et al. 2005; Nayyar et al. 2014; Russell et al. 2016) and *Trichomonas vaginalis* infection (Gray et al. 2009).

Sexual networks and core groups

Sexual networks are a type of social network that are defined by sexual activity among a group of individuals. Sexual networks can be determined by social, economic, cultural, migratory, or political forces, in addition to the physical or virtual spaces in which those networks can take place (Doherty et al. 2005). STI transmission within a sexual network depends on the number of sexual linkages within a sexual network and temporality of partnerships (concurrently versus sequentially). Bathhouses are an example of a physical space in which a sexual network can form. In the 1970s and 1980s, it was common for some men who have sex with men to find sex partners in bathhouses. Epidemiologic studies on the sexual networks formed in those bathhouses were used to trace the spread of HIV and other STIs. Now, the Internet has given rise to new virtual spaces facilitating sexual contacts through mobile applications and social media sites.

Future directions and challenges

New global initiatives have been enacted with the goal to end STI epidemics as a public health concern. Recently, new methods to prevent HIV infection have been developed, the use of HPV vaccines has increased, and more countries have successfully eliminated of MTCT of HIV and syphilis. However, multi-drug resistant *Neisseria gonorrhoeae* infection continues to spread and challenge STI control programmes.

Global and regional initiatives

In 2016, the WHO debuted a new Global Health Sector Strategy on Sexually Transmitted Infections 2016–2021: Towards Ending STIs, which sets a vision, goals, targets, guiding principles, and priority actions to end STIs epidemics as a public health problem (WHO 2016). The strategy emphasized the importance of accelerating and focusing comprehensive prevention efforts through scaling up evidence-based behavioural, biomedical, and structural approaches, facilitating individual's access to information of their STI status, improving access to treatment and comprehensive long-term care, and challenging pervasive stigmatization and discrimination. The 2030 targets for the strategy were (1) a 90% reduction of *Treponema pallidum* infection incidence globally from a 2018 global baseline, (2) a 90% reduction in *Neisseria gonorrhoeae* infection incidence globally from a 2018 global baseline, (3) ≤ 50 cases of congenital syphilis per 100,000 live births in 80% of counties, and (4) to sustain 90% coverage and at least 80% in every district in countries with HPV vaccine in their national immunization programme.

The WHO also made the dual elimination of MTCT of HIV and syphilis a global health priority, with a few countries also aiming to eliminate MTCT hepatitis B infection. So far, 11 countries have been validated by the WHO for eliminating MTCT of HIV or syphilis (WHO 2014; WHO 2017b). Criteria required for validation include 95% of pregnant women receiving antenatal care, 95% of pregnant women receiving HIV and syphilis testing in pregnancy, and 95% of pregnant women diagnosed with HIV or syphilis receiving treatment, with targets of ≤50 new paediatric HIV infections per 100,000 live births and a transmission rate of either <5% in breastfeeding populations or <2% in non-breastfeeding populations and ≤50 cases of congenital syphilis per 100,000 live births. The WHO also requires that this

process be conducted within a context of human rights, gender equality, and community engagement.

Antimicrobial resistance

Since the introduction of effective antibiotic treatments for *Neisseria gonorrhoeae* infection with sulfa drugs in the 1930s, the bacterium has become resistant to every antibiotic used for treatment (Hook and Kirkcaldy 2018). *Neisseria gonorrhoeae* developed resistance to sulfonamides, penicillins, early-generation cephalosporins, tetracyclines, macrolides, and fluoroquinolone drugs (Alirol et al. 2017). Most recently, *Neisseria gonorrhoeae* has been developing resistance to extended-spectrum cephalosporins, which is the last remaining empiric monotherapy for *Neisseria gonorrhoeae* infection. Furthermore, there has been a report of *Neisseria gonorrhoeae* antibiotic resistance to dual therapy (ceftriaxone and azithromycin). In 2015, the WHO adopted a Global Action Plan on Antimicrobial Resistance, which launched the Global Antibiotic Research and Development Partnership. This group aims to accelerate the development of new antibiotics, evaluate the potential of existing antibiotics and various combinations of them for more effective treatment, develop fixed-dose drug combinations, and develop simplified treatment guidelines for antibiotic-resistant bacterium. There are also three new drugs that are being developed to treat antibiotic-resistant *Neisseria gonorrhoeae* infection.

Advancement in technology

While use of mobile geosocial networking applications used for meeting sex partners may increase the risk of STI acquisition (Beymer et al. 2014), other advances like pre-exposure prophylaxis for HIV infection have been shown to be effective in preventing HIV acquisition (WHO 2016). Increased use of STI self-testing and rapid testing has also been shown to be helpful in increasing linkage to care, quality assurance, and partner services (Tucker et al. 2013). Additionally, digital innovations like mobile health-based messaging, Internet-based outreach, and combined interventions have been found to be acceptable, feasible, and impactful. Those interventions have also been useful in engaging harder to reach traditionally healthcare adverse populations (Daher et al. 2017). Finally, innovations in diagnostics may make it feasible to perform home or clinic-based STI testing for multiple

organisms including the characterization of antimicrobial resistance in some organisms in as few as 20 minutes.

Conclusions

STIs continue to be a major global cause of morbidity, mortality, and increased healthcare costs. This chapter outlined major issues regarding STI transmission, prevention, and programmes aimed to reduce global STI burden. While there are new challenges in the field of STIs, advancements in technology and global initiatives set by WHO could help reduce STIs as a public health epidemic in the near future. More effort is needed in research, implementation, and global collaboration and financial investment to eliminate many STIs as major public health problems.

FURTHER READING

Centers for Disease Control and Prevention (2015). *Sexually Transmitted Diseases Treatment Guidelines*. [Online] Available at: https://www.cdc.gov/std/tg2015/default.htm

World Health Organization (WHO) (2015). *Report on Global Sexually Transmitted Infection Surveillance*. [Online] Available at: https://www.who.int/reproductivehealth/publications/rtis/stis-surveillance-2015/en/

World Health Organization (WHO). *Global Health Sector Strategy on Sexually Transmitted Infections 2016–2021*. [Online] Available at: https://www.who.int/reproductivehealth/publications/rtis/ghss-stis/en/

REFERENCES

Alirol, E., et al. (2017). Multidrug-resistant gonorrhea: a research and development roadmap to discover new medicines. *PLoS Medicine*, 14, e1002366.

American Cancer Society (2015). American Cancer Society. The Global Economic Cost of Cancer. [Online] Available at: http://www.cancer.org/acs/groups/content/@internationalaffairs/documents/document/acspc-026203.pdf

Auvert, B., et al. (2009). Effect of male circumcision on the prevalence of high-risk human papillomavirus in young men: results of a randomized controlled trial conducted in Orange Farm, South Africa. *Journal of Infectious Disease*, **199**, 14–19.

Auvert, B., Taljaard, D., Lagarde, E., Sobngwi-Tambekou, J., Sitta, R., and Puren, A. (2005). Randomized, controlled intervention trial of male circumcision for reduction of HIV infection risk: the ANRS 1265 Trial. *PLoS Medicine*, 2, e298.

Bailey, R.C., et al. (2007). Male circumcision for HIV prevention in young men in Kisumu, Kenya: a randomised controlled trial. *The Lancet*, **369**, 643–56.

Baud, D., Gubler, D.J., Schaub, B., Lanteri, M.C., and Musso, D. (2017). An update on Zika virus infection. *The Lancet*, **390**, 2099–109.

Beymer, M.R., et al. (2014). Sex on demand: geosocial networking phone apps and risk of sexually transmitted infections among a cross-sectional sample of men who have sex with men in Los Angeles County. *Sexually Transmitted Infections*, **90**, 567–72.

Blohm, G.M., et al. (2018). Evidence for mother-to-child transmission of Zika virus through breast milk. *Clinically Infectious Diseases*, **66**, 1120–1.

Borkakoty, B., Biswas, D., Walia, K., and Mahanta, J. (2010). Potential impact of spouse's circumcision on herpes simplex virus type 2 prevalence among antenatal women in five northeastern states of India (14th International Congress on Infectious Diseases (ICID) Abstracts). *International Journal of Infectious Diseases*, 14, e411.

Bruni, L., Diaz, M., Castellsague, X., Ferrer, E., Bosch, F.X., and De Sanjose, S. (2010). Cervical human papillomavirus prevalence in 5 continents: meta-analysis of 1 million women with normal cytological findings. *Journal of Infectious Diseases*, **202**, 1789–99.

Buchacz, K., et al. (2004). Syphilis increases HIV viral load and decreases CD4 cell counts in HIV-infected patients with new syphilis infections. *AIDS*, **18**, 2075–9.

Castellsague, X., et al. (2005). *Chlamydia trachomatis* infection in female partners of circumcised and uncircumcised adult men. *American Journal of Epidemiology*, **162**, 907–16.

Centers for Diseases Control and Prevention (CDC) (2017). *Manual for the Surveillance of Vaccine-Preventable Diseases*. Atlanta, GA: Centers for Disease Control and Prevention.

Centers for Diseases Control and Prevention (CDC) (2018a). *Information for Providers to Share with Male Patients and Parents Regarding Male Circumcision and the Prevention of HIV Infection, Sexually Transmitted Infections, and other Health Outcomes*. Atlanta, GA: Centers for Disease Control and Prevention.

Centers for Diseases Control and Prevention (CDC) (2018b). *Outcomes of Pregnancies with Laboratory Evidence of Possible Zika Virus Infection, 2015–2018*. Atlanta, GA: Centers for Disease Control and Prevention.

Centers for Diseases Control and Prevention (CDC) (2018d). *Viral Hepatitis*. Atlanta, GA: Centers for Disease Control and Prevention.

Centers for Diseases Control and Prevention(CDC) (2018c). *STDs & Infertility*. Atlanta, GA: Centers for Disease Control and Prevention.

Crosbie, E.J., Einstein, M.H., Franceschi, S., and Kitchener, H.C. (2013). Human papillomavirus and cervical cancer. *The Lancet*, **382**, 889–99.

Cunningham, S. D., Kerrigan, D. L., Jennings, J. M. & Ellen, J. M. 2009. Relationships between perceived STD-related stigma, STD-related shame and STD screening among a household sample of adolescents. *Perspectives on Sexual & Reproductive Health*, **41**, 225–30.

Daher, J., et al. (2017). Do digital innovations for HIV and sexually transmitted infections work? Results from a systematic review (1996–2017). *British Medical Journal Open*, 7, e017604.

Davis, S., et al. (2017). Association between HIV and sexually transmitted infections and partner circumcision among women in uMgungundlovu District, South Africa: a cross-sectional analysis of HIPSS baseline data. *IAS2017*. [Online] Available at: http://programme.ias2017.org/Abstract/Abstract/2833

Doherty, I.A., Padian, N.S., Marlow, C., and Aral, S.O. (2005). Determinants and consequences of sexual networks as they affect the spread of sexually transmitted infections. *Journal of Infectious Diseases*, 191 Suppl 1, S42–54.

Drain, P.K., Halperin, D.T., Hughes, J.P., Klausner, J.D., and Bailey, R.C. (2006). Male circumcision, religion, and infectious diseases: an ecologic analysis of 118 developing countries. *BMC Infectious Diseases*, 6(172), 1–10.

EL-Serag, H.B. (2012). Epidemiology of viral hepatitis and hepatocellular carcinoma. *Gastroenterology*, **142**, 1264–73.e1.

Fettig, J., Swaminathan, M., Murrill, C.S., and Kaplan, J.E. (2014). Global epidemiology of HIV. *Infectious Disease Clinics of North America*, **28**, 323–37.

Galvin, S.R. and Cohen, M.S. (2004). The role of sexually transmitted diseases in HIV transmission. *Nature Reviews Microbiology*, **2**, 33–42.

Gomez, G.B., Kamb, M.L., Newman, L.M., Mark, J., Broutet, N., and Hawkes, S.J. (2013). Untreated maternal syphilis and adverse outcomes of pregnancy: a systematic review and meta-analysis. *Bulletin of the World Health Organization*, **91**, 217–26.

Grabowski, M.K., et al. (2016). Partner human papillomavirus viral load and incident human papillomavirus detection in heterosexual couples. *Journal of Infectious Diseases*, **213**, 948–56.

Gray, R.H., et al. (2007). Male circumcision for HIV prevention in men in Rakai, Uganda: a randomised trial. *The Lancet*, **369**, 657–66.

Gray, R.H., et al. (2009). The effects of male circumcision on female partners' genital tract symptoms and vaginal infections in a randomized trial in Rakai, Uganda. *American Journal of Obstetric Gynecology*, *200*(42), e1–7.

Hook, E.W., 3rd and Kirkcaldy, R.D. (2018). A brief history of evolving diagnostics and therapy for gonorrhea: lessons learned. *Clinical Infectious Diseases*, *67*(8), 1294–9.

Jemal, A., et al. (2013). Annual Report to the Nation on the Status of Cancer, 1975–2009, featuring the burden and trends in human papillomavirus (HPV)-associated cancers and HPV vaccination coverage levels. *Journal of the National Cancer Institute*, **105**, 175–201.

Korenromp, E.L., Wi, T., Resch, S., Stover, J., and Broutet, N. (2017). Costing of national STI program implementation for the global STI control strategy for the health sector, 2016–2021. *PLoS One*, 12, e0170773.

Looker, K.J., et al. (2017). First estimates of the global and regional incidence of neonatal herpes infection. *Lancet Global Health*, 5, e300–9.

Looker, K.J., Magaret, A.S., Turner, K.M., Vickerman, P., Gottlieb, S.L., and Newman, L.M. (2015). Global estimates of prevalent and incident herpes simplex virus type 2 infections in 2012. *PLoS One*, 10, e114989.

Madeleine, M.M., et al. (2007). Risk of cervical cancer associated with Chlamydia trachomatis antibodies by histology, HPV type and HPV cofactors. *International Journal of Cancer*, **120**, 650–5.

Marsico, C. and Kimberlin, D.W. (2017). Congenital cytomegalovirus infection: advances and challenges in diagnosis, prevention and treatment. *Italian Journal of Pediatrics*, **43**, 38.

May, R.M. and Anderson, R.M. (1987). Transmission dynamics of HIV infection. *Nature*, **326**, 137–42.

Mehta, S.D., Moses, S., Parker, C.B., Agot, K., Maclean, I., and Bailey, R.C. (2012). Circumcision status and incident herpes simplex virus type 2 infection, genital ulcer disease, and HIV infection. *AIDS*, **26**, 1141–9.

Mugo, N., et al. (2011). Prevalence of herpes simplex virus type 2 infection, human immunodeficiency virus/herpes simplex virus type 2 coinfection, and associated risk factors in a national, population-based survey in Kenya. *Sexually Transmitted Diseases*, **38**, 1059–66.

Mujugira, A., Magaret, A. S., Baeten, J. M., Celum, C., and Lingappa, J. (2011). Risk factors for HSV-2 infection among sexual partners of HSV-2/HIV-1 co-infected persons. *BMC Research Notes*, **4**, 64.

Nayyar, C., Chander, R., Gupta, P., and Sherwal, B.L. (2014). Co-infection of human immunodeficiency virus and sexually transmitted infections in circumcised and uncircumcised cases in India. *Indian Journal of Sexually Transmitted Diseases*, **35**, 114–17.

Pintye, J., et al. (2014). Association between male circumcision and incidence of syphilis in men and women: a prospective study in HIV-1 serodiscordant heterosexual African couples. *Lancet Global Health*, 2, e664–71.

Russell, A.N., et al. (2016). Analysis of factors driving incident and ascending infection and the role of serum antibody in Chlamydia trachomatis genital tract infection. *Journal of Infectious*

Diseases, **213**, 523–31.
Serwadda, D., et al. (2010). Circumcision of HIV-infected men: effects on high-risk human papillomavirus infections in a randomized trial in Rakai, Uganda. *Journal of Infectious Diseases*, **201**, 1463–9.
Svare, E.I., Kjaer, S.K., Worm, A.M., Osterlind, A., Meijer, C.J., and Van Den Brule, A.J. (2002). Risk factors for genital HPV DNA in men resemble those found in women: a study of male attendees at a Danish STD clinic. *Sexually Transmitted Infections*, **78**, 215–18.
Tobian, A.A., et al. (2009). Male circumcision for the prevention of HSV-2 and HPV infections and syphilis. *New England Journal of Medicine*, **360**, 1298–309.
Tucker, J.D., Bien, C.H., and Peeling, R.W. (2013). Point-of-care testing for sexually transmitted infections: recent advances and implications for disease control. *Current Opinion on Infectious Disease*, **26**, 73–9.
UNAIDS (2013a). *AIDS by the Numbers*. Geneva: Joint United Nations Programme on HIV/AIDS.
UNAIDS (2013b). *Global Report on the Global AIDS Epidemic. Joint United Nations Program on HIV/AIDS (UNAIDS)*. Geneva: UNAIDS.
UNAIDS (2018). *UNAIDS Data 2018*. Geneva: UNAIDS.
Ward, H. and Ronn, M. (2010). Contribution of sexually transmitted infections to the sexual transmission of HIV. *Current Opinion in HIV & AIDS*, **5**, 305–10.
Wilson, L.E., et al. (2013). Male circumcision reduces penile high-risk human papillomavirus viral load in a randomised clinical trial in Rakai, Uganda. *Sexually Transmitted Infections*, **89**, 262–6.
Workowski, K.A., Bolan, G.A., and Centers for Disease, Control & Prevention (2015). Sexually transmitted diseases treatment guidelines, 2015. *MMWR Recommended Reports*, **64**, 1–137.
World Health Organization (WHO) (2014). *Global Guidance on Criteria and Processes for Validation: Elimination of Mother-to-Child Transmission of HIV and Syphilis*. Geneva: WHO. [Online] Available at: https://www.who.int/reproductivehealth/publications/emtct-hiv-syphilis/en/
World Health Organization (WHO) (2015). *Report on Global Sexually Transmitted Infection Surveillance 2015*. Geneva: WHO. [Online] Available at: https://www.who.int/reproductivehealth/publications/rtis/stis-surveillance-2015/en/
World Health Organization (WHO) (2016). *Global Health Sector Strategy on Sexually Transmitted Infections 2016–2021*. Geneva: WHO.
World Health Organization (WHO) (2017a). *Global Hepatitis Report 2017*. Geneva: WHO.
World Health Organization (WHO) (2017b). *Six Caribbean Islands Eliminate Mother-to-Child Transmission of HIV and Syphilis*. Geneva: WHO.
World Health Organization (WHO) (2018). *Human Papillomavirus (HPV) and Cervical Cancer*. Geneva: WHO.
Yarchoan, R. and Uldrick, T.S. (2018). HIV-associated cancers and related diseases. *New England Journal of Medicine*, **378**, 1029–41.

8.13

Acquired immunodeficiency syndrome (AIDS)

Quarraisha Abdool Karim, Urisha Singh, Cheryl Baxter, and Salim S. Abdool Karim

Introduction

In May 2016, about 35 years since the first reported cases of AIDS, United Nations (UN) member states adopted the Declaration of ending AIDS by 2030, at a specially convened high level meeting in New York. The impetus for this remarkable announcement emanated from the confluence of new human immunodeficiency virus (HIV) knowledge, centrally around the use of antiretrovirals for treatment and prevention.

Given that a vaccine or cure, much needed tools for elimination and eradication of HIV respectively are still in development, what does 'ending AIDS' mean? Epidemiologically it can be defined as the reduction of disease incidence, prevalence, morbidity, or mortality to a locally acceptable level as a result of deliberate intervention measures. While an epidemiological metric for achieving epidemic control has yet to be defined, it is defined mathematically as the reproductive rate of infection (R_0) reaching below one. There is some consensus that this goal can be reached at an HIV incidence rate of 1/1,000 per annum.

In order to track progress on achieving epidemic control, the Joint United Nations Programme on HIV/AIDS (UNAIDS) developed the Fast Track Strategy that monitors progress towards elimination of AIDS-related stigma and discrimination; AIDS-related mortality and reduction in HIV incidence rates. The intermediate 2020 targets include: (a) 90% of those who are infected know their status; 90% of those who are infected are initiated on antiretroviral (ARV) treatment and 90% of those on ARV treatment are virally suppressed; also referred to as 90–90–90; (b) a 90% reduction of incidence rates compared to 2010 and iii. Reduction in stigma and discrimination.

This chapter traces the history of HIV from its origins, the remarkable scientific advances, and unprecedented global responses through to the current state of the epidemic, progress towards ending AIDS and remaining challenges.

In the beginning

Acquired immune deficiency syndrome (AIDS) was first recognized in 1981 when a cluster of five previously healthy homosexual men from Los Angeles presented with *Pneumocystis carinii* pneumonia (PCP) (now *Pneumocystis jirovecii*) and additional clusters of homosexual men in California and New York presented with Kaposi's sarcoma (CDC 1981; Gottlieb et al. 1981; Greene 2007). Within the next 18 months, additional cases of previously healthy homosexual men with PCP and Kaposi's sarcoma were reported in various large cities in the United States (CDC 1982a). The main presenting symptoms of these men included severely depressed cell-mediated immunity coupled with opportunistic infections and rare malignancies (CDC 1981). The prognosis of the illness was poor with the outcome being rapid death following the manifestation of these symptoms. The predominance of this new disease entity in young homosexual men and linked cases, which suggested it was sexually related, led to the belief that it was a lifestyle-based illness and it was termed gay-related immune deficiency (GRID) or 'gay plague'. A month later, the illness was detected in Haitians and then haemophiliacs which initially led to the belief that it originated in Haiti (CDC 1982b). However, the identification of cases of a new wasting illness in Uganda, which mimicked symptoms of AIDS, termed 'slim disease' placed doubt on the origin of the illness (Serwadda et al. 1985). Between 1982 and 1983, the first case of mother-to-child transmission of AIDS and the first AIDS-related death of an infant, aged 20 months, was reported. Additionally, reports of AIDS infection in haemophiliacs, injecting drug users (IDUs), recipients of blood transfusions and female partners of men with AIDS were released, providing the first indication that AIDS infection was likely also spread through blood and heterosexual sex (CDC 1982b; Greene 2007). At this time the new disease entity was renamed AIDS, all major routes of transmission were identified, however, its cause, diagnosis and treatment remained elusive (Marx 1982; Gottlieb et al. 1983).

The lack of information on the causative agent of AIDS coupled with a lack of treatment and testing methods for the illness imparted great fear, uncertainty, and stigma towards those infected. Many people believed that AIDS could spread through holding hands or sharing a toilet seat. Parents with infected adult children refused to see them, nurses and auxiliary staff refused to take meals to infected patients, doctors argued whether treating infected patients was a moral obligation, and faith communities titled those infected as ‘abominations’ with the belief that AIDS was a form of ‘divine punishment for those with sinful lifestyles’ (Fee and Krieger 1993; Smith and Whiteside 2010). People living with AIDS (PLWA) were; physically attacked, unfairly dismissed from their jobs, unlawfully evicted from their homes and denied various health services. This drove the development of several organizations and activist groups aimed at advocating for and filling the gaps in care for PLWA and initiating civil rights protection policies (Gostin 1990; Herek and Capitanio 1993). The gay rights movement campaigned for AIDS to be viewed as a human rights issue with the argument that, other than the risk of infection with AIDS, there was also a high risk of harmful discrimination against those found to be infected (Bayer and Fairchild 2006). A multitude of public health officials supported this argument and raised additional concerns of stigma being a barrier to testing for the illness and for those with AIDS from accessing health services (Smith and Whiteside 2010). Media coverage of AIDS remained low with social movements and activism propagating the early response to AIDS and encouraging the discovery of its causative agent and diagnostic tests.

The discovery and spread of HIV

The identification and isolation of the causative agent of AIDS was challenging due to the manifestation of clinical signs and symptoms years after infection by which time patients also presented with various additional infections.

In 1983, Luc Montagnier and Francoise Barre-Sinoussi from the Pasteur Institute in France announced the isolation of a new retrovirus; lymphadenopathy-associated virus (LAV), and suggested that it was the cause of AIDS, however, proof of causality was not established. Eighteen months later, Robert Gallo’s group from the US National Cancer Institute announced the identification and isolation of a second retrovirus, the human

T-lymphotropic virus type-III (HTLV-III), and provided sufficient evidence of causality of AIDS (Popovic et al. 1984). Three months later Jay Levy's group in San Francisco isolated an 'AIDS-associated retrovirus' from 22 AIDS infected patients confirming the link between a retrovirus and AIDS (Levy et al. 1984). By 1985, LAV, HTLV-III and the AIDS-associated retrovirus were confirmed to be variants of the same retrovirus (Ratner et al. 1985; Marx 1988), which was then established as the cause of AIDS and renamed HIV (Coffin et al. 1986). A second, less prevalent and virulent strain of HIV was isolated from patients in West Africa in 1986, providing further confirmation of causality. This virus was named HIV-type 2 (HIV-2) with HIV subsequently renamed HIV-type 1 (HIV-1) (Clavel et al. 1986).

Following the identification of the causative agent, its origins were investigated using phylogenetics and statistical modelling of the earliest available HIV sequences obtained between 1950 and 1960. Studies showed that HIV-1 and HIV-2 originated in central West Africa from separate zoonotic transfer of simian immunodeficiency virus (SIV) from primates to humans where primates were butchered and consumed as bush meat (Fig. 8.13.1) (Sharp and Hahn 2008; Tebit and Arts 2011). Cross-species transfer of SIV from old world monkeys to chimpanzees then from chimpanzees to gorillas resulted in variations of SIV that gave rise to HIV-1 in humans. The HIV-1 groups M and N have been associated with SIV from *Pan troglodytes*, a subspecies of chimpanzee, HIV-1 group P and O have been associated with western gorillas and HIV-2 has been associated with SIV from *Cerocebus atys*, a subspecies of the sooty mangabey (Nerrienet et al. 2005; Santiago et al. 2005; Keele et al. 2006; Sharp and Hahn 2008; Plantier et al. 2009).

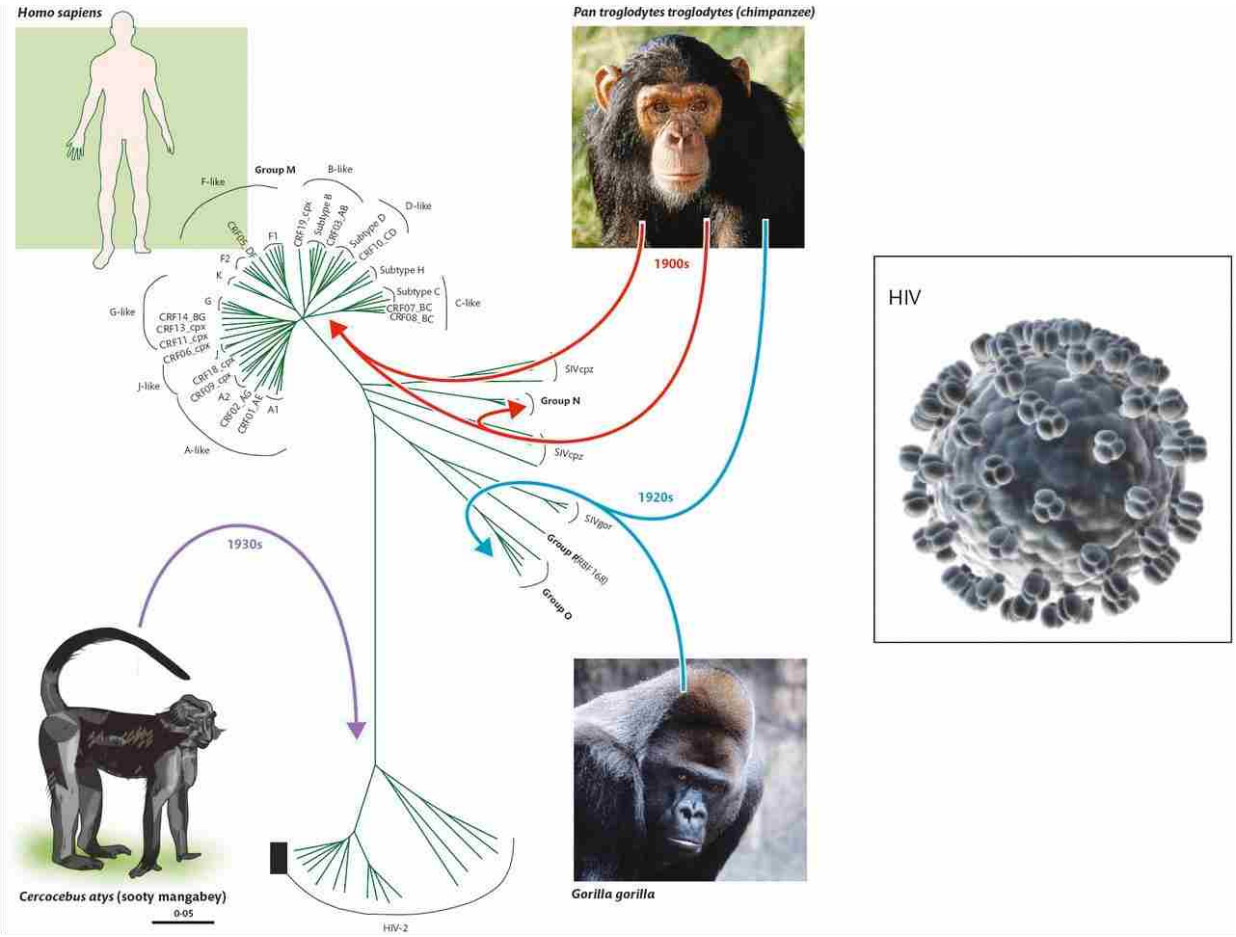


Fig. 8.13.1 Origins of HIV.
Reprinted from Denis M Tebit, Eric J Arts. Tracking a century of global expansion and evolution of HIV to drive understanding and to combat disease. *The Lancet Infectious Diseases*, 11(1):45–56 with permission from Elsevier.

It has been estimated that HIV-1 group M, N and O originated in 1908 (range: 1884–1924), 1963 (range: 1948–1977), and 1920 (range: 1890–1940) respectively, while the origin of HIV-2 group A and B has been estimated to be 1932 (range: 1906–1955) and 1935 (range: 1907–1961) respectively (Korber et al. 2000; Wertheim and Worobey 2009; Tebit and Arts 2011; Sharp and Hahn 2008, 2011). Faria et al. (2014) demonstrated that HIV group M, the oldest lineage of HIV-1, originated in Kinshasa and attributed its spread through central Africa to extensive use of railway transport that linked various regions and enabled interactions between individuals from these regions (Faria et al. 2014). Additional studies cite the growth of cities, migration, urbanization, and improved travel as the main contributing factors towards the exponential spread of HIV and its uneven

distribution of subtypes and circulating recombinant forms of HIV (Tebit and Arts 2011; Sharp and Hahn 2008; Faria et al. 2014).

The development of the HIV antibody test and HIV/AIDS surveillance

Shortly after the discovery of HIV, diagnostic tools started to be developed to identify individuals infected with HIV. By March 1985, an enzyme linked immunosorbent assay (ELISA) was approved as the first HIV diagnostic test. This test recognized the IgG antibody produced in response to HIV-1 only. Blood banks in the USA and elsewhere in the world began routine screening of donated blood (Roberts 1994). Although sensitive and useful in the screening of blood products, the ELISA test had an antibody-negative window of at least 12 weeks post-infection and had a high rate of false negative results in response to other infections, pregnancy, autoimmune diseases, and various unknown conditions (Chappel et al. 2009). By 1986, two confirmatory tests, the western blot immunoassay (CDC 1989) and the HTLV-III immunofluorescence assay were approved with an antibody-negative window of approximately 6 weeks (Hedenskog et al. 1989; Chappel et al. 2009). These confirmatory tests were used in conjunction with the ELISA test and required that tests be performed in duplicate. Second generation HIV tests developed in the late 1980s, were designed to further reduce the antibody-negative window period to 4 weeks and improve the specificity of testing by including: the HIV-1 specific P24 antigen, an HIV-2 specific protein and an HIV-1 group O antigen to the antigen milieu in order to detect antibodies to these antigens. Third and fourth generation tests were subsequently developed, each with reduced antibody-negative windows, however, each of these tests were still used in combination with the confirmatory western blot immunoassays or the HTLV-III immunofluorescence assay. Fifth generation tests developed in 2015, eliminated the need for confirmatory testing as they provided results for the P24 antigen and the HIV antibody and reported separate results for HIV-1 and HIV-2 antibodies (Alexander 2016). Rapid HIV test kits became available in 2002 with 99.6% accuracy and results available within 20 minutes, it is now the predominant test of choice for an initial diagnosis in most settings.

These diagnostic tests enabled surveillance systems to be established to monitor the evolving HIV epidemic. By the end of 1985, every region in the world reported at least one case of AIDS. HIV surveillance (Detels et al. 2012) has remained essential and critical to quantify the magnitude of the epidemic; monitor trends in new infections, identify risk factors associated with infection, target interventions; and importantly, compare clinical manifestations of each case of AIDS that lacked aetiological explanations. In the USA, reporting by name of each newly diagnosed case of AIDS became a state health-department requirement shortly after the initial Centre for Disease Control (CDC) case reports of AIDS in 1981. While AIDS reporting was widely accepted and implemented with ease, HIV case reporting sparked community protest due to the use of name-based reporting of HIV infection (Goston et al. 1997). Gay rights organizations and civil libertarians expressed concern of invasion of privacy and the possibility of information reaching entities that would institute prejudice on those infected concerning employment, insurance, and housing (Gostin et al. 1997). Health officials justified the use of name-based reporting on the basis that it would assist health departments to: monitor those infected and ensure they remain linked to care, enable contact between health departments and those infected to enable counselling on prevention of further transmission of HIV and to monitor the prevalence and incidence of HIV. However, continued protest against the use of names for reporting on HIV incidence resulted in the implementation of confidential case-report systems that used code-identifiers instead of names for reports early in 2003 (Institute of Medicine (US) Committee on the Ryan White CARE Act 1994). These systems laid the basis for unlinked, anonymous HIV surveillance globally.

Each year, UNAIDS releases a report on the global status of the HIV epidemic based on data gathered from member states. Today HIV is found in every part of the world. About 77.3 million people have become infected with HIV and 35.4 million people have died from AIDS-related illnesses since the start of the epidemic. In 2017, there were about 37 million people living with HIV; 960,000 AIDS-related deaths occurred and 1.8 million new HIV infections. Sub-Saharan Africa continues to bear a disproportionate (65%) burden of infection. About three-quarters of all HIV infections are in 15 countries with the majority in sub-Saharan Africa (Fig. 8.13.2).

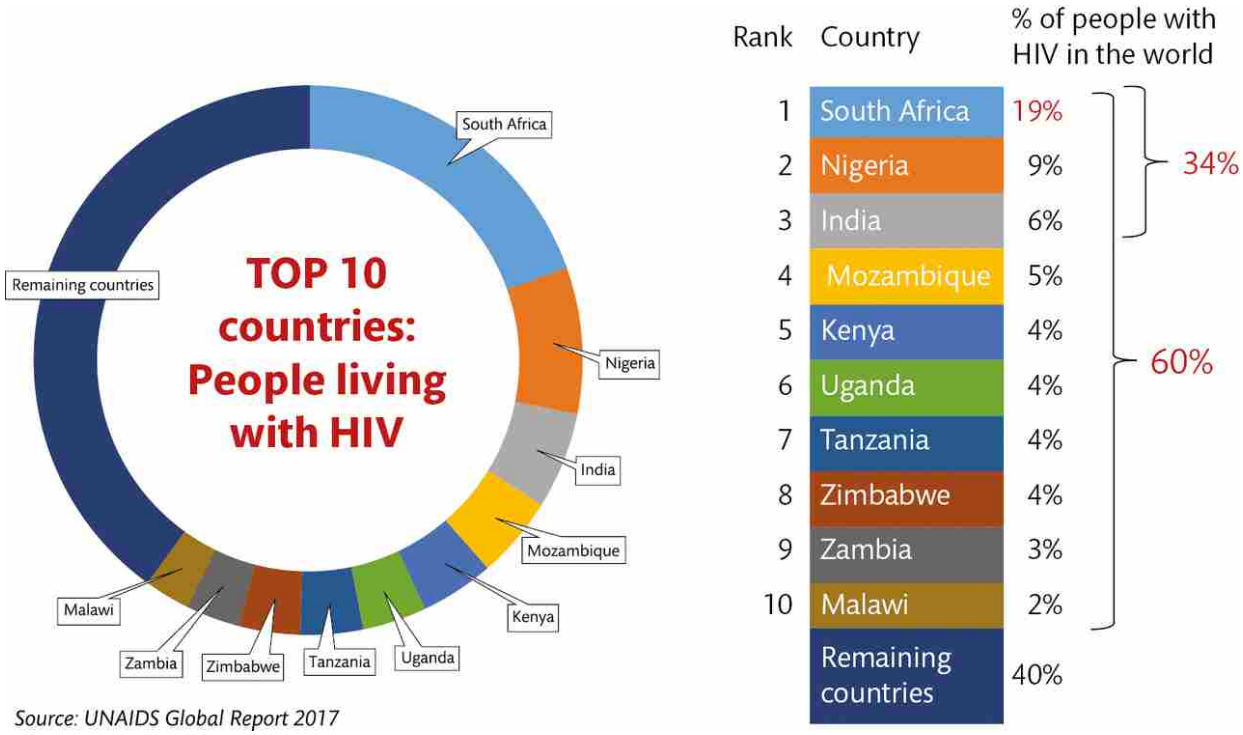


Fig. 8.13.2 Overview of burden of HIV globally.

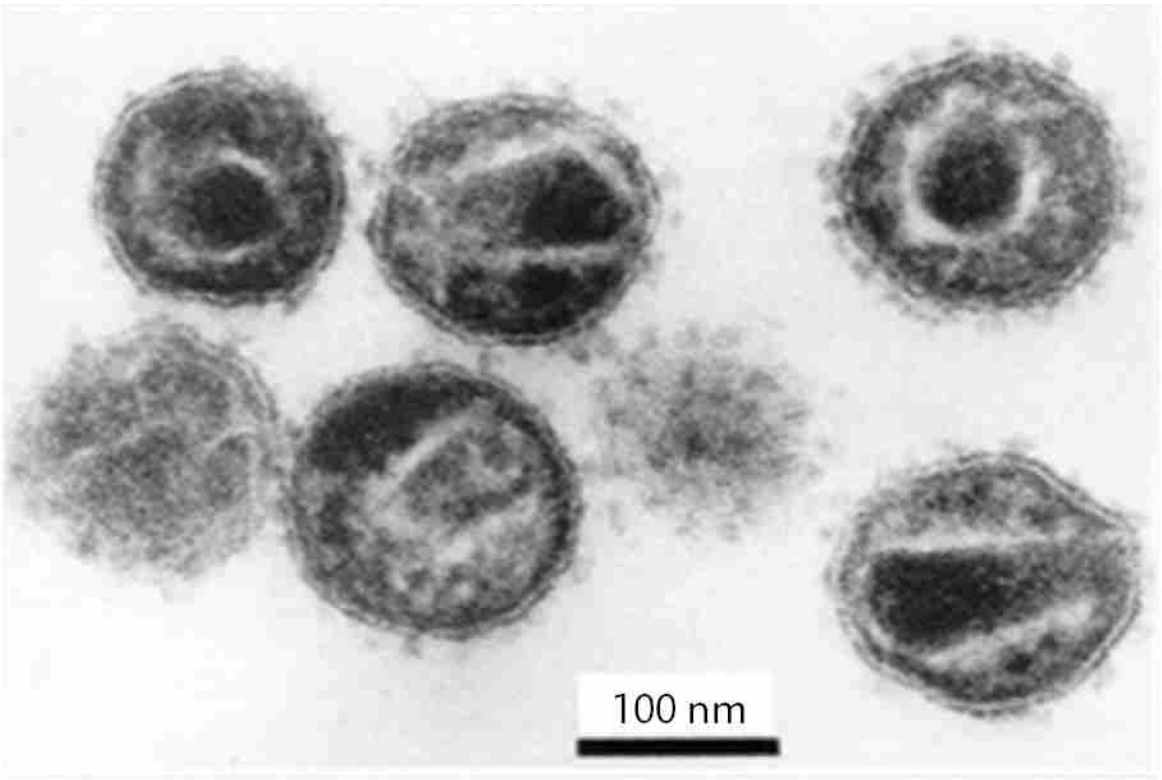
These global data mask the diversity of epidemics within and between countries. Southern and Eastern Africa for example account for 52% of all new HIV infections that occur globally. Globally about a third of all new infections are taking place in young people between 15 and 24 years of age. In sub-Saharan Africa, adolescent girls and young women account for about 40% of all new HIV infections. Globally, key and priority populations such as men who have sex with men (MSM); transgendered persons, IDUs, and sex workers bear a disproportionately high burden of HIV infection and experience substantial challenges in access to services particularly in settings where they are marginalized or worse if criminalized.

Viral structure

HIV is a retrovirus, so named because it encodes the enzyme reverse transcriptase. This unique enzyme allows a DNA copy to be made from viral RNA. Reverse transcription is a highly error prone process and is the major source of genetic variability that typifies HIV.

Viral particles are spherical with a diameter of approximately 100 nm and as such can only be viewed under an electron microscope (Fig. 8.13.3a). The outer lipid membrane is derived from the human host and contains the two major envelope glycoproteins, gp120, and gp41 (‘gp’ stands for glycoprotein and the number is the molecular weight in kilodaltons) (Fig. 8.13.3b). These proteins are linked together and exist as trimers on the viral surface facilitating binding and entry to the host cell. The envelope protein is referred to as ‘Env’. Other major structural proteins are the group-associated antigens or Gag proteins. The matrix protein (p17), which lies just beneath the envelope, provides structure to the particle, while the capsid protein (p24) encloses the viral RNA. The capsid protein is clearly visible under the electron microscope as a dark conical structure in the centre of the particle (Fig. 8.13.3a). HIV encodes three unique enzymes that are major targets for antiretroviral therapies. These are reverse transcriptase (RT), integrase, and protease, collectively called polymerases or ‘Pol’ for short. All three enzymes are carried inside the viral particle where they remain active. In addition to the major structural proteins, HIV also produces a number of regulatory and accessory proteins that are only produced once the virus infects cells and are not present inside viral particles. These include Tat and Rev, which enhance levels of gene expression and Vif, Vpr, Vpu, and Nef, which function to increase viral production and infectivity.

(a)


100 nm

(b)

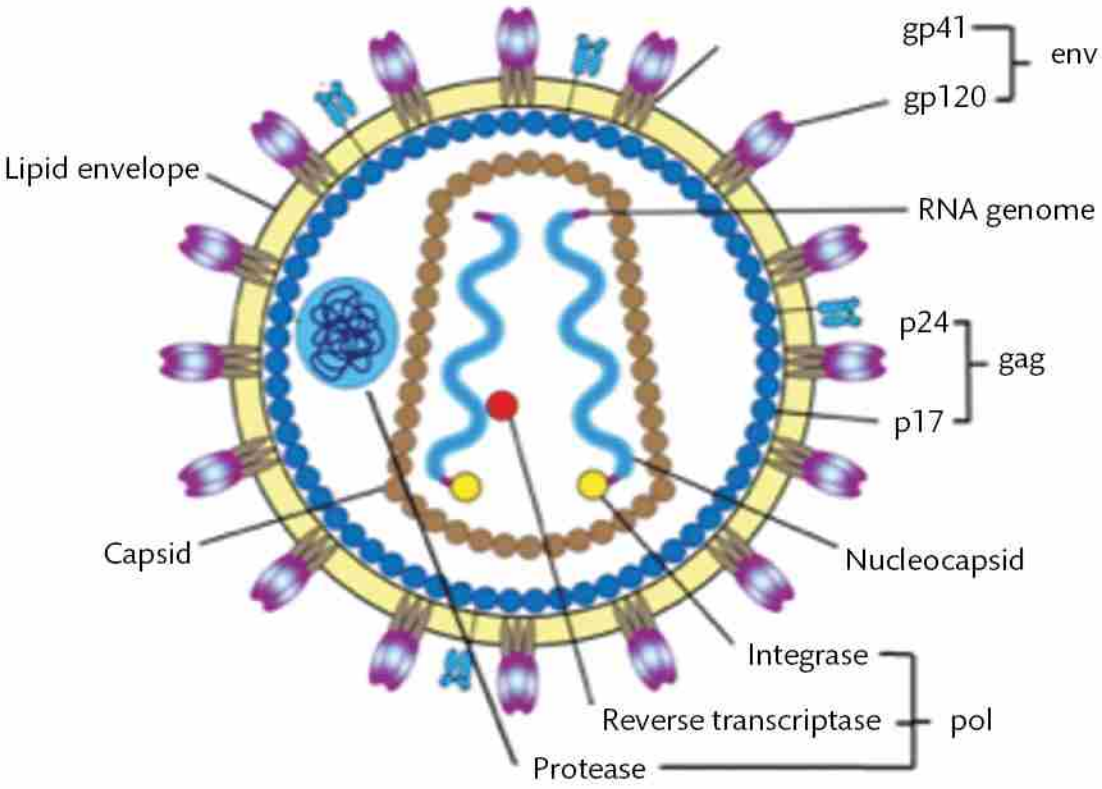

gp41
env
gp120
Lipid envelope
RNA genome
p24
gag
p17
Capsid
Nucleocapsid
Integrase
Reverse transcriptase
pol
Protease

Fig. 8.13.3 (a) Electron micrograph image showing mature HIV particles; (b) schematic representation of HIV-1 particle.

Treating AIDS

Prior to 1986, the clinical management of AIDS included prophylactic treatment against AIDS-related illnesses and opportunistic infections with patients usually experiencing slow, painful lingering deaths usually within two years of presentation at health facilities. The discovery of HIV, characterization of the virus and the development of diagnostic assays to detect HIV led to a turning point for AIDS patients in 1986 through the discovery of azidothymidine (AZT), the first drug able to suppress HIV replication in culture that emanated from an unprecedented collaboration between the United States National Cancer Institute and the Burroughs Wellcome pharmaceutical company (Mitsuya et al. 1985; Broder 2010). This nucleoside reverse transcriptase inhibitor (NRTI), originally synthesized in 1964 for cancer treatment, was found to act as a DNA polymerase chain terminator with selectivity to the HIV RT (Furman et al. 1986; De Clercq 2005). It was approved for use in treatment of patients with HIV by the USA Food and Drug Administration (FDA) in 1987 and provided the initial proof-of-concept for HIV treatment (Broder 2010). Monotherapy with AZT was the primary treatment for HIV from 1987 to 1995. The discovery of AZT was followed by the development of various additional nucleoside analogues including: didanosine (ddI) (1991), zalcitabine (ddc) (1992), stavudine (d4T) (1994), and lamivudine (3TC) (1995) (Table 8.13.1). By 1995, the first protease inhibitor (PI), saquinavir (SQV), was approved and in 1996 the first non-nucleoside RT inhibitor (NNRTI), nevirapine (NVP) (De Clercq 2009) was approved.

The FDA approval of additional nucleoside analogues, PIs, and NNRTIs following AZT approval, changed the face of HIV treatment, which transitioned from monotherapy (1987–1995) to dual therapy (1992–1996) to the game-changing triple therapy/highly active antiretroviral therapy (HAART), now referred to as combination antiretroviral therapy (cART) (1996-present) (Arts and Hazuda 2012). cART demonstrated significant reductions in plasma viral load in patients treated for at least six months and became the standard of care since 1996 (Carpenter et al. 1997).

Table 8.13.1 Antiretroviral drugs used in the treatment of HIV infection

Brand name	Generic name	Abbreviation	Approved
Nucleoside reverse transcriptase inhibitors (NRTIs)			
Retrovir	zidovudine, azidothymidine	AZT, ZDV	March 1987
Videx	didanosine, dideoxyinosine	ddI	October 1991
Hivid	zalcitabine dideoxycytidine (no longer marketed)	ddC	June 1992
Zerit	stavudine	d4T	June 1994
Epivir	lamivudine	3TC	November 1995
Combivir	lamivudine and zidovudine	3TC/ZDV	September 1997
Ziagen	abacavir sulphate	ABC	December 1998
Trizivir	abacavir, zidovudine, and lamivudine	ABC/ZDV/3TC	November 2000
Videx EC	enteric coated didanosine	ddI EC	October 2000
Viread	tenofovir disoproxil fumarate	TDF	October 2001
Emtriva	emtricitabine	FTC	July 2003
Epzicom	abacavir and lamivudine	ABC/3TC	August 2004
Truvada	tenofovir disoproxil fumarate and emtricitabine	TDF/FTC	August 2004
Nonnucleoside reverse transcriptase inhibitors (NNRTIs)			
Viramune	nevirapine (immediate release)	NVP	June 1996
Rescriptor	delavirdine	DLV	April 1997

Brand name	Generic name	Abbreviation	Approved
Sustiva	efavirenz	EFV	September 1998
Intelence	etravirine	RTR	January 2008
Edurant	rilpivirine	RPV	May 2011
Viramune XR	nevirapine (extended release)	NVP	March 2011
Protease inhibitors (PIs)			
Invirase	saquinavir mesylate	SQV	December 1995
Norvir	ritonavir	RTV	March 1996
Crixivan	indinavir	IDV	March 1996
Viracept	nelfinavir mesylate	NFV	March 1997
Fortovase	saquinavir (no longer marketed)	SQV	November 1997
Agenerase	amprenavir, (no longer marketed)	APV	April 1999
Kaletra	lopinavir and ritonavir	LPV/RTV	September 2000
Reyataz	atazanavir sulphate	ATV	June 2003
Lexiva	fosamprenavir calcium	FOS-APV	October 2003
Aptivus	tipranavir	TPV	June 2005
Prezista	darunavir	DRV	June 2006
Fusion inhibitors			
Fuzeon	enfuvirtide	T-20	March 2003
Entry inhibitors—CCR5 co-receptor antagonist			

Brand name	Generic name	Abbreviation	Approved
Selzentry	maraviroc	MVC	August 2007
HIV integrase strand transfer inhibitors			
Isentress	raltegravir	RAL	October 2007
Tivicay	dolutegravir	DTG	August 2013
Vitekta	elvitegravir	EVG	September 2014
Multiclass combination products			
Atripla	efavirenz, emtricitabine, and tenofovir disoproxil fumarate	EFV/FTC/TDF	July 2006
Complera	emtricitabine, rilpivirine, and tenofovir disoproxil fumarate	FTC/RPV/TDF	August 2011
Stribild	elvitegravir, cobicistat, emtricitabine, tenofovir disoproxil fumarate	EVG/COBI/FTC/TDF	August 2012
Evotaz	atazanavir sulphate, cobicistat	ATV/COBI	January 2015
Prezcobix	cobicistat, darunavir ethanolate	COBI/DRV	January 2015

To combat HIV drug resistance, the development of new drugs continues to progress with the first fusion inhibitor/gp41 inhibitor (enfuviritide (T-20)) developed in 2003, the first CCR5 inhibitor/entry inhibitor (maraviroc (MVC)) developed in 2007 and the first integrase inhibitor (raltegravir (RAL)) developed in 2007. These drugs provide alternate treatment options for patients with complications (i.e. toxicity or incompatibility with other medications) or drug resistance (Arts and Hazuda 2012).

Following the initiation of cART as the standard of care for HIV patients globally, the development of fixed-dose combination (FDC) drugs containing two or three-drug combinations emerged as methods to improve adherence and thus reduce treatment failure and the development of drug resistance. Combivir™, a single pill containing a combination of AZT and

3TC was the first FDC pill developed in 1997. This was followed by the development of additional two-drug combination pills; Kaletra (lopinavir (LPV) and ritonavir (RTV)) in 2002, Epzicom/Kivexa (abacavir (ABC) and 3TC) in 2004 and Truvada (tenofovir disoproxil fumarate (TDF) and emtricitabine (FTC)) in 2004. The first three-drug combination pill, Trizivir, was developed in 2000 and contained a combination of ABC, 3TC, and AZT while Atripla, a combination of TDF, FTC and efavirenz (EFV) was developed in 2006 (Table 8.13.1). Atripla was the first complete FDC regimen for treatment naïve patients and has been the most commonly used FDC treatment (De Clercq 2009; Este and Cihlar 2010).

The face of AIDS in the United States (US) between 1996 and 1997 changed subsequent to the introduction of cART from an inevitably fatal condition to one that is chronic and manageable, however, AIDS-related deaths in resource-constrained settings only started to decline in 2007 (UNAIDS 2013). In sub-Saharan Africa, home to about 70% of the global burden of HIV disease, access to cART was limited due to the high prices of HIV drugs enforced through trade law agreements preventing generic manufacture of patented drugs. In South Africa, for example, the mean income in 1986 was $2,600 per annum while the cost of ARVs at the time was $1,000 per month. Countries like Thailand and India that were excluded from such trade law agreements imposed by the World Trade Organization (WTO) were able to develop and market generic ARVs at a fraction of the price of the patented ARVs enabling their citizens access to affordable cART and realize its life-saving benefits. In contrast to the estimated $10,000 per annum per person cost of patented drugs, the average cost of generic ARVs were $365 per annum per person (approximately $1 per day) (Fisher and Rigamont 2005). Africa still could not gain access to generic ARVs, due to the rights to exclusivity, resulting in millions of people dying from AIDS with substantial reductions in life expectancy and hundreds of thousands of orphans. By the year 2000, a combination of global solidarity, activism, and the establishment of the Global Fund to Fight AIDS, Tuberculosis and Malaria and the US President’s Emergency Plan for AIDS Relief (PEPFAR) facilitated access to life-saving ARVs for millions of Africans.

Through the discovery of ARVs and increased global access to ARVs, AIDS has evolved globally from an inevitably fatal condition to one that is chronic and manageable. In 2017, approximately 22 million HIV-infected people globally were on cART. An estimated 40% still need to have access

to treatment. While mortality rates have been reduced substantially, we still continue to see approximately 1 million AIDS-related deaths each year. Continued stigma and discrimination; limited access to healthcare services and untreated tuberculosis remain critical barriers to realizing the survival benefits of cART for many and ongoing efforts to increase access to cART need to pay more careful attention to these factors.

HIV-infected individuals on cART are living longer, and are experiencing new health challenges associated with ageing. Wing et al. (2016) reported that HIV-positive patients on cART experience more rapid development of dementia, chronic neurological complications, frailty, and geriatric syndromes than HIV-negative individuals (Wing 2016). Long-term ARV use has also been associated with hyperactivation of the immune system, premature immunological ageing, and chronic inflammation (Appay and Kelleher 2016). Various studies suggest that these immune and inflammatory changes drive more rapid ageing, this however remains controversial (Justice and Falutz 2014; Pathai et al. 2014).

The gradual deterioration of immune function, increased occurrence of geriatric syndromes and frailty, increased rates of metabolic syndrome (i.e. chronic comorbidities) and continuous increase in inflammatory markers (referred to as inflammaging), all of which occur at younger ages in HIV positive vs. negative patients, supports the argument that HIV accelerates the ageing process (Franceschi et al. 2000). In contrast, some researchers suggest that adherence to cART translates to HIV-positive patients having suppressed viral loads and CD4 counts of >500cells/µl which supports life expectancy no different to that of HIV-negative individuals and thus HIV does not accelerate ageing (Althoff et al. 2015; Rasmussen et al. 2015). Further research is required to provide clarification.

In addition to the possible association with accelerated ageing, the long-term use of ARVs has also been associated with an increased risk of non-AIDS defining illnesses (or AIDS comorbidities) at all ages (Smit et al. 2015). These comorbidities include diabetes (Kalra et al. 2011; Medapalli et al. 2012), obesity (Freitas et al. 2012) and cardiovascular disease and stroke (Shahbaz et al. 2015). Like with ageing, studies on the association between diabetes and HIV have also been controversial with early studies showing a clear link but later studies demonstrating variations in this link (Ford et al. 2010; Rasmussen et al. 2012; Jarrett et al. 2013; Monroe et al. 2015). While the association between abacavir use and cardiovascular disease also

remains controversial (Martin-Iguacel et al. 2015) the Veteran Aging Cohort Study (Freiberg et al. 2013; Sico et al. 2015) and additional studies (Obel et al. 2007; Klein et al. 2015) reported a non-age-dependant increased risk of myocardial infarction and stroke in HIV-positive individuals even among those with supressed viral loads. Recently, a link between diet and ARVs has also been elucidated whereby ARV use coupled with a high-fat diet was found to impair energy balance, glucose metabolism, inflammatory transcriptional pathways and cause metabolic dysfunction (Pepin et al. 2018).

Collectively these data highlight that notwithstanding the benefits of cART in the treatment of AIDS; given its requirement for lifelong use and long-term effects, the need for a cure remains paramount.

Preventing mother-to-child transmission of HIV

The prevention of mother-to-child transmission (PMTCT) has been a significant milestone in the HIV epidemic and represents one of the greatest successes in the field of human therapeutics. Mother-to-child transmission (MTCT) was first described in 1982 and was attributed to infection *in utero*, intrapartum or post-partum through breastfeeding (CDC 1982a; CDC 1983; Greene 2007). The first PMTCT recommendations were issued by the US public health services late in 1985 and suggested counselling for HIV-positive pregnant women on how to live with HIV and take care of a child with HIV. These guidelines also advised against breastfeeding to prevent postnatal transmission of HIV due to early reports suggesting breastmilk as a route of transmission (CDC 1985; Ziegler et al. 1985). By 1987, the World Health Organization (WHO) confirmed breastfeeding as a route of HIV transmission (WHO 1987) and perinatal transmission of HIV was classified as the main cause of HIV in children. Prevention efforts focused on trying to prevent HIV infection in women of reproductive age through education and advising HIV-positive women against breastfeeding. This, however, did not address HIV acquisition *in utero* or intrapartum for children born to HIV-positive mothers or lack of access to clean water supplies in some regions of the world and, not surprisingly, the incidence of HIV in children continued to increase.

By April 1991, the National Institutes of Health (NIH) in collaboration with the National Institute of Health and Medical Research and the National

Agency of Research on AIDS in France sponsored the paediatrics AIDS Clinical Trials Group (ACTG) Protocol 076 (ACTG 076), a randomized double-blinded clinical trial investigating the efficacy of AZT (also referred to as zidovudine) in PMTCT, based in the USA and France (CDC 1985). Oral AZT was administered from week 14–34 of gestation throughout the pregnancy, intravenous AZT was administered during labour and oral AZT was administered to the baby for six weeks following delivery (CDC 1994). Interim results released in December 1993 demonstrated a 67.5% reduction in risk of perinatal HIV transmission among the selected HIV pregnant women and their infants without significant toxicity (CDC 1994; Connor et al. 1994; National Institute of Allergy and Infectious Diseases 1994). These results led to the termination of the placebo arm enrolment and existing placebo patients were offered AZT. By 1994, the use of AZT for HIV PMTCT was approved with an understanding that the long-term effects of AZT were still unknown and the research on which this recommendation was based was incomplete and required completion with time (CDC 1993). It soon became a standard of care in industrialized countries however, the low-income regions struggled to implement this regimen due to poor infrastructure, drug costs, and local policies among other issues.

Four additional AZT/ZDV clinical trials investigating AZT/ZDV against placebo were conducted; one in Thailand (Shaffer et al. 1999), two in Africa (Wiktor et al. 1999) and one in Cote d'Ivoire/Burkina Faso (i.e. the ANRS 049 DITRAME trial) (Dabis et al. 1999; Msellati et al. 1999). Variations in trial protocols made direct comparison of trials difficult. However, one African study (based in Cote d'Ivoire) and the Thai study had identical protocols and showed efficacies of 37% and 50% respectively. Irrespective of variations in protocols, all AZT/ZDV studies demonstrated a reduction in MTCT with AZT/ZDV. The PHPT-1 trial, a PMTCT trial in which mothers did not breastfeed, investigated the impact on varied durations of AZT/ZDV therapy to the mother and infant and found that longer antenatal treatment with AZT/ZDV was more effective than the shorter treatment, however, longer postnatal treatment provided no added benefit to the infant (Lallemant et al. 2000).

The HIVNET 012 perinatal HIV prevention trial, which began in 1999 in Uganda, aimed to investigate the efficacy and safety of neonatal and intrapartum single-dose NVP and AZT/ZDV (Guay et al. 1999). Additionally, the trial also aimed to compare the efficacy of single-dose

NVP (sdNVP) to short-course AZT/ZDV treatment. The trial reported lower rates of vertical transmission with NVP than AZT without added side-effects or changes in mortality (Guay et al. 1999; Jackson et al. 2003). Aspects of the trial including record-keeping, methodologies and data analysis presented some controversy resulting in the NIH requesting the US Institute of Medicine (IOM) to conduct an independent review of the HIVNET 012 trial which was completed by 2004 and confirmed that the results which were reported were accurate (Delvaux 2004).

By this time, a short course of AZT or a single-dose of NVP administered to mother and infant demonstrated success in reducing vertical transmission of HIV by 50% and presented cost-effective methods for PMTCT in resource-constrained settings, with few side-effects (Guay et al. 1999; Shaffer et al. 1999). Although utilized, at the time, additional clinical trials to elucidate efficacious methods for PMTCT were under investigation, particularly to address drug resistance in HIV, such as that associated with NVP monotherapy (Wainberg and Friedland 1998).

The ANRS 075 trial (no breastfeeding) (Mandelbrot et al. 2001) and the PETRA trial (included breastfeeding) conducted in South Africa, Tanzania, and Uganda (Petra Study Team 2002) assessed the efficacy of combination therapy (AZT/ZDV and 3TC) on PMTCT. While results showed efficacy in PMTCT, both trials demonstrated the development of drug resistance, specifically to 3TC, with the ANRS 075 trials having higher levels of drug resistance due to longer treatment duration with 3TC. Interestingly, the South African SAINT trial, which compared the efficacy of AZT/ZDV+3TC (administered intrapartum and for one week following birth) to sdNVP (one dose given intrapartum and one dose given during labour to the mother and within 72 hours of birth to the infant) on PMTCT, found no significant difference in outcomes (Moodley et al. 2003).

The PHPT-2 Thai trial (Lallement et al. 2004), ANRS1201 DITRAME PLUS study in Cote d'Ivoire (Dabis et al. 2005) and the Malawi based open-label randomized trial (Taha et al. 2004) all investigated the effects of using NVP in addition to AZT/ZDV regimens for PMTC and found no added benefit.

While making comparisons between trials is difficult, the trend showed that a longer course of treatment was associated with improved outcomes and that sdNVP was the most efficacious and cost-effective regimen for prevention of HIV in the infant. Further, the addition of sdNVP to a

AZT/ZDV inclusive regimen only showed added benefits in the case where infants were breastfed.

Until 2009, the most commonly used regimen in developing countries was short-course AZT/ZDV with sdNVP given to mother and newborn during labour and 72 hours after birth respectively, with mothers from high-income settings advised to refrain from breastfeeding. In 2009, the WHO released guidelines advising the administration of cART to all pregnant women with a CD4 count <350 cells/ml (from 14 weeks of gestation) and the daily administration of NVP to infants from birth to 6 weeks if they are to be breastfed, or a course of daily AZT or NVP would be used for infants that were not breastfed. The guidelines also advised 12 months of daily NVP therapy for infants if they are to be breastfed for the duration, as is commonly the case in developing regions. These guidelines changed a few years later, advising the initiation of all pregnant women with CD4 <500 cells/ml (from 14 weeks of gestation onward) on cART. Currently the option B+ treatment strategy states that all pregnant women, irrespective of CD4 count, be placed on lifelong cART as soon as possible to reduce their viral load and prevent MTCT and all breastfeeding women should also receive lifelong cART (Chappell and Cohn 2014).

The use of antiretroviral drugs in PMTCT has led to substantial reductions in MTCT of HIV, however; the goal to eliminate MTCT (EMTCT) is still paramount. The Global Plan launched by the WHO in 2011 targets a reduction in MTCT by 90% (WHO 2011) with the WHO developing guidelines in 2014 on integrated approaches for EMTCT and methods for validation of EMTCT (WHO 2014). The EMTCT of HIV is defined by the WHO as less than 50 paediatric infections per 100,000 live births and a transmission rate of either <5% for breastfed babies or <2% for non-breastfed babies. Additionally, 95% of pregnant women must be receiving antenatal care, HIV testing, and subsequent treatment if HIV positive. Cuba was the first country to receive validation for the EMTCT of HIV in 2015 Caffe et al. 2016), followed by Belarus, Thailand, and Armenia in 2016 (WHO 2016). Showing positive results in PMTCT, were Botswana, Mozambique, Namibia, South Africa, Swaziland, and Uganda (six priority countries identified by the WHO), which all achieved the Global Plan target of reducing PMTCT by 90% in 2015 (UNAIDS 2016a). An additional 14 priority countries still require achievement of the Global Plan and/or EMTCT goals. Interestingly, all these countries, except Nigeria, rolled out

option B+ for PMTCT in 2015 and consequently experienced a decline in AIDS-related deaths among females of reproductive age between 2009–2015 (UNAIDS 2016a).

In addition to a decline in PMTCT by 90% in 2015, the Global Plan also set a target for countries to achieve a reduction in MTCT rate to <5% among women that breast fed and <2% among non-breastfeeding women. The 21 Global Plan priority countries collectively demonstrated a decrease in MTCT among breastfeeding women from 22.4% to 8.9% between 2009 and 2015 (UNAIDS 2014) showing decline in the direction of the target.

There are still various challenges that need to be addressed in order to achieve EMTCT, particularly in low- and middle-income settings. The main issues hindering EMTCT in these regions include: delays in seeking pre-natal care, misperceptions over exclusive breastfeeding, HIV discrimination and stigma, lack of resources at a clinic and country level, long waiting times for HIV results for infants, side-effects of treatment, gender dynamics and cultural beliefs (UNAIDS 2014, 2016). These challenges highlight the need for interventions that go beyond the implementation of therapeutics and speaks to the need for a vaccination or cure. Furthermore, it highlights the fact that to achieve EMTCT a decline in incidence of HIV in adults is essential.

Preventing sexual transmission of HIV

There is no risk of HIV infection among those who practice sexual abstinence or lifelong mutual monogamy. Serial monogamy, where there are multiple sequential individual short-lived monogamous partnerships, is associated with an increased risk of HIV, but not to the same extent as the substantial increase in risk of transmission emanating from multiple concurrent sexual partnerships. Reduction in the number of concurrent sexual partnerships and the use of condoms are key components of HIV prevention messages, widely promoted as part of ‘ABC’ campaigns promoting Abstinence, Be faithful, and Condom use. Since 2010, this prevention mantra has been expanded to include knowledge of HIV status; medical male circumcision, pre-exposure prophylaxis (PrEP), and test and treat.

Male and female condoms

Condoms are a pivotal component of the fight against HIV/AIDS. They are inexpensive and relatively easy to use and provide protection against acquisition and transmission of HIV, a wide range of other sexually transmitted infections as well as pregnancy (i.e. an available multipurpose technology).

When used correctly and consistently, the male and female condom is highly effective in preventing the sexual transmission of HIV. The strongest evidence for the role of condoms in preventing the transmission of HIV comes from sero-discordant couple studies, which uniformly show that increased condom use is associated with a substantially reduced risk of HIV transmission.

Despite their increasing accessibility and proven effectiveness in prevention of HIV, a wide range of factors have been implicated as barriers to condom use; the most common being the widespread perception that condoms reduce sexual pleasure and that suggesting the use of condoms represents self-acknowledgement of HIV infection or a lack of trust in the partner. In the context of a marital relationship or stable partnership where pregnancy is desired, or where subordination of women limits their ability to negotiate safer sex practices, attempts to introduce or promote condom use have had limited success.

Several studies have demonstrated that alcohol consumption is associated with inconsistent condom use; this poses a challenge to individuals who meet high-risk sexual partners in social settings where alcohol is available. Condom use is lower in partnerships where an effective form of contraception is being used. This points to the need for interventions that promote dual method use (the simultaneous use of condoms with another form of contraception) among high-risk women.

Sexually transmitted infections

HIV transmission and acquisition during intercourse is enhanced in the presence of sexually transmitted infections, particularly ulcerative infections such as syphilis, chancroid and herpes simplex type 2 virus (HSV-2) infection (Wald and Link 2002; Freeman et al. 2006). Genital ulceration or inflammation caused by sexually transmitted infections increase the infectiousness of HIV-positive individuals and the susceptibility of HIV-

negative individuals. Although only one randomized controlled trial has been able to demonstrate a significant reduction in HIV incidence rates following treatment of sexually transmitted infections (Grosskurth et al. 2000), the significant sexual and reproductive health challenge posed by the high burden of curable sexually transmitted infections needs to be addressed in any HIV prevention effort.

Medical male circumcision

Medical male circumcision is a proven HIV prevention option for men, reducing the risk of female-to-male transmission of HIV by up to 60% (Siegfried et al. 2009). Medical male circumcision has become a priority component of HIV prevention globally. Introduced in 2007, the number of medical male circumcisions has increased from just 5,190 in 2008 to approximately 14.5 million in 2016. Evidence of a population level impact of circumcision in South Africa has recently become available from Orange Farm, where an increase in the male circumcision rate from 12% in 2007 to 53% in 2010 was associated with a 57% reduction in HIV incidence rates (Auvert et al. 2013).

Use of antiretrovirals as PrEP for HIV prevention

In addition to treatment benefits, the prophylactic use of antiretroviral drugs has generated new hope in the HIV prevention landscape. Since July 2010, when ARVs were first shown by the CAPRISA 004 trial (Abdool Karim et al. 2010) to prevent sexual transmission of HIV, the HIV prevention landscape has been transformed, principally through oral tenofovir (TFV)-containing PrEP (Grant et al. 2010; Baeten et al. 2012; Thigpen et al. 2012; Van Damme et al. 2012a; Marrazzo et al. 2015; McCormack et al. 2015).

Although daily oral TDF/FTC has been demonstrated to be consistently effective in MSM and transgender women globally (Grant et al. 2010; McCormack et al. 2015), the results of clinical trials in African women have been inconsistent, most likely due to varying levels of adherence (Thigpen et al. 2012; Baeten et al. 2012; Van Damme et al. 2012a; Marrazzo et al. 2015). In the FEM-PREP trial (Van Damme et al. 2012b) for example, only 24% of the women allocated to the daily oral TDF/FTC group had detectable drug levels. Similarly, in the VOICE trial, only 25%, 29%, and 30% of women

allocated to the daily tenofovir gel, daily oral TDF/FTC and daily oral TDF groups respectively had detectable drug levels (Marrazzo et al. 2015). The challenge of adherence, however, is not simply a behavioural issue. Women are more likely than men to require near-perfect adherence to oral PrEP in order to be protected against HIV (Donnell et al. 2014), indicating that biological factors may also be important to consider. Genital inflammation has been shown to be an independent risk factor for HIV acquisition (Masson et al. 2015) and affects PrEP effectiveness in women (McKinnon et al. 2018). Deep sequencing techniques have shed light on the cause of genital inflammation showing that anaerobic vaginal bacteria, which is common in a dysbiotic vaginal microbiome, is associated with both genital inflammation and HIV risk (Williams and Passmore 2016). Using proteomics, it was recently shown that a dysbiotic vaginal microbiome, comprising <50% *Lactobacillus* spp., directly influenced topical PrEP efficacy (Klatt et al. 2017).

Despite the adherence challenges, oral PrEP is an important HIV prevention option and is now being implemented as part of a comprehensive HIV prevention package in several countries. In industrialized countries, high drug costs remain a major barrier. An increase in sexually transmitted infections has been observed among PrEP users underscoring the importance of barrier protection in addition to PrEP use. Important research is currently underway with novel ARVs and delivery mechanisms that are less user dependent.

Treatment as prevention

There is now compelling evidence from the landmark HPTN 052 trial that antiretroviral treatment of HIV-infected individuals can prevent sexual transmission of HIV (Quinn et al. 2000). Early antiretroviral treatment initiation in the HIV-positive partner showed that HIV transmission was reduced by 96% in HIV discordant couples from nine countries (HPTN052) (Cohen et al. 2016). Taken in conjunction with other observational data, which show significant reductions in HIV transmission among serodiscordant couples receiving ART, onward transmission of HIV can be significantly reduced if viral load is suppressed through ART. This added benefit of cART and viral suppression for preventing HIV infection in

uninfected persons is also increasingly being used as a prevention option captured as ‘undetectable equals uninfectious’ or abbreviated to U=U.

Modern technologies

In recognition of stigma and discrimination as a barrier to test and treat for HIV, modern communication strategies have been introduced to encourage HIV testing, treatment, and support ARV adherence. These strategies include the eHealth and mHealth technologies (Muessig et al. 2015). eHealth is designed to deploy educational messages, to people at risk, either electronically or via a message on a website which the person frequents. To date, eHealth has targeted the MSM population with gay organizations utilizing the tool to develop websites frequented by gay individuals to deliver both educational information and referrals to gay-friendly test and treat sites (Sullivan et al. 2013; Muessig et al. 2015).

Similarly, mHealth, utilizes social media applications (apps) to deliver tailored HIV/AIDS messages to users via their mobile devices or tablets. This technology has for example been incorporated into the Blued app (Blued app 2009), a Chinese gay dating app, which allows the delivery of HIV/AIDS educational messages and messages encouraging testing for HIV via the use of ‘non-judgemental community mobilization campaigns’. In addition to this, mHealth also offers options to encourage ARV adherence, by reminding patients to take pills and asking patients to indicate if pills were taken. The technology also provides reminders to patients for clinic visits.

Phylogenetics, the use of viral sequences from recently infected individuals in a common social network (Villandré et al. 2019) or geographically defined area (de Oliveira et al. 2017), is being used to understand transmission dynamics better and enhancing surveillance efforts and informing more targeted and efficient approaches for responding to HIV.

Microbiome analysis (Williams and Passmore 2016) and proteomics (Klatt et al. 2017) is advancing our understanding of the role of biology in rendering some women more vulnerable to HIV compared to others or the role the microbiome plays in reducing the potency of prophylactic use of ARVs.

Injecting drug use

IDU represents a major risk factor for HIV infection, often of the opiate heroin. It has been estimated that one in ten new HIV infections is caused by sharing of contaminated needles (AIDS Alliance 2016). Approximately one-third of all new HIV infections in regions outside of sub-Saharan Africa occur among people who inject drugs (PWID) (Harm Reduction International 2016a; UNODC 2016). Collectively, China, Russia, and the USA account for 50% of all PWID worldwide and contribute the most to HIV infections among this group (Harm Reduction International 2016a). Eastern Europe and central Asia also have high burdens of PWID, with half of all new HIV infections in these regions occurring among PWID (Harm Reduction International 2016b). Despite PWID categorized as a priority group for HIV prevention and treatment, they represent a group with the least access to treatment, prevention, and testing programmes. This is driven by the legislation, criminalization, and stigma associated with drug use, which detracts PWID from utilizing healthcare and harm reduction programmes (UNAIDS 2016b). The high susceptibility to HIV in this group, is further driven by the criminalization associated with distribution or possession of syringes or needles for non-medical purposes in some regions, which encourages sharing of used needles (Beletsky et al. 2012a, 2012b).

Despite PWID representing a key population independently, there is also bridging between sex workers and drug addiction. Mongolia, Afghanistan, and central Asia have reported a 20 times increased risk of HIV infection in sex workers who inject drugs compared to those who do not inject drugs (Baral et al. 2013). This overlap demonstrates the importance of HIV treatment and prevention programmes for this key population as they place other populations at risk for HIV acquisition as well.

To date, interventions aimed at reducing the risk of HIV acquisition in PWID include needle and syringe programmes (NSPs), opioid substitution therapy (OST) and the use of PrEP, often referred to as harm reduction initiatives, with evidence suggesting a combination of these approaches should be used in PWID (Degenhardt et al. 2010).

NSPs aim not only to reduce the transmission of HIV, but also that of other blood borne infections such as hepatitis C and B. Some programmes provide 200 sterile needles per drug user per year, some sites also provide equipment to prepare drugs such as sterile water, filters and mixing containers. In addition to this, some programmes employ interventions to reduce harm associated with IDU such as; provision of advice on how to

inject drugs safely, information on how to prevent and manage overdoses, information on how to handle and safely dispose of needles and syringes and provision of referrals for HIV testing and treatment and other health services (NICE 2014). Some NSPs are delivered via; fixed sites, mobile units, peer-NSP programmes, street outreach, backpacking services, home delivery, syringe vending machines and pharmacies (Obadia et al. 1999; Riley et al. 2000; WHO 2007; Bryant et al. 2010; Jain et al. 2014). Despite demonstrating success in reduction of HIV incidence among PWID (Jones et al. 2010; Palmateer et al. 2010; Abdul-Quader et al. 2013; Aspinall et al. 2014; MacArthur et al. 2014), the NSP has only been implemented in 90 countries out of the 158 in which it is required in (Harm Reduction International 2016b).

While all 29 regions in eastern Europe and central Asia have NSPs, the success of these programmes are hindered by hostile political reactions and inadequate government support which has forced some NSPs to shut down (Harm Reduction International 2016b). Within Asia and the Pacific, only 17 countries have NSPs with the lack of NSPs in this region driven by harsh drug policies, inflexible working hours, reduction in funding and human resource issues (Harm Reduction International 2016b). The coverage of NSPs is high in western and central Europe and North America, particularly in Canada were an estimated 94.5% of PWID reported to have used a clean needle on last injection. Many countries within Europe though are seeing declines in the number of operational NSPs and have various PWID that are not reached due to people <18 years of age being denied access to services and immigrants not being reached due to underdeveloped services in areas where they live. Latin America, the Middle East, and North Africa all have limited NSPs, which are not adequate for reaching the required number of PWID in the region. Here, criminalization of drug use, restriction of operating hours, inadequate resources, and poorly trained staff contribute to the inadequacy of the existing programmes. Within sub-Saharan Africa, NSPs only exist in five out of 54 countries due to a lack of political will and government support and funding (Harm Reduction International 2016b).

In addition to NSP, OST, a medication-assisted treatment (MAT), is also used to manage injecting drug use and reduce associated harm among PWID. OST provides illegal drug users with replacement medication (i.e. opioids), most often methadone maintenance therapy (MMT) and sometimes buprenorphine, administered orally in a supervised clinical environment.

Various studies have demonstrated the success of OST in reducing the frequency of injection use, improving health and social functioning, improving ARV use, improving linkage to health services, and reducing crime and unsafe injecting practices, particularly in developed countries (Lawrinson et al. 2008; Weber et al. 2009; Gowing et al. 2011; Vorma et al. 2013). Studies also reported a 54% reduction in HIV transmission among PWID with the use of OST (Riley et al. 2000; Aspinall et al. 2014). Studies have also shown that the more time PWID spend on opioid therapy the more positive their health outcomes and less likely they are to engage in risky behaviour (Kumar et al. 2009). Despite the success of OST, it is estimated that only 8% of PWID have access to this intervention. A recent systematic review of coverage of interventions to prevent and manage HIV among PWID shows that there is a high coverage of both NSP and OST in Australia, the Netherlands, Austria, and Norway, moderate coverage in Canada, New Zealand and Vietnam and low coverage in China, Indonesia, Tanzania and several other developing regions while places like the Russian federation have very limited access to NSP and do not provide OST (LArney et al. 2017). This has been attributed to social and structural discrimination, stringent laws that criminalize drug use and the lack of political will to implement interventions like OST due to the belief that abstinence of drug use is the only successful method to treat drug addiction (Connock et al. 2007; Jurgens et al. 2010; Kermode et al. 2011).

Oral PrEP, a daily course of ARVs, which protects HIV-negative individuals from HIV acquisition prior to exposure to the virus, is also a recommended method to reduce HIV transmission among PWID. Oral PrEP typically comprises of a combination of TDF and FTC combined as a single pill Truvada. The success of oral PrEP in reducing HIV transmission among PWID has been reported, however, the high cost of oral PrEP is a major factor hindering its large-scale implementation for PWID (Choopanya et al. 2013; Fu et al. 2018).

Although there is strong evidence to show that NSPs and OST reduce injecting use behaviours and HIV transmission by 48% and 54% respectively, there are still barriers to their implementation in most regions (Aspinall et al. 2014; MacArthur et al. 2014) predominantly driven by the criminalization and stigma associated with drug use and the lack of political support and government funding. While oral PrEP offers an alternate route that is somewhat less controversial and possibly more accessible than OST

and NSP, it comes with its own setback of high costing and does not lower the prevalence of HCV and drug-related infections unless additional drugs are included. It must be recognized that prevention of HIV among PWID has far-reaching consequences on sex workers and on the general population as these individuals overlap and thus reducing incidence of HIV in one group could likely translate to reducing its incidence in various other groups.

Conclusion

Notwithstanding the enormous scientific progress and programmatic success in initiating almost 22 million infected persons on cART and reducing vertical transmission rates to <1%; in 2017, globally there were about 1 million AIDS-related deaths, 1.8 million new HIV infection and about 37 million people living with HIV. These data underscore the challenges that continue to face us. Much more needs to be done to prevent new HIV infections and reach and facilitate access to services by key and priority populations.

While globally, we are on track with meeting the 2020 intermediate targets for treatment; prevention of new infections remains a major challenge with only about 15% reduction in new HIV infections we are way off the 2020 target of 500,000 new infections. Stigma and discrimination, real or perceived, remains a major barrier to knowledge of HIV status or seeking prevention or care services by key and priority populations.

Several countries have reached their 2020 targets and are well on their way to reaching their 2030 ending AIDS goal. Most countries are at different stages of progress towards reaching this 2020 intermediate goal. There is increasing recognition that 90–90–90 treatment targets alone will not be sufficient to get us to the 2030 UN goal of ending AIDS.

Despite huge progress on prevention science research, preventing HIV infection remains a major challenge with little impact on reductions in incidence rates in the past 6 years. With approximately 5,000 new HIV infections occurring each day globally, ending AIDS will remain an elusive goal. Combinations of stigma and discrimination, marginalization, moral and judgemental attitudes, and criminalization of populations on the basis of sexual orientation, occupation or age remain barriers to knowledge of HIV status and access to HIV testing and much needed treatment and prevention services. According to the 2018 UNAIDS report, it is estimated that as much

as 9.4 million HIV-infected individuals globally and much higher proportions in some countries do not know they are infected and, therefore, get neither tested nor treated. It is recognized that transmission persists because of these untested infected individuals and inadequately treated individuals continuing to engage in high risk activities. Peer-led interventions that engage and partner with key and priority populations and communities is critical to reach those we are not reaching with typical healthcare delivery systems and approaches. Given the diversity of epidemics within and between countries in terms of magnitude, modes of transmission and populations at risk no single approach will address all the challenges and implementation science and programming needs to identify optimal combinations of interventions and strategies needed at a local level for prevention success.

Investments in novel prevention approaches need to continue including for the development of a vaccine and a cure to realize the 2030 ending AIDS goal as well as eventually eliminating and eradicating HIV. At the heart of our response and success to date are courageous advocates and people whose rights have to be respected and who have to continue to be part of our responses for success.

REFERENCES

Abdool Karim, Q., et al. (2010). Effectiveness and safety of tenofovir gel, an antiretroviral microbicide, for the prevention of HIV infection in women. *Science*, **329**, 1168–74.

Abdul-Quader, A.S., et al. (2013). Effectiveness of structural-level needle/syringe programs to reduce HCV and HIV infection among people who inject drugs: a systematic review. *AIDS & Behavior*, **17**, 2878–92.

AIDS Alliance (2016). The facts speak for themselves: so why is no-one listening? [Online] Available at: https://aidsalliance.org/blog/719-the-facts-speak-for-themselves-so-why-is-noone-listening

Alexander, T.S. (2016). Human immunodeficiency virus diagnostic testing: 30 years of evolution. *Clinical and Vaccine Immunology*, **23**, 249–53.

Althoff, K.N., et al. (2015). Comparison of risk and age at diagnosis of myocardial infarction, end-stage renal disease, and non-AIDS-defining cancer in HIV-infected versus uninfected adults. *Clinical Infectious Diseases*, **60**, 627–38.

Appay, V. and Kelleher, A.D. (2016). Immune activation and immune aging in HIV infection. *Current Opinion in HIV & AIDS*, **11**, 242–9.

Arts, E.J,. and Hazuda, D.J. (2012). HIV-1 antiretroviral drug therapy. *Cold Spring Harbor Perspectives in Medicine*, 2, a007161.

Aspinall, E.J., et al. (2014). Are needle and syringe programmes associated with a reduction in HIV transmission among people who inject drugs: a systematic review and meta-analysis. *International Journal of Epidemiology*, **43**, 235–48.

Auvert, B., et al. (2013). Association of the ANRS-12126 male circumcision project with HIV levels among men in a South African township: evaluation of effectiveness using cross-sectional surveys. *PLoS Medicine*, 10, e1001509.

Baeten, J.M., et al. (2012). Antiretroviral prophylaxis for HIV prevention in heterosexual men and women. *New England Journal of Medicine*, **367**, 399–410.

Baral, S., et al. (2013). HIV among female sex workers in the Central Asian Republics, Afghanistan, and Mongolia: contexts and convergence with drug use. *Drug & Alcohol Dependency*, 132 S13–16.

Bayer, R. and Fairchild, A.L. (2006). Changing the paradigm for HIV testing--the end of exceptionalism. *New England Journal of Medicine*, **355**, 647–9.

Beletsky, L., et al. (2012a). Mexico's northern border conflict: collateral damage to health and human rights of vulnerable groups. *Revista Panamericana de Salud Pública*, **31**, 403–10.

Beletsky, L., et al. (2012b). Policy reform to shift the health and human rights environment for vulnerable groups: the case of Kyrgyzstan's Instruction 417. *Health & Human Rights*, **14**, 34–48.

Blued app (2019). Blued. [Online] Available at: https://www.blued.com/en/index.html

Broder, S. (2010). The development of antiretroviral therapy and its impact on the HIV-1/AIDS pandemic. *Antiviral Research*, **85**, 1–18.

Bryant, J., et al. (2010). Is point of access to needles and syringes related to needle sharing? Comparing data collected from pharmacies and needle and syringe programs in south-east Sydney. *Drug & Alcohol Review*, **29**, 364–70.

Caffe, S., et al. (2016). Cuba validated as the first country to eliminate mother-to-child transmission of human immunodeficiency virus and congenital syphilis: lessons learned from the implementation of the global validation methodology. *Sexually Transmitted Diseases*, **43**, 733–6.

Carpenter, C.C., et al. (1997). Antiretroviral therapy for HIV infection in 1997. Updated recommendations of the International AIDS Society-USA panel. *Journal of the American Medical Assocation*, **277**, 1962–9.

Center for Disease Control (CDC), Morbidity & Mortality Weekly Report (1981). Kaposi's sarcoma and pneumocystis pneumonia among homosexual men—New York City and California. *MMWR Weekly*, **30**, 305–8.

Centers for Disease Control (CDC) (1989). Interpretation and use of the Western blot assay for serodiagnosis of human immunodeficiency virus type 1 infections. *MMWR Morbidity & Mortality Weekly Report*, **38**, 1–7.

Centers for Disease Control and Prevention (CDC) (1983). Current trends acquired immunodeficiency syndrome (AIDS) update—United States. *MMWR Morbidity & Mortality Weekly Report*, **32**, 309–11.

Centers for Disease Control and Prevention (CDC) (1985). Current trends recommendations for assisting in the prevention of perinatal transmission of human T-lymphotropic virus type III/lymphadenopathy-associated virus and acquired immunodeficiency syndrome. *MMWR Morbidity & Mortality Weekly Report*, **34**, 721–6, 731–2.

Centers for Disease Control and Prevention (CDC) (1993). Recommendations of the U.S. Public Health Service Task Force on the use of zidovudine to reduce perinatal transmission of human immunodeficiency virus. *MMWR Morbidity & Mortality Weekly Report*, **43**, 1–20.

Centers for Disease Control and Prevention (CDC) (1994). Zidovudine for the prevention of HIV transmission from mother to infant. *MMWR Morbidity & Mortality Weekly Report*, **43**, 285–7.

Centers for Disease Control and Prevention (1982b). Opportunistic infections and Kaposi's Sarcoma among Haitians in the United States. *MMWR Morbidity & Mortality Weekly Report*, **31**, 353–4, 360–1.

Centre for Disease Control (CDC) (1982a). Epidemiologic notes and reports: possible transfusion-associated acquired immune deficiency syndrome, AIDS—California. *MMWR Morbidity & Mortality Weekly Report*, **31**, 652–4.

Chappel, R.J., Wilson, K.M., and Dax, E.M. (2009). Immunoassays for the diagnosis of HIV: meeting future needs by enhancing the quality of testing. *Future Microbiology*, **4**, 963–82.
Chappell, C.A. and Cohn, S.E. (2014). Prevention of perinatal transmission of human immunodeficiency virus. *Infectious Disease Clinics of North America*, **28**, 529–47.
Choopanya, K., et al. (2013). Antiretroviral prophylaxis for HIV infection in injecting drug users in Bangkok, Thailand (the Bangkok Tenofovir Study): a randomised, double-blind, placebo-controlled phase 3 trial. *The Lancet*, **381**, 2083–90.
Clavel, F., et al. (1986). Isolation of a new human retrovirus from West African patients with AIDS. *Science*, **233**, 343–6.
Coffin, J., et al. (1986). What to call the AIDS virus? *Nature*, **321**, 10.
Cohen, M.S., et al. (2016). Antiretroviral therapy for the prevention of HIV-1 transmission. *New England Journal of Medicine*, **375**, 830–9.
Connock, M., et al. (2007). Methadone and buprenorphine for the management of opioid dependence: a systematic review and economic evaluation. *Health Technology Assessment*, **11**, 1–171, iii-iv.
Connor, E.M., et al. (1994). Reduction of maternal-infant transmission of human immunodeficiency virus type 1 with zidovudine treatment. Pediatric AIDS Clinical Trials Group Protocol 076 Study Group. *New England Journal of Medicine*, **331**, 1173–80.
Dabis, F., et al. (2005). Field efficacy of zidovudine, lamivudine and single-dose nevirapine to prevent peripartum HIV transmission. *AIDS*, **19**, 309–18.
Dabis, F., et al. (1999). 6-month efficacy, tolerance, and acceptability of a short regimen of oral zidovudine to reduce vertical transmission of HIV in breastfed children in Cote d'Ivoire and Burkina Faso: a double-blind placebo-controlled multicentre trial. DITRAME Study Group. DIminution de la Transmission Mere-Enfant. *The Lancet*, **353**, 786–92.
De Clercq, E. (2009). Anti-HIV drugs: 25 compounds approved within 25 years after the discovery of HIV. *International Journal of Antimicrobial Agents*, **33**, 307–20.
De Clercq, E. (2005). Emerging anti-HIV drugs. *Expert Opinion in Emergerging Drugs*, **10**, 241–73.
De Clercq, E. (2009). The history of antiretrovirals: key discoveries over the past 25 years. *Reviews in Medical Virology*, **19**, 287–99.
de Oliveira, T., et al. (2017). Transmission networks and risk of HIV infection in KwaZulu-Natal, South Africa: a community-wide phylogenetic study. *Lancet HIV*, 4, e41–e50.
Degenhardt, L., et al. (2010). Prevention of HIV infection for people who inject drugs: why individual, structural, and combination approaches are needed. *The Lancet*, **376**, 285–301.
Delvaux, T. (2004). Nevirapine sustains reduction of mother-to-child HIV transmission at 18 months. *Evidence-based Healthcare*, **8**, 87–8.
Detels, R., et al. (2012). The multicenter AIDS Cohort Study, 1983 to … *Public Health*, **126**, 196–8.
Donnell, D., et al. (2014). HIV protective efficacy and correlates of tenofovir blood concentrations in a clinical trial of PrEP for HIV prevention. *Journal of Acquired Immune Deficiency Syndrome*, **66**, 340–8.
Este, J.A. and Cihlar, T. (2010). Current status and challenges of antiretroviral research and therapy. *Antiviral Research*, **85**, 25–33.
Faria, N.R., et al. (2014). HIV epidemiology. The early spread and epidemic ignition of HIV-1 in human populations. *Science*, **346**, 56–61.
Fee, E. and Krieger, N. (1993). Understanding AIDS: historical interpretations and the limits of biomedical individualism. *American Journal of Public Health*, **83**, 1477–86.
Fisher, W.W. and Rigamont, C.P. (2005). *The South Africa AIDS Controversy: A Case Study in Patent Law and Policy.* Cambridge, MA: Harvard Law School.
Ford, E.S., Li, C., and Zhao, G. (2010). Prevalence and correlates of metabolic syndrome based on a harmonious definition among adults in the US. *Journal of Diabetes*, **2**, 180–93.

Franceschi, C., et al. (2000). Inflammaging. An evolutionary perspective on immunosenescence. *Annals of the New York Academy of Sciences*, **908**, 244–54.
Freeman, E.E., et al. (2006). Herpes simplex virus 2 infection increases HIV acquisition in men and women: systematic review and meta-analysis of longitudinal studies. *AIDS*, **20**, 73–83.
Freiberg, M.S., et al. (2013). HIV infection and the risk of acute myocardial infarction. *JAMA Internal Medicine*, **173**, 614–22.
Freitas, P., et al. (2012). Prevalence of obesity and its relationship to clinical lipodystrophy in HIV-infected adults on anti-retroviral therapy. *Journal of Endocrinology Investigation*, **35**, 964–70.
Fu, R., Owens, D.K., and Brandeau, M.L. (2018). Cost-effectiveness of alternative strategies for provision of HIV preexposure prophylaxis for people who inject drugs. *AIDS*, **32**, 663–72.
Furman, P.A., et al. (1986). Phosphorylation of 3'-azido-3'-deoxythymidine and selective interaction of the 5'-triphosphate with human immunodeficiency virus reverse transcriptase. *Proceeding of the National Academy of Sciences U S A*, **83**, 8333–7.
Gostin, L.O., Ward, J.W., and Baker, A.C. (1997). National HIV case reporting for the United States. A defining moment in the history of the epidemic. *New England Journal of Medicine*, **337**, 1162–7.
Gostin, L.O. (1990). The AIDS litigation project. A national review of court and human rights commission decisions, part II: discrimination. *Journal of the American Medical Association*, **263**, 2086–93.
Gottlieb, M.S., et al. (1983). The acquired immunodeficiency syndrome. *Annals of Internal Medicine*, **99**, 208–20.
Gottlieb, M.S., et al. (1981). Pneumocystis carinii pneumonia and mucosal candidiasis in previously healthy homosexual men: evidence of a new acquired cellular immunodeficiency. *New England Journal of Medicine*, **305**, 1425–31.
Gowing, L., et al. (2011). Oral substitution treatment of injecting opioid users for prevention of HIV infection. *Cochrane Database of Systematic Reviews*, 10, CD004145.
Grant, R.M., et al. (2010). Preexposure chemoprophylaxis for HIV prevention in men who have sex with men. *New England Journal of Medicine*, **363**, 2587–99.
Greene, W.C. (2007). A history of AIDS, looking back to see ahead. *European Journal of Immunology*, 37, S94–102.
Grosskurth, H., et al. (2000). Control of sexually transmitted diseases for HIV-1 prevention: understanding the implications of the Mwanza and Rakai trials. *Lancet*, **355**, 1981–7.
Guay, L.A., et al. (1999). Intrapartum and neonatal single-dose nevirapine compared with zidovudine for prevention of mother-to-child transmission of HIV-1 in Kampala, Uganda: HIVNET 012 randomised trial. *Lancet*, **354**, 795–802.
Harm Reduction International (2016a). *The Case for a Harm Reduction Decade: Progress, Potential and Paradigm Shifts*. [Online] Available at: https://www.hri.global/files/2016/03/10/Report_The_Case_for_a_Harm_Reduction_Decade.pdf
Harm Reduction International (2016b). *Global State of Harm Reduction*. [Online] Available at: https://www.hri.global/global-state-of-harm-reduction-reports
Hedenskog, M., et al. (1986). Testing for antibodies to AIDS-associated retrovirus (HTLV-III/LAV) by indirect fixed cell immunofluorescence: specificity, sensitivity, and applications. *Journal of Medical Virology*, **19**, 325–34.
Herek, G.M. and Capitanio, J.P. (1993). Public reactions to AIDS in the United States: a second decade of stigma. *American Journal of Public Health*, **83**, 574–77.
Institute of Medicine (US) Committee on the Ryan White CARE Act (1994). *Data for Resource Allocation, Planning, and Evaluation. Measuring What Matters: Allocation, Planning, and Quality Assessment for the Ryan White CARE Act*. Washington, DC: National Academies Press (US).
Jackson, J.B., et al. (2003). Intrapartum and neonatal single-dose nevirapine compared with zidovudine for prevention of mother-to-child transmission of HIV-1 in Kampala, Uganda: 18-

month follow-up of the HIVNET 012 randomised trial. *The Lancet*, **362**, 859–68.
Jain, B., et al. (2014). Effect of peer-led outreach activities on injecting risk behavior among male drug users in Haryana, India. *Harm Reduction Journal*, 11. doi.org/10.1186/1477-7517-11-3.
Jarrett, O.D., et al. (2013). Metabolic syndrome predicts all-cause mortality in persons with human immunodeficiency virus. *AIDS Patient Care STDS*, **27**, 266–71.
Jones, L., et al. (2010). Optimal provision of needle and syringe programmes for injecting drug users: a systematic review. *International Journal of Drug Policy*, **21**, 335–42.
Jurgens, R., et al. (2010). People who use drugs, HIV, and human rights. *The Lancet*, **376**, 475–85.
Justice, A. and Falutz, J. (2014). Aging and HIV: an evolving understanding. *Current Opinion in HIV & AIDS*, **9**, 291–3.
Kalra, S., et al. (2011). Understanding diabetes in patients with HIV/AIDS. *Diabetology & Metabolic Syndrome*, **3**, 2.
Keele, B.F., et al. (2006). Chimpanzee reservoirs of pandemic and nonpandemic HIV-1. *Science*, **313**, 523–6.
Kermode, M., et al. (2011). Opioid substitution therapy in resource-poor settings. *Bulletin of the World Health Organization*, **89**, 243.
Klatt, N.R., et al. (2017). Vaginal bacteria modify HIV tenofovir microbicide efficacy in African women. *Science*, **356**, 938–45.
Klein, D.B., et al. (2015). Declining relative risk for myocardial infarction among HIV-positive compared with HIV-negative individuals with access to care. *Clinical Infectious Diseases*, **60**, 1278–80.
Korber, B., et al. (2000). Timing the ancestor of the HIV-1 pandemic strains. *Science*, **288**, 1789–96.
Kumar, M.S., et al. (2009). Opioid substitution treatment with sublingual buprenorphine in Manipur and Nagaland in Northeast India: what has been established needs to be continued and expanded. *Harm Reduction Journal*, **6**, 4.
Lallemant, M., et al. (2000). A trial of shortened zidovudine regimens to prevent mother-to-child transmission of human immunodeficiency virus type 1. Perinatal HIV Prevention Trial (Thailand) Investigators. *New England Journal of Medicine*, **343**, 982–91.
Lallemant, M., et al. (2004). Single-dose perinatal nevirapine plus standard zidovudine to prevent mother-to-child transmission of HIV-1 in Thailand. *New England Journal of Medicine*, **351**, 217–28.
Larney, S., et al. (2017). Global, regional, and country-level coverage of interventions to prevent and manage HIV and hepatitis C among people who inject drugs: a systematic review. *Lancet Global Health*, 5, e1208–20.
Lawrinson, P., et al. (2008). Key findings from the WHO collaborative study on substitution therapy for opioid dependence and HIV/AIDS. *Addiction*, **103**, 1484–92.
Levy, J.A., et al. (1984). Isolation of lymphocytopathic retroviruses from San Francisco patients with AIDS. *Science*, **225**, 840–2.
MacArthur, G.J., et al. (2014). Interventions to prevent HIV and Hepatitis C in people who inject drugs: a review of reviews to assess evidence of effectiveness. *International Journal of Drug Policy*, **25**, 34–52.
Mandelbrot, L., et al. (2001). Lamivudine-zidovudine combination for prevention of maternal-infant transmission of HIV-1. *Journal of the American Medical Association*, **285**, 2083–93.
Marrazzo, J.M., et al. (2015). Tenofovir-based preexposure prophylaxis for HIV infection among African women. *New England Journal of Medicine*, **372**, 509–18.
Martin-Iguacel, R., Llibre, J.M., and Friis-Moller, N. (2015). Risk of cardiovascular disease in an aging HIV population: where are we now? *Current HIV/AIDS Reports*, **12**, 375–87.
Marx, J.L. (1985). A virus by any other name. *Science*, **227**, 1449–51.
Marx, J.L. (1982). New disease baffles medical community. *Science*, **217**, 618–21.

Masson L, et al. (2015). Genital inflammation and the risk of HIV acquisition in women. *Clinical Infectious Disease*, **61**, 260–9.
McCormack, S., et al. (2015). Pragmatic open-label randomised trial of preexposure prophylaxis: the PROUD study abstract 22LB. In: *Conference on Retroviruses and Opportunistic Infections (CROI)*. Seattle, USA.
McKinnon, L.R., et al. (2018). Genital inflammation undermines the effectiveness of tenofovir gel in preventing HIV acquisition in women. *Nature Medicine*, **24**, 491–6.
Medapalli, R.K., et al. (2012). Comorbid diabetes and the risk of progressive chronic kidney disease in HIV-infected adults: data from the Veterans Aging Cohort Study. *Journal of Acquired Immune Deficiency Syndrome*, **60**, 393–9.
Mitsuya, H., et al. (1985). 3'-Azido-3'-deoxythymidine (BW A509U): an antiviral agent that inhibits the infectivity and cytopathic effect of human T-lymphotropic virus type III/lymphadenopathy-associated virus *in vitro*. *Proceedings of the National Academy of Sciences U S A*, **82**, 7096–100.
Monroe, A.K., Glesby, M.J., and Brown, T.T. (2015). Diagnosing and managing diabetes in HIV-infected patients: current concepts. *Clinical Infectious Diseases*, **60**, 453–62.
Moodley, D., et al. (2003). A multicenter randomized controlled trial of nevirapine versus a combination of zidovudine and lamivudine to reduce intrapartum and early postpartum mother-to-child transmission of human immunodeficiency virus type 1. *Journal of Infectious Disease*, **187**, 725–35.
Msellati, P., et al. (1999). Zidovudine and reduction of vertical transmission of HIV in Africa. ANRS 049 Trial Group. *American Journal of Public Health*, **89**, 947–8.
Muessig, K.E., et al. (2015). A systematic review of recent smartphone, Internet and Web 2.0 interventions to address the HIV continuum of care. *Current HIV/AIDS Reports*, **12**, 173–90.
National Institute for Health and Care Excellence (NICE) (2014). *Needle and Syringe Programmes*. *Public health guideline [PH52]*. [Online] Available at: https://www.nice.org.uk/guidance/ph52
National Institute of Allergy and Infectious Diseases (1994). *Clinical Alert: Important Therapeutic Information on the Benefit of Zidovudine for the Prevention of the Transmission of HIV from Mother to Infant*. In. Bethesda, MD: National Institutes of Health, National Institute of Allergy and Infectious Diseases.
Nerrienet, E., et al. (2005). Simian immunodeficiency virus infection in wild-caught chimpanzees from cameroon. *Journal of Virology*, **79**, 1312–19.
Obadia, Y., et al. (1999). Syringe vending machines for injection drug users: an experiment in Marseille, France. *American Journal of Public Health*, **89**, 1852–4.
Obel, N., et al. (2007). Ischemic heart disease in HIV-infected and HIV-uninfected individuals: a population-based cohort study. *Clinical Infectious Diseases*, **44**, 1625–31.
Palmateer, N., et al. (2010). Evidence for the effectiveness of sterile injecting equipment provision in preventing hepatitis C and human immunodeficiency virus transmission among injecting drug users: a review of reviews. *Addiction*, **105**, 844–59.
Pathai, S., et al. (2014). Is HIV a model of accelerated or accentuated aging? *Journals of Gerontology: Series A Biological Sciences & Medical Science*, **69**, 833–42.
Pepin, M.E., et al. (2018). Antiretroviral therapy potentiates high-fat diet induced obesity and glucose intolerance. *Molecular Metabolism*, **12**, 48–61.
Petra Study Team (2002). Efficacy of three short-course regimens of zidovudine and lamivudine in preventing early and late transmission of HIV-1 from mother to child in Tanzania, South Africa, and Uganda (Petra study): a randomised, double-blind, placebo-controlled trial. *The Lancet*, **359**, 1178–86.
Plantier, J.C., et al. (2009). A new human immunodeficiency virus derived from gorillas. *Nature Medicine*, **15**, 871–2.

Popovic, M., et al. (1984). Detection, isolation, and continuous production of cytopathic retroviruses (HTLV-III) from patients with AIDS and pre-AIDS. *Science*, **224**, 497–500.
Quinn, T.C., et al. (2000). Viral load and heterosexual transmission of human immunodeficiency virus type 1. *New England Journal of Medicine*, **342**, 921–9.
Rasmussen, L.D., et al. (2012). Risk of diabetes mellitus in persons with and without HIV: a Danish nationwide population-based cohort study. *PLoS One*, 7, e44575.
Rasmussen, L.D., et al. (2015). Time trends for risk of severe age-related diseases in individuals with and without HIV infection in Denmark: a nationwide population-based cohort study. *Lancet HIV*, 2, e288–98.
Ratner, L., Gallo, R.C., and Wong-Staal, F. (1985). HTLV-III, LAV, ARV are variants of same AIDS virus. *Nature*, **313**, 636–7.
Riley, E.D., et al. (2000). Comparing new participants of a mobile versus a pharmacy-based needle exchange program. *Journal of Acquired Immune Deficiency Syndrome*, **24**, 57–61.s
Roberts, B.D. (1994). HIV antibody testing methods: 1985–1988. *Journal of Insurance Medicine*, **26**, 13–14.
Santiago, M.L., et al. (2005). Simian immunodeficiency virus infection in free-ranging sooty mangabeys (Cercocebus atys atys) from the Tai Forest, Cote d'Ivoire: implications for the origin of epidemic human immunodeficiency virus type 2. *Journal of Virology*, **79**, 12515–27.
Serwadda, D., et al. (1985). Slim disease: a new disease in Uganda and its association with HTLV-III infection. *The Lancet*, **2**, 849–52.
Shaffer, N., et al. (1999). Short-course zidovudine for perinatal HIV-1 transmission in Bangkok, Thailand: a randomised controlled trial. Bangkok Collaborative Perinatal HIV Transmission Study Group. *The Lancet*, **353**, 773–80.
Shahbaz, S., et al. (2015). Cardiovascular disease in human immunodeficiency virus infected patients: a true or perceived risk? *World Journal of Cardiology*, **7**, 633–44.
Sharp, P.M. and Hahn, B.H. (2008). AIDS: prehistory of HIV-1. *Nature*, **455**, 605–6.
Sharp, P.M. and Hahn, B.H. (2011). Origins of HIV and the AIDS pandemic. *Cold Spring Harbor Perspectives in Medicine*, 1, a006841.
Sico, J.J., et al. (2015). HIV status and the risk of ischemic stroke among men. *Neurology*, **84**, 1933–40.
Siegfried, N., et al. (2009). Male circumcision for prevention of heterosexual acquisition of HIV in men. *Cochrane Database of Systematic Reviews*, (2), CD003362.
Smit, M., et al. (2015). Future challenges for clinical care of an ageing population infected with HIV: a modelling study. *Lancet Infectious Diseases*, **15**, 810–18.
Smith, J.H. and Whiteside, A. (2010). The history of AIDS exceptionalism. *Journal of the International AIDS Society*, **13**, 47.
Sullivan, P.S., Grey, J.A., and Simon Rosser, B.R. (2013). Emerging technologies for HIV prevention for MSM: what we have learned, and ways forward. *Journal of Acquired Immune Deficiency Syndrome*, 63 Suppl 1, S102–7.
Taha, T.E., et al. (2004). Nevirapine and zidovudine at birth to reduce perinatal transmission of HIV in an African setting: a randomized controlled trial. *Journal of the American Medical Association*, **292**, 202–9.
Tebit, D.M. and Arts, E.J. (2011). Tracking a century of global expansion and evolution of HIV to drive understanding and to combat disease. *Lancet Infectious Diseases*, **11**, 45–56.
Thigpen, M.C., et al. (2012). Antiretroviral preexposure prophylaxis for heterosexual HIV transmission in Botswana. *New England Journal of Medicine*, **367**, 423–34.
UNAIDS (2013). *How Africa Turned AIDS Around*. In: African Union Summit: Joint United Nations Programme for HIV/AIDS.
UNAIDS (2014). *The Gap Report*. Geneva: Joint United Nations Programme for HIV/AIDS.
UNAIDS (2016a). *On the Fast-Track to an AIDS-free Generation*. Geneva: Joint United Nations Programme for HIV/AIDS.

UNAIDS (2016b). *Prevention Gap Report*. Geneva: Joint United Nations Programme for HIV/AIDS.
United Nations Office on Drugs and Crime (UNODC) (2016). *World Drug Report*. [Online] Available at: https://www.unodc.org/wdr2016/
Van Damme, L., et al. (2012a). Preexposure Prophylaxis for HIV Infection among African Women. *New England Journal of Medicine*, **367**, 411–22.
Van Damme, L., et al. (2012b). The FEM-PrEP trial of emtricitabine/tenofovir disoproxil fumarate (truvada) among African women [Abstract 32LB]. Seattle, WA: Conference of Retroviruses and Opportunistic Infections.
Villandré, L., et al. (2019). Assessing the role of transmission chains in the spread of HIV-1 among men who have sex with men in Quebec, Canada. *PLoS One*, 14, e0213366.
Vorma, H., et al. (2013). Participation in opioid substitution treatment reduces the rate of criminal convictions: evidence from a community study. *Addictive Behaviors*, **38**, 2313–16.
Wainberg, M.A. and Friedland, G. (1998). Public health implications of antiretroviral therapy and HIV drug resistance. *Journal of the America Medical Association*, **279**, 1977–83.
Wald, A. and Link, K. (2002). Risk of human immunodeficiency virus infection in herpes simplex virus type 2-seropositive persons: a meta-analysis. *Journal of Infectious Diseases*, **185**, 45–52.
Weber, R., et al. (2009). Uptake of and virological response to antiretroviral therapy among HIV-infected former and current injecting drug users and persons in an opiate substitution treatment programme: the Swiss HIV Cohort Study. *HIV Medicine*, **10**, 407–16.
Wertheim, J.O. and Worobey, M. (2009). Dating the age of the SIV lineages that gave rise to HIV-1 and HIV-2. *PLoS Computational Biology*, 5, e1000377.
Wiktor, S.Z., et al. (1999). Short-course oral zidovudine for prevention of mother-to-child transmission of HIV-1 in Abidjan, Cote d'Ivoire: a randomised trial. *The Lancet*, **353**, 781–5.
Williams, B. and Passmore, J.-A. (2016). Role of vaginal microbiota in genital inflammation and enhancing HIV acquisition in women. Durban: International AIDS Conference 2016.
Wing, E.J. (2016). HIV and aging. *International Journal of Infectious Diseases*, **53**, 61–8.
World Health Organization (WHO) (2014). *Global Guidance on Criteria and Processes for Validation: Elimination of Mother-to-Child Transmission of HIV and Syphilis*. Geneva: WHO.
World Health Organization (WHO) (2011). *Global Monitoring Framework and Strategy for the Global Plan Towards the Elimination of new HIV Infections Among Children by 2015 and Keeping their Mothers Alive*. Geneva: WHO.
World Health Organization (WHO) (2007). *Guide to Starting and Managing Needle and Syringe Programmes*. Geneva: WHO.
World Health Organization (WHO) (1987). *Summary Statement on Breast-Feeding, Breast Milk and HIV*. Geneva: WHO.
World Health Organization (WHO) (2016). *WHO Validates Elimination of Mother-To-Child Transmission of HIV and Syphilis in Thailand, Armenia, Belarus and the Republic of Moldova*. Geneva: WHO.
Ziegler, J.B., et al. (1985). Postnatal transmission of AIDS-associated retrovirus from mother to infant. *The Lancet*, **1**, 896–8.

8.14

Tuberculosis

Roxana Rustomjee

Introduction

Tuberculosis (TB) is the greatest of human scourges, leaving no population in the world unaffected and wreaking havoc since time immemorial. It has exacted a death toll of over a billion human lives since the 1900s, more than any other disease, despite being a treatable and preventable condition (Paulson 2013). This global health catastrophe represents the biggest failure in public health delivery, together with one of the slowest efforts in discovery of new drugs and a paucity in the fundamental understanding of various disease processes. As a result, TB research has suffered from the lack of novel interventions, with a single new diagnostic introduced in 1998 and only two entirely new drugs since the 1950s. Reaching elimination targets, as set out by the World Health Organization (WHO) End TB Strategy therefore presents as a formidable challenge (WHO End TB Strategy http://www.who.int/tb/strategy/en/).

Over the past two decades, addressing the impact of the HIV epidemic on TB disease burden; together with a redoubled effort at TB control; including the implementation of new diagnostics and treatments, has led to a reduction in the global burden of disease (Maher et al. 2007). Nevertheless, the road ahead is not for the faint-hearted. Solutions for ending TB are complex and rooted as much in closing the knowledge gap in understanding pathogenesis as they are in the rapid discovery and deployment of diagnostics drugs and vaccines. In addition, remedying and reorienting the health systems and addressing the economic, social, and behavioural drivers of disease are imperative. It is time to dispel dogma and to take bold and unchartered measures in effecting change, starting with appropriately allocating resources and rapidly translating research findings into policy and practice.

TB is an ‘indicator disease’ of global poverty and inequality (Fox et al. 2017). Prior to the discovery of antibiotics, the ‘Papworth Experiment’ (1918 to 1943) found that improving the social context through improvements in employment, nutrition, and housing reduced TB incidence in children. In contrast the TB incidence rate rapidly, increased post World Wars when populations faced widespread poverty (Fox et al. 2017). The WHO’s ‘End TB Strategy’, which aims to end the global TB epidemic by 2035 while reducing TB deaths by 95% and incidence rates by 90% between 2015 and 2030, emphasizes the role of poverty alleviation, human rights and equity within an effective biomedical and public health response (WHO End TB Strategy http://www.who.int/tb/strategy/en/; Lonnroth et al. 2010; Lonnroth and Raviglione 2016; Fox et al. 2017). This chapter while focussing on biomedical intervention does not minimize the relative contribution of improved socioeconomic status and behavioural interventions in bringing TB to an end.

Burden of TB and global trends

The WHO reported 6.3 million newly notified cases of TB worldwide in 2016 and estimated that TB incidence and mortality appear to be on the decline by about 3% and 2% per year, respectively (WHO 2017). TB incidence is falling at about 2% per year and 16% of TB cases die from the disease. These estimates fall short of the projected 4–5% per year and 10%, respectively, required to reach the 2020 milestones of the End TB Strategy. While the fastest decline in TB incidence has been in the WHO European Region (4.6% from 2015 to 2016); notably, the decline since 2010 has also exceeded 4% per year in several high-TB burden countries, including Ethiopia, Kenya, Lesotho, Namibia, the Russian Federation, Tanzania, Zambia, and Zimbabwe. Twenty-two high-burden, low- to middle-income countries, collectively account for 80% of the global TB burden. The annual number of incident TB cases was under 10 per 100 000 population in most high-income countries; to 150–300 in most of the high-TB burden countries; and above 500 in a few countries including the Democratic People’s Republic of Korea, Lesotho, Mozambique, the Philippines, and South Africa. Alarmingly, almost 39% of cases remain unreported bringing the estimated incidence of TB in the world to approximately 10.4 million; of which 90% were adults and 10% were people living with HIV (PLHIV). In

southern Africa, HIV stands out as the key driver of the TB epidemic as it creates a large pool of susceptible individuals who are either vulnerable to infection or progression from latent infection to active disease (WHO 2017).

Drug-resistant (DR) TB is a continuing threat with 600,000 new cases of resistance to rifampicin (RRTB), the most effective first-line drug, of which 490,000 had multidrug-resistant TB (MDR-TB) and 6.2% of those cases identified as extensively drug-resistant TB (XDR-TB) (WHO 2017). Almost half (47%) of DR cases were in India, China, and the Russian Federation, the majority of whom (72%) remain untreated. A total of 129 689 individuals are initiated on treatment for drug-resistant TB, representing only 22%, with treatment success remaining at a 54% global low. Under-reporting and linkages to care remain a key challenge, particularly for countries in Asia and Africa (WHO 2017).

TB infection

A further 1.7 billion people (close to a quarter of the world's population) are estimated to be infected with *Mycobacterium tuberculosis*, representing a huge reservoir of potential TB disease (Houben and Dodd 2016). **ORCID** http://orcid.org/0000-0001-5825-9347 This substantial number of infected cases highlights ongoing transmission of TB in households and communities and underscores an urgent need for transmission blocking strategies. Progression to active TB occurs in 5–15% of individuals (http://www.who.int/mediacentre/factsheets/fs104/en/). The precise mechanism of disease progression and the complex host-pathogen interactions that are associated with this remain unknown. Traditionally, we have distinguished 'latent infection' from 'active TB' and responded distinctly to these two compartmentalized states both in individual treatment and public health responses. We prefer to describe the disease process as a spectrum of five different states ranging from immunity to TB disease to lack of immune control and active disease (Barry et al. 2009). Following exposure to *M. tuberculosis*, infection may either be eliminated by the innate immune response (without the need for T-cell priming) or eliminated following development of an acquired immune response. Among those who are unable to prevent or eliminate infection following exposure, the majority establish and maintain immune control. Such individuals typically have evidence of T-cell priming and maintain low bacillary numbers that are

difficult to detect or quantify using current diagnosis. A subsequent shift in the host-pathogen balance may permit active mycobacterial replication, leading to subclinical active infection (or incipient TB). The relationship between ‘latent’ and active infection is likely to be a dynamic one over time with bidirectional shifts between the two. Loss of immune control and escalating mycobacterial load, may subsequently lead to the development of symptoms and overt clinical disease as we outline in Figure 8.14.1. Thus, what has been referred to as latent TB infection may represent a dynamic balance between the organisms in a range of metabolic states and variations in immune control. Understanding ‘latent TB infection’ as a spectrum of immune responses, mycobacterial metabolic activity, and organism load redefines approaches to screening, diagnosis, and prevention of TB. An appreciation of these complexities directs the goals for the development of new vaccines and biomarker discovery and emphasizes the effect of HIV on TB reactivation. Also important is the pivotal role of antiretrovirals in regaining immune control of active infection in HIV-infected people. HIV infection is the most potent risk factor for the development of active TB disease. Recurrent TB, due to exogenous re-exposure to *M. tuberculosis* in high-TB prevalence settings is also very likely to play a key role, further increasing bacillary numbers and enhancing the likelihood of progression to disease. Here we consider implications of latent infection for disease control strategies because any strategy aimed at TB elimination needs to address this reservoir of potential infection and disease. Central to this is increasing the fundamental understanding of host factors, including host genetics as well as where bacteria are located during the early stages of TB infection, and whether different locations and environments affect bacterial behaviour and treatment response. Another important question that arises is how similar is the bacterial metabolism along the spectrum between infection and disease? If there are differences, how do we define vulnerable drug targets for latent TB infection? Answering these questions is limited by the inability to meaningfully sample bacteria from latently infected individuals. These knowledge gaps should be addressed simultaneously, as interventions to prevent TB progression evolve.

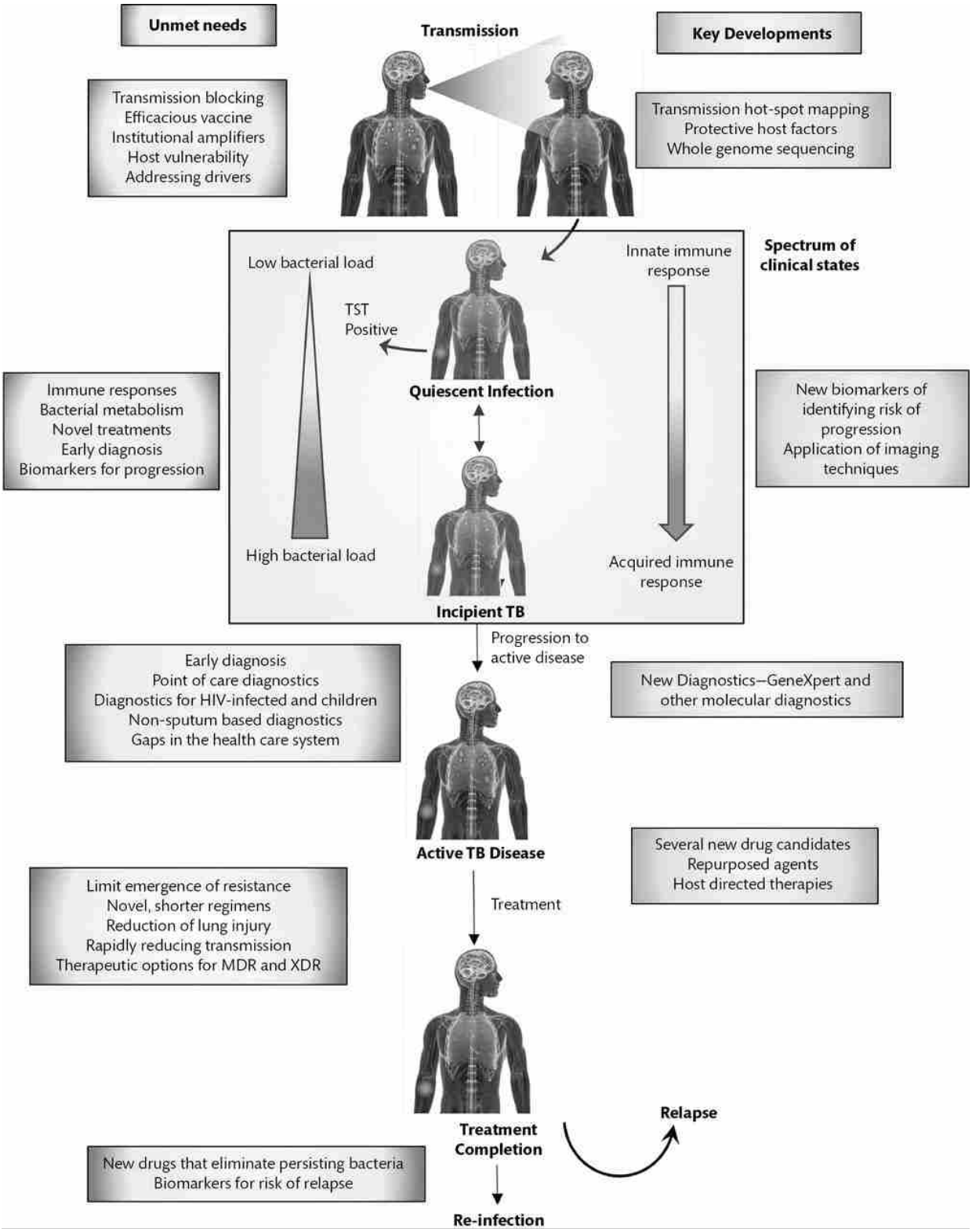


Fig. 8.14.1 Unmet needs and key developments for TB eradication. Shown are the unmet needs and key developments for specific stages of disease. TB transmission is a complex process and blocking transmission is critical to eliminating TB. After infection, a spectrum of infected states is expected to

prevail. These vary in bacterial load and immune responses with increasing bacterial loads reflecting a progression to active disease. Early identification of those individuals that will progress to active disease, using biomarkers, will allow for targeting preventive treatment to the most vulnerable populations. Diagnosis of active TB and rapid engagement into care requires novel diagnostics and addressing gaps in the healthcare system. New TB drugs that shorten duration of therapy and achieve non-relapsing cure are urgently needed.

Interventions to prevent TB progression

Isoniazid preventive therapy (IPT) has been used successfully since the 1950s as part of a comprehensive control strategy to halt transmission at the population level and reduces the risk of active TB by 60%, at the individual level. The Thibela TB trial (2006–2011), where South African mining workforces were randomized to mass IPT or standard of care (mine health services provided comprehensive free medical care. This included tuberculosis diagnostic and treatment services, HIV testing, and antiretroviral therapy), was aimed at interrupting TB transmission by mass screening linked to treatment of either active disease or latent infection (Churchyard et al. 2014). However, the intervention did not reduce TB incidence or prevalence, most likely due to multiple factors including uptake and adherence. Several studies evaluating prevention of drug-susceptible TB infection are underway and use novel drugs or drug combinations (Tiberi et al. 2018). These innovative therapeutic approaches propose to either decrease treatment duration or improve on drug delivery mechanisms. A recent trial (A5279 study; NCT01404312) demonstrated that a 1-month course of weekly isoniazid/rifapentine (INH/RPT), was found to be safe and effective when compared with the usual 9-month course of isoniazid alone in people with HIV infection (Swindells et al. 2018). These results offer HIV-positive individuals who are at risk of developing active TB an additional, shorter-duration prevention option that is safe, effective, and more convenient. Results from other trials simplifying preventive therapy are expected in 2018 including a three-arm comparison between 3-month HP (H-isoniazid, P-pyrazinamide), 9-month H and intermittent 3HP (the WHIP3TB study; NCT02980016) among PLHIV; as well as, a 4 months of rifampicin (4R) vs. 9 months of isoniazid (9H) in adults and children with TB infection (NCT00931736). Further, supervised weekly 3HP is being compared to isoniazid among pregnant and post-partum women (P2001;

NCT02651259). The results of these studies are long awaited and if successful, will greatly simplify preventive therapeutic options.

Nevertheless, the main limitation of widespread TB prophylaxis remains the scalability to entire populations where more than 80% of the adult population is likely to be infected with only a small proportion (<10%) are at risk for progression to active disease. Recent advances in the discovery of correlates of risk, identify inflammatory signals in blood cells from healthy persons that reveal the process of progression from infection to active TB enabling a targeted 'test and treat' approach to TB infection. The CORTIS (Correlate of Risk Targeted Intervention Study) is evaluating the effectiveness of 3HP for high-risk individuals, with the treatment decision based upon a 16-gene, blood-based transcriptional (RNA) biomarker (NCT02735590) (Zak et al. 2016). This approach enables the prospective evaluation of treatment to be targeted to those at elevated risk of developing active TB. If successful, it may result in a point-of-care tool that would have potential to majorly impact the global burden of TB.

With the high prevalence of strains resistant to first-line anti-TB drugs, where isoniazid and rifampin derivatives are of limited utility, different approaches will be needed especially in the absence of international prophylaxis guidelines. If an index case has MDR or XDR-TB, testing of household contacts has revealed significant numbers already infected with latent or active TB infection. Not all household contacts are infected by the same strain as the index case however, significant numbers (10–80% depending on the study) may have been infected by a discordant (sometimes sensitive) strain in the community (Moore 2016). Several drugs or combinations of drugs including fluoroquinolones, ethambutol, pyrazinamide, and ethionamide have been assessed in small trials but none can, at present, be recommended. Most are toxic and even 6 months of treatment cannot be completed by the majority of patients (Moore 2016). For drug-resistant infection, three randomized controlled trials are underway. Six months of daily levofloxacin is being compared to placebo for individuals with presumptive multidrug-resistant TB in the VQUIN MDR Trial (ACTRN12616000215426), in adults, and the TB CHAMP study (ISRCTN92634082), in children under 5 years). In the PHOENIX MDR-TB Trial (https://www.actgnetwork.org/newsletter/oct2015/phoenix), six months of delamanid plus isoniazid will be compared with 6 months of isoniazid alone for patients with MDR- or extensively drug-resistant TB. Results are

expected around 2021. In planning interventions to interrupt transmission through preventive therapy, the epidemiological and health system context must be considered. Population-level impact may not follow unless other interventions (including active case finding linked to effective treatment) are deployed.

The assessment of complete sterilization of TB infection is a far greater challenge because unlike in active disease, where the bacterial burden reduces measurably from high detectable levels to little or no bacteria, the levels of bacteria and antigen present in latent TB infection are very low and discerning clearance is more challenging. Identifying a signature of persistence versus bacterial clearance could aid in identifying those individuals at highest risk for progressing to TB and direct them for TB infection treatment.

Interventions to halt the transmission of TB

Ultimately, it is not simply enough to decrease TB incidence by preventing cases through prophylaxis targeted at close contacts, it is also necessary to understand the strategy and type of interventions most likely to avert the largest proportion of transmission events at the population level. There should be a specific focus on those events that result in secondary cases of infectious TB. As an example, in settings where transmission originates largely from specific and identifiable populations (e.g. young adult men) (Dodd et al. 2016), it may be feasible to target case-finding efforts in those groups. Where reactivation is common in high-risk groups, e.g. HIV-positive individuals (Pretorius et al. 2014), people with risk factors such as diabetes or malnutrition (Odone et al. 2014), elderly populations (Prince et al. 2015), it may be critical to target preventive therapy to those populations. In congregate settings (e.g. prisons or healthcare institutions), infection control and modification of the environment can have major impact in reducing transmission. The key issues include addressing drivers of transmission, measuring transmission, and identifying novel transmission blocking strategies (Shah et al. 2017). The variable success of preventive therapy interventions, alongside a growing number of studies demonstrating that close contacts only account for a minority of transmission events (Verver et al. 2004; Mathema et al. 2006; Chamie et al. 2015; Glynn et al. 2015; Shah et al. 2017) indicate that our current tools and approach for interrupting

transmission are inadequate. Additional gaps in our understanding of the biology and local determinants of transmission underscore the need for more effort in this area.

Drivers of TB transmission

Recent evidence from endemic countries, with varying levels of HIV and drug-resistant TB, implicate ongoing transmission as the driving force maintaining high levels of TB incidence (Li et al. 2007; Becerra et al. 2011; Shah et al. 2017). Drivers of transmission can be seen as a cascade in three stages: (1) sharing of air between infectious and susceptible individuals; (2) infectiousness of an index case; and (3) susceptibility to disease of other exposed individuals (Dowdy et al. 2014). These stages are heterogeneous at regional and national levels and the relative role of each stage is heavily dictated by local attributes including access to healthcare, population age structure, population density and migration (Wood et al. 2010; Liu et al. 2012). Variations in individual host susceptibility, in line with the local prevalence of HIV, diabetes mellitus, and silicosis, can also dramatically alter the transmission cascade (Murray et al. 1996; Hanifa et al. 2009). The strength of local TB control programmes in terms of case finding and treatment strategies is yet another factor that can affect local transmission dynamics. A larger proportion of undiagnosed or untreated patients will contribute to ongoing transmission in a community. This heterogeneity of transmission poses a substantial challenge to TB control. To design efficacious and efficient TB control strategies, it will be necessary to delineate the relative contribution of the various drivers of TB in different settings. Ultimately, drivers of transmission, including, but not limited to infectiousness of index cases, susceptibility of contacts as well poverty, overcrowding, health seeking behaviour, and institutional drivers must be studied and understood at a local level.

Measuring levels of TB transmission

The prolonged natural history of TB means that observed cases of disease reflect a combination of recent transmission and recurrent disease, the latter made up of reactivation and exogenous reinfection. In addition, most cases observed via passive case finding systems (self-referral of patients to

healthcare providers), underestimate the number of undiagnosed, but diagnosable, infectious individuals in the community. Current methods to measure transmission include examining case notification rates and trends in tuberculin skin test (TST) or interferon-gamma release assays (IGRA). However, real-time, definitive diagnosis of recent exposure to *M. tuberculosis* at the individual level is not possible with any of the currently available tests (e.g. TST, IGRA, chest X-ray, sputum analysis). A biomarker of recent exposure as described, but not limited to, the CORTIS trial, would enable precisely targeted preventive therapy and accelerate identification of TB 'hotspots'—areas of high-TB incidence—for earlier and more effective implementation of infection control measures (Wood et al. 2014). Such hotspot detection and elimination represent an immediate measure to interrupt transmission and reduce the global burden of TB while new drugs, diagnostics, and vaccines are progressing through the product development pipeline. These hotspot communities could also serve as priority areas for piloting new transmission interventions, as they become available (Lee et al. 2015).

More recently, molecular epidemiology utilizing whole-genome sequencing has greatly facilitated studies of patient- and population-level transmission (Verver et al. 2004; Mathema et al. 2006; Glynn et al. 2015) Several studies have aimed to estimate the proportion of transmission that occurs within households by comparing genotypes of cases occurring within households or community settings. These studies have found that in medium- and high-TB burden settings, the majority of TB transmission occurs outside of households or with known close contacts (Chamie et al. 2015; Guerra-Assuncao et al. 2015). As such, the yield of contact investigations as a TB control activity may be lower than traditionally anticipated in such settings. Congregate locations that facilitate air exchange between infectious cases with non-infected individuals, such as transport hubs, churches, school classrooms, and prisons have been implicated as 'hotspots' or institutional amplifiers of transmission and may offer opportunities for targeted interventions (Dowdy et al. 2012; Zelner et al. 2016). However, it is unknown whether other, less obvious sites of congregation, may also be driving ongoing transmission. An integrated approach that combines traditional epidemiology with methods drawn from spatial, demography, and whole-genome analysis will undoubtedly be required to identify such drivers (Davies et al. 2015; Vynnycky et al. 2015).

Halting TB transmission need not be limited to biomedical interventions. Historical data from western Europe, where TB incidence declined 10% annually after the Second World War, demonstrates the impact of socioeconomic development on reducing TB transmission. A broader approach regarding how socioeconomic development may interrupt transmission is desperately need together with systematic testing of interventions for impact, including engaging other government sectors equipped with the power and knowledge needed to accelerate economic development for TB-affected communities.

In addition to the devastating impact of TB on families and communities, ongoing nosocomial TB transmission in clinics and hospitals represents a life-threatening risk to the global healthcare workforce. TB has become an occupational lung disease for providers from high-burden countries, including frontline community health workers, nurses, medical students, and physicians (Naidoo and Jinabhai 2006; Barnagarwala 2014). TB as a personal risk associated with a professional choice is unacceptable in the twenty-first century and must be addressed by national and international professional associations, lest their workers succumb to TB.

Ultimately, both established and new interventions will be required to end TB. Given the variable and possibly non-durable success of preventive therapy, particularly in settings where transmission and the force of infection is very high (Comstock et al. 1967; Hanson et al. 1967; Comstock et al. 1979; Ayles et al. 2013; Churchyard et al. 2014); it is conceivable that any new interventions, including a preventive vaccine, which may protect against a single exposure or a low inoculum; may be ineffective against repeated, high-dose exposures. TB control programmes should therefore reduce ongoing transmission using all available methods. Synergy in reducing TB transmission may be attainable by combining interventions that shrink the reservoir of latent TB infection (preventive therapy), shorten time to treatment (case finding and diagnosis), and prevent transmission in key settings.

Ultimately, there is cause for optimism. There are effective interventions available to target each component of TB transmission (and the analytical tools to demonstrate their effectiveness). Furthermore, the lack of an animal reservoir facilitates focus on human populations, behaviours, and environments. The increasing commitment from researchers, funders, and

policymakers to address transmission of TB is supported by a focus on developing innovative approaches to meet the end TB targets.

New approaches to TB diagnosis

Rapid, early, and accurate diagnosis is critical to timeous anti-TB treatment initiation, monitoring, and ultimately TB control. However, 39% of the estimated 10.4 million estimated cases were either undiagnosed or unreported (WHO 2017). Diagnosis of extrapulmonary TB and TB in children, PLHIV, and pregnant women remains particularly problematic. Furthermore, only 41% of TB cases notified globally were tested for rifampicin resistance and just 22% of the estimated drug-resistant TB incidence were initiated on appropriate treatment (WHO 2017). Diagnostic tools should ideally: (1) identify more TB with highly accurate point-of-care tests and replace smear microscopy; (2) include a simple, low-cost point-of-care triage test to exclude TB; (3) enable universal, rapid DST that informs regimen selection (Denkinger 2017); (4) offer point-of-care (POC) alternatives to sputum-based technologies with capability to detect all forms of TB especially paediatric (Kik et al. 2014). Nevertheless, addressing all these critical gaps in current diagnostic tools, without identifying and treating individuals with TB infection who may later progress to active TB disease (Pai et al. 2016) will greatly hamper progress in reaching TB elimination targets. Therefore, we add a further stipulation (5) to identify and differentiate TB across the spectrum of TB infection.

Sophisticated bioinformatic computational tools and systems biology approaches are being applied to the discovery and validation of biomarkers, resulting in substantial progress in the understanding of T-cell responses and T-cell function, protein, and gene expression studies. Alternative diagnostic strategies under investigation as potential screening and triaging tools include non-sputum-based detection with breath-based tests and automated digital radiography. Reviewed herein are current TB diagnostics for disease and infection their application and performance, newer technologies in the pipeline and identify weaknesses and areas of progress that need to be addressed if we are to enhance TB control programmes and transition efforts towards TB elimination.

Microscopy, culture, and nucleic acid amplification tests (NAAT) are currently the three validated diagnostic methods for the detection of active

TB while chest radiography is sometimes used for screening of active TB (Pai et al. 2016). Major drawbacks associated with smear and culture are well described and will not be discussed here except to note the substantially reduced capability to diagnose TB in a large proportion of people living with HIV has enormous public health implications further contributing to disease transmission, morbidity and mortality (Dorman 2010). Because culturing tubercle bacilli is a pre-requisite for detailed phenotypic DST and genotyping (Pai et al. 2016), the protracted time interval required to detect bacterial resistance, technical expertise needed, biosafety requirements and high cost per test are some of the limitations of this method of diagnosis (Dorman 2010).

NAATs are molecular assays that detect *M. tuberculosis* through the amplification of particular regions of DNA (Denkinger and Pai 2017). NAATs are both highly sensitivity and specific, detecting about 100% of smear positive pulmonary TB while differentiating between *M. tuberculosis* and non-tuberculosis mycobacteria (Flores et al. 2005). Sensitivity is however considerably reduced in smear-negative patients, extrapulmonary specimens, and in HIV prevalence settings and as a result, NAAT tests can therefore not be utilized as a rule out test (Denkinger and Pai 2017). There are several NAAT tests on the market and others in development (see https://www.finddx.org/tb/).

Of the NAATs, GeneXpert MTB/RIF (Cepheid, Sunnyvale, CA, USA) allows for rapid and simultaneous detection of *M. tuberculosis* complex DNA and rifampin resistance in just 2 hours directly from clinical specimens and exhibits both high sensitivity and specificity and is the most widely used. GeneXpert MTB/RIF is considerably more sensitive at identifying *M. tuberculosis*, requiring 131 colony forming units per millilitre of sputum in contrast to 10 000 bacilli per millilitre of sputum required for smear microscopy (Steingart et al. 2014). In late 2010, the WHO endorsed the GeneXpert MTB/RIF assay for use in TB endemic settings (WHO 2014). Despite high sensitivity overall, GeneXpert sensitivity is only 60–80% in smear-negative TB. Sensitivity is also reduced in children (66%) and PLHIV (79%) and is variable in extrapulmonary specimens (Detjen et al. 2015; Pai et al. 2016). Despite its revolutionary approach, GeneXpert MTB/RIF was not designed to penetrate all levels of the healthcare system nor can it meet all the diagnostic demands for TB elimination (Pai et al. 2016). The cost of GeneXpert MTB/RIF coupled with the inability to detect latent TB and

resistance to multiple drugs are just some of its notable shortcomings. Moreover, the inability to distinguish between live and dead bacilli, sensitivity to high temperatures and need for a continuous power supply are some of the important limitations of this diagnostic tool that hampers widespread roll-out (Pai et al. 2016).

A new molecular test, GeneXpert MTB/RIF Ultra, exhibits improved sensitivity to GeneXpert MTB/RIF, requiring just 16 bacilli per millilitre of sputum, and similar sensitivity to liquid culture. In March 2017 the WHO recommended its use as a replacement for the current GeneXpert MTB/RIF cartridge (WHO 2017). The Ultra assay should significantly increase TB detection in smear-negative patients and provide more reliable rifampin resistance detection. In addition, GeneXpert Omni, a portable, and POC device is expected to become available in 2018 (Pai et al. 2016). Unlike the other GeneXpert platforms, the Omni is battery operated and can withstand significant temperature variations, enabling it to be used as point of care in clinics for the rapid diagnosis of drug-sensitive and drug-resistant TB, allowing patients to be promptly initiated on effective treatment. This has the potential to improve patient outcomes and, ultimately, reduce TB transmission. The rapid uptake of such new tools in the highest burden countries, where health systems are often under strain, is essential. Ongoing, multicountry evaluations of the accuracy and feasibility are underway (https://www.finddx.org/news/find-ultra-study/).

Line probe assays (LPAs) like that of the GenoType MTBDR (Hain Lifescience, Germany) allow for rapid differentiation of mycobacteria and simultaneous detection of drug-resistant mutations from culture using polymerase chain reaction and reverse hybridization methods (Denkinger and Pai 2017). The MTBDR*sl* test allows for rapid detection of mutations associated with resistance to fluoroquinolone, aminoglycosides, and cycloserine, drugs used to treat MDR-TB. LPAs have demonstrated excellent sensitivity and specificity for rifampin resistance but variable results for isoniazid resistance (Ling et al. 2008; Bwanga et al. 2009). LPAs are more rapid than conventional phenotypic DST methods and offer a lower biosafety risk but require skilled laboratory personnel, good quality control and rely on culture methods thereby limiting its role in resource-poor settings (Pai et al. 2016; Denkinger and Pai 2017). The GenoType MTBDR is endorsed by WHO and with rapid scale-up in high-burden countries through the EXPAND-TB project (Denkinger and Pai 2017). Given the high

proportion of drug-resistant TB attributable to ongoing transmission (Kendall et al. 2015) and the substantial delays often experienced in diagnosing and treating drug-resistant TB, the impact of improved case finding, and diagnosis should be even greater. Broader implementation of a rapid molecular test—even only for passive case detection—has been projected to avert substantial morbidity and mortality due to rifampin-resistant TB in settings like India (Salje et al. 2014).

The massive reservoir of latently infected adults generates an enormous pool for future reactivation cases. In fact, it is estimated that even if prevention of all new TB infections was achieved, the incidence of TB, arising from the 1.7 billion individuals in the global reservoir, would be 16.6 per 100,000 person-years in 2035, making the 2050 target to eradicate TB unattainable. The TST, and interferon-γ release assays (IGRAs) developed in the early 1900s and 2005, respectively, are the two diagnostic tools currently in use for latent TB infection. Both tests evaluate the presence of *M. tuberculosis*-specific T-cell responses to recent sensitization and are therefore unable to distinguish between latent, active, or past TB, or the progression thereof, relapse after treatment or protection after vaccination (except for newer generation IGRAs) (Denkinger and Pai 2017). There is therefore a critical need for the early identification of TB cases with the use of novel non-sputum-based diagnostics. A blood-based triage test, as described in the CORTIS trial, that enables targeted investigation of individuals with active and subclinical TB disease, including asymptomatic individuals with the highest risk of progression from latent infection to active disease status, has the potential to reduce the time to treatment initiation as well as prevent incipient TB disease prior to the onset of symptoms. Recent studies performed in latent TB infection subjects showed that *M. tuberculosis*-specific CD4 T-cells have a characteristic chemokine expression signature with increased numbers of these cells in latently infected subjects compared to healthy participants suggesting a role for subsets of CD4 T-cells in containing TB (Arlehamn et al. 2014). These signatures may have important implications in the identification of individuals with the greatest need for preventive therapy and to determine if targeted short-course preventive therapy can prevent TB disease among individuals found to exhibit the RNA signature (Penn-Nicholson et al. 2016).

While more rapid diagnosis should intuitively result in more rapid treatment initiation—and hence reduction in disease transmission—the

current paradigm of passive case finding constrains the potential transmission impact of improved diagnostic tools (Dowdy et al. 2017). TB patients are generally symptomatic for weeks to months prior to presentation to healthcare providers (Dheda et al. 2016), and they may have detectable bacilli in their sputum for many additional months before developing symptoms (Dowdy et al. 2013). A randomized controlled trial, while shortening the time to diagnosis, did not demonstrate a significant treatment benefit of GeneXpert MTB/RIF over smear microscopy (Theron et al. 2014). By contrast, active case finding (ACF) can detect infectious cases much earlier and hence potentially interrupt disease transmission. A recent randomized trial (XACT I) evaluated the feasibility and impact of GeneXpert MTB/RIF for ACF in South Africa and Zimbabwe, using mobile vans with on-board GeneXpert capacity and staffed by three healthcare workers each (Calligaro et al. 2017). Relative to sputum smear microscopy, GeneXpert MTB/RIF was associated with a 53% increase in the number of patients initiating anti-TB therapy. This trial demonstrates that ACF using newer molecular tools is feasible and substantially increases the proportion of patients initiating treatment in high-burden settings. Importantly, as demonstrated in another cluster-randomized trial in South Africa (Churchyard et al. 2015), scale-up of GeneXpert MTB/RIF is unlikely to impact mortality—or transmission—unless also accompanied by better linkage to care. Multiple ongoing studies are evaluating diagnostic tests of higher sensitivity for example, GeneXpert Ultra and/or greater portability in the context of ACF (https://www.finddx.org/news/find-ultra-study/). Targeted ACF interventions are needed.

In summary, early data suggest that newer molecular diagnostic tools—if combined with ACF—may help reduce TB transmission in high-burden settings. Additional needs include trials to evaluate the effectiveness of more sensitive and/or scalable diagnostic tests (especially in the context of ACF), studies of comparative cost-effectiveness and feasibility, new tools to identify patients with highest transmission risk, and conclusive studies demonstrating that earlier diagnosis and treatment initiation can reduce TB burden at the population level. Despite these promising developments, there remain several unanswered questions regarding interventions to halt TB transmission through diagnosis and case finding. These include cost-effectiveness, feasibility, and impact in different settings, optimal combination of screening tools, and characterization of TB transmission

from immunological, clinical, and healthcare-seeking behaviour perspectives. Ultimately, new diagnostic tools capable of detecting TB at earlier stages will likely only have transformative impact on transmission when used in the context of broader screening and ACF (Dowdy et al. 2017).

TB diagnostics pipeline

Resistance mutations are not fully characterized for all TB drugs, and they are particularly deficient for the second-line and newer drugs used to treat MDR and XDR-TB. Next-generation whole-genome sequencing (NGS) remains the most attractive tool for TB drug resistance testing, with the potential to generate antibiotic susceptibility profiles for all drugs in a single analysis (Dookie et al. 2018) and might eventually become the reference standard for drug resistance identification. Reduced costs and the establishment of high-throughput sequencing centres and easy-to-use analytical tools have greatly increased access to NGS, which is being implemented as a routine service in several countries. However, in most high-burden countries it remains only a means for research with the analysis undertaken overseas. Proof of principle for NGS directly from sputum has been shown, but for comprehensive analysis the need to first isolate and culture the organism to obtain sufficient bacterial DNA hinders its application to patient management. Recent studies have shown some success in this regard (Votintseva et al. 2015; Votintseva et al. 2017; Dookie et al. 2018).

The TB diagnostic pipeline is at its most robust and focused on novel technologies equipped with a better understanding of the variable host immune response and non-sputum-based options (Walzl et al. 2018, https://www.finddx.org/news/find-launches-diagnostic-pipeline-tracker/). Novel approaches being investigated include testing breath for TB metabolites and the application of nanotechnology and microfluidics to increase detection capacity and reduce hands-on sample manipulation (Walzl et al. 2018). Promising results have been reported on the use of new, more sensitive technology to detect lipoarabinomannan and circulating antigen peptides. New platforms capable of assessing molecules of variable size and composition or that detect multiple proteins are under development to support the measurement of host-response biomarkers. Competitors for the GeneXpert MTB/RIF assay are undergoing assessment for regulatory

approval and are closest to launch (FIND). These competitors are devices designed for low-technology microscopy centres or high-throughput instruments with extended drug resistance detection intended for the reference laboratory. Most tests are sputum-based but some aim to use blood or urine, which might detect extrapulmonary TB and TB in children. Ending the epidemic will demand both innovative technologies and an enabling environment. It is hoped that by 2020 the following measures will be implemented: (1) a triage test and highly sensitive detection at the first point of contact; (2) TB confirmation with rapid integrated DST for critical drugs to drive treatment regimen decisions at a dedicated unit, and; (3) comprehensive DST at a reference level compatible with health system strengthening (Denkinger 2017).

The need for new drugs against tuberculosis

Currently treatment for TB is long and toxic, ranging from a 4-month standard regimen for patients who are sputum smear and culture negative with minimal pulmonary disease in the USA, to the WHO standard 6 month short-course treatment for drug-susceptible TB. Short-course chemotherapy for drug-susceptible TB comprises an initial 2-month intensive phase (isoniazid, rifampin, pyrazinamide, and ethambutol daily), continued by isoniazid and rifampin for the last 4 months. In the case of drug-resistant TB, 18–24 months treatment with second-line, less potent and more toxic drugs is required, involving an intensive phase of 6–9 months with aminoglycosides (amikacin/kanamycin) and cycloserine, terizidone, ethionamide, protionamide, capreomycin, aminosalicylic acid, and fluoroquinolones. Treatment success is approximately 85% for drug-sensitive TB, 50% for MDR-TB, and 30% with XDR-TB. To improve on these outcomes, an ideal new regimen should be safe, rapidly bactericidal and possess potent sterilizing activity against subpopulations of persisting bacteria that are drug tolerant in the face of continued chemotherapy. The global vision is to introduce an ultra-short, simple, and affordable TB regimen that works in virtually all people with TB, including children. This requires a multidrug regimen comprised entirely of new drugs. Such a regimen would unify the treatment of all current forms of TB under one cure. There is therefore a shift in focus from single drug development to

regimen building. Drug combinations are needed to eradicate various bacterial subpopulations and prevent the development of resistance.

To develop new drug candidates in order to have a sufficient number and combination of drugs for a new regimen, a multipronged approach has been adopted, which entails the discovery of new antitubercular agents, together with repurposing of previously approved antimicrobials and the development of host-directed therapies. An integrated approach with high-throughput screening of chemical libraries has been used to identify new lead compounds, followed by deconvolution of the drug target, assessing the effects on metabolism and identifying possible mechanisms of resistance (Evans and Mizrahi 2018). This approach, in combination with others, has allowed for repopulation of the TB drug development pipeline, with 14 chemical compounds having progressed along the clinical development pathway, including new chemical classes in preclinical (eight compounds) and clinical (nine compounds) phases (Tiberi et al. 2018). Novel mechanisms of action with activity against both replicating and non-replicating bacteria, reduced drug-drug interactions particularly with P450 enzymes; oral bioavailability with acceptable pharmacokinetic and pharmacodynamic safety and tolerability profiles are desirable attributes of new TB drug candidates. However, many more candidate drugs are needed to populate a pipeline robust enough to build entirely new regimens. An ideal drug combination should consist of at least three drugs that are active against MDR and XDR tuberculosis, and have potent, synergistic, and complementary activities against various subpopulations of *M. tuberculosis*. Such a combination should be equally effective against drug-susceptible and drug-resistant tuberculosis and produce a stable cure in a much shorter period than does the standard treatment. New antituberculosis drugs that do not interact with protease inhibitor-based treatments are also needed for the effective treatment of the HIV co-infected population. For this to be successful innovative trials need identify suitable drug candidates, to be inclusive of children and supported by sufficiently resourced trial capacities and regulatory guidance.

Lead optimization and clinical evaluation

Bedaquiline, the first novel TB drug developed in five decades, with a novel mode of action by binding to the F_oF_1 ATP synthase, which catalyses the

production of ATP using the proton motive force generated by the mycobacterial electron transport chain (Andries et al. 2005). Bedaquiline is highly potent against drug-susceptible and drug-resistant strains of *M tuberculosis* and is approved for treatment of drug-resistant TB. For drug-sensitive TB, bedaquiline will be combined with rifampin, delamanid, and linezolid in five different combinations in the TRUNCATE-TB28 (NCT03474198), Phase 2c open-label trial (Diacon et al. 2009; Belard et al. 2015; Tiberi et al. 2018). Bedaquiline also features in Next-5001, STREAM, and end-TB trials. The success of bedaquiline also highlighted the mycobacterial electron transport chain as a rich source of possible new drug targets, as evidenced by the developed of other similar agents such as Q203 (Bald et al. 2017; Iqbal et al. 2018).

Delaminid and pretomanid are nitroimidazoles that undergo conversion by a F_{420}-dependent nitroreductase to form metabolites that inhibit the synthesis of mycolic acids in the mycobacterial cell wall (Xavier and Lakshmanan 2014). Both these agents are currently in Phase 3 trials for drug-resistant TB and to assess the potential for treatment shortening, for example, through combination with bedaquiline (Tiberi et al. 2018). SQ109, a 1,2-ethylene small molecule that is structurally related to ethambutol, inhibits the incorporation of mycolic acids into the *M. tuberculosis* cell wall (Protopopova et al. 2005; Tahlan et al. 2012; Boeree et al. 2017). Encouraging data from the early bactericidal activity assessment confirm safety, tolerability, and no adverse cardiac effects. In the presence of an optimized background regimen, SQ109 has demonstrated promise in the treatment of drug-resistant TB (Tiberi et al. 2018). Benzothiazinones represent another interesting class of compounds that target the mycobacterial cell wall, in this particular case through inhibition of arabinogalactan biosynthesis (Makarov et al. 2009). One molecule, PBTZ169, a piperazinobenzothiazinone derivative of BTZ043, which has a relatively easy chemical synthesis and promising pharmacodynamics, has been advanced to Phase 2 clinical trials.

Oxazolidinones represent a new class of orally bioavailable inhibitors of protein synthesis that act by binding to the 23s subunit of the ribosome and interrupting protein synthesis., linezolid, is recommended by the WHO to treat drug-resistant tuberculosis as a medicine with ‘unclear efficacy’. Two randomized controlled trials in patients with XDR-TB demonstrated improved rates of sputum culture conversion but linezolid was poorly

tolerated which has limited its widespread application (Lee et al. 2012; Maartens and Benson 2015; Tang et al. 2015). Sutezolid follows in this class, with favourable pharmacokinetic properties (Rausch et al. 2018). However, advancing sutezolid to Phase 2 clinical trials (Wallis et al. 2014) is pending review after recent safety concerns arose from a 6-Month Toxicity Study of sutezolid in rats. Delpazolid (LCB01-0371), a new oxazolidinone with a cyclic amidrazone moiety, and contezolid (also known as MRX) are currently being advanced in the clinical development pipeline.

GSK070, a new candidate drug with an entirely novel model of action, inhibits mycobacterial Leucyl-tRNA synthetase (LeuRS), an enzyme required for protein synthesis. The LeuRS system allows for aminoacylation of tRNAs, together with a proof reading site to remove incorrectly charged amino acids and inhibition of this process affects translation (need ref). GSK070 is currently entering Phase 1 clinical trials.

Regimen building

Various combinations of new agents, together with existing drugs are being tested to develop short-course regimens with low toxicity. Based on promising results from a Phase 2b trial, the NC-005 regimen, BPaMZ (bedaquiline + pretomanid + moxifloxacin + pyrazinamide) will be assessed in the Phase 3 trial (NCT02193776) to provide a more effective, shorter regimen for both drug-susceptible and drug-resistant TB. The Nix-TB trial (NCT02333799) will assess a shorter, orally administered treatment for XDR-TB with pretomanid, bedaquiline, and linezolid. This combination promises to provide a new alternative for drug-resistant TB. Similarly, the ZeNix trial (NCT03086486) will also assess the potential for improved safety with a shorter, oral regimen consisting of pretomanid, bedaquiline, and lower dose linezolid (BPaL).

Clofazimine (CFZ). CFZ, a riminophenazine, has been historically used for the treatment of lepromatous leprosy (Sharma et al. 2017). The drug has several advantages including intracellular accumulation within macrophages, a low prevalence of resistance, anti-inflammatory properties and contributes substantially to the sterilizing activity of TB regimens in mice and to the relapse-free cure of MDR-TB regimens in humans. However optimal dosing, pharmacokinetics/pharmacodynamics, safety and tolerability rates, and treatment duration have not been definitively tested. Reversible but

potentially stigmatizing skin discoloration along with known QT prolongation has tempered enthusiasm. Nevertheless, CFZ has demonstrated significant *M. tuberculosis* killing activity and treatment-shortening potential in murine models. This will be investigated in a 3- and 4-month regimen of standard TB treatment (rifampin/isoniazid/pyrazinamide/ethambutol) with the addition of CFZ (ACTG trial A5362) early efficacy (12-week culture conversion) and treatment shortening in dug-susceptible TB (Treatment Action Group 2017).

Host-directed therapies

Host-directed therapies (HDT) are novel approaches in the field of anti-infectives for overcoming antimicrobial resistance and restoring cellular pathways that are perturbed by the causative pathogen, *M. tuberculosis* (Wallis and Hafner 2015; Yang 2017; Kaufmann et al. 2018). HDT for TB, aims to (i) modulate host cell factors that are required by a pathogen for replication or persistence; (ii) enhance protective immune responses against a pathogen; (iii) reduce exacerbated inflammation; and (iv) balance immune reactivity at sites of pathology. In TB, HDT aims to enhance the antimicrobial activities of phagocytes and to curtail inflammation through modulation of secreted factors (such as eicosanoids and cytokines) or cell membrane proteins (such as co-stimulatory molecules) and modulate granulomas to allow the access of antimicrobials or to restrict tissue damage (Kaufmann et al. 2018). By directly targeting the host, rather than the pathogen, HDTs may be less susceptible to the development of resistance to these therapies. Most HDT approaches are not considered to be stand-alone therapies but are combined with conventional therapy. Several HDTs have been directed at each stage of the macrophage life cycle of *M. tuberculosis* to overcome resistance to microbial killing by: (i) Triggering activation: IFN and TNF modulators and phosphodiesterase (PDE) inhibitors; (ii) inducing antimicrobial peptides: (vitamin d; histone deacetylase [hdac] inhibitors; and phenylbutyrate); (iii) inhibiting lipid formation: (statins and vitamin D); (iv) triggering autophagy (gefitinib, vitamin D, mTOR inhibitors, metformin, statins, verapamil, fluoxetine, nitazoxanide, carbamazepine, valproic acid); and (v) promoting phagosome maturation (metformin, kinase inhibitors, and statins) (Parihar et al. 2014; Singhal et al. 2014; Pauwels et al. 2017). Thus far few have progressed to clinical evaluation with successful results.

Notably the clinical evaluation of metformin has not yielded promising results (Dutta et al. 2017). Results from a large Phase 2 trial designed to examine the safety and preliminary efficacy of multiple adjunctive host-directed TB therapies CC-11050 (PDE-4 inhibitor), auranofin, everolimus (mTOR inhibitor), and vitamin D, to assess their potential to shorten TB treatment and/or prevent permanent lung damage are expected in 2019 (https://clinicaltrials.gov/ct2/show/NCT02968927). Preclinical evaluation of anti-IL-1 blockade; autophagy targets; statins and Imatinib have concluded, and the statin and Imatinib projects are now proceeding to proof-of-concept clinical studies for *M. tuberculosis* treatment in the context of HIV co-infection (Mayer-Barber et al. 2010; Napier et al. 2011; Dutta et al. 2016; Gupta et al. 2016).

While progress is slow and difficult, HDT will undoubtedly feature in the future of TB therapy. Further refinement in the identification of potential genetic predispositions in patients, discovery and validation of host or pathogen biomarkers that may indicate disease stage; treatment responses; type of inflammatory response; will potentially permit targeted HDT. Significant barriers include the potential for some of these agents to be cytotoxic or induce a worsened inflammatory response; lack of supportive preclinical and clinical data in assessing potential efficacy and selecting suitable candidate; and not least the potentially higher cost of adding these agents to already expensive treatments for MDR and XDR-TB.

Vaccines for TB

Vaccination is the most effective intervention for the prevention, control, and/or eradication of an infectious disease. Mathematical modelling estimates that without new interventions, particularly those that blocks transmission such as protective vaccines, it will not be possible to reduce the incidence of TB to less than 10 per 100,000 by 2035. The growing burden of extensively resistant and untreatable TB as well a compelling argument for the return on investment emphasizes the need, and incentive for the development of an effective vaccine to prevent TB. This section highlights key challenges and recent advances in TB vaccine research. The current discovery, preclinical, and clinical pipeline have been described in detail elsewhere. Recent advances offering a broader diverse repertoire of TB vaccine candidates together with an enthusiasm to leap-frog TB vaccine

development and include broadening both antigen selection and the range of vaccine-mediated immune responses as well as experimental design in both in animal models and smaller human clinical trials offer an unprecedented opportunity to expand demonstration of vaccine efficacy for the prevention and possibly the treatment of TB.

Based on animal and human studies of TB, a robust cellular immune response is required for protection against Mtb infection and disease (Evans et al. 2016) whereas, most successful, licensed vaccines available today induce neutralizing antibodies that provide protective immunity. For this reason, the majority of current clinical TB vaccine candidates are based on a variety of vectors, adjuvants and antigens that induce CD4 or CD8 T-cell responses with TH1 cytokines such as IFN-γ or TNF-α. These clinical candidates are based on a variety of vaccine approaches, such as inactivated whole cell or whole cell extracts (*Mycobacterium indicus pranii*, *Mycobacterium vaccae*, DAR-901 (*Mycobacterium obuense*), RUTI and *Mycobacterium smegmatis*), viral-vectored candidates (vaccinia-based MVA85A, influenza, and human adenovirus 5 and 35 and chimpanzee adenovirus), fusion protein subunits with TH1-inducing adjuvants (M72/AS01, Hybrid 1/CAF01, Hybrid 1/IC31, H4/IC31, H56/IC31 and ID93/GLA-SE), and live recombinant BCG or attenuated TB vaccines (VPM 1002, Aeras 422, rBCG30, and MTBVAC). DNA vaccines are being developed in different countries, notably emerging economies, but have not yet entered into human clinical trials. To date, clinical trials characterizing these candidate vaccines have included studies of safety and immunogenicity in diverse populations, including healthy, TB-naïve adults, and infants, latently infected adolescents, and adults, HIV-infected adults, as well as patients undergoing drug treatment for TB (Evans et al. 2016).

Vaccination against TB can be divided into one of three categories: (1) preventive pre-exposure vaccines (or priming vaccines) typically administered to neonates for prevention of infection (POI); (2) preventive post-exposure vaccines (or boost vaccines) that can be administered to adolescents and adults with latent TB infection or those that have previous vaccination and (3) therapeutic vaccines which are administered together with TB drugs to shorten and ameliorate the course of therapy and prevent recurrent disease. Different vaccines may be required to induce immune responses in diverse populations, such as infants, young adults, those already

latently infected with *M. tuberculosis* or on treatment for TB, and those co-infected with HIV.

The first candidate to be tested for efficacy, MVA85A (a modified vaccinia Ankara virus vector expressing the antigen 85A), failed to show protection (Tameris et al. 2013), other candidates remain in the clinical pipeline, with efficacy results expected (2018) from a Phase 2b prevention of disease trial with M72/AS01, the GSK adjuvanted fusion protein vaccine in 3,600 HIV-uninfected, adults latently infected with *M tuberculosis*.

Iterative approaches focused on prevention of disease by viral-vectored vaccine candidates are yet to yield promising results. Innovation in trial design, vaccine approaches, constructs and routes of administration aimed at demonstrating biologic activity of vaccine candidates using more focused populations specifically selected to reduce sample size, are providing encouraging new evidence and pathways for vaccine developers. In a prevention-of-infection Phase 2 trial conducted in South Africa, revaccination with BCG significantly reduced sustained TB infections in high-risk adolescents. An experimental vaccine candidate, H4:IC31, a fusion protein of the immunodominant antigens TB10.4 and Ag85B, formulated with IC31 adjuvant, also reduced sustained infections, although not at statistically significant levels. However, the trend observed for H4:IC31 is the first time a subunit vaccine has shown any indication of ability to protect against TB infection or disease in humans. These results indicating that sustained QFT-GIT conversions are preventable by BCG revaccination, provide novel insights into possible mechanisms of protection in humans. Results from a concluded Phase 3 trial investigating the efficacy of killed *M. vaccae* lysate vaccine to prevent TB in adults with latent disease from Guanxi province China are expected while a ‘Prevention of Recurrence’ trial with the vaccine candidate ID93 is designed to determine whether the vaccine can prevent the 4–6% relapse and/or reinfection rate typically observed following successful treatment of active TB. All these experimental designs follow a non-traditional clinical development path, but attempt increase efficiencies by reducing sample size requirements of demonstrating biological signals earlier than conventionally phased product development plan.

Recently a novel vaccine approach using a weakened form of a common herpes virus, cytomegalovirus (CMV) as a vector expressing *M. tuberculosis* Ag85A demonstrated complete protection in 41% and reduced overall TB

disease by 68% in vaccinated rhesus macaques. Cytomegaloviruses persist lifelong *in vivo* and induce powerful adaptive and innate immune and increasing ('inflationary') responses (Beverley et al. 2014).

The BCG control group showed no evidence of protection. It is critical to rapidly understand the mechanisms by which this re-engineered CMV creates and maintains a high state of immunity against the pathogen in vaccinated monkeys especially because this approach has the potential to work better than standard vaccines. The results demonstrate that a small Ag-specific response may be highly effective in synergy with the correct innate activation, in this case involving IL-21 and natural killer (NK) cells and hold promise for future development of this candidate vaccine. Human trials are expected by 2019.

The development of appropriate preclinical models to prioritize the best vaccine candidate for advancement in human studies represents a key unmet need. Many TB vaccine animal challenge models have a narrow dynamic range of responses that do not allow for easy differentiation among candidates. This limitation is being addressed by refinement of the mouse, guinea pig, and macaque models to better approximate natural infection by *M tuberculosis* and better mimic human disease. The use of low dose challenge, sophisticated PET/CT imaging techniques, and novel vaccine candidates (such as H56, *Cytomegalovirus* [CMV] approaches, aerosolized adenovirus vaccines and IV administration of BCG) have recently shown that the macaque model may potentially be useful to delineate correlates of vaccine-induced protection.

The lack of known correlates of protection and surrogate biomarkers of immune responses associated with different stages of TB infection and disease has been a barrier against the clinical evaluation of the vaccine candidates (Abubakar et al. 2016). We have discussed the implications of gene expression patterns of inflammatory biosignatures correlated with risk for TB disease progression, in both active and latent TB cases (Evans et al. 2016). Similarly, a broad unbiased systems biology approach in which gene expression signatures are compared in samples from various time points (Evans et al. 2016) and then correlated to either specific measures of immunogenicity, or to protection in efficacy studies are likely to yield potential correlates of vaccine-induce protection.

Given no validated immunological correlates of protection, significant limitations of animal models and the strained capacity to fund and perform

efficacy trials, a TB human challenge model would facility vaccine development by offering an option to investigate candidate selection, route of vaccination, dose optimization, and the identification of immune correlates despite the need to overcome significant safety, measurement/reporter, ethical, and regulatory challenges. If successful potential TB vaccines are tested in healthy donors challenged with attenuated *M. tuberculosis* strains that are viable, capable of replicating but do not cause infection or disease or be transmittable. Experiments in non-human primates are scheduled to start in mid-2019. A safer option, using BCG challenge has been pursued, while vaccines that target antigens unique to M tuberculosis would not be suitable for the BCG challenge model promising results have been demonstrated.

Exponential advances in TB vaccine development have occurred over the past 5 years. Results from efficacy trials will provide insights into what is needed from a vaccine to deliver protective immunity and new tools will accelerate TB vaccine development, not only by providing a less expensive and rapid method for assessing potential vaccine efficacy, but also by permitting more rapid progress towards the identification of an immune correlate of protection. While the challenges in developing TB vaccines are great, the research community is embracing this challenge.

Closing the Gap—addressing global threats

Financing for TB care and prevention has been increasing for more than 10 years, but a crippling funding gap still exists, estimated at US$ 2.3billion in 2017. Total health spending also falls short of the resources needed to achieve universal health coverage. TB kills more people than any other bacterial agent and when this fact is contrasted with the reducing investments in TB research, it becomes clear that the governments of endemic countries, together with those of donor nations, have yet to commit appropriate resources to realize the End TB Strategy. Closing these gaps requires more resources from both domestic sources (especially in middle-income countries) and international donors (especially in low-income countries). Given this complex funding landscape, TB is particularly susceptible to sudden changes in the global political landscape. The UK referendum to leave the European Union has created a risk that the sixth largest contributor to TB research will most likely adopt a trade-focused

approach and withdraw support from global health efforts. Similar concerns prevail in the USA where the low incidence of TB makes it an issue of international development and foreign aid, rather than a concern for public health. TB also ranks low in terms of the research and development agenda globally, especially in the context of emerging infectious diseases such as Ebola, which grab international attention and resources. In this context, political will from endemic countries become mission critical to nucleate international support and must therefore feature centrally in efforts to eliminate TB.

Profits from TB research and development are perceived to be minimal, owing largely to the fact that the disease is mostly prevalent in the poorest settings, which are unable to provide a financially lucrative market to drive drug discovery and development. TB research and development n is being currently driven by collaborative efforts involving public–private partnerships. Although the total funding for TB research and development increased from US $ 358.1 million in 2005 to 620.6 million in 2015, this has been insufficient to meet costs of discovery. Since 2009, approximately 2.8 billion US$ has been spent on TB research by public institutions, with the National Institutes of Health (NIH) contributing more than half this amount. Other sources of funding for TB research include Japanese pharmaceutical company Otsuka contributing 60% of industry spending, the Gates Foundation contributing 89% of philanthropic funding, and Unitaid contributing 47% of total multilateral funding. It is noteworthy that the largest multilateral supporter of TB research and development to date has been the Global Fund to Fight AIDS, Tuberculosis, and Malaria, contributing US$ 120.2 million from 2002 to 2015 on TB operational research There have been substantial advances in HIV/AIDS within three-four decades, from huge investments by governments and donor agencies. Relative this this, investments in TB research and development has remained vastly inadequate and should be prioritized.

Conclusion

TB has defined the fabric of modern medicine in the developing world and continues to claim the lives of the economically and socially disadvantaged. To address this, the WHO has set bold goals in the End TB Strategy and to achieve these, a higher level of annual decline in incidence is required.

While a bold approach is needed to eradicate TB, it is questionable whether achieving these milestones places an unreasonable burden of performance on the crippled healthcare systems of endemic countries, which may require a slower, but steady, rate of sustainable decline. This can already be achieved through more effective deployment of current tools, together with safe-guarding the fidelity of available antibiotics using medical practices that limit the emergence of resistance. That said, new tools for diagnosis and treatment of TB are urgently required, in addition to an integrated approach for bolstering the healthcare system. Identifying and addressing gaps in the cascade of care for TB is critical and central to this is strong political will, which is required to create resources, build infrastructure, and attract much needed funding. Another challenge that needs to be addressed is the implementation of TB control programmes in the context of limited laboratory capacity and treatment barriers, which threaten to undermine any gains that the introduction of new interventions may bring. Hence, strengthening healthcare delivery and linkage to care is pivotal for TB eradication. As with many other infectious diseases, a cross-disciplinary approach to the development of health interventions promises success and, in this regard, it is critical to support research and development efforts across the value chain to advance new drugs and vaccines and to lay the foundation for future elimination efforts. Furthermore, considering comorbidities such as HIV infection and diabetes is required to ensure future control programmes consider those most at risk for progressing to active disease. While the treatment of drug-sensitive disease is affordable for most endemic countries, treating drug-resistant TB is prohibitive. This creates a research and development environment which calls on all stakeholders, including academic institutions, governments, and big pharma to shift cost paradigms and invest in the development of novel, cheaper regimens. For success in these endeavours, coordination between donors is essential to leverage the best value for money in terms of prioritizing research questions and for attracting further funding to the field. Finally, the world can no longer afford to ignore the growing menace of antimicrobial resistance. Considering that more than half of the recorded cases of bacterial resistance involve drug-resistant *M. tuberculosis*, the inability to eliminate TB threatens to plunge humanity back into the pre-antibiotic era. As a society, we simply cannot let this happen.

REFERENCES

Abubakar, I., Lipman, M., McHugh, T.D., and Fletcher H. (2016). Uniting to end the TB epidemic: advances in disease control from prevention to better diagnosis and treatment. *BMC Medicine*, **14**, 47.

Andries, K., et al. (2005). A diarylquinoline drug active on the ATP synthase of Mycobacterium tuberculosis. *Science*, 307(5707), 223–7.

Arlehamn, C.L., et al. (2014). Transcriptional profile of tuberculosis antigen-specific T cells reveals novel multifunctional features. *Journal of Immunology*, 193(6), 2931–40.

Ayles, H., et al. (2013). Effect of household and community interventions on the burden of tuberculosis in southern Africa: the ZAMSTAR community-randomised trial. *The Lancet*, 382(9899), 1183–94.

Bald, D., Villellas, C., Lu, P., and Koul, A. (2017). Targeting energy metabolism in mycobacterium tuberculosis, a new paradigm in antimycobacterial drug discovery. *mBio*, 8(2), e00272–17.

Barnagarwala, T. (2014). TB hospital staff live under shadow of dreaded disease. *The Indian Express*. Uttar Pradesh, India: IE Online Media Services. [Online] Available at: https://indianexpress.com/article/cities/mumbai/tb-hospital-staff-live-under-shadow-of-dreaded-disease/

Barry, C.E., et al. (2009). The spectrum of latent tuberculosis: rethinking the biology and intervention strategies. *Nature Reviews Microbiology*, 7(12), 845–55.

Becerra, M.C., et al. (2011). Tuberculosis burden in households of patients with multidrug-resistant and extensively drug-resistant tuberculosis: a retrospective cohort study. *The Lancet*, 377(9760), 147–52.

Belard, S., Heuvelings, C.C., Janssen, S., and Grobusch, M.P. (2015). Bedaquiline for the treatment of drug-resistant tuberculosis. *Expert Review of Anti-Infective Therapy*, 13(5), 535–53.

Beverley, P.C., et al. (2014). A novel murine cytomegalovirus vaccine vector protects against *Mycobacterium tuberculosis*. *Journal of Immunology*, 193(5), 2306–16.

Boeree, M.J., et al. (2017). High-dose rifampicin, moxifloxacin, and SQ109 for treating tuberculosis: a multi-arm, multi-stage randomised controlled trial. *Lancet Infectious Diseases*, 17(1), 39–49.

Bwanga, F., Hoffner, S., Haile, M., and Joloba, M.L. (2009). Direct susceptibility testing for multi drug resistant tuberculosis: a meta-analysis. *BMC Infectious Diseases*, **9**, 67.

Calligaro, G.L., et al. (2017). Effect of new tuberculosis diagnostic technologies on community-based intensified case finding: a multicentre randomised controlled trial. *Lancet Infectious Diseases*, 17(4), 441–50.

Chamie, G., et al. (2015). Identifying locations of recent TB transmission in rural Uganda: a multidisciplinary approach. *Tropical Medicine & International Health*, 20(4), 537–45.

Churchyard, G.J., Fielding, K.L., and Grant, A.D. (2014). A trial of mass isoniazid preventive therapy for tuberculosis control. *New England Journal of Medicine*, 370(17), 1662–3.

Churchyard, G.J., et al. (2014). A trial of mass isoniazid preventive therapy for tuberculosis control. *New England Journal of Medicine*, 370(4), 301–10.

Churchyard, G. J., et al. (2015). Xpert MTB/RIF versus sputum microscopy as the initial diagnostic test for tuberculosis: a cluster-randomised trial embedded in South African roll-out of Xpert MTB/RIF. *Lancet Global Health*, 3(8), e450–7.

Comstock, G.W., Baum, C., and Snider, D.E. Jr. 1979. Isoniazid prophylaxis among Alaskan Eskimos: a final report of the bethel isoniazid studies. *American Review of Respiratory Disease*, 119(5), 827–30.

Comstock, G.W., Ferebee, S.H., and Hammes, L.M. (1967). A controlled trial of community-wide isoniazid prophylaxis in Alaska. *American Review of Respiratory Disease*, 95(6), 935–43.

Davies, G.R., Phillips, P.P., and Jaki, T. (2015). Adaptive clinical trials in tuberculosis: applications, challenges and solutions. *International Journal of Tuberculosis & Lung Disease*, 19(6), 626–34.

Denkinger, C.M. (2017). *The TB Diagnostic Pipeline*. Guadalajara, Mexico: The 48th Union World Conference on Lung Health.

Denkinger, C.M. and Pai, M. (2017). *Handbook of Global Tuberculosis Control—Practices and Challenges*. New York: Springer.

Detjen, A.K., et al. (2015). Xpert MTB/RIF assay for the diagnosis of pulmonary tuberculosis in children: a systematic review and meta-analysis. *Lancet Respiratory Medicine*, 3(6), 451–61.

Dheda, K., Barry, C.E. 3rd, and Maartens, G. (2016). Tuberculosis. *The Lancet*, 387(10024), 1211–26.

Diacon, A.H., et al. (2009). The diarylquinoline TMC207 for multidrug-resistant tuberculosis. *New England Journal of Medicine*, 360(23), 2397–405.

Dodd, P. J., et al. (2016). Age- and sex-specific social contact patterns and incidence of mycobacterium tuberculosis infection. *American Journal of Epidemiology*, 183(2), 156–66.

Dookie, N., Rambaran, S., Padayatchi, N., Mohomed, S., and Naidoo, K. (2018). Evolution of drug resistance in *Mycobacterium tuberculosis*: a review of the molecular determinants of resistance and implications for personalised care. *Journal of Antimicrobial Chemotherapy*, 73(5), 1138–51.

Dorman, S.E. (2010). New diagnostic tests for tuberculosis: bench, bedside, and beyond. *Clinical Infectious Diseases*, 50 Suppl 3, S173–7.

Dowdy, D.W., Azman, A.S., Kendall, E.A., and Mathema, B. (2014). Transforming the fight against tuberculosis: targeting catalysts of transmission. *Clinical Infectious Diseases*, 59(8), 1123–9.

Dowdy, D.W., Basu, S., and Andrews, J.R. (2013). Is passive diagnosis enough? The impact of subclinical disease on diagnostic strategies for tuberculosis. *American Journal of Respiratory & Critical Care Medicine*, 187(5), 543–51.

Dowdy, D.W., Golub, J.E., Chaisson, R.E., and Saraceni, V. (2012). Heterogeneity in tuberculosis transmission and the role of geographic hotspots in propagating epidemics. *Proceeding of the National Academies of Science U S A*, 109(24), 9557–62.

Dowdy, D.W., Grant, A.D., Dheda, K., Nardell, E., Fielding, K., and Moore, D.A.J. (2017). Designing and Evaluating Interventions to Halt the Transmission of Tuberculosis. *Journal of Infectious Diseases*, 216(suppl_6), S654–61.

Dutta, N.K., et al. (2016). Statin adjunctive therapy shortens the duration of TB treatment in mice. *Journal of Antimicrobial Chemotherapy*, 71(6), 1570–7.

Dutta, N.K., Pinn, M.L., and Karakousis, P.C. (2017). Metformin adjunctive therapy does not improve the sterilizing activity of the first-line antitubercular regimen in mice. *Antimicrobial Agents in Chemotherapy*, 61(8), PMID: 28559262.

Evans, J.C. and Mizrahi, V. (2018). Priming the tuberculosis drug pipeline: new antimycobacterial targets and agents. *Current Opinion in Microbiology*, **45**, 39–46.

Evans, T.G., Schrager, L., and Thole, J. (2016). Status of vaccine research and development of vaccines for tuberculosis. *Vaccine*, 34(26), 2911–14.

Flores, L.L., Pai, M., Colford, J.M., Jr., and Riley, L.W. (2005). In-house nucleic acid amplification tests for the detection of Mycobacterium tuberculosis in sputum specimens: meta-analysis and meta-regression. *BMC Microbiology*, **5**, 55.

Fox, G.J., Dobler, C.C., Marais, B.J., and Denholm, J.T. (2017). Response to: socio-political prescriptions for latent tuberculosis infection are required to prevent reactivation of tuberculosis. *International Journal of Infectious Diseases*, **58**, 117–18.

Glynn, J.R., et al. (2015). Whole genome sequencing shows a low proportion of tuberculosis disease is attributable to known close contacts in rural Malawi. *PLoS One*, 10(7), e0132840.

Guerra-Assuncao, J.A., et al. (2015). Large-scale whole genome sequencing of *M. tuberculosis* provides insights into transmission in a high prevalence area. *eLife*, 4, e05166.

Gupta, A., Misra, A., and Deretic, V. (2016). Targeted pulmonary delivery of inducers of host macrophage autophagy as a potential host-directed chemotherapy of tuberculosis. *Advances in Drug Delivery Reviews*, **102**, 10–20.

Hanifa, Y., Grant, A.D., Lewis, J., Corbett, E.L., Fielding, K., and Churchyard, G. (2009). Prevalence of latent tuberculosis infection among gold miners in South Africa. *International Journal of Tuberculosis & Lung Disease*, 13(1), 39–46.

Hanson, M.L., Comstock, G.W., and Haley, C.E. (1967). Community isoniazid prophylaxis program in an underdeveloped area of Alaska. *Public Health Reports*, 82(12), 1045–56.

Houben, R.M. and Dodd, P.J. (2016). The global burden of latent tuberculosis infection: a re-estimation using mathematical modelling. *PLoS Medicine*, 13(10), e1002152.

Iqbal, I.K., Bajeli, S., Akela, A.K., and Kumar, A. (2018). Bioenergetics of *Mycobacterium*: an emerging landscape for drug discovery. *Pathogens*, 7(1), 24.

Kaufmann, S.H.E., Dorhoi, A., Hotchkiss, R.S., and Bartenschlager, R. (2018). Host-directed therapies for bacterial and viral infections. *National Reviews Drug Discoveries*, 17(1), 35–56.

Kendall, E.A., Fofana, M.O., and Dowdy, D.W. (2015). Burden of transmitted multidrug resistance in epidemics of tuberculosis: a transmission modelling analysis. *Lancet Respiratory Medicine*, 3(12), 963–72.

Kik, S.V., Denkinger, C.M., Casenghi, M., Vadnais, C., and Pai, M. (2014). Tuberculosis diagnostics: which target product profiles should be prioritised? *European Respiratory Journal*, 44(2), 537–40.

Lee, M., et al. (2012). Linezolid for treatment of chronic extensively drug-resistant tuberculosis. *New England Journal of Medicine*, 367(16), 1508–18.

Lee, R.S., et al. (2015). Population genomics of *Mycobacterium tuberculosis* in the Inuit. *Proceedings of the National Academy of Sciences*, 112(44), 13609–14.

Li, X., et al. (2007). Transmission of drug-resistant tuberculosis among treated patients in Shanghai, China. *Journal of Infectious Diseases*, 195(6), 864–9.

Ling, D.I., Zwerling, A.A., and Pai, M. (2008). GenoType MTBDR assays for the diagnosis of multidrug-resistant tuberculosis: a meta-analysis. *European Respiratory Journal*, 32(5), 1165–74.

Liu, L., Wu, J., and Zhao, X.Q. (2012). The impact of migrant workers on the tuberculosis transmission: general models and a case study for China. *Mathematical Biosciences & Engineering*, 9(4), 785–807.

Lonnroth, K., et al. (2010). Tuberculosis control and elimination 2010–50: cure, care, and social development. *The Lancet*, 375(9728), 1814–29.

Lonnroth, K. and Raviglione, M. (2016). The WHO's new End TB Strategy in the post-2015 era of the Sustainable Development Goals. *Transactions of the Royal Society Tropical Medicine & Hygience*, 110(3), 148–50.

Maartens, G. and Benson, C.A. (2015). Linezolid for treating tuberculosis: a delicate balancing act. *EBioMedicine*, 2(11), 1568–9.

Maher, D., et al. (2007). Planning to improve global health: the next decade of tuberculosis control. *Bulletin of the World Health Organization*, 85(5), 341–7.

Makarov, V., et al. (2009). Benzothiazinones kill *Mycobacterium tuberculosis* by blocking arabinan synthesis. *Science*, 324(5928), 801–4.

Mathema, B., Kurepina, N.E., Bifani, P.J., and Kreiswirth, B.N. (2006). Molecular epidemiology of tuberculosis: current insights. *Clinical Microbiology Reviews*, 19(4), 658–85.

Mayer-Barber, K.D., et al. (2010). Caspase-1 independent IL-1beta production is critical for host resistance to Mycobacterium tuberculosis and does not require TLR signaling *in vivo*. *Journal of Immunology*, 184(7), 3326–30.

Moore, D.A. (2016). What can we offer to 3 million MDR-TB household contacts in 2016? *BMC Medicine*, **14**, 64.

Murray, J., Kielkowski, D., and Reid, P. (1996). Occupational disease trends in black South African gold miners. An autopsy-based study. *American Journal of Respiratory & Critical Care Medicine*, 153(2), 706–10.

Naidoo, S. and Jinabhai, C.C. (2006). TB in health care workers in KwaZulu-Natal, South Africa. *International Journal of Tuberculosis & Lung Disease*, 10(6), 676–82.

Napier, R.J., et al. (2011). Imatinib-sensitive tyrosine kinases regulate mycobacterial pathogenesis and represent therapeutic targets against tuberculosis. *Cell Host Microbe*, 10(5), 475–85.

Odone, A., Houben, R.M., White, R.G., and Lonnroth, K. (2014). The effect of diabetes and undernutrition trends on reaching 2035 global tuberculosis targets. *Lancet Diabetes Endocrinology*, 2(9), 754–64.

Pai, M., Nicol, M.P., and Boehme, C.C. (2016). Tuberculosis diagnostics: state of the art and future directions. *Microbiology Spectrum*, 4(5), doi: 10.1128/microbiolspec.TBTB2-0019-2016.

Parihar, S.P., et al. (2014). Statin therapy reduces the *Mycobacterium tuberculosis* burden in human macrophages and in mice by enhancing autophagy and phagosome maturation. *Journal of Infectious Diseases*, 209(5), 754–63.

Paulson, T. (2013). Epidemiology: a mortal foe. *Nature*, 502(7470), S2–3.

Pauwels, A.M., Trost, M., Beyaert, R., and Hoffmann, E. (2017). Patterns, receptors, and signals: regulation of phagosome maturation. *Trends Immunology*, 38(6), 407–22.

Penn-Nicholson, A., Scriba, T.J., Hatherill, M., White, R.G., and Sumner, T. (2016). A novel blood test for tuberculosis prevention and treatment. *South Africa Medical Journal*, 107(1), 4–5.

Pretorius, C., et al. (2014). The potential effects of changing HIV treatment policy on tuberculosis outcomes in South Africa: results from three tuberculosis-HIV transmission models. *AIDS*, 28 Suppl 1, S25–34.

Prince, M.J., et al. (2015). The burden of disease in older people and implications for health policy and practice. *The Lancet*, 385(9967), 549–62.

Protopopova, M., et al. (2005). Identification of a new antitubercular drug candidate, SQ109, from a combinatorial library of 1,2-ethylenediamines. *Journal of Antimicrobial Chemotherapy*, 56(5), 968–74.

Rausch, L., McFarlane, C., Iyer, L., Thai, D., and Mirsalis, J.C. (2018). *6-Month Safety Evaluation of Sutezolid in Sprague Dawley Rats*. Menlo Park, CA: *Society of Toxicology Annual Meeting*, Biosciences Division, SRI International.

Salje, H., Andrews, J.R., Deo, S., Satyanarayana, S., Sun, A.Y., Pai, M., and Dowdy, D.W. (2014). The importance of implementation strategy in scaling up Xpert MTB/RIF for diagnosis of tuberculosis in the Indian health-care system: a transmission model. *PLoS Med*, 11(7), e1001674.

Shah, N.S., et al. (2017). Transmission of extensively drug-resistant tuberculosis in South Africa. *New England Journal of Medicine*, 376(3), 243–53.

Shah, N.S., Kim, P., Kana, B.D., and Rustomjee, R. (2017). Getting to zero new tuberculosis infections: insights from the National Institutes of Health/US Centers for Disease Control and Prevention/Bill & Melinda Gates Foundation workshop on research needs for halting tuberculosis transmission. *Journal of Infectious Diseases*, 216(suppl 6), S627–8.

Sharma, D., Dhuriya, Y.K., Deo, N., and Bisht, D. (2017). Repurposing and revival of the drugs: a new approach to combat the drug resistant tuberculosis. *Frontiers of Microbiology*, **8**, 2452.

Singhal, A., et al. (2014). Metformin as adjunct antituberculosis therapy. *Science Translational Medicine*, 6(263), 263ra159.

Steingart, K.R., Schiller, I., Horne, D.J., Pai, M., Boehme, C.C., and Dendukuri, N. (2014). Xpert(R) MTB/RIF assay for pulmonary tuberculosis and rifampicin resistance in adults. *Cochrane Database of Systematic Reviews*, (1), CD009593.

Swindells, S., et al. (2018). *One Month of Rifapentine/Isoniazid to Prevent TB in People with HIV: BRIEF-TB/A5279. Conference on Retroviruses and Opportunistic Infections*. Boston, MA. Abstract number: 37LB.

Tahlan, K., et al. (2012). SQ109 targets MmpL3, a membrane transporter of trehalose monomycolate involved in mycolic acid donation to the cell wall core of Mycobacterium tuberculosis. *Antimicrobial Agents & Chemotherapy*, 56(4), 1797–809.
Tameris, M.D., et al. (2013). Safety and efficacy of MVA85A, a new tuberculosis vaccine, in infants previously vaccinated with BCG: a randomised, placebo-controlled Phase 2b trial. *The Lancet*, 381(9871), 1021–8.
Tang, S., et al. (2015). Efficacy, safety and tolerability of linezolid for the treatment of XDR-TB: a study in China. *European Respiratory Journal*, 45(1), 161–70.
Theron, G., et al. (2014). Feasibility, accuracy, and clinical effect of point-of-care Xpert MTB/RIF testing for tuberculosis in primary-care settings in Africa: a multicentre, randomised, controlled trial. *The Lancet*, 383(9915), 424–35.
Tiberi, S., et al. (2018). Tuberculosis: progress and advances in development of new drugs, treatment regimens, and host-directed therapies. *Lancet Infectious Diseases*, 18(7), e183–98.
Treatment Action Group (2017). *2017 Pipeline Report*. [Online]. Available at: https://www.treatmentactiongroup.org/resources/pipeline-report/2017-pipeline-report/
Verver, S., et al. (2004). Proportion of tuberculosis transmission that takes place in households in a high-incidence area. *The Lancet*, 363(9404), 212–14.
Votintseva, A.A., et al. (2017). Same-day diagnostic and surveillance data for tuberculosis via whole-genome sequencing of direct respiratory samples. *Journal of Clinical Microbiology*, 55(5), 1285–98.
Votintseva, A.A., et al. (2015). Mycobacterial DNA extraction for whole-genome sequencing from early positive liquid (MGIT) cultures. *Journal of Clinical Microbiology*, 53(4), 1137–43.
Vynnycky, E., et al. (2015). Tuberculosis control in South African gold mines: mathematical modeling of a trial of community-wide isoniazid preventive therapy. *American Journal of Epidemiology*, 181(8), 619–32.
Wallis, R.S., et al. (2014). Mycobactericidal activity of sutezolid (PNU-100480) in sputum (EBA) and blood (WBA) of patients with pulmonary tuberculosis. *PLoS One*, 9(4), e94462.
Wallis, R.S. and Hafner, R. (2015). Advancing host-directed therapy for tuberculosis. *Nature Reviews Immunology*, 15(4), 255–63.
Walzl, G., et al. (2018). Tuberculosis: advances and challenges in development of new diagnostics and biomarkers. *Lancet Infectious Diseases*, 18(7), e199–e210.
Wood, R., et al. (2010). Changing prevalence of tuberculosis infection with increasing age in high-burden townships in South Africa. *International Journal of Tuberculosis and Lung Disease*, 14(4), 406–12.
Wood, R., et al. (2014). Quantification of shared air: a social and environmental determinant of airborne disease transmission. *PLoS One*, 9(9), e106622.
World Health Organization (WHO) (2014). *Global Tuberculosis Report 2014*. Geneva: WHO.
World Health Organization (WHO) (2017). *Global Tuberculosis Report 2017*. Geneva: WHO.
World Health Organization (WHO) (2017). *Next-Generation Xpert® MTB/RIF Ultra Assay Recommended by WHO*. [Online] Available at: https://www.who.int/tb/features_archive/Xpert-Ultra/en/
Xavier, A.S. and Lakshmanan, M. (2014). Delamanid: a new armor in combating drug-resistant tuberculosis. *Journal of Pharmacology & Pharmacotherapy*, 5(3), 222–4.
Yang, C.S. (2017). Advancing host-directed therapy for tuberculosis. *Microbiol Cell*, 4(3), 105–7.
Zak, D.E., et al. (2016). A blood RNA signature for tuberculosis disease risk: a prospective cohort study. *The Lancet*, 387(10035), 2312–22.
Zelner, J.L., et al. (2016). Identifying hotspots of multidrug-resistant tuberculosis transmission using spatial and molecular genetic data. *Journal of Infectious Diseases*, 213(2), 287–94.

8.15

Malaria

Frank Baiden, Keziah L. Malm, and Fred Binka

Introduction

Malaria is a life-threatening infectious tropical and subtropical disease of major global health concern. It is caused by parasites that are transmitted through the bites of infective female Anopheles mosquito. According to the 2018 World Malaria Report released by the World Health Organization, there were 219 million cases of malaria in 2017. These resulted in 435,000 deaths, nearly 61% of which were in children below 5 years of age. The World Health Organization (WHO) Africa Region carries the highest burden of malaria cases (92%) and deaths (93%).

In 2017, five countries accounted for nearly half of all malaria cases. The countries are Nigeria (25%), the Democratic Republic of the Congo (11%), Mozambique (5%), India (4%), and Uganda (4%) (Fig. 8.15.1). Malaria is an important cause of underdevelopment and poverty in these and other countries in malaria-endemic sub-Saharan Africa. It reduces gross domestic product in sub-Saharan Africa by about 1.3% per year. In continued high transmission settings, malaria could be reducing life expectancy at birth by more than six years (Bawah and Binka 2007; WHO 2018a).

Per global population, the number of cases and deaths from malaria has seen steady decline since the 1990s. The period from 2000 to 2015 witnessed year-on-year reduction in the number of cases. Since 2015 however, this trend has stalled and there is a real threat that gains made over the past two decades could be reversed if urgent steps are not taken. Malaria thus remains an important threat to the achievement of the Sustainable Development Goals (SDGs) (Rowe 2017).

Malaria: past to present

The earliest evidence of malaria parasites was found in fossils of mosquitoes that dates to approximately 30 million years old. Hippocrates described the manifestations of malaria, and related them to where patients lived, and time of the year. His description is preceded by references to malaria and its symptoms by Nei Ching (Chinese Canon of Medicine) in 2700 BC and in the Cuneiform tablets of 6th century BC (Ross 1923; Neghina et al. 2010).

The pioneering work of Alphone Laveran in Algeria in 1879 led to vastly improved understanding of the infectious basis of the disease. His discovery of black pigment in the blood of people suffering from malaria ultimately led to the knowledge of plasmodium as the infectious agent (Sundberg 2007; Cox 2010). Experiments conducted by Donald Ross in the late 1800s led to an understanding of the life cycle of the parasite. The use of the microscope to identify the plasmodium parasite and the deployment of interventions to break the chain in the life cycle of the parasite have remained the basis for the diagnosis, treatment, and control of the disease (Dobson 1999; Cox 2017).

Interventions for the prevention and treatment of malaria preceded a full understanding of the plasmodium parasite and its life cycle. The perception of association between malaria and stagnant waters (ultimately to be appreciated as the breeding grounds for mosquitoes) led the Romans in 1930s to undertake drainage programmes (Majori 2012). Derivatives of the *Artemisia annua*, used in early traditional Chinese Medicine, remain an importance part of malaria treatment to date (Krishna et al. 2008). Peruvian Indians used the bitter bark of the Cinchona tree, later processed as quinine, in the treatment of fever. Quinine continues to be an important drug in the treatment of malaria. Between 1955 and 1969, a Global Malaria Eradication Programme (GMEP), resulted in the elimination of malaria from many regions in south Europe and parts of the Americas. There was however very little success from the rather limited efforts in sub-Saharan Africa (Bruce-Chwatt 1974; Krishna et al. 2008; Najera et al. 2011; Cohen et al. 2012; Majori 2012). The launch of Roll Back Malaria (RBM) in 1998 by Dr Gro Harlem Brundtland, then Director-General of the WHO set in motion a momentum of increased investments and support to endemic countries (Nabarro 1999; Remme et al. 2001).

Compared to the 1980s and 90s, malaria transmission has substantially reduced in many countries; to a point there is now a push for country-by-

country elimination and eradication (Mendis et al. 2009; Alonso et al. 2011) (Fig. 8.15.1).

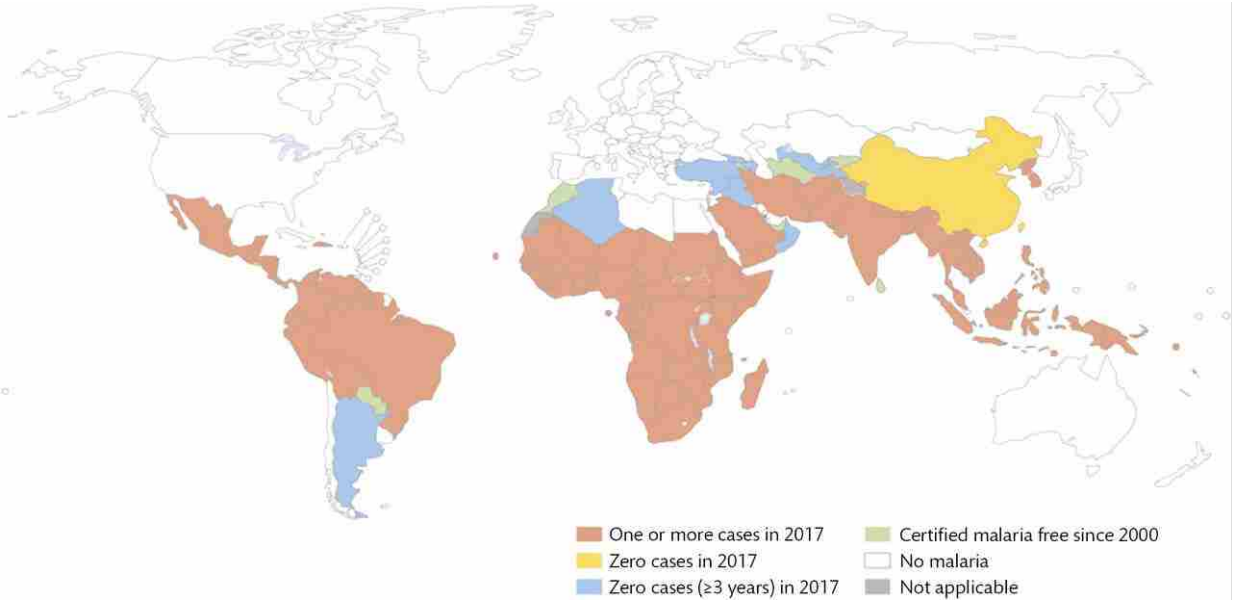


Fig. 8.15.1 Countries and territories with indigenous cases in 2000 and their status by 2017. Reprinted from World Malaria Report 2018. Geneva: World Health Organization; 2018. License: CC BY-NC-SA 3.0 IGO, with permission from the World Health Organization.

Countries with zero indigenous cases over at least the past 3 consecutive years are eligible to request certification of malaria-free status from WHO. All countries in the WHO European Region reported zero indigenous cases in 2016. Kyrgyzstan and Sri Lanka were certified malaria free in 2016.

Human *Plasmodium* parasites

Malaria is caused by the protozoa that belongs to the genus Plasmodium. Out of the over 100 species of *Plasmodium* which infect mammals, five are so far known to cause disease in humans. These are the *Plasmodium falciparum*, *Plasmodium vivax*, *Plasmodium malariae*, *Plasmodium ovale*, and *Plasmodium knowlesi*.

Plasmodium falciparum

Plasmodium falciparum (*P. falciparum*) is responsible for over 90% of cases and deaths worldwide. It is the most dominant form of malaria in sub-Saharan Africa. Children, pregnant women, and non-immune adults are vulnerable to the severe forms of the disease. According to the 2018 World Malaria Report, children under 5 years of age are the most affected, and account for 61% of the total deaths (WHO 2018a). Repeated infections in people who live in high transmission settings lead to the development of some partial immunity. This is the reason adults living in high transmission are less vulnerable to the severe forms of *P. falciparum* malaria. *P. falciparum* malaria is associated with the development of Burkitt's lymphoma, a B-lymphocyte cancer (Thorley-Lawson et al. 2016). *P. falciparum* lacks a fixed structure. It undergoes continuous change during its life cycle as part of survival strategies (Dhangadamajhi et al. 2010). The incubation period for *P. falciparum* malaria ranges from nine to 14 days.

Plasmodium vivax

Plasmodium vivax infection is the most geographically widespread of the five human malarias, with about 35% of the world's population being at risk. It infects 19–50 million people each year mostly in the Asia Pacific region, Latin America, and parts of Africa. It causes significant morbidity and can be associated with severe malaria and death. Clinically, *P. vivax* malaria presents as an acute debilitating illness that is characterized by recurrent paroxysmal fever and relapses from dormant liver stages (hypnozoites). These symptoms can occur even when blood parasite density is low. *P.* vivax can be present and be infectious to mosquito vectors, while infected persons show no symptoms. The liver stage of *P. vivax* can lie dormant for weeks to years after first infection before causing relapses of symptomatic malaria. Primaquine is the only licensed antimalarial drug that targets the liver stage of the lifecycle. It used to be thought *Plasmodium vivax* depended on the Duffy antigen to invade red blood cells and individuals who do not express the Duffy antigen were refractory to *P. vivax* infection. Reports of *P. vivax* infections in Duffy negative patients raise the possibility of an alternative invasion mechanism (Menard et al. 2010; Howes et al. 2016). The incubation period for *P.* vivax malaria ranges from 12 to 18 days.

P. vivax presents a particular challenge to malaria elimination for the following reasons: (1) It can survive in cooler climates hence its wider

geographical coverage; (2) it is more difficult to detect using current diagnostic techniques; (3) a single infection can give rise to multiple episodes of malaria; (4) treatment of liver-stage parasites requires a 14-day course of primaquine; and (5) primaquine treatment can produce serious side effects in persons with glucose-6-phosphate dehydrogenase (G6PD) deficiency (WHO 2015a; Baird et al. 2016).

Plasmodium malariae

Plasmodium malariae is a widely distributed *Plasmodium* species. It is endemic in parts of Africa, South America, Asia, and the Western Pacific, and is responsible for a mild form of malaria. Malaria caused by *P. malariae* is often referred to a benign, quartan malaria. This is because it rarely results in severe disease, and the associated fever tends to occur at three-day intervals unlike the others which are usually tertian (i.e. 2-day interval for the occurrence of fever). *P. malariae* has a tendency for low-dose chronic, asymptomatic infection, which when untreated can lead to nephrotic syndrome and albuminuria (Gilles and Hendrickse 1963; Ehrich and Eke 2007). There are no animal reservoirs for *P. malariae*. A large case series reported from southern Papua in 2015 shows that *P. malariae* is responsible for a minority (2.6%) of clinical malaria cases but has potential to cause a severer form of anaemia. The mean haemoglobin in patients with *P. malariae* infection was 0.3–0.6g/dl lower than the other locally prevalent Plasmodium species, with an associated 2.3-fold greater odds of severe anaemia compared to individuals without malaria (Langford et al. 2015). The incubation period for *P. malariae* malaria is highly variable and ranges from 16 to 59 days (Collins and Jeffery 2007). Late recrudescence may occur with *P. malariae* malaria.

Plasmodium ovale

Plasmodium ovale is primarily found in sub-Saharan Africa (Collins and Jeffery 2005). Molecular methods have led to the elucidation of two distinct sympatric species termed *P. ovale curtisi* and *P. ovale wallikeri*. *P. ovale* also causes mild (*benign*) malaria and is relatively unimportant as a cause of severe malaria morbidity and mortality. A systematic review of clinical presentation, treatment, and relapse characteristics of *P. ovale* malaria

reported in 2017 identified 22 severe cases, five fatalities, and two cases of congenital malaria. Most cases were from sub-Saharan Africa, Indonesia, Papua New Guinea, and India (Groger et al. 2017). Comparatively, *P. ovale* has received little attention in medical research and knowledge gaps exist in the optimal drug for its treatment. The incubation period ranges from 18 to 40 days.

Plasmodium knowlesi

Plasmodium knowlesi is a malaria parasite found in monkeys which can infect humans via mosquito bites (Moyes et al. 2014). Human infection was first reported in 1964. It was however not until 2004 that reports of a large focus of naturally acquired human infections in Sarawak, Malaysia brought this species to greater global attention (Singh et al. 2004). *P.* knowlesi causes life-threatening malaria yet it is grossly under-reported and often misreported as *P. malariae*. It is the most common cause of malaria in Malaysia. Cases have also been reported in Indonesia, Thailand, Vietnam, and Myanmar. Infection with *P. knowlesi* is associated with high levels of parasitaemia. The risk of severe disease in adults appears as high as that of *P.* falciparum. For many years, misdiagnoses were associated with delayed administration of treatment and high fatality rates. Increased awareness of pathogenic potential has led to improvement in treatment and reduction in case fatalities (Rajahram et al. 2012; Jeremiah et al. 2014). Local transmission of *P. knowlesi* in sub-Saharan Africa is unknown and this may be due to lack of a monkey reservoir. Intravenous artesunate is highly effective in severe knowlesi malaria while chloroquine and artemisinin-combination therapy (ACT) are highly effective for uncomplicated *P. knowlesi* malaria (Barber et al. 2017). There is uncertainty about human-to-human *P. knowlesi* infection.

Life cycles of the malaria parasite and vector

Two life cycles are important in discussing how the bite of an infected female Anopheles mosquito (vector) transmits *Plasmodium* (parasite) and causes malaria. These are the life cycles of the *Plasmodium* parasite, and of the female Anopheles mosquito. A good knowledge of the essential aspects of these life cycles is important for an appreciation of how preventive and

therapeutic interventions have emerged and how they have been applied to date, their limitations and prospect for the development of new tools (Acharya et al. 2017).

Life cycle of the *Plasmodium* parasite

Malaria infection is initiated when an infected female Anopheles mosquito bites a human host. It injects into the host saliva that contain anophelin, an anticoagulant. This material counteracts the host's haemostatic response and prevents the sucked blood from clotting. In the process, the female Anopheles mosquito ingests sporozoites of *Plasmodium* parasite into the blood stream of the victim. The stage of the parasite life cycle that is transmitted in the bite is the sporozoite. Within 30 to 60 minutes of ingestion, sporozoites migrate through the venous circulation to the liver. (https://www.ncbi.nlm.nih.gov/books/NBK215619/). While within liver cells, sporozoites undergoes a series of asexual multiplications that lead to the development of schizonts that contain the next stage of the parasite called merozoites. This process is called sporogonic schizogony and because it takes place in liver, it is also called the hepatic or pre-erythrocytic stage (Mota et al. 2002; Mota and Rodriguez 2002) (Fig. 8.15.2).

Upon their maturation, merozoites burst out of schizonts and into the bloodstream where they attack red blood cells (RBCs) to begin the erythrocytic phase of development. Within the RBCs, merozoites differentiate further into ring form called trophozoites. Trophozoites undergo further asexual multiplication and develop into schizonts. This process is called erythrocytic schizogony. Schizonts burst on maturation to release more merozoites into the bloodstream. The rupture of erythrocytic-stage schizonts out of RBCs coincides with the emergence of symptoms. While within RBCs, the developing parasite digests haemoglobin and converts the waste product into hemozoin (also called malaria pigment) which are packaged in vesicular compartments called digestive vacuoles. The burst of schizonts leads to the release of digestive vacuoles (DVs) into the bloodstream, a triggering of immune response and the emergence of symptoms. While DVs get phagocytosed, merozoite largely escape the attack of the immune system, attack more RBCs, and further aggravate the illness.

Typically, it takes 10–14 days from the time of sporozoite injection to the emergence of symptoms. This is called the incubation period. The cycles of

release from RBCs and DVs, attack on new cells, and further RBC destruction underline bouts of fever, chills, and other symptoms of the disease (Mueller et al. 2015; Tannous and Ghanem 2018).

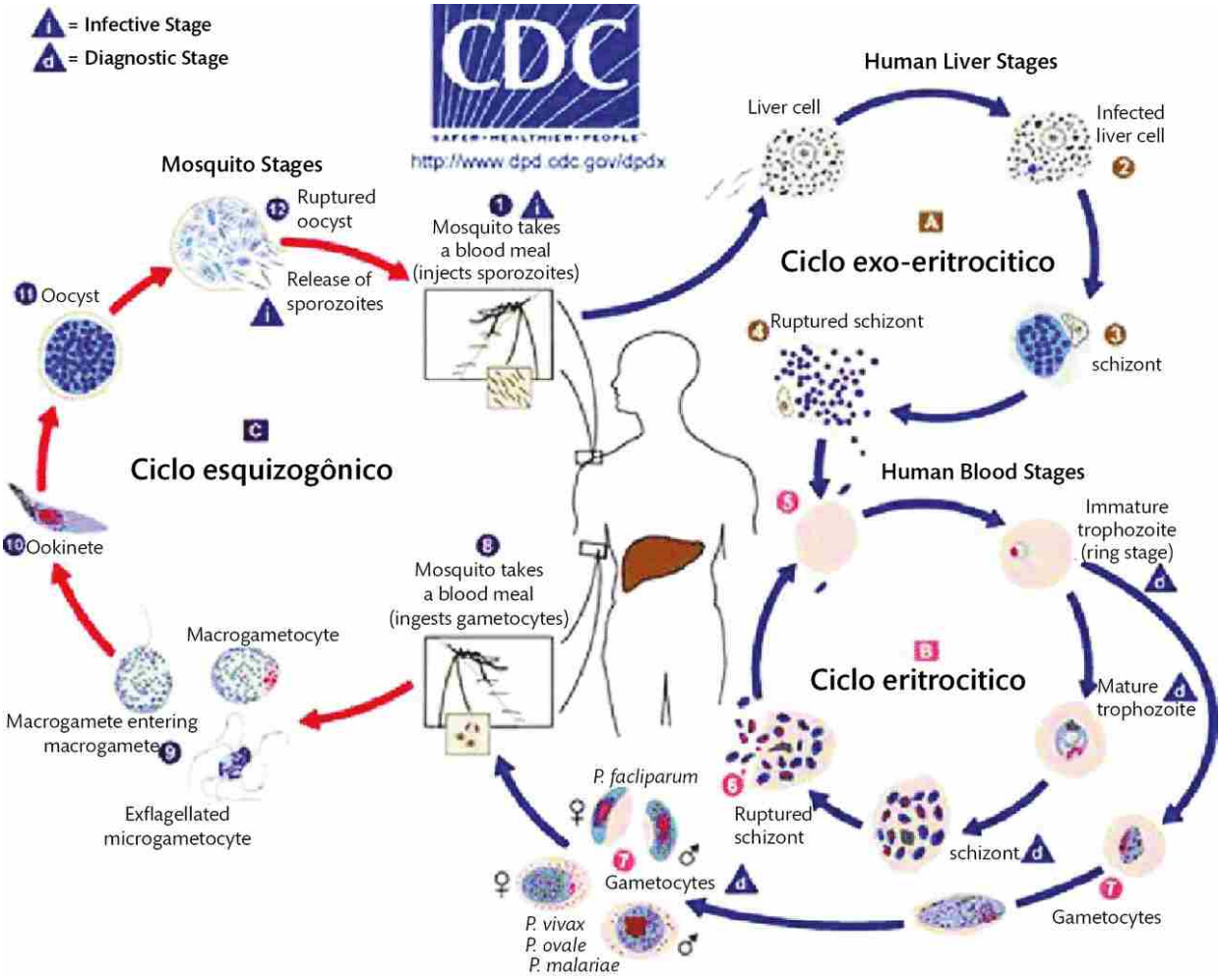


Fig. 8.15.2 The lifecycle of the malaria parasite.

Source: CDC, available freely at https://www.cdc.gov/malaria/about/biology/mosquitoes/. This does not constitute its endorsement or recommendation by the U.S. Government, Department of Health and Human Services, or Centers for Disease Control and Prevention.

This does not constitute its endorsement or recommendation by the U.S. Government, Department of Health and Human Services, or Centers for Disease Control and Prevention During the intraerythrocytic stage, some merozoites differentiate into male (microgametocytes) and female (macrogametocytes) gametocytes. Gametocytes are the sexual forms of the

parasite and they are the stage of development that can infect mosquitoes. During a blood meal, the female Anopheles mosquito ingests gametocytes and this reinitiates the sporogonic stage of parasite development. In the gut of the mosquito, male and female gametocytes develop into gametes which merge, and over 2 to 3 weeks develop into sporozoites in the salivary glands. Higher climatic temperatures accelerate parasite development within the mosquito (Afrane et al. 2012) (https://www.cdc.gov/malaria/about/biology/mosquitoes/). This is one of the anticipated effects of climate change on malaria.

While the aforementioned is the classical presentation of the life cycle of the *Plasmodium* parasite, it important to appreciate that important difference exist in the life cycles of the different *Plasmodia*. In the case of *P. vivax* and *P. ovale*, a dormant stage (hypnozoites) can persist in the liver and cause relapses by invading the bloodstream weeks, or even years later after infection (Olliaro et al. 2016). Once in the bloodstream, *P.* vivax tends to attack reticulocytes (young RBCs) and remain in the ring form. The ability of *P. vivax* to attack RBCs is due to the presence of Duffy blood group antigens on the red cells of susceptible populations. For many years, it was thought that the absence of this antigen is the reason populations in sub-Saharan Africa are less susceptible to *P. vivax* infection. This is now firmly disputed.

Generally, the level of parasitaemia with *P. vivax* infection is much lower than occurs with *P. falciparum* infection. Gametocytaemia also tends occur earlier in the course of *P. vivax* malaria. This makes patients infectious almost at the onset of symptoms.

The different stages in the life cycle of *plasmodium* present a range of antigenic challenges to the host immune system. Antigen expression has been implicated in the aetiology of severe malaria (Pal et al. 2016, 2017). An improved understanding of the nature of antigen expression, facilitated by emerging genetic technology, should lead to the development of new diagnostic and therapeutic options for the management and control of the disease.

Life cycle of the *Anopheles* mosquito

Anopheles is a vector that belongs to the class of mosquitoes named by J.W. Meigen in 1818. While over 100 of the *Anopheles* species can transmit human malaria, only 30–40 commonly do so. *Anopheles* gambiae is one of

the best studied because of its role in malaria transmission (Afrane et al. 2012).

The anopheline mosquito develops through four stages: egg, larva, pupa, and imago. This begins with the laying of eggs in fresh or stagnant water by the female anopheline mosquito. Cans, buckets, discarded tyres, clogged rain gutters that hold stagnant water are common sites for the laying of eggs. The eggs develop into larva which further develops into the pupa. All three stages are aquatic and last about 5 to 14 days. This length of time depends on the species of anopheline, the level of humidity and ambient temperature (Beck-Johnson et al. 2013).

The adult stage emerges from the pupa on the surface of the breeding water. Female anopheline mosquitoes seek blood meals to derive vitally needed nutrients to fertilize its eggs. In the process it acts as vector for malaria and transmits the parasite into man.

Although the adult female has the potential to live for up to a month, it typically lives for no more than two weeks. For malaria transmission to be possible therefore, the *Anopheles* mosquito needs to live long enough to enable the sporogonic stage of *Plasmodium* development to be completed. One of the postulated effect of global warming is that it will shorten the sporogonic phase of parasite development and enhance the ability of the mosquito to transmit malaria (Afrane et al. 2012; Caminade et al. 2014).

Pathogenesis of malaria

All the mechanisms through which infection with *P. falciparum* leads to malaria are associated with the erythrocytic stage of parasite development (Buffet et al. 2011; Sanyal et al. 2012). When RBCs are infected by malaria parasites, they (RBCs) undergo configurational changes. These include loss of the normal discoid shape, increased rigidity of the membrane, elevated permeability to a wide variety of ionic and other species and increased adhesiveness (Mohandas and An 2012).

Sticky RBCs more easily adhere to other normal RBCs (rosetting) and to the inside walls of capillaries (cytoadhesion). Clumping, rosetting, and cytoadhesion occur, blocking small vessels and causing ischaemic cell injury in several organs of the body (Moxon et al. 2011). Most notable are the lung, heart, brain, and kidney. This explains why malaria is a multisystem disease with wide range of unspecific clinical manifestations. The effect in the brain

forms the basis of cerebral malaria. The rupture of parasitized RBCs and clearance of deformed parasitized and unparasitized erythrocytes by cells of the immune system including the spleen is the principal cause of malarial anaemia (Autino et al. 2012; Safeukui et al. 2015; White 2017). The release of toxins (packaged as digestive vacuoles referred to earlier) from ruptured RBCs into the bloodstream leads to cytokine expression and unspecific symptoms like fever, headache, loss of appetite, and vomiting.

In pregnant women, parasitized RBCs adhere to chondroitin sulphate A in the placenta. The accumulation of RBCs compromises the efficiency of the placenta and this leads to clinical manifestations such as abortion, intrauterine growth retardation, low birth weight, and stillbirths.

Cells that are attacked by RBCs express antigen on the surfaces as part of activation of the immune system. It is the repeated triggering of the immune system through antigen expression that forms the basis of immunity and tolerance to malaria (Matuschewski 2017). It is the reason the severity of malaria generally declines with increasing age and gravidity (in the case of pregnancy) in persons who live in malaria-endemic countries (Mendonca and Barral-Netto 2015). It is also the reason why individuals who have lived in endemic settings loss some immunity when they move into non-endemic settings.

An important aspect of the immune response of the body is the mobilization of all possible capacity to destroy parasitized RBCs. The major organs in the body responsible for the destruction of parasitized and damaged RBCs are the spleen and liver. Hyperactivity due to repeated attacks leads to enlargement of these organs, most particularly the spleen. Community-based prevalence of palpable spleen enlargement was in the past used as a measure of the endemicity of malaria (Snow et al. 2012).

Clinical presentation

The clinical presentation of malaria is closely related to the stages in its pathogenesis as outlined earlier. The *liver phase* of the *malaria* lifecycle occurs without clinical symptoms**.** Symptoms emerge following invasion and rupture of RBCs. In an ‘interruptible’ continuum, there are two forms of clinical presentations. These are uncomplicated and complicated or severe malaria.

Uncomplicated malaria

Uncomplicated malaria is characterized by symptoms such as fever, headache, vomiting, and general malaise. In most high transmission settings, these symptoms often trigger suspicion of malaria and the initiation of treatment, when available. Unfortunately, the symptoms of uncomplicated malaria are so unspecific that they could as well be caused by other infections, particularly viral infections. In the absence of a test, the false diagnosis of malaria leads to treatment with antimalarials. If correctly diagnosed and effectively treated at this stage, good and complete recovery is nearly always assured. Most cases of malaria are uncomplicated and do not progress to severe or complicated malaria unless treatment is delayed or wrongfully administered.

Severe or complicated malaria

When treatment of malaria is delayed or ineffectively administered, a cascade of pathophysiological events leads to progression to severe or complicated malaria. For epidemiological and research purposes, severe malaria is defined as one or more of the conditions defined in Table 8.15.1, occurring in the absence of an identified alternative cause, and in the presence of *P. falciparum* asexual parasitaemia (WHO 2014) (Table 8.15.1).

Sustained RBC destruction and multiorgan damage leads to severe or complicated malaria. Although all organs of the body are affected, the most sensitively affected at early complicated stage is the brain. Brain involvement manifests a syndrome of presentations, the most common of which is cerebral malaria. Cerebral malaria refers to the syndrome in which altered consciousness is associated with a malarial infection, and cannot be attributed to convulsions, sedative drugs, or hypoglycaemia alone, or to a non-malarial cause. Commonly associated symptoms are severe loss of appetite, inability to tolerate oral feeds, repeated vomiting, disorientation, and convulsions.

Sustained red cell destruction results in severe weakness, pallor, jaundice (yellowing of the eyes and skin), and kidney failure. At this stage, urgent and effective treatment is needed to prevent death. The highest mortalities occur in children aged less than 5 years, pregnant women, and non-immune adults.

In these persons, malaria takes a rapid course from uncomplicated to complicated state. The state of consciousness commonly assessed in children

in sub-Saharan Africa with the Blantyre coma score (BCS) is a good indicator of prognosis.

In Gambia, a BCS of 2 or less was found in a review of 2,901 children with severe malaria to be the most sensitive and specific predictor of death or development of neurological sequelae (Jallow et al. 2012). Children who survive severe malaria may develop long-term neurological deficits that impact development and performance in school (Fernando et al. 2003; Brooker et al. 2008; Clarke et al. 2008; Gomes et al. 2009; Dondorp et al. 2010). Multiorgan failure including renal failure occurs more commonly in non-immune adults than in children in highly-endemic areas.

Table 8.15.1 Epidemiological and research definition of severe falciparum malaria

Impaired consciousness	A Glasgow Coma Score <11 in adults or a Blantyre coma score <3 in children
Acidosis	A base deficit of >8 meq/L or, if unavailable, a plasma bicarbonate of <15 mm or venous plasma lactate >5 mm. Severe acidosis manifests clinically as respiratory distress—rapid, deep, and laboured breathing
Hypoglycaemia	Blood or plasma glucose <2.2 mm (<40 mg/dl)
Severe malarial anaemia	A haemoglobin concentration <5 g/dl or a haematocrit of <15% in children <12 years of age (<7 g/dl and <20%, respectively, in adults) together with a parasite count >10 ,000/µl
Renal impairment (acute kidney injury)	Plasma or serum creatinine >265 µm (3 mg/dl) or blood urea >20 mm
Jaundice	Plasma or serum bilirubin >50 µm (3 mg/dl) together with a parasite count >100, 000/µl
Pulmonary oedema	Radiologically confirmed, or oxygen saturation <92% on room air with a respiratory rate >30/min, often with chest indrawing and crepitations on auscultation
Significant bleeding	Including recurrent or prolonged bleeding from nose gums or venipuncture sites; haematemesis or melaena
Shock	Compensated shock is defined as capillary refill ≥3 s or temperature gradient on leg (mid to proximal limb), but no hypotension. Decompensated shock is defined as systolic blood pressure <70 mm Hg in children or <80 mmHg in adults with evidence of impaired perfusion (cool peripheries or prolonged capillary refill)
Hyperparasitaemia	P. falciparum parasitaemia >10%

Reprinted from WHO et al (2014) 'Severe Malaria' *Tropical Medicine & International Health* 19(s1):123, © WHO 2014, with permission from John Wiley & Sons.

Malaria in pregnancy

Malaria in pregnancy is a major cause of maternal morbidity and perinatal mortality in high transmission countries. Up to 9.5 million pregnant women in sub-Saharan Africa are estimated to be at risk of *Plasmodium falciparum* infection. In the absence of interventions, malaria in pregnancy results in

900,000 (95% CI 530,000–1,240,000) babies who are born with low birth weight (Walker et al. 2014).

The state of pregnancy increases the risk of malaria infection and an estimated 125 million pregnant women get infected annually. The risk of severe malaria is particularly high in primigravidae in areas of higher transmission. The development of partial immunity leads to an attenuation of the risk in successive pregnancies. This pattern is less marked in areas of low transmission. The risk of malaria is highest in the second trimester. In the first trimester, malaria increases the risk of miscarriage by 60% and the rate of mortality from severe malaria in pregnancy can be as high as 50%. Younger women are particularly at risk of malaria during pregnancy (WHO 2015b; Tarning 2016).

Pathogenesis, clinical presentation, and diagnosis

The increased susceptibility of pregnant women to malaria is due to the sequestration of malaria-infected erythrocytes in the maternal placenta. This sequestration is highest with *P. falciparum* infections. The affinity of parasitized RBCs for placenta is mediated by the expression of the parasite-derived protein VAR2CSA on the surface of infected RBCs. These cells bind to chondroitin sulphate A (CSA) in the placenta and induce placental inflammation. This inflammatory process leads to the aggregation of monocytes in the placenta, with compromise to the efficiency of the placenta as a medium of feto-maternal transfer.

Parasite sequestration in the placenta can occur in early pregnancy and lead to poor placenta development. This is evident in poor villous vasculature and small placenta size. This early effect can lead to miscarriages, intrauterine growth retardation, and death. In high transmission settings, placenta parasite sequestration can occur in the absence of malaria symptoms during early pregnancy. Concurrent parasite sequestration in the brain however results in severe malaria with possible dire consequences.

A challenge to the diagnosis of malaria in pregnancy is the fact that it may present without symptoms and blood smear microscopy will be negative even in the presence of a high parasite load in the placenta (Geelhoed et al. 2001; Kakuru et al. 2016). The sensitivity of microscopy to detect malaria parasites in peripheral blood in pregnant women can be as low as 42%. The

use of antigen-based diagnosis and Polymerase Chain Reaction increases sensitivity to about 80% and 97%, respectively (Mockenhaupt et al. 2002).

Diagnosis of malaria

Practically considered, there are three approaches to the diagnosis of malaria. These are

The presumptive approach
Rapid diagnostic test
Blood smear microscopy

The presumptive approach

The presumptive approach relies on the clinical judgement of an attending clinician, and it is based on the symptoms and signs that a patient presents with. For many years, fever was used as proxy indicator of malaria infection in high transmission settings. This was justified because transmission was high and that made malaria either a direct or underlying cause of most fevers. It was also because it was impractical to have microscopy in many settings where malaria was treated and most of the antimalarials in common use at the time were also relatively inexpensive.

The symptoms of malaria are very unspecific, and they could be caused by virtually many other acute infectious diseases. The presumptive approach was therefore associated with considerable overdiagnosis, inappropriate use of antimalarials, and selection of resistance strains. The approach also led to delayed diagnosis and management of the definitive causes of fevers. A shift away from the presumptive approach only began in the late 2000s and this was made conceivable by the development, global validation, and deployment of rapid diagnostic test kits. Today the management of malaria solely based on symptoms and signs is frowned upon and considered inconsistent with the WHO recommended strategy of '*TEST-TREAT-TRACK*' (next). The presumptive approach should therefore be used only when parasitological testing is not available.

Rapid diagnostic test

The dipstick technology for the rapid diagnosis of malaria was developed in the early 1990s. It was however not until the 2000s that they came into widespread use (Bell et al. 2006a, 2006b; Baiden et al. 2012). Rapid diagnostic tests for malaria (mRDT) are a point-of-care device that can be used to confirm parasite infection in about 15 to 30 minutes (Perkins and Bell 2008). Their availability has revolutionized the diagnosis of malaria and improved the targeting of antimalarials. The technology depends on the detection of antigens produced by the plasmodium parasite. These are histidine-rich protein-2 (HRP-2), lactate dehydrogenase (LDH), and adolase. Whereas the LDH and aldolase are expressed by all *Plasmodium* species, HRPS-2 antigen is specific to *P. falciparum*. This has made it possible to develop RDTs that distinguish *P. falciparum* from non-*P. falciparum* infections.

Important advantages with the use of mRDTs include the fact that non-laboratory technicians can use it with minimal training. It can also be deployed in all settings irrespective of infrastructural capacity. A drawback to the use of mRDTs is the fact that it does not afford a quantitative measure of the degree of parasitaemia. Additionally, antigen persistence in the bloodstream means HRP-2-based mRDT positive results can remain positive for up to 3–4 weeks after successful treatment (Baiden et al. 2012; Mbabazi et al. 2015; de Oliveira et al. 2009). HRP-2-based mRDTs can therefore not be used to monitor therapeutic response and, therefore blood smear microscopy should be used in the diagnosis and management of severe malaria.

An emerging challenge is the deletion of *Plasmodium falciparum hrp2* and *hrp3* genes from the genome of some parasites. This can affect the performance of HRP2-based mRDTs. These have been reported in South America, India, Eritrea and Kenya and are likely to spread as parasites with *hrp2* and *hrp3* genes will escape detection during test (Nima et al. 2017). Although the prevalence of HRP2 gene deletions in most high transmission countries remains low, effective surveillance is needed (Gamboa et al. 2010; Beshir et al. 2017; WHO 2018a).

Blood smear microscopy

Microscopy has been the mainstay for the definitive diagnosis of malaria since the work of Alphone Laveran. It involves taking venous blood,

preparing a blood smear on a glass slide, processing it, and reading under a light microscope. A technically competent technician is required to read the slide. The process requires much sophistication in terms of infrastructure, logistics, and technical competence. In most parts of the malaria-endemic world, the capacity to perform malaria microscopy could not be sustained beyond regional and district levels. Even in facilities that have the capacity to perform malaria microscopy, results may become available after treatment has been initiated due to the time it takes to perform this test. In the hands of a well-trained microscopist, microscopy is the gold standard for the diagnosis of malaria.

A major advantage of microscopy is the fact that it affords species differentiation and quantification of parasite load. It can therefore be used to monitor parasite load as part of response to treatment. A major reason why the capacity for blood smear microscopy needs to be sustained is the fact that the possibility of diagnosis of other infectious diseases adds to the utility and cost-effectiveness of blood smear microscopy.

Polymerase chain reaction (PCR) in malaria diagnosis

Another approach to detecting malaria infection is the use of PCR-based assays. PCR-based assays for the detection of malaria parasites were developed in the 1990s. They have over a 100-fold greater sensitivity than microscopy for the detection of malaria parasitaemia. High cost and technical sophistication however limit its use as a point-of-care device. They are extremely useful for surveys in areas where malaria is approaching elimination and where low levels of parasitaemia make microscopy and mRDT less useful. A setting where PCR-based assays could be extremely useful is detecting malaria infection during pregnancy.

T3: Test-Treat-Track

On World Malaria Day 2012, WHO launched *T3: Test. Treat. Track*, with a call on malaria-endemic countries, donors, and the global malaria community to scale-up diagnostic testing, treatment, and surveillance for malaria. The T3 strategy was premised on the vastly improved availability of quality-assured mRDTs, the substantially reduced levels of malaria

transmission and the need for prudent use of the available range of antimalarials (WHO 2010, 2015b).

Consistent with the strategy, it is now required that all cases of suspected malaria be confirmed through parasitological testing before treatment is initiated. In various trials and operational research, this approach has been shown to lead to significant reduction in the use of antimalarials. Reductions of between 30% and 60% have been reported (Skarbinski et al. 2009; D'Acremont et al. 2010; Kyabayinze et al. 2010; Thiam et al. 2011; Yukich et al. 2012). In a trial in a high transmission setting in Ghana however, the reduction in antimalarial use was a rather modest 12%. It was found in this trial that relying on the results of mRDTs to manage under-5 children presenting with fever (compared with the presumptive approach) improved the targeted use of antimalarials without compromise to clinical outcomes (Baiden et al. 2016). The use of mRDTs in the management of uncomplicated malaria has also been found to be cost-effective in diverse transmission settings (Yukich et al. 2010; Tawiah et al. 2016).

Improved targeted use of antimalarials has come at some considerable cost. In the absence of tools to establish the aetiology of other fevers, there has been a shift towards increased presumptive use of antibiotics in the management of non-malaria febrile illnesses (Baiden et al. 2011; Iroh Tam et al. 2016). An improved understanding of the aetiology of fevers in malaria-endemic countries is needed if clinicians in sub-Saharan Africa are to be better-guided (Baiden et al. 2012; D'Acremont et al. 2014; Hopkins et al. 2017).

Treatment of malaria

The primary goals of malaria treatment are:

To achieve cure and complete recovery in patients
To reduce the risk of transmission of the infection to others, by reducing the infectious reservoir and
To prevent the emergence and spread of resistance to antimalarial medicines

The effectiveness and utility of antimalarials are assessed on their intrinsic potential to achieve these goals.

Antimalarial drugs

Artemisinins

The remarkable story of the discovery of artemisinin and the establishment of its antimalarial activity by Chinese scientists represents one of the great discoveries in medicine in the latter half of the twentieth century (Krishna et al. 2008). The artemisins are a class of antimalarials derived from the Chinese plant, *Artemisia anua*, or sweet wormwood.

It is a herb that has been part of traditional medicine in China for over 1,000 years. In 1972, Youyou Tu, a scientist at the China Academy of Traditional Chinese Medicine, extracted the artemisinin from *Artemisia anua* (Liu and Liu 2016). Today, treatments containing artemisinin derivatives are part of the standard treatment for malaria.

In 2015, Youyou Tu was awarded the Nobel Prize in Medicine for her pioneering work. The artemisinins and its semi-synthetic derivatives are described as a gift of Chinese traditional medicine to the world. The account about how the artemisinins have today emerged as the most important class of antimalarials is a classic in global health. It is one that calls for respect for oriental alternative and complementary medicine (Guo 2016; Ikram and Simonsen 2017).

Artemisinin is a sesquiterpene, lactone-containing a distinctive peroxide bridge that is hypothesized to be responsible for its antimalarial properties. Once in the bloodstream, artemisinin derivatives kill the majority of parasites within a few minutes while getting rapidly metabolized. The poor level of artemisinin bioavailability limits its effectiveness. Semi-synthetic derivatives have been developed to overcome this limitation. Dihydroartemisinin is the active metabolite of all the artemisinin compounds. It is now available as a drug in the form of dihydorartemisinin-piperaquine (Baiden et al. 2015; Adjei et al. 2016; Oduro et al. 2017). The side effects of the artemisinins are minor and include nausea, vomiting, persistent cough, and dizziness. Very little is known about the teratogenicity of the artemisinins. They are therefore not recommended for routine use in first trimester pregnancy. This is likely to change soon.

It was initially thought that the short half-lives of the artemisinins will limit the possibility of resistance development. Recent reports however confirm the emergence of artemisinin resistance in the Southeast Asia region. Artemisinin resistance is now a major global public health concern

(Mbengue et al. 2015; Woodrow and White 2017). Derivatives of the artemisinins that are in common use are artesunate, artemether, and arteether.

Artesunate is a water-soluble ester derivative of the artemisinin. It is used in combination with antimalarials like amodiaquine, sulphadoxine-pyrimethanine and mefloquine in the treatment of uncomplicated malaria. Artesunate in combination therapy is first-line drug for the treatment of uncomplicated malaria in many countries in sub-Saharan Africa and Southeast Asia. As monotherapy, it is administered intravenously or intramuscularly to treat severe malaria. In a series of clinical trials conducted among Asian and African populations, the superior efficacy of artesunate (compared to quinine and artemether) was established (Dondorp et al. 2005, 2010; Esu et al. 2014). Rectal artesunate is a particularly useful formulation for the home and community-level management of severe malaria in children who are so sick they cannot tolerate oral medications. Where reaching the nearest health facility is likely to be delayed because of remoteness, rectal artesunate has been demonstrated to reduce the risk of death and neurological sequelae in children less than five years of age (Gomes et al. 2009).

Artemeter is a semi-synthetic, lipid-soluble methyl ether of dihydroartemisinin. When administered intramuscularly, peak plasma concentration is achieved within 6 hours. Artemeter has very rapid schizontocidal activity against blood forms of *P.* falciparum and *P.* vivax. It is combined with lumefantrine in the popular antimalarial brand Coartem (Artemeter-Lumefantrine). It is also among the first-line drugs for the management of uncomplicated malaria in most malaria-endemic countries in sub-Saharan Africa. Intramuscular injection of Artemeter is used in the treatment of severe malaria.

Arteether comes in two forms: artemotil and alpha/beta arteether, both of which are given as an intramuscular injection once a day. It is the least commonly used among the artemisinin derivatives.

Artemisinin-combinations are the most effective antimalarial medicines available today. Five ACTs are currently recommended by WHO for use against *P.* falciparum malaria. Individual country choice of ACT should be based on the knowledge of the local strains of *P.* falciparum and ACT efficacy. Unfortunately, artemisinin is produced in small quantities and this leads to a shortage of global supply and increase in price (Peplow 2016; Ikram and Simonsen 2017).

Chloroquine

Chloroquine is a 4-aminiquinolone that was discovered in 1934 and had over 60 years of widespread use in the treatment of malaria. Chloroquine is administered through oral and intramuscular routes. A bitter taste and itching are its major side effects. Once in the bloodstream however, chloroquine kills asexual forms of the parasite. It has no activity against mature gametocytes and does not have effect on either sporozoites or hypnozoites of *Plasmodium vivax*. Its terminal elimination half-life is 45–55 days. Its major plasma metabolite desethylchloroquine has a longer half-life of 59–67 days. Chronic use (as in the treatment of skin conditions) can result in retinopathies. Ease of production, monotherapy and sustained presumptive use led to the emergence of parasite resistance which by the late 1990s had surpassed the WHO recommended threshold of 25% in many malaria-endemic countries (Abuaku et al. 2005; Oduro et al. 2005; Takala-Harrison and Laufer 2015).

The WHO no longer recommends chloroquine for the treatment of malaria in sub-Saharan Africa. It is however retained (in combination with primaquine) for treatment for *vivax* or *ovale* malaria in travellers that aim at radical cure. Recent reports however suggest chloroquine efficacy is re-establishing in some areas in sub-Saharan Africa where resistance was initially high, and where chloroquine has not been used for many years (Kublin et al. 2003; Frosch et al. 2014; Galatas et al. 2017; Ndam et al. 2017). It is possible that chloroquine will regain efficacy and be reintroduced for malaria prevention and treatment.

Amodiaquine

Amodiaquine is a 4-aminoquinolone that is similar in structure and activity to chloroquine. It has been used as an antimalarial since 1948 and remains active against many of the chloroquine-resistance strains of the parasite. It is considered safe for use in post-first trimester pregnancy although tolerability precludes its use in intermittent presumptive treatment (Clerk et al. 2008). In combination with artesunate, amodiaquine is among current first-line drugs for the treatment of malaria in some sub-Saharan African countries. Like chloroquine, amodiaquine can be produced synthetically and is considered an affordable antimalarial. It has a bitter taste and the main side effects are

headache, drowsiness, and visual disturbances. Cardiac arrest may occur with overdose (Olliaro and Mussano 2003).

Sulfadoxine-pyrimethamine

This is a combination drug that contains sulfadoxine (a sulfonamide) and pyrimethamine (an antiprotozoal). It has been used for decades in the treatment and prevention of malaria. The combination acts synergistically to inhibitor protozoan folate synthesis. While the sulfadoxine competes with p-aminobenzoic acid, pyrimethamine serves as selective inhibitor of dihydrofolate. For many years sulfadoxine-pyrimethamine was used in the treatment of uncomplicated malaria, and over a brief period in the early 2000s it was used in combination with artesunate for the treatment of uncomplicated malaria in India, and some countries in east and southern Africa. Growing parasite resistance has lowered the efficacy of sulfadoxine-pyrimethamine in many parts of sub-Saharan Africa. Currently sulfadoxine-pyrimethamine is almost exclusively used in intermittent preventive treatment (IPT) during pregnancy (Owusu-Boateng and Anto 2017). Its use for this indication is however threatened by resistance mutations in the parasite dhps gene. Hypersensitivity can result in Stevens–Johnson syndrome. Sulfadoxine-pyrimethamine is synthetically produced and is considered an affordable antimalarial.

Primaquine

Discovered in 1946, primaquine is an 8-aminoquinoline used in the treatment and prevention of malaria caused by *P. vivax* and *P. ovale*. Primaquine is recommended for primary prophylaxis prior to travel to *P.* vivax-endemic regions. In low transmission settings, a single dose of 0.25 mg/kg body weight of primaquine combined with ACT is recommended to patients with *P. falciparum* malaria except pregnant women, infants aged less than 6 months and women breastfeeding infants less than 6 months of age. This is intended to reduce the risk of transmission.

Until recently primaquine was the only antimalarial drug capable of eliminating hypnozoites, the dormant liver form of Plasmodium vivax. It was therefore used to prevent malaria relapse in persons who travel to *P.* vivax-endemic regions. Primaquine has rapid and potent ability to kill the gametocytes of *P. falciparum* and *P. vivax*. In early 2018, the Food and

Drugs Agency (FDA) of the United States of America approved *Krintafel* (tafenoquine), as a single dose antimalarial for radical cure (prevention of relapse) of *P. vivax* malaria. It is approved for use in patients aged older than 15 years and who are receiving appropriate antimalarial therapy for acute *P. vivax* infection.

Primaquine kills gametocytes and because of this, WHO in 2012 recommended that a single dose of 0.25 mg/kg primaquine be added to malaria treatment schedules in areas with artemisinin resistance and where the intent is to stop malaria transmission. The side effects of primaquine include nausea, vomiting, and stomach cramps. Primaquine is contraindicated in persons with glucose-6-phosphate dehydrogenase deficiency (G6PD). In such persons, primaquine cause dose-dependent oxidant haemolysis in individuals with glucose-6-phosphate dehydrogenase (G6PD) deficiency (Watson et al. 2018). It is similarly for this reason primaquine is contraindicated in pregnant women and lactating mothers. There is synthetic production of primaquine and it is considered an affordable antimalarial. A case report in 2018 has highlighted the important role CYP-4502D6 mutation plays in primaquine treatment failure (Dijanic et al. 2018).

Quinine

Quinine is a basic amine that was first isolated in 1820 from the bark of a cinchona tree and has remained an effective antimalarial drug to date. Prior to its laboratory isolation, the bark of the cinchona tree was used in the treatment of fever for well over a century. Quinine is hypothesized to act against *P. falciparum* by interfering with the parasite's ability to dissolve and metabolize haemoglobin in RBCs. Quinine in its various salts is administered orally, intravenously, and intramuscularly. Until recently, quinine was the drug of choice for the treatment of severe and complicated malaria. Although it remains among drugs listed in many national guidelines for the management of severe malaria, intramuscular artemether, intravenous and intramuscular artesunate are preferred (see next). The common side effects of quinine are headache, palpation, nausea, tinnitus, hearing impairment, dizziness, blurred vision, and disturbance in colour perception. An important side effect of intravenously administered quinine is the risk of hypoglycaemia. Quinine toxicity can lead to cardiac arrest. Although quinine

can be synthetically produced, the cinchona tree remains the most economically-viable source of it (Shanks 2016; Gachelin et al. 2017).

Treatment of malaria during pregnancy

Except in the first trimester, the drugs recommended for use in malaria in non-pregnancy adults can also be use pregnant women. This includes artemisinin-based derivatives. The use of the artemisinin-based derivatives in the first trimester however remains restricted to situations where the benefits of its use outweighs the risk or where safer options are not available. There are increasing calls on WHO to reverse the current restriction and recommend the use of artemisinin derivatives for the treatment of malaria during the first trimester (Moore et al. 2016). The antimalarial medicines considered safe in the first trimester of pregnancy are quinine, chloroquine, clindamycin, and proguanil. WHO recommends the combination of oral quinine and clindamycin for the treatment of malaria in the first trimester of pregnancy (Table 8.15.2).

Table 8.15.2 Effects of *P. falciparum* malaria in pregnancy and the drugs used in the management

	Adverse effects		Drug options for treatment	
	Pregnant woman	**Foetus and baby**	**Uncomplicated**	**Severe or complicated**
1st Trimester • Psychological effects of miscarriage • Anaemia	• Miscarriage • Intrauterine growth retardation	• Oral quinine + clindamycin • Oral quinine	• IV/IM Quinine • IV/IM Artesunate • IM Artemeter
2nd and 3rd Trimesters • Premature delivery • Anaemia • Hypoglycaemia (if on quinine) • Pulmonary oedema • Death • Increased risk of other infections	• Intrauterine growth retardation • *In utero* death	• Artesunate-amodiaquine • Artemeter-lumefantrine • Dihydroartemisinin-piperaquine	• IV/IM Artesunate • IM Artemeter • IV/IM Quinine
Delivery and postpartum • Premature labour • Bacterial infection	• Stillbirth • Low infant birth weight • Neonatal death		

Vector control

An important part of malaria control involves killing the mosquitoes and or destroying breeding sites. This approach is referred to as vector control. Vector control interventions that have proven to be efficacious and are in common use are sleeping under an insecticide-treated net (ITNs) and in-door residual spraying (IRS). Both interventions rely on the use of chemical insecticides (Alonso et al. 2011; Williams et al. 2018).

Insecticide-treated nets

Investigation into the efficacy of ITNs began in the late 1980s and early 1990s with series of studies in different parts of sub-Saharan Africa that assessed the effect of ITN use on morbidity and all-cause child mortality. ITN use has been demonstrated to reduce overall childhood mortality by between 17 and 33% (Alonso et al. 1991; D'Alessandro et al. 1995; Binka et al. 1996; Nevill et al. 1996; Habluetzel et al. 1997; Lengeler et al. 1998; Phillips-Howard et al. 2003; Binka and Akweongo 2006). The use of ITNs was subsequently shown to provide community-wide protection by reducing the number of infective mosquitoes in the vicinity where ITNs are used (Binka et al. 1998).

An important basis for the development of ITNs was the fact that most biting by the vector mosquitoes occurred at hours of the night when most people were in bed. The peak-biting activity of *Plasmodium*-transmitting mosquitoes in sub-Saharan Africa was known to be between 22.00 hrs and 05.00 hrs. Sleeping under an ITN therefore reduced the risk of bite and enhanced the killing of mosquitoes that are likely to transmit parasites. At about the mid-2000s however, evidence emerged of change in the biting pattern of *Anopheles funestus* following massive deployment of long-lasting insecticide nets (LLINs). Evidence is now firmly established, and this adaptive behaviour of the vector now poses a challenge to the use of ITN in malaria vector control and elimination.

Notwithstanding the aforementioned, sleeping under an ITN remains one of the most cost-effective interventions in malaria control. In rural northern Ghana, the estimated cost per discounted healthy life-year gained is US $73.5 (Binka et al. 1997). In other areas, the cost per death averted could be as high as US $1,449 (Yukich et al. 2009).

To promote widespread use of ITNs, free distribution is undertaken in several endemic countries. Between 2014 and 2016, a total of 582 million ITNs (505 million in sub-Saharan Africa) were delivered by manufacturers globally. In Africa 75% of ITNs were distributed through mass distribution campaigns. Household ownership of at least one ITN has increased from 50% in 2010 to 80% in 2016. More people at risk of malaria in Africa are sleeping under an ITN. In 2016, 54% of the population was protected by this intervention, an increase from 30% in 2010 (WHO 2018a). The scale-up in the distribution of ITNs is reported to account for the substantial reduction in malaria transmission, and related morbidity and mortality in sub-Saharan Africa (Bhatt et al. 2015).

Two major challenges confront the continued effectiveness of ITN use in malaria control. First is the emerging gap between ITN possession and use (Lengeler and Snow 1996). In rural northern Ghana, ITN use is found to depend on cost and time of the year (Binka and Adongo 1997). Free mass distribution of ITNs has overcome the problem of cost in many places and has led to high levels of household possession though there is still inadequate access to ITNs even in the household with these ITNs. These high levels of possession are however not matched with usage (Doda et al. 2018).

A second major challenge is the spread of insecticide resistance to the chemicals used to impregnate nets (Ranson and Lissenden 2016; Kisinza et al. 2017; Rakotoson et al. 2017). Pyrethroids are the most commonly used insecticides. Resistance to pyrethroids is now widespread in *An. Gambiae* and *An. Funestus* mosquitoes, the major malaria vectors. A breakthrough in efforts to overcome ITN insecticide resistance was recently reported in a trial led by the London School of Hygiene & Tropical Medicine. The trial involved 15,000 children in Tanzania and it found that compared with nets impregnated with pyrethroid, nets impregnated with piperonyl butoxide reduced malaria infections by 44% (Protopopoff et al. 2018). This finding has prompted WHO to revise its recommendations on ITN production and use. It now recommends the use of nets sprayed with piperonyl butoxide in areas in Africa where mosquitoes are pyrethroid-resistant. A trial that is evaluating the effects of using combinations of nets in currently underway in Burkina Faso.

Indoor residual spraying

IRS is one of the oldest interventions in malaria control and elimination. Literally put, it is a ‘recycled’ and ‘repackaged’ malaria control intervention. With dichlorodiphenyltrichloroethane (DDT) as insecticide, IRS was the principal method by which malaria was eliminated in many countries between the 1940s and 1960s. An estimated 700 million people mainly in Europe, Asia and Latin America were freed of the risk of malaria within a 20-year period (Mabaso et al. 2004; Africa WHOROf 2007). IRS involves spraying the inside walls of dwellings with insecticides. The technical basis is that the inner walls of dwellings retain doses of insecticide that are sufficiently lethal to kill mosquito that rest on these walls.

In the years that followed the WHO-led efforts in the 1940s through the 1960s, IRS got used sparingly amid concerns that DDT use in sub-Saharan Africa would cause damage to ecology and poisoning within the food chain. Eventually DDT was banned from use as an IRS insecticide. The development of pyrethroids as alternative re-ignited interest in IRS but this was short-lived as insecticide resistance reduced effectiveness. An epidemic of malaria in South Africa in 1999–2000 however compelled authorities in the country to obtain international exemption to use DDT in an IRS-based malaria elimination programme. The success of this programme re-ignited interest in the use of IRS and DDT in countries in the southern Africa region (Mabaso et al. 2004; Channa et al. 2012a, 2012b). Insecticides that are currently being used in IRS include DDT, pyrethroids (lambdacyahlothrin, deltamethrin, alphacypermethrin), organophosphates (malathion, pirimiphosmethyl) and carbamates (bendiocarb). In the southern Africa Region DDT remains effective and populations-at-risk in Botswana, Namibia, South Africa, and Swaziland are being protected with this intervention (Smith Gueye et al. 2016a, 2016b). Resistance to cheaper insecticides and the need to shift to more expensive chemicals is slowing IRS implementation in the rest of sub-Saharan Africa (WHO 2018a). According to the 2018 World Malaria Report, IRS protection declined from a peak of 5% in 2010 to 3% in 2017, with decreases seen across all WHO regions. In the WHO African Region, coverage dropped from 80 million people at risk in 2010 to 45 million in 2016 (WHO 2018a). The intervention is also particularly popular with households because it can be disruptive of normal life.

Larval source management

Larval source management (LSM) refers to the targeted management of mosquito breeding sites, with the objective of reducing the number of mosquito larvae and pupae. It is considered by WHO to be a supplementary measure in malaria vector control. The four types of LSM are:

Habitat modification, i.e. permanent alteration to the environment, e.g. land reclamation;
Habitat manipulation, i.e. a recurrent activity, e.g. flushing of streams;
Larviciding: the regular application of biological or chemical insecticides to water bodies;
Biological control: the introduction of natural predators into water bodies.

For reason that are not always clear, LSM receives very little mention in global malaria control and elimination. This is even though LSM was a major intervention in malaria eradication efforts in southern Europe in the early 1970s, Cuba, other parts of Latin America and parts of north Africa. Its effectiveness has been demonstrated in areas in Africa where it has been well-implemented. In 2009, the Governments of Cuba and Venezuela entered into agreement with countries in the Economic Community of West African States (ECOWAS), to expand larviciding activities through the provision of financial technical support, including the establishment of African based factories for larviciding products (Newman et al. 2013). Beyond the current efforts spearheaded by Cuba, there appears to be very little global interest in the implementation of LSM. There is only one mention of LSM in the 2018 World Malaria Report.

Chemoprophylaxis

Antimalarials play an important role in malaria control when they are used in chemoprophylaxis. This is targeted at the most vulnerable including non-immune travellers to malaria-endemic countries. Current major interventions in this regard are intermittent preventive treatment in pregnant women (IPTp) and seasonal malaria chemoprophylaxis (SMC).

Intermittent preventive treatment in pregnancy

The risk malaria poses to mother and child during pregnancy has always been appreciated and chemoprophylaxis has been part of the management of pregnancies in endemic countries for many years. In the era when chloroquine was effective, weekly subtherapeutic doses was recommended to prevent malaria during pregnancy (Geelhoed et al. 2001). The mass

presumptive use of subtherapeutic doses and limited adherence contributed to parasite resistance to chloroquine.

Trials conducted in the late 1990s to early 2000 to explore alternative drugs led to the choice of sulfadoxine-pyrimethamine (SP). In clinical trials, the use of SP in IPTp reduced the risk of moderate-to-severe anaemia, antenatal parasitaemia, placental parasitaemia, spontaneous abortions and low birthweight by 40%, 62%, 55%, 39%, and 19%, respectively (Wilson et al. 2011; Desai et al. 2018). Currently WHO recommends that in areas of moderate-to-high malaria transmission, IPTp with SP is given to all pregnant women at each scheduled antenatal care visit after the first trimester. The first IPTp-SP dose should be administered as early as 13 weeks' gestation and each subsequent dose should be at least one month apart. The last dose may be administered up to the time of delivery (WHO 2015b).

Logistic and other reasons continue to hamper effective implementation of ITPp-SP in many countries in sub-Saharan Africa. According to the 2018 World Malaria Report, an estimated 22% of eligible pregnant women received the recommended three or more doses of IPTp in 2017. This was up from 17% in 2015, 19% in 2016, and 0% in 2010. Coverage varies considerably by country. It is only Zambia that has coverage of the third ITPp dose (IPTp3) was 50% or more as at 2017 (WHO 2018a).

With continued concern about parasite resistance to SP, dihydroartemisinin-piperaquine looks the most promising as potential alternative or replacement antimalarial. Trials have demonstrated that it is well tolerated, effective, and acceptable, although it is also vulnerable to development of artemisinin resistance (Chico and Moss 2015; Kakuru et al. 2016; Tarning 2016; Rogerson et al. 2018).

Seasonal malaria chemoprevention

In many parts of the world where malaria is a major cause of morbidity and mortality, transmission is highly seasonal. This pattern offers the opportunity to implement interventions that are targeted at the time of the year when incidence of malaria is highest. This principle underlines SMC. This is an intervention in malaria control where young children are provided with a treatment course of an effective antimalarial combination at monthly intervals for 3 or 4 months during the peak period of malaria transmission.

In clinical trials, this approach reduced the incidence of uncomplicated malaria by 82% and of severe malaria by 82%. The WHO in 2012 recommended that SMC with SP and amodiaquine (AQ) be used in children aged ages 3–59 months in areas of sub-Saharan Africa where malaria transmission is highly seasonal, and parasites are sensitive to these drugs (Cairns et al. 2012; Matangila et al. 2015; Coldiron et al. 2017; Greenwood 2017) (Fig. 8.15.3). There is growing evidence of the need to extend this intervention to older children due to the benefits being shown in that particular age group and the observed increasing burden of malaria in children older than 5 years. A similar principle has been proposed for the deployment of malaria vaccines that have high initial efficacy, but which provide only a short period of protection. An on-going trial in Burkina Faso and Mali is exploring this hypothesis (Greenwood 2017; Greenwood et al. 2017).

Intermittent preventive treatment in infants

Intermittent preventive Treatment for Infants (IPTi) is the administration of a full therapeutic course of SP delivered through the Expanded Programme on Immunization at defined intervals corresponding to immunization schedules (usually at 10 weeks, 14 weeks, and nine months of age) to infants at risk of falciparum malaria in areas of high-to-medium transmission (Gosling et al. 2009).

IPTi reduces clinical malaria, anaemia, and severe malaria in the first year of life. Despite WHO recommending IPTi in 2009, only Sierra Leone has incorporated it into policy and implemented the intervention; albeit on a limited scale.

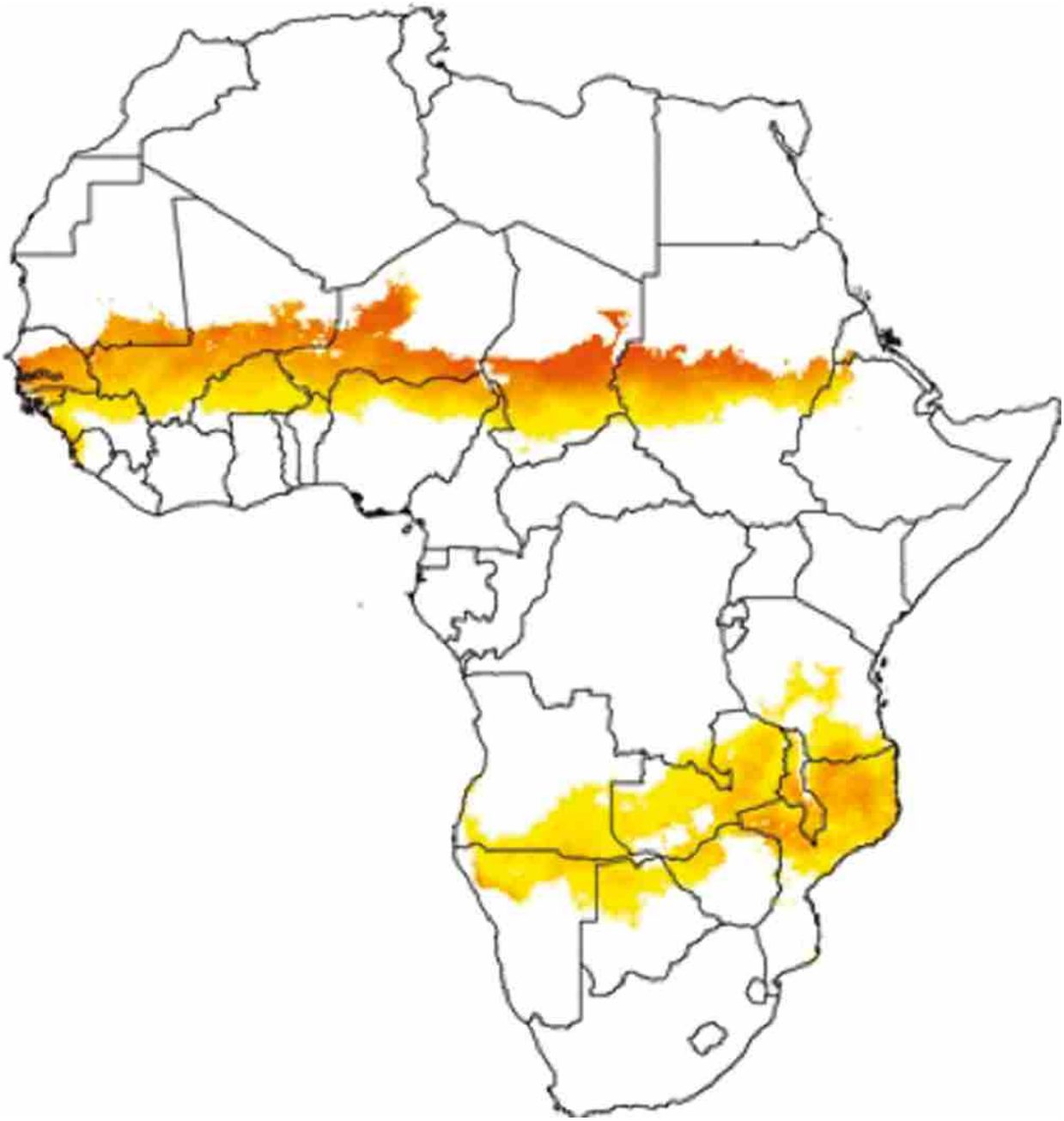


Fig. 8.15.3 A map of sub-Saharan Africa showing the areas where malaria transmission is likely to be highly seasonal.

Reprinted from Greenwood B. et al (2017) ‘Seasonal vaccination against malaria: a potential use for an imperfect malaria vaccine’. *Malar. J.* 18(162) under the Creative Commons Attribution 4.0 International License (CC BY 4.0).

Mass drug administration

Mass drug administration (MDA) consists in the administration of a full dose of antimalarial treatment, irrespective of the knowledge of symptoms or presence of infection, to an entire population in a given area, except those in whom the medicine is contraindicated. In the context of transmission reduction, MDA aims to provide therapeutic concentrations of antimalarial drugs to as large a proportion of the population as possible in order to cure asymptomatic infections and to prevent re-infection during the period of post-treatment prophylaxis. To impact on transmission, MDA requires high coverage of the target population which, in turn, demands a high level of community participation and engagement.

MDA has played an important role in the control and elimination of neglected tropical diseases (NTDs) such as trachoma, lymphatic filariasis, and onchocerciasis. WHO recommends MDA to achieve the following: (1) Interrupt falciparum malaria transmission in areas where elimination is the target; (2) To reduce the risk for spread of multidrug resistance in the Greater Mekong area; (3) Reduction in malaria burden during malaria epidemics and exceptional complex emergencies. Examples of exceptional complex emergencies was the MDA undertaken during the Ebola outbreak in West Africa (2014–2016) and the outbreak that followed in the Democratic Republic of Congo.

Malaria and travel medicine

There are 91 countries where international travellers may be at risk of malaria (WHO 2018a). For non-immune travellers to tropical regions, the risk of malaria is an ever-present danger. Non-immune here refers to persons who are ordinarily resident in temperate, non-malaria-endemic countries. Persons who were originally resident in malaria-endemic countries but have travelled and stayed out in a non-endemic country for more than six months are also considered as non-immune. For such persons, little or no immunity exist against malaria and they are at risk of malaria developing in the very severe forms. The disease can take a rapidly deteriorating course and may develop into cerebral malaria with a few hours to a day.

Non-immune travellers to tropical countries need to seek expert advice. They may be required to begin antimalarial chemoprophylaxis while at their countries of origin, continue during the period of stay in the tropical region and then for a period when they are back to their countries of regular

residence. *Malarone*, containing atovaquone and proguanil is one of the most commonly used antimalarial prophylaxis. Other antimalarials used in antimalarial prophylaxis are mefloquine and doxycycline. An emerging alternative to malaria chemoprophylaxis is 'standby emergency treatment' (SBET). This consist of providing travellers with emergency malaria treatment for self-administration in case of fever when no medical attention is available, or eventually for use under medical supervision after a confirmed malaria diagnosis (Genton and D'Acremont 2017).

As much as feasible pregnant women and children aged less than 5 years should put off non-essential travel to malaria-transmission areas. Where travel is unavoidable, appropriate advice should be sought on the best forms of malaria protection and chemoprophylaxis before, during and after travel (Irvine et al. 2011). This includes avoiding mosquito bites through the use of ITNs, protective clothes, and mosquito repellents.

Malaria vaccines

The current highly invigorated efforts to control, eliminate, and possibly eradicate malaria will have a greater likelihood of success if an effective vaccine were to become available to complement other interventions. The anticipation that this will be achieved has never been greater as gene technology is leading to identification of more target *plasmodium* antigens. The three main mechanisms being explored in malaria vaccine development are: (1) anti-infection; (2) transmission-blocking; and (3) antidisease. The most advanced malaria vaccine to date is the anti-infection RTS,S/AS01 vaccine developed by GlaxoSmithKline (GSK).

RTS,S/AS01 vaccine

The RTS,S/AS01 vaccine is an example of an anti-infective mechanism in malaria vaccine development. It is a monovalent recombinant protein vaccine that targets a fragment of the circumsporozoite protein parasite antigen. It has been evaluated in large phase 3 trials in 11 sites across Africa. These trials took place between 2009 and 2013 and showed the vaccine to offer partial protection against clinical malaria in the older age group (5–17 months). Efficacy in children who received three doses plus a booster on months 0, 1, 2, and 20 was 36.3% between month 0 and study end (median

48 months per child). This was a decline from 50.4% achieved 14 months from the first dose. No significant efficacy against severe malaria was noted in the 6–12 week age group; not even with a booster over the duration of the trial (Moorthy and Okwo-Bele 2015; Rts 2015). The RTS,S/AS01 vaccine has received a positive scientific opinion from the European Medicines Agency in July 2015 and WHO has recommended limited, pilot deployment in Ghana, Kenya, and Malawi. The deployment is being undertaken under the banner of the Malaria Vaccine Implementation Programme (MVIP) and is targeted at children who are at least 5 months of age. The objective is to assess the feasibility of administering the needed four doses of the vaccine in children, and whether administrative will be safe and effective in reducing the incidence of malaria among vaccinated populations. The programmes in Malawi and Ghana were launched in April 2019.

Transmission-blocking vaccines (TBV)

TBV are designed to target the sexual stages of parasite development in the mosquito. They utilize antigens expressed by gametocytes, gametes, zygotes and ookinetes to induce antibodies capable of attacking the parasite in the mosquito and impairing its viability and further development. TBVs that have reached human trials include Pfs25 and its Plasmodium vivax ortholog Pvs25, and Pfs230 (Ogobara et al. 2018).

The antigen Pfs25 has been the most evaluated. It is made up of a glycosylphosphatidylinositol-linked protein which is expressed on the surface of ookinetes. Its configuration is complex and eluding full elucidation. An improved understanding of the structural conformation of the antigen and the finding of a good matching adjuvants will likely make Pfs25 and other TBVs important additions to existing tools for malaria elimination and eradication.

Pre-erythrocytic vaccines

Pre-erythrocytic vaccines target the sporozoite and liver stages of Plasmodium and are designed to prevent disease and interrupt transmission. The leading approach in this mechanism is the vaccination using whole sporozoites which are genetically attenuated parasites, radiation attenuated sporozoites, and/or wild-type sporozoites administered in combination with

chemoprophylaxis. The principle of immune response following vaccination with whole sporozoites have been established (Coelho et al. 2017). As with the TBVs, the principal challenge is improving upon the level of immune response that is induced.

The PfSPZ is a non-replicating irradiated whole sporozoite vaccine which has entered human trials. In Mali, it proved to be efficacious against naturally transmitted parasites. A phase 3 programme is currently being developed.

It has been proposed that an imperfect malaria vaccine may still be useful if used in seasonal vaccination against malaria (Greenwood et al. 2017). While the findings of the RTS,S/AS01 vaccine moves the world closer to the goal of an effective enough malaria vaccine, the rapid waning of immunity and the lack of effect in children aged 6–12 weeks proved that it did not get the human immune system to do more than it will naturally do. Proving the principle is nevertheless a significant breakthrough and there is optimism that the world is not too far away from achieving the target of a truly effective and practical malaria vaccine.

Surveillance

Public health surveillance is the continuous, systematic collection, analysis, and interpretation of health-related data needed for the planning, implementation, and evaluation of public health practice. An effective surveillance aims to achieve the following:

Provide early warning for impending public health emergencies;
Benchmark set goals and track progress in the outcome of intervention(s); and
Accurately define the epidemiology of the health problem to enable priority-setting and informed public health policy and strategies.

The Global Technical Strategy (GTS) for Malaria 2016–2030 emphasizes the need for universal coverage of core malaria interventions for all populations at risk and highlights the importance of using high-quality surveillance data for decision-making. Surveillance is one of the three pillars of the GTS for Malaria 2016–2030, transformed from a routine activity of

the health system into a core intervention in all settings of malaria endemicity. The new 12 core principles of malaria surveillance are:

Accurate parasitological diagnosis of a malaria case should be the foundation of a malaria surveillance system.
Major components of a malaria surveillance system should be integrated within routine health management information systems (HMIS).
National standard operating procedures should be based on country needs and WHO recommendations.
Front-line staff involved in the detection, recording, and reporting of cases should be the first users of data from the malaria surveillance system.
Surveillance systems should address the within-country heterogeneity of malaria.
Necessary investments in surveillance and system transition in response to changing malaria burden should be made.
All malaria surveillance data should be linked to a decision at appropriate levels in the health system.
Concerted effort should be made to include data from all sectors, including the private sector.
Malaria surveillance may become the broad responsibility of general health services following the interruption of transmission. Malaria programme surveillance systems should however be supported to ensure monitoring.
Innovation and advances in technology should be incorporated in malaria surveillance based on empirical evidence of additional benefits, cost, and sustainability.
Malaria surveillance should incorporate regular monitoring of the efficacy of antimalarial drugs used in treatment.
Malaria surveillance systems should be assessed routinely for accuracy, reliability, completeness, precision, timeliness, integrity, and actions taken based on the results of surveillance.

The sources of data for malaria surveillance are be routine (health facility based and/or community based) or sentinel sites. Cases can be identified through an active or passive approach. The intensity of either or a mix of both is determined by the malaria transmission setting. Active approaches are increasingly adopted as transmission levels decline.

A set of 46 indicators have been recommended. These indicators consider varying levels of resources availability for malaria control (programme

financing, commodities), levels of service provision (intervention coverage), performance of systems for surveillance, monitoring, and evaluation; and trends in transmission, infection, and disease. Twelve (one for input, six for outcome, and five for impact) out of the 46 indicators are regarded as core. These are outlined in Table 8.15.3.

With surveillance a core malaria intervention, WHO in March 2018 published a thoroughly detailed reference manual for malaria surveillance, monitoring, and evaluation. This manual addresses the evidence-base for making malaria surveillance a core intervention, how malaria surveillance systems should be established, new concepts and practices of malaria surveillance, surveillance for antimalarial drug efficacy and drug resistance, approaches to entomological surveillance, and vector control monitoring and evaluation. It also addressed issues of malaria early warning, detection, and response to outbreaks, monitoring, and evaluation of national programmes. This manual is a highly-recommended text for all health professionals whose work impact the continuum of malaria control through elimination (WHO 2018b).

Table 8.15.3 The 12 recommended core indicators for monitoring malaria programmes and implementation of the Global Technical Strategy for malaria (2016–2030)

Indicator			Transmission intensity			Data source		
High	**Low and very low**	**Elimination or prevention of re-establishment**	**Routine reporting system**	**Health facility survey**	**Household survey**			
Inputs	Financing	Expenditure per capita for malaria control and elimination	**X**	**X**	**X**	**X**		
Outcome	Vector control	Proportion of population at risk sleeping under an ITN or living in house sprayed by IRS in the previous 12 months	**X**	**X**		**C**		**C**
	Chemoprevention	Proportion of pregnant women who received three or more doses of IPTp	**X**			**X**		**X**
	Case detection	Proportion of children <5 years with fever in the previous 2 weeks for whom advice or treatment was sought	**X**	**D**				**X**
	Diagnostic testing	Proportion of patients with suspected malaria who received a parasitological test	**X**	**X**		**X**	**X**	

Indicator			Transmission intensity			Data source		
High	Low and very low	Elimination or prevention of re-establishment	Routine reporting system	Health facility survey	Household survey			
	Treatment	Proportion of patients with confirmed malaria who received first-line antimalarial treatment according to national policy	**X**	**X**	**X**	**X**	**X**	
	Surveillance	Proportion of malaria cases detected by surveillance systems	**X**	**X**	**X**	**C**		**C**
Impact	Prevalence	Parasite prevalence: proportion of population with evidence of infection with malaria parasites	**X**	**D**				**X**
	Incidence	Malaria case incidence: number and rate per 1,000 people per year	**X**	**X**	**X**	**X**		
	Mortality	Malaria mortality: number and rate per 100,000 people per year	**X**	**D**	**D**	**X**		
	Elimination	Number of areas and countries that have eliminated malaria since 2015			**X**	**X**		

Indicator			Transmission intensity			Data source		
High	Low and very low	Elimination or prevention of re-establishment	Routine reporting system	Health facility survey	Household survey			
	Prevention of re-establishment	Number of areas and countries that were malaria-free in 2015 in which malaria has been re-established			**X**	**X**		

X Indicator highly relevant to the setting; **C** requires data from both routine systems and household surveys; **D** Indicator potentially relevant to the setting.

Reprinted from Malaria surveillance, monitoring & evaluation: a reference manual. Geneva: World Health Organization; 2018. Licence: CC BY-NC-SA 3.0 IGO, with permission from the World Health Organization.

Measures of malaria frequency

The number of malaria cases per at-risk population (incidence) and the proportion of a population carrying malaria parasite in the blood (prevalence) are the most commonly used proxy indicators of the intensity of malaria endemicity. The advantage these two indicators have over other indicators are that they are relatively simple to compute and inexpensive to determine, and report within a short time. The input for the two indicators can be obtained from routine health information systems and population-based surveys. For reasons elaborated further in this section, both indicators are imprecise measures of malaria transmission (Oduro et al. 2011, 2016).

Incidence of malaria

Although absolute numbers of malaria cases, inpatients and deaths can be used to estimate trends over time and to identify places in which the problem of malaria is greatest, these are less useful for assessing which populations are at highest risk for acquiring malaria because of different population sizes. Incidence of malaria gives indication of risk of acquiring malaria and it is given by dividing the number of new cases of malaria over a given

period by the population at risk. The ascertainment of both measures has challenges in most malaria-endemic countries.

Coverage in diagnostic testing is still not universal and, in many places, there is a considerable mix of confirmed and unconfirmed cases counted among malaria cases. It is not always clear from routine health service records the number of cases of confirmed malaria and what proportion of unconfirmed cases could be estimated to be truly malaria. Another challenge is the so-called iceberg phenomenon. Access to health facilities remains a challenge in many rural and urban communities in malaria-endemic countries. Access is often limited by financial, geographical, and cultural constraints and only a fraction of people who fall sick seek care from health facilities. Health service data that are used to compute the incidence of malaria therefore tends to be biased towards underestimating the risk of disease.

To determine the population that is risk, WHO recommends the use of estimates of population size published by a relevant government department such as a statistics office, planning bureau, or census office. However, the breakdown in vital registration and the lack of systems for continuously tracking populations makes it difficult to report with certainty that figures obtained from such official sources truly represent at-risk populations. Despite these limitation, incidence of malaria when applied over a reasonable length of time and over large populations gives an accurate measure of the magnitude of risk of infection. Trend data also conveys a strong enough impression of the direction of risk.

Prevalence of malaria

The second measure of malaria frequency that is readily determined is the prevalence of parasitaemia in the population. Traditionally this represents the proportion of children aged 2–10 years who are found to have malaria parasites in their blood during surveys. National malaria programmes routinely use this to monitor the effect of interventions. The procedures have been incorporated into various demographic and health surveys, multiple indicators clusters surveys, and malaria indicator surveys. The determination of parasitemia is made at point-of-interview by trained, unskilled field workers, and the availability of mRDTs has made the procedures less cumbersome and less sophisticated.

A challenge with survey-determined malaria prevalence however is that it fails to adequately measure transmission and the burden of disease. In many high malaria transmission settings, a large proportion of children carry malaria parasites in their blood without being ill (White 2017). This phenomenon makes imprecise inferences between levels of parasitaemia and risk of disease. Another challenge is the inherent limitation in the use of mRDTs. The persistence of HRP2 antigen in the blood similarly undermines any inference that a positive mRDT connote risk of malaria. There have been attempts to establish modelling-based link between prevalence and infection rates. However, differences in model structure and calibration data has meant that no consensus yet exists on the optimal application in disease-burden estimation (Patil et al. 2009). With these caveats, the ease of application and ability to establish a trend through year-on-year consistency in procedures make the determination of parasite prevalence an attractive and pragmatic proxy measure of endemicity and impact of interventions.

Entomological surveillance indicators

The main entomological indicators are categorized into five groups:

Adult vector composition (species occurrence and density);
Adult vector behaviour (human blood index, human biting rate, biting time, biting location, resting location);
Adult vector resistance to insecticides (resistance frequency, status, intensity, and mechanisms);
Immature vector aquatic habitats (habitat availability and occupancy, larval density); and
Proxies for transmission (sporozoite rate, entomological inoculation rate (EIR), receptivity).

Two key indicators with direct implication on malaria transmission are the EIR and vectorial capacity.

Entomological inoculation rate

The most direct measure of malaria transmission is the EIR. The EIR measures the average number of infectious bites that a person receives per

year in a specified population. To compute it, two parameters need to be established. These are the human landing rate (HLR) and sporozoite carriage rate. Determining the HLR involves an elaborate process of capturing all mosquitoes that land on a person (human biting rate) throughout the night and expressing this per person per night. All captured mosquitoes are sent to the laboratory where the salivary glands of each is dissected to determine the proportion of mosquitoes that carry sporozoites (sporozoite rate—SR) (Morrow and Moss 2017). The EIR is the arithmetic product of HLR and SR. The process of determining the EIR in a given location is expensive and technically demanding. It is therefore not part of routine malaria surveillance in endemic countries. Often EIRs are only available in the context of research, or specific well-funded programmes. The day-biting characteristics of some species of the vector is becoming an increasing problem.

Vectorial capacity

Vectorial capacity (VC) is another basic indicator of malaria transmission. It is the number of new infections that the population of a given vector would induce per case per day at a given place and time, if the human population is and remains fully susceptible to malaria. It measures the rate of potentially infective contact and depends on mosquito longevity and susceptibility to *Plasmodium* infection.

It is critical for malaria transmission for the female *Anopheles* mosquito to live long enough to enable the sporogonic phase of the life cycle of the Plasmodium parasite to be completed and for sporozoites to be injected during a blood meal. Like EIR, VC provides a good measure of malaria transmission, but it is very technical to compute and therefore not routinely used.

Gametocytes carriage and its implications

The gametocyte is the stage in the life cycle of the Plasmodium parasite that is infective to the mosquito. Gametocytes are therefore vital to malaria transmission. However, because gametocyte carriage is asymptomatic (may be symptomatic in *P. vivax* malaria), this stage of arrested development in humans serves as reservoir for sustained malaria infection and re-infection. Developing antimalarial drugs that target gametocytes is an important part of

the agenda to eliminate malaria. Of recent interest is the gametocidal activity role of methylene blue (Lu et al. 2018). Adding a single dose of 0·25 mg/kg primaquine to amodiaquine and SP or 3 days of 15 mg/kg per day methylene blue to dihydroartemisinin-piperaquine has been found to prevent *P. falciparum* transmission (Dicko et al. 2018). The 'tolerant' existence of the gametocyte stage in human provides insight into how natural immunity evolves in malaria. It has been hypothesized that some level of gametocytaemia is required to sustain natural immunity (Stone et al. 2018).

Control, elimination, and eradication: the continuum

There are four distinct phases in the path towards country and global malaria-free status. These are control, pre-elimination, elimination, and prevention of reintroduction. A well-detailed process exists at WHO to define countries or regions as having transited from one phase into the other. This calls for clear, correct, and uniform use of the related descriptive terms. These have been supplied by WHO as follows:

Control: This is the reduction of disease incidence, prevalence, morbidity, or mortality to a locally acceptable level as a result of deliberate efforts. Continued intervention is required to sustain control.

Elimination: It is the interruption of local transmission (that is, reducing the rate of malaria cases to zero) of a specified parasite in a defined geographic area. Continued measures are required to prevent the re-establishment of transmission.

Eradication: It is the permanent reduction to zero of the worldwide incidences of infection caused by human malaria parasites as a result of deliberate efforts. Once eradication has been achieved, intervention measures would no longer be needed. The interchangeable use of the terms, elimination and eradication is no longer acceptable.

Malaria elimination

In 2016, WHO identified 21 countries (so-called E-2020 countries) with the potential to eliminate malaria by the year 2020 (Table 8.15.4). By 2018, 11 of these countries remained on track to achieve their elimination goals, while 10 had reported increases in indigenous malaria cases in 2017. Globally

however the elimination net keeps widening, with more countries moving towards zero indigenous cases. In 2017, 46 countries reported fewer than 10,000 such cases, up from 44 countries in 2016 and 37 countries in 2010. Between 2016 and mid-2019, the WHO Europe region, Paraguay, Algeria, and Argentina have been certified as malaria-free.

Table 8.15.4 Countries with the potential to eliminate local transmission of malaria by 2020, by WHO region (as at 2016)

WHO region	Countries
Africa region	Algeria, Botswana, Cabo Verde, Comoros, South Africa, Swaziland
Region of the Americas	Belize, Costa Rica, Ecuador, El Salvador, Mexico, Paraguay, Suriname
Eastern Mediterranean region	Iran (Islamic Republic of), Saudi Arabia
Southeast Asia region	Bhutan, Nepal, Timor-Leste
Western Pacific region	China, Malaysia, Republic of Korea

Certification

This is the WHO certification of a country as malaria-free. This certification is granted when local transmission of malaria by *Anopheles* mosquitoes has been halted for at least three consecutive years and an adequate system for preventing re-establishment of the disease has been established. The process of certification is managed at WHO by the Global Malaria Programme. It is a process that countries voluntarily submit to through application and providing the necessary proof. It is the WHO Director-General who grants the certification of malaria elimination (WHO 2016).

Malaria eradication

Over the past few years, discussions about malaria on the global scale has been dominated by issues related elimination and possible eradication, and

what it would take to achieve this feat. For many years after the failed attempt of the 1960s, the eradication of malaria was considered technically impossible. Of recent however, and particularly after the American billionaire-philanthropist Bill Gates declared eradication as a target of his foundation's malaria programme, a surge in momentum has brought back malaria eradication on the global agenda (Star 2015). Discussions have dramatically shifted from a blanket declaration of non-feasibility to which tools, if available, could make this target achievable. There is also a roll-back to an aspiration that malaria eradication can be achieved through elimination of the disease, country-by-country, region-by-region. This aspiration is given impetus in the vision of the GTS for Malaria 2016–2030, 'World-Free of Malaria'. The strategy targets elimination of malaria in 10, 20, and 35 countries by 2020, 2025, and 2030, respectively (WHO 2015c).

The Malaria Eradication Research Agenda (malERA) initiative has brought together some of the world's best malariologists to design a path in malaria research that is targeted at eradication. In the opinion of this group, malaria eradication is not feasible with current tools and state of knowledge (Alonso et al. 2011; Drugs mCGo 2011; Strategies mCGoI 2011). Among others, they advocate the need for drugs with the capacity to achieve Single Encounter Radical Cure and Prophylaxis (SERCaP). The target product profile (TPP) of such a drug should include radical cure, defined as elimination of all parasites (including the long-lived hypnozoites of *P. vivax* or *P. ovale* in the liver), suitability for mass administration (including administration to healthy subjects and the subsequent need of a very good safety profile), and prophylaxis for at least 1 month after treatment, to outlast the typical development period of *Plasmodia* parasites in anopheline mosquitoes. There is also the need to develop vaccines that will reduce and contribute to interruption of transmission, so-called TBVs. The vaccines in current development do not meet this standard as they aim at reducing morbidity and mortality, and not at blocking transmission (Drugs mCGo 2011; Vaccines mCGo 2011). To address in-country heterogenicity as the incidence of malaria declines will required new well-targeted tools for control and surveillance (Greenwood and Targett 2011). The drive towards malaria eradication is further buoyed by the formation of the Lancet Commission on Malaria Eradication in April 2018. This is a Commission of 24 experts from around the world, put together to develop a roadmap for malaria eradication (Chen et al. 2018).

Less-used terminologies

For many years, four levels of classification are used to describe malaria endemicity. These are holoendemic, hyperendemic, mesoendemic, and hypoendemic (coinciding with reducing levels of intensity and EIRs). Neither the 2018 World Malaria Report nor the WHO manual on malaria surveillance refers to this approach to classification. Currently, these terminologies are so much less-used that they are best to be avoided.

Global programme and strategy

Global Malaria Programme

The Global Malaria Programme (GMP) is WHO's lead agency in the fight against malaria. It is responsible for coordinating global efforts to control and eliminate malaria. The GMP supports Member States to achieve the GTS targets through:

Setting and disseminating global guidance and policies on malaria control and elimination;
Supporting countries as they adapt, adopt, and implement WHO global norms and standards for malaria control and elimination;
Assisting countries as they formulate national malaria strategic plans;
Helping countries develop robust malaria surveillance systems;
Responding to requests from countries to address biological and operational emergencies and bottleneck issues.

The GMP publishes the World Malaria Report each year. This is the most comprehensive report on the status of malaria in the world. It provides global and, country-by-country account of progress made to achieve set global targets. The GMP also produces a newsletter every two months that highlight latest news and developments in the work of the GMP. Subscription is available at https://www.who.int/malaria/news/sign_up_form/en/

Malaria Policy Advisory Committee

The Malaria Policy Advisory Committee (MPAC) was established in 2011 following an extensive review of the policy-setting process for malaria control and elimination, initiated by the WHO GMP (D'Souza and Newman 2012). It provides independent strategic advice to WHO on the following:

Appropriate malaria policies and standards based on data from malaria programme implementation by Member States and malaria control partners as well as reviews of the best available evidence;
Engagement of WHO in malaria-related initiatives;
Major issues and challenges to achieving global malaria goals; and
The identification of priority activities to address identified challenges.

The MPAC convenes twice a year in Geneva and is made of some of the world's foremost experts on malaria. It is supported by the GMP Guidelines Development Group and ad hoc Evidence Review Groups which support the policy making process. Among important policy and initiatives over which the MPAC has had profound influence has been those related to updated policy recommendations on the use of sulphadoxine-pyrimethamine for ITPp, as well as the use of single dose primaquine as a *Plasmodium falciparum* gametocytocide. It is on the recommendation of the MPAC that the current limited deployment of the RTS,S/AS01 malaria vaccine has been implemented in Ghana, Kenya, and Malawi.

GTS for Malaria 2016–2030

The GTS for Malaria 2016–2030 is a document produced by the GMP and adopted by the World Health Assembly in May 2015. It provides a comprehensive framework to guide countries and agencies in efforts to accelerate progress towards malaria elimination. The strategy has set the most ambitious targets for reductions in malaria cases and deaths since the malaria eradication era began. (Table 8.15.5) The timeline of 2016–2030 is aligned with the 2030 Agenda for Sustainable Development, endorsed in 2015 by all United Nations Member States. The path towards achieving these targets are anchored on three pillars and two supporting elements. The pillars are

Pillar 1. Ensure universal access to malaria prevention, diagnosis, and treatment
Pillar 2. Accelerate efforts towards elimination and attainment of malaria-free status
Pillar 3. Transform malaria surveillance into a core intervention

The two supporting elements of the GTS are (1) innovation and research and (2) a strong enabling environment.

Table 8.15.5 Goals, milestones, and targets of the Global Technical Strategy for malaria 2016–2030

Goals	Milestones		Targets
	2020	2025	2030
1. Reduce malaria mortality rates globally compared with 2015	>40%	>75%	>90%
2. Reduce malaria case incidence globally compared with 2015	>40%	>75%	>90%
3. Eliminate malaria from countries in which malaria is transmitted in 2015	At least 10 countries	At least 20 countries	At least 25 countries
4. Prevent re-establishment of malaria in all countries that are malaria-free	Re-establishment prevented	Re-establishment prevented	Re-establishment prevented

REFERENCES

Abuaku, B.K., Koram, K.A., and Binka, F.N. (2005). Antimalarial prescribing practices: a challenge to malaria control in Ghana. *Medical Principles & Practice*, 14(5), 332–7.

Acharya, P., Garg, M., Kumar, P., Munjal, A., and Raja, K.D. (2017). Host-parasite interactions in human malaria: clinical implications of basic research. *Frontiers of Microbiology*, **8**, 889.

Adjei, A., et al. (2016). Treatment outcomes in a safety observational study of dihydroartemisinin/piperaquine (Eurartesim((R))) in the treatment of uncomplicated malaria at public health facilities in four African countries. *Malaria Journal*, **15**, 43.

Afrane, Y.A., Githeko, A.K., and Yan, G. (2012). The ecology of Anopheles mosquitoes under climate change: case studies from the effects of deforestation in East African highlands. *Annals of the New York Academy of Sciences*, **1249**, 204–10.

Africa WHOROf (2007). *Implementation of Indoor Residual Spraying of Insecticides for Malaria Control in the WHO African Region Report World Health Organization Regional Office for*

Africa. Geneva: WHO.
Alonso, P.L., et al. (1991). The effect of insecticide-treated bed nets on mortality of Gambian children. *The Lancet*, 337(8756), 1499–502.
Alonso, P.L., et al. (2011). A research agenda to underpin malaria eradication. *PLoS Medicine*, 8(1), e1000406.
Autino, B., Corbett, Y., Castelli, F., and Taramelli, D. (2012). Pathogenesis of malaria in tissues and blood. *Mediterranean Journal of Hematology & Infectious Diseases*, 4(1), e2012061.
Baiden, F., et al. (2016). Effect of test-based versus presumptive treatment of malaria in under-five children in rural Ghana—a cluster-randomised trial. *PLoS One*, 11(4), e0152960.
Baiden, F., Webster, J., Owusu-Agyei, S., and Chandramohan, D. (2011). Would rational use of antibiotics be compromised in the era of test-based management of malaria? *Tropical Medicine & International Health*, 16(2), 142–4.
Baiden, F., et al. (2012). Accuracy of rapid tests for malaria and treatment outcomes for malaria and non-malaria cases among under-five children in rural Ghana. *PLoS One*, 7(4), e34073.
Baiden, R., et al. (2015). Prospective observational study to evaluate the clinical safety of the fixed-dose artemisinin-based combination Eurartesim(R) (dihydroartemisinin/piperaquine), in public health facilities in Burkina Faso, Mozambique, Ghana, and Tanzania. *Malaria Journal*, **14**, 160.
Baird, J.K., Valecha, N., Duparc, S., White, N.J., and Price, R.N. (2016). Diagnosis and treatment of plasmodium vivax malaria. *American Journal of Tropical & Medical Hygiene*, 95(6 Suppl), 35–51.
Barber, B.E., Grigg, M.J., William, T., Yeo, T.W., and Anstey, N.M. (2017). The treatment of plasmodium knowlesi malaria. *Trends in Parasitology*, 33(3), 242–53.
Bawah, A.A. and Binka, F.N. (2007). How many years of life could be saved if malaria were eliminated from a hyperendemic area of northern Ghana? *American Journal of Tropical & Medical Hygiene*, 77(6 Suppl), 145–52.
Beck-Johnson, L.M., Nelson, W.A., Paaijmans, K.P., Read, A.F., Thomas, M.B., and Bjornstad, O.N. (2013). The effect of temperature on Anopheles mosquito population dynamics and the potential for malaria transmission. *PLoS One*, 8(11), e79276.
Bell, D., Peeling, R.W., and Pacific/TDR WH-ROftW (2006a). Evaluation of rapid diagnostic tests: malaria. *Nature Reviews Microbiology*, 4(9 Suppl), S34–8.
Bell, D., Wongsrichanalai, C., Barnwell, J.W. (2006b). Ensuring quality and access for malaria diagnosis: how can it be achieved? *Nature Reviews Microbiology*, 4(9 Suppl), S7–20.
Beshir, K.B., et al. (2017). Plasmodium falciparum parasites with histidine-rich protein 2 (pfhrp2) and pfhrp3 gene deletions in two endemic regions of Kenya. *Science Reports*, 7(1), 14718.
Bhatt, S., et al. (2015). The effect of malaria control on Plasmodium falciparum in Africa between 2000 and 2015. *Nature*, 526(7572), 207–11.
Binka, F. and Akweongo, P. (2006). Prevention of malaria using ITNs: potential for achieving the millennium development goals. *Current Molecular Medicine*, 6(2), 261–7.
Binka, F.N. and Adongo, P. (1997). Acceptability and use of insecticide impregnated bednets in northern Ghana. *Tropical Medicine & International Health*, 2(5), 499–507.
Binka, F.N., Indome, F., and Smith, T. (1998). Impact of spatial distribution of permethrin-impregnated bed nets on child mortality in rural northern Ghana. *American Journal of Tropical & Medical Hygiene*, 59(1), 80–5.
Binka, F.N., et al. (1996). Impact of permethrin impregnated bednets on child mortality in Kassena-Nankana district, Ghana: a randomized controlled trial. *Tropical Medicine & International Health*, 1(2), 147–54.
Binka, F.N., Mensah, O.A., and Mills, A. (1997). The cost-effectiveness of permethrin impregnated bednets in preventing child mortality in Kassena-Nankana district of Northern Ghana. *Health Policy*, 41(3), 229–39.

Brooker, S., Clarke, S., Snow, R.W., and Bundy, D.A. (2008). Malaria in African schoolchildren: options for control. *Transactions of the Royal Society of Tropical Medicine & Hygiene*, 102(4), 304–5.

Bruce-Chwatt, L.J. (1974). Resurgence of malaria and its control. *Journal of Tropical Medicine & Hygiene*, 77(4), s62–6.

Buffet, P.A., et al. (2011). The pathogenesis of Plasmodium falciparum malaria in humans: insights from splenic physiology. *Blood*, 117(2), 381–92.

Cairns, M., et al. (2012). Estimating the potential public health impact of seasonal malaria chemoprevention in African children. *Nature Communications*, **3**, 881.

Caminade, C., et al. (2014). Impact of climate change on global malaria distribution. *Proceedings of the National Academy of Sciences U S A*, 111(9), 3286–91.

Channa, K.R., et al. (2012a). Regional variation in pesticide concentrations in plasma of delivering women residing in rural Indian Ocean coastal regions of South Africa. *Journal of Environmental Monitoring*, 14(11), 2952–60.

Channa, K., Rollin, H.B., Nost, T.H., Odland, J.O., and Sandanger, T.M. (2012b). Prenatal exposure to DDT in malaria endemic region following indoor residual spraying and in non-malaria coastal regions of South Africa. *Science of the Total Environment*, **429**, 183–90.

Chen, I., Cooney, R., Feachem, R.G.A., Lal, A., and Mpanju-Shumbusho, W. (2018). The Lancet Commission on malaria eradication. *The Lancet*, 391(10130), 1556–8.

Chico, R.M. and Moss, W.J. (2015). Prevention of malaria in pregnancy: a fork in the road? *The Lancet*, 386(10012), 2454–6.

Clarke, S.E., et al. (2008). Effect of intermittent preventive treatment of malaria on health and education in schoolchildren: a cluster-randomised, double-blind, placebo-controlled trial. *The Lancet*, 372(9633), 127–38.

Clerk, C.A., et al. (2008). A randomized, controlled trial of intermittent preventive treatment with sulfadoxine-pyrimethamine, amodiaquine, or the combination in pregnant women in Ghana. *Journal of Infectious Diseases*, 198(8), 1202–11.

Coelho, C.H., Doritchamou, J.Y.A., Zaidi, I., and Duffy, P.E. (2017). Advances in malaria vaccine development: report from the 2017 malaria vaccine symposium. *NPJ Vaccines*, **2**, 34.

Cohen, J.M., et al. (2012). Malaria resurgence: a systematic review and assessment of its causes. *Malaria Journal*, **11**, 122.

Coldiron, M.E., Von Seidlein, L., and Grais, R.F. (2017). Seasonal malaria chemoprevention: successes and missed opportunities. *Malaria Journal*, 16(1), 481.

Collins, W.E. and Jeffery, G.M. (2007). Plasmodium malariae: parasite and disease. *Clinical Microbiology Reviews*, 20(4), 579–92.

Collins, W.E. and Jeffery, G.M. (2005). Plasmodium ovale: parasite and disease. *Clinical Microbiology Reviews*, 18(3), 570–81.

Cox, F.E. (2010). History of the discovery of the malaria parasites and their vectors. *Parasite Vectors*, 3(1), 5.

Cox, F.E.G. (2017). The Golden Age of parasitology—1875–1925: the Scottish contributions. *Parasitology*, 144(12), 1567–81.

D'Acremont, V., et al. (2014). Beyond malaria—causes of fever in outpatient Tanzanian children. *New England Journal of Medicine*, 370(9), 809–17.

D'Souza, B.J. and Newman, R.D. (2012). Strengthening the policy setting process for global malaria control and elimination. *Malaria Journal*, **11**, 28.

D'Acremont, V., Lengeler, C., and Genton, B. (2010). Reduction in the proportion of fevers associated with Plasmodium falciparum parasitaemia in Africa: a systematic review. *Malaria Journal*, **9**, 240.

D'Alessandro, U., et al. (1995). A comparison of the efficacy of insecticide-treated and untreated bed nets in preventing malaria in Gambian children. *Transactions of the Royal Society of Tropical Medicine & Hygiene*, 89(6), 596–8.

de Oliveira, A.M., et al. (2009). Performance of malaria rapid diagnostic tests as part of routine malaria case management in Kenya. *American Journal of Tropical Medicine & Hygiene*, 80(3), 470–4.

Desai, M., et al. (2018). Prevention of malaria in pregnancy. *Lancet Infectious Diseases*, 18(4), e119–e32.

Dhangadamajhi, G., Kar, S.K., and Ranjit, M. (2010). The survival strategies of malaria parasite in the red blood cell and host cell polymorphisms. *Malaria Research & Treatment*, **2010**, 973094.

Dicko, A., et al. (2018). Efficacy and safety of primaquine and methylene blue for prevention of Plasmodium falciparum transmission in Mali: a phase 2, single-blind, randomised controlled trial. *Lancet Infectious Diseases*, 18(6), 627–39.

Dijanic, C., Nickerson, J., Shakya, S., Dijanic, A., and Fabbri, M. (2018). Relapsing malaria: a case report of primaquine resistance. *Case Reports in Infectious Disease*, **2018**, 9720823.

Dobson, M.J. (1999). The malariology centenary. *Parassitologia*, 41(1–3), 21–32.

Doda, Z., Solomon, T., Loha, E., Gari, T., and Lindtjorn, B. (2018). A qualitative study of use of long-lasting insecticidal nets (LLINs) for intended and unintended purposes in Adami Tullu, East Shewa Zone, Ethiopia. *Malaria Journal*, 17(1), 69.

Dondorp, A., Nosten, F., Stepniewska, K., Day, N., White, N., South East Asian Quinine Artesunate Malaria Trial (2005). Artesunate versus quinine for treatment of severe falciparum malaria: a randomised trial. *The Lancet*, 366(9487), 717–25.

Dondorp, A.M., et al. (2010). Artesunate versus quinine in the treatment of severe falciparum malaria in African children (AQUAMAT): an open-label, randomised trial. *The Lancet*, 376(9753), 1647–57.

Drugs mCGo (2011). A research agenda for malaria eradication: drugs. *PLoS Medicine*, 8(1), e1000402.

Ehrich, J.H. and Eke, F.U. (2007). Malaria-induced renal damage: facts and myths. *Pediatric Nephrology*, 22(5), 626–37.

Esu, E., Effa, E.E., Opie, O.N., Uwaoma, A., and Meremikwu, M.M. (2014). Artemether for severe malaria. *Cochrane Database of Systematic Reviews*, (9), CD010678.

Fernando, S.D., et al. (2003). The impact of repeated malaria attacks on the school performance of children. *American Journal of Tropical Medicine & Hygiene*, 69(6), 582–8.

Frosch, A.E., et al. (2014). Return of widespread chloroquine-sensitive Plasmodium falciparum to Malawi. *Journal of Infectious Diseases*, 210(7), 1110–14.

Gachelin, G., Garner, P., Ferroni, E., Trohler, U., and Chalmers, I. (2017). Evaluating Cinchona bark and quinine for treating and preventing malaria. *Journal of the Royal Society of Medicine*, 110(2), 73–82.

Galatas, B., et al. (2017). In-vivo efficacy of chloroquine to clear asymptomatic infections in Mozambican adults: a randomized, placebo-controlled trial with implications for elimination strategies. *Science Reports*, 7(1), 1356.

Gamboa, D., et al. (2010). A large proportion of *P. falciparum* isolates in the Amazon region of Peru lack pfhrp2 and pfhrp3: implications for malaria rapid diagnostic tests. *PLoS One*, 5(1), e8091.

Geelhoed, D.W., Visser, L.E., Addae, V., Asare, K., Schagen van Leeuwen, J.H., and van Roosmalen, J. (2001). Malaria prophylaxis and the reduction of anemia at childbirth. *International Journal of Gynaecological Obstetrics*, 74(2), 133–8.

Genton, B. and D'Acremont, V. (2017). Standby emergency treatment of malaria in travellers (SBET): So Be Eager to Test. *Journal of Travel Medicine*, 24(5).

Gilles, H.M. and Hendrickse, R.G. (1963). Nephrosis in Nigerian children. Role of Plasmodium malariae, and effect of antimalarial treatment. *British Medical Journal*, 2(5348), 27–31.

Gomes, M.F., et al. (2009). Pre-referral rectal artesunate to prevent death and disability in severe malaria: a placebo-controlled trial. *The Lancet*, 373(9663), 557–66.

Gosling, R.D., Carneiro, I., and Chandramohan, D. (2009). Intermittent preventive treatment of malaria in infants: how does it work and where will it work? *Tropical Medicine & International Health*, 14(9), 1003–10.

Greenwood, B., et al. (2017). Seasonal vaccination against malaria: a potential use for an imperfect malaria vaccine. *Malaria Journal*, 16(1), 182.

Greenwood, B. (2017). New tools for malaria control—using them wisely. *Journal of Infection*, 74(Suppl 1), S23–S6.

Greenwood, B.M. and Targett, G.A. (2011). Malaria vaccines and the new malaria agenda. *Clinical Microbiology & Infection*, *17*(11), 1600–7.

Groger, M., Fischer, H.S., Veletzky, L., Lalremruata, A., and Ramharter, M. (2017). A systematic review of the clinical presentation, treatment and relapse characteristics of human Plasmodium ovale malaria. *Malaria Journal*, 16(1), 112.

Guo, Z. (2016). Artemisinin anti-malarial drugs in China. *Acta Pharmaceutica Sinica B*, 6(2), 115–24.

Habluetzel, A., et al. (1997). Do insecticide-treated curtains reduce all-cause child mortality in Burkina Faso? *Tropical Medicine & International Health*, 2(9), 855–62.

Hopkins, H., et al. (2017). Impact of introduction of rapid diagnostic tests for malaria on antibiotic prescribing: analysis of observational and randomised studies in public and private healthcare settings. *British Medical Journal*, 356, j1054.

Howes, R.E., et al. (2016). Global epidemiology of *Plasmodium vivax*. *American Journal of Tropical & Medical Hygiene*, 95(6 Suppl), 15–34.

Ikram, N. and Simonsen, H.T. (2017). A review of biotechnological artemisinin production in plants. *Frontiers in Plant Science*, *8*, 1966.

Iroh Tam, P.Y., Obaro, S.K., and Storch, G. (2016). Challenges in the etiology and diagnosis of acute febrile illness in children in low- and middle-income countries. *Journal of the Pediatric Infectious Diseases Society*, 5(2), 190–205.

Irvine, M.H., Einarson, A., and Bozzo, P. (2011). Prophylactic use of antimalarials during pregnancy. *Canadian Family Physician*, 57(11), 1279–81.

Jallow, M., et al. (2012). Clinical features of severe malaria associated with death: a 13-year observational study in the Gambia. *PLoS One*, 7(9), e45645.

Jeremiah, S.S., Balaji, V., Anandan, S., and Sahni, R.D. (2014). A possible alternative to the error prone modified Hodge test to correctly identify the carbapenemase producing Gram-negative bacteria. *Indian Journal of Medical Microbiology*, 32(4), 414–18.

Kakuru, A., et al. (2016). Dihydroartemisinin-piperaquine for the prevention of malaria in pregnancy. *New England Journal of Medicine*, 374(10), 928–39.

Kisinza, W.N., et al. (2017). Multiple insecticide resistance in Anopheles gambiae from Tanzania: a major concern for malaria vector control. *Malaria Journal*, 16(1), 439.

Krishna, S., Bustamante, L., Haynes, R.K., and Staines, H.M. (2008). Artemisinins: their growing importance in medicine. *Trends in Pharmacological Science*, 29(10), 520–7.

Kublin, J.G., et al. (2003). Reemergence of chloroquine-sensitive Plasmodium falciparum malaria after cessation of chloroquine use in Malawi. *Journal of Infectious Diseases*, 187(12), 1870–5.

Kyabayinze, D.J., Asiimwe, C., Nakanjako, D., Nabakooza, J., Counihan, H., and Tibenderana, J.K. (2010). Use of RDTs to improve malaria diagnosis and fever case management at primary health care facilities in Uganda. *Malaria Journal*, **9**, 200.

Langford, S., et al. (2015). Plasmodium malariae infection associated with a high burden of anemia: a hospital-based surveillance study. *PLoS Neglected Tropical Diseases*, 9(12), e0004195.

Lengeler, C., Armstrong-Schellenberg, J., D'Alessandro, U., Binka, F., and Cattani, J. (1998). Relative versus absolute risk of dying reduction after using insecticide-treated nets for malaria control in Africa. *Tropical Medicine & International Health*, 3(4), 286–90.

Lengeler, C. and Snow, R.W. (1996). From efficacy to effectiveness: insecticide-treated bednets in Africa. *Bulletin of the World Health Organization*, 74(3), 325–32.
Liu, W. and Liu, Y. (2016). Youyou Tu: significance of winning the 2015 Nobel Prize in physiology or medicine. *Cardiovascular Diagnosis & Therapy*, 6(1), 1–2.
Lu, G., et al. (2018). Efficacy and safety of methylene blue in the treatment of malaria: a systematic review. *BMC Medicine*, 16(1), 59.
Mabaso, M.L., Sharp, B., and Lengeler, C. (2004). Historical review of malarial control in southern African with emphasis on the use of indoor residual house-spraying. *Tropical Medicine & International Health*, 9(8), 846–56.
Majori, G. (2012). Short history of malaria and its eradication in Italy with short notes on the fight against the infection in the Mediterranean basin. *Mediterranean Journal of Hematology & Infectious Diseases*, 4(1), e2012016.
Matangila, J.R., Mitashi, P., Inocencio da Luz, R.A., Lutumba, P.T., and Van Geertruyden, J.P. (2015). Efficacy and safety of intermittent preventive treatment for malaria in schoolchildren: a systematic review. *Malaria Journal*, **14**, 450.
Matuschewski, K. (2017). Vaccines against malaria-still a long way to go. *FEBS Journal*, 284(16), 2560–8.
Mbabazi, P., Hopkins, H., Osilo, E., Kalungu, M., Byakika-Kibwika, P., and Kamya, M.R. (2015). Accuracy of two malaria rapid diagnostic tests (RDTS) for initial diagnosis and treatment monitoring in a high transmission setting in Uganda. *American Journal of Tropical Medicine & Hygiene*, 92(3), 530–6.
Mbengue, A., et al. (2015). A molecular mechanism of artemisinin resistance in Plasmodium falciparum malaria. *Nature*, 520(7549), 683–7.
Menard, D., et al. (2010). Plasmodium vivax clinical malaria is commonly observed in Duffy-negative Malagasy people. *Proceedings of the National Academy of Sciences U S A*, 107(13), 5967–71.
Mendis, K., Rietveld, A., Warsame, M., Bosman, A., Greenwood, B., and Wernsdorfer, W.H. (2009). From malaria control to eradication, The WHO perspective. *Tropical Medicine & International Health*, 14(7), 802–9.
Mendonca, V.R. and Barral-Netto, M. (2015). Immunoregulation in human malaria: the challenge of understanding asymptomatic infection. *Memórias do Instituto Oswaldo Cruz*, 110(8), 945–55.
Mockenhaupt, F.P., Ulmen, U., von Gaertner, C., Bedu-Addo, G., and Bienzle, U. (2002). Diagnosis of placental malaria. *Journal of Clinical Microbiology*, 40(1), 306–8.
Mohandas, N. and An, X. (2012). Malaria and human red blood cells. *Medical Microbiology & Immunology*, 201(4), 593–8.
Moore, K.A., et al. (2016). Safety of artemisinins in first trimester of prospectively followed pregnancies: an observational study. *Lancet Infectious Diseases*, 16(5), 576–83.
Moorthy, V.S. and Okwo-Bele, J.M. (2015). Final results from a pivotal phase 3 malaria vaccine trial. *The Lancet*, 386(9988), 5–7.
Morrow, R.H. and Moss, W.J. (2017). Malaria. In: Detels, R.B., Lansang, M.A., Gulliford, M. (eds.) *Oxford Textbook of Public Health* (5th ed.) Oxford: Oxford University Press, p. 1728.
Mota, M.M., Hafalla, J.C., and Rodriguez, A. (2002). Migration through host cells activates Plasmodium sporozoites for infection. *Nature Medicine*, 8(11), 1318–22.
Mota, M.M. and Rodriguez, A. (2002). Invasion of mammalian host cells by Plasmodium sporozoites. *Bioessays*, 24(2), 149–56.
Moxon, C.A., Grau, G.E., and Craig, A.G. (2011). Malaria: modification of the red blood cell and consequences in the human host. *British Journal of Haematology*, 154(6), 670–9.
Moyes, C.L., et al. (2014). Defining the geographical range of the *Plasmodium knowlesi* reservoir. *PLoS Neglected Tropical Diseases*, 8(3), e2780.
Mueller, I., Shakri, A.R., and Chitnis, C.E. (2015). Development of vaccines for *Plasmodium vivax* malaria. *Vaccine*, 33(52), 7489–95.

Nabarro, D. (1999). Roll Back Malaria. *Parassitologia*, 41(1–3), 501–4.
Najera, J.A., Gonzalez-Silva, M., and Alonso, P.L. (2011). Some lessons for the future from the Global Malaria Eradication Programme (1955–1969). *PLoS Medicine*, 8(1), e1000412.
Ndam, N.T., et al. (2017). Reemergence of chloroquine-sensitive pfcrt K76 Plasmodium falciparum genotype in southeastern Cameroon. *Malaria Journal*, 16(1), 130.
Neghina, R., Neghina, A.M., Marincu, I., and Iacobiciu, I. (2010). Malaria, a journey in time: in search of the lost myths and forgotten stories. *American Journal of Medical Science*, 340(6), 492–8.
Nevill, C.G., et al. (1996). Insecticide-treated bednets reduce mortality and severe morbidity from malaria among children on the Kenyan coast. *Tropical Medicine & International Health*, *1*(2), 139–46.
Newman, R.D., Mnzava, A., and Szilagyi, Z. (2013). Mosquito larval source management: evaluating evidence in the context of practice and policy. *Cochrane Database of Systematic Reviews*, (8), ED000066.
Nima, M.K., et al. (2017). Case report: a case of plasmodium falciparum hrp2 and hrp3 gene mutation in Bangladesh. *American Journal of Tropical Medicine & Hygiene*, 97(4), 1155–8.
Oduro, A.R., et al. (2005). A randomized comparative study of chloroquine, amodiaquine and sulphadoxine-pyrimethamine for the treatment of uncomplicated malaria in Ghana. *Tropical Medicine & International Health*, 10(3), 279–84.
Oduro, A.R., Bojang, K.A., Conway, D.J., Corrah, T., Greenwood, B.M., and Schellenberg, D. (2011). Health centre surveys as a potential tool for monitoring malaria epidemiology by area and over time. *PLoS One*, 6(11), e26305.
Oduro, A.R., Maya, E.T., Akazili, J., Baiden, F., Koram, K., and Bojang, K. (2016). Monitoring malaria using health facility based surveys: challenges and limitations. *BMC Public Health*, **16**, 354.
Oduro, A.R., et al. (2017). Post-licensure safety evaluation of dihydroartemisinin piperaquine in the three major ecological zones across Ghana. *PLoS One*, 12(3), e0174503.
Ogobara, K., Doumbo, K.N., Healy, S.A., Sagara, I., and Duffy, P.E. (2018). Malaria transmission-blocking vaccines: present status and future perspectives. In: Manguin, S. and Dev, V. (eds.) *Towards Malaria Elimination—A Leap Forward*. IntechOpen. [Online] Available at: https://www.intechopen.com/books/towards-malaria-elimination-a-leap-forward/malaria-transmission-blocking-vaccines-present-status-and-future-perspectives
Olliaro, P. and Mussano, P. (2003). Amodiaquine for treating malaria. *Cochrane Database of Systematic Reviews*, (2), CD000016.
Olliaro, P.L., et al. (2016). Implications of *Plasmodium vivax* biology for control, elimination, and research. *American Journal of Tropical Medicine & Hygiene*, 95(6 Suppl), 4–14.
Owusu-Boateng, I. and Anto, F. (2017). Intermittent preventive treatment of malaria in pregnancy, a cross-sectional survey to assess uptake of the new sulfadoxine-pyrimethamine five dose policy in Ghana. *Malaria Journal*, 16(1), 323.
Pal, P., Balaban, A.E., Diamond, M.S., Sinnis, P., Klein, R.S., and Goldberg, D.E. (2017). Plasmodium falciparum histidine-rich protein II causes vascular leakage and exacerbates experimental cerebral malaria in mice. *PLoS One*, 12(5), e0177142.
Pal, P., Daniels, B.P., Oskman, A., Diamond, M.S., Klein, R.S., and Goldberg, D.E. (2016). Plasmodium falciparum Histidine-Rich Protein II Compromises Brain Endothelial Barriers and May Promote Cerebral Malaria Pathogenesis. *MBio*, 7(3).
Patil, A.P., et al. (2009). Defining the relationship between Plasmodium falciparum parasite rate and clinical disease: statistical models for disease burden estimation. *Malaria Journal*, **8**, 186.
Peplow, M. (2016). Synthetic biology's first malaria drug meets market resistance. *Nature*, 530(7591), 389–90.
Perkins, M.D. and Bell, D.R. (2008). Working without a blindfold: the critical role of diagnostics in malaria control. *Malaria Journal*, 7 Suppl 1, S5.

Phillips-Howard, P.A., et al. (2003). Impact of permethrin-treated bed nets on the incidence of sick child visits to peripheral health facilities. *American Journal of Tropical & Medical Hygiene*, 68(4 Suppl), 38–43.

Protopopoff, N., et al. (2018). Effectiveness of a long-lasting piperonyl butoxide-treated insecticidal net and indoor residual spray interventions, separately and together, against malaria transmitted by pyrethroid-resistant mosquitoes: a cluster, randomised controlled, two-by-two factorial design trial. *The Lancet*, 391(10130), 1577–88.

Rajahram, G.S., Barber, B.E., William, T., Menon, J., Anstey, N.M., and Yeo, T.W. (2012). Deaths due to *Plasmodium knowlesi* malaria in Sabah, Malaysia: association with reporting as *Plasmodium malariae* and delayed parenteral artesunate. *Malaria Journal*, **11**, 284.

Rakotoson, J.D., et al. (2017). Insecticide resistance status of three malaria vectors, *Anopheles gambiae* (s.l.), *An. funestus* and *An. mascarensis*, from the south, central and east coasts of Madagascar. *Parasitic Vectors*, 10(1), 396.

Ranson, H. and Lissenden, N. (2016). Insecticide resistance in African Anopheles mosquitoes: a worsening situation that needs urgent action to maintain malaria control. *Trends in Parasitology*, 32(3), 187–96.

Remme, J.H., Binka, F., and Nabarro, D. (2001). Toward a framework and indicators for monitoring Roll Back Malaria. *American Journal of Tropical & Medical Hygiene*, 64(1–2 Suppl), 76–84.

Rogerson, S.J., Desai, M., Mayor, A., Sicuri, E., Taylor, S.M., and van Eijk, A.M. (2018). Burden, pathology, and costs of malaria in pregnancy: new developments for an old problem. *Lancet Infectious Diseases*, 18(4), e107–e18.

Ross, R. (1923). *Memoirs, With a Full Account of the Great Malaria Problem and its Solution*. London: J. Murray.

Rowe, A.K. (2017). Assessing the health impact of malaria control interventions in the MDG/sustainable development goal era: a new generation of impact evaluations. *American Journal of Tropical & Medical Hygiene*, 97(3 Suppl), 6–8.

Rts, S.C.T.P. (2015). Efficacy and safety of RTS,S/AS01 malaria vaccine with or without a booster dose in infants and children in Africa: final results of a phase 3, individually randomised, controlled trial. *The Lancet*, 386(9988), 31–45.

Safeukui, I., et al. (2015). Malaria induces anemia through CD8+ T cell-dependent parasite clearance and erythrocyte removal in the spleen. *MBio*, 6(1):e02493–14.

Sanyal, S., et al. (2012). Plasmodium falciparum STEVOR proteins impact erythrocyte mechanical properties. *Blood*, 119(2), e1–8.

Shanks, G.D. (2016). Historical review: problematic malaria prophylaxis with quinine. *American Journal of Tropical & Medical Hygiene*, 95(2), 269–72.

Singh, B., et al. (2004). A large focus of naturally acquired Plasmodium knowlesi infections in human beings. *The Lancet*, 363(9414), 1017–24.

Skarbinski, J., et al. (2009). Effect of malaria rapid diagnostic tests on the management of uncomplicated malaria with artemether-lumefantrine in Kenya: a cluster randomized trial. *American Journal of Tropical & Medical Hygiene*, 80(6), 919–26.

Smith Gueye, C., et al. (2016b). Strategies and approaches to vector control in nine malaria-eliminating countries: a cross-case study analysis. *Malaria Journal*, **15**, 2.

Smith Gueye, C., Newby, G., Tulloch, J., Slutsker, L., Tanner, M., and Gosling, R.D. (2016a). The central role of national programme management for the achievement of malaria elimination: a cross case-study analysis of nine malaria programmes. *Malaria Journal*, 15(1), 488.

Snow, R.W., Amratia, P., Kabaria, C.W., Noor, A.M., and Marsh, K. (2012). The changing limits and incidence of malaria in Africa: 1939–2009. *Advances in Parasitology*, **78**, 169–262.

Star, T. (2015). Bill Gates and UN say malaria could be eradicated by 2040. [Online] Available at: https://www.thestar.com/life/health_wellness/2015/09/28/bill-gates-and-un-say-malaria-could-be-eradicated-by-2040.html

Stone, W.J.R., et al. (2018). Unravelling the immune signature of Plasmodium falciparum transmission-reducing immunity. *Nature Communications*, 9(1), 558.
Strategies mCGoI (2011). A research agenda for malaria eradication: cross-cutting issues for eradication. *PLoS Medicine*, *8*(1), e1000404.
Sundberg, C. (2007). Alphonse Laveran: the Nobel Prize for Medicine 1907. *Parassitologia*, 49(4), 257–60.
Takala-Harrison, S. and Laufer, M.K. (2015). Antimalarial drug resistance in Africa: key lessons for the future. *Annals of the New York Academy of Sciences*, **1342**, 62–7.
Tannous, S. and Ghanem, E. (2018). A bite to fight: front-line innate immune defenses against malaria parasites. *Pathologies of Global Health*, 112(1), 1–12.
Tarning, J. (2016). Treatment of malaria in pregnancy. *New England Journal of Medicine*, 374(10), 981–2.
Tawiah, T., et al. (2016). Cost-effectiveness analysis of test-based versus presumptive treatment of uncomplicated malaria in children under five years in an area of high transmission in Central Ghana. *PLoS One*, 11(10), e0164055.
Thiam, S., et al. (2011). Major reduction in anti-malarial drug consumption in Senegal after nation-wide introduction of malaria rapid diagnostic tests. *PLoS One*, 6(4), e18419.
Thorley-Lawson, D., Deitsch, K.W., Duca, K.A., and Torgbor, C. (2016). The link between plasmodium falciparum malaria and endemic Burkitt's lymphoma-new insight into a 50-year-old enigma. *PLoS Pathogens*, 12(1), e1005331.
Vaccines mCGo (2011). A research agenda for malaria eradication: vaccines. *PLoS Medicine*, 8(1), e1000398.
Walker, P.G., ter Kuile, F.O., Garske, T., Menendez, C., and Ghani, A.C. (2014). Estimated risk of placental infection and low birthweight attributable to Plasmodium falciparum malaria in Africa in 2010: a modelling study. *Lancet Global Health*, 2(8), e460–7.
Watson, J., Taylor, W.R.J., Bancone, G., Chu, C.S., Jittamala, P., and White, N.J. (2018). Implications of current therapeutic restrictions for primaquine and tafenoquine in the radical cure of vivax malaria. *PLoS Neglected Tropical Diseases*, 12(4), e0006440.
White, N.J. (2017). Does antimalarial mass drug administration increase or decrease the risk of resistance? *Lancet Infectious Diseases*, 17(1), e15–e20.
White, N.J. (2017). Malaria parasite clearance. *Malaria Journal*, 16(1), 88.
Williams, Y.A., et al. (2018). Expanding the vector control toolbox for malaria elimination: a systematic review of the evidence. *Advances in Parasitology*, **99**, 345–79.
Wilson, N.O., et al. (2011). Intermittent preventive treatment with sulfadoxine-pyrimethamine against malaria and anemia in pregnant women. *American Journal of Tropical & Medical Hygiene*, 85(1), 12–21.
Woodrow, C.J. and White, N.J. (2017). The clinical impact of artemisinin resistance in Southeast Asia and the potential for future spread. *FEMS Microbiology Reviews*, 41(1), 34–48.
World Health Organization (WHO) (2010). *Guidelines for the Treatment of Malaria* (2nd ed.). Geneva: WHO.
World Health Organization (WHO) (2014). Severe malaria. *Tropical Medicine & International Health*, 19 Suppl 1, 7–131.
World Health Organization (WHO) (2015a). *Confronting Plasmodium Vivax Malaria*. Geneva: WHO.
World Health Organization (WHO) (2015b). *Guidelines for the Treatment of Malaria*. Geneva: WHO.
World Health Organization (WHO) (2015c). *Global Technical Strategy for Malaria 2016–2030*. Geneva: WHO.
World Health Organization (WHO) (2016). *Eliminating Malaria*. Geneva: WHO.
World Health Organization (WHO) (2018a). *World Malaria Report*. Geneva: WHO.

World Health Organization (WHO) (2018b). *Malaria Surveillance, Monitoring & Evaluation: A Reference Manual*. Geneva: WHO. Licence: CC BY-NC-SA 3.0 IGO, 2018.

Yukich, J., D'Acremont, V., Kahama, J., Swai, N., and Lengeler, C. (2010). Cost savings with rapid diagnostic tests for malaria in low-transmission areas: evidence from Dar es Salaam, Tanzania. *American Journal of Tropical & Medical Hygiene*, 83(1), 61–8.

Yukich, J.O., et al. (2012). Reductions in artemisinin-based combination therapy consumption after the nationwide scale up of routine malaria rapid diagnostic testing in Zambia. *American Journal of Tropical Medicine & Hygiene*, 87(3), 437–46.

Yukich, J.O., Zerom, M., Ghebremeskel, T., Tediosi, F., and Lengeler, C. (2009). Costs and cost-effectiveness of vector control in Eritrea using insecticide-treated bed nets. *Malaria Journal*, **8**, 51.

8.16

Viral alcoholic and fatty liver diseases

Ehud Zigmond and Daniel Shouval

Introduction

The burden of liver diseases continues to rise worldwide. The aetiology of acute and chronic liver disease (CLD) varies among continents. Viral hepatitis B and C comprise the main causes for CLD in Africa and Asia, while alcoholic and non-alcoholic fatty liver diseases (NAFLDs) dominate in Europe and the United states. However, the ongoing wave of immigrants from Asia and Africa towards European countries is expected to have an impact on the epidemiology of these liver diseases worldwide. Prevalence and mortality data suggest that the increasing numbers of patients with cirrhosis and HCC in Europe, the US, Africa and East Asia are linked to increase in alcohol consumption and fatty liver disease (in the Western hemisphere) and to chronic viral hepatitis C and B in Asia and Africa (El-Serag and Kanwal 2014; Gower et al. 2014; Beste et al. 2015; Pimpin et al. 2018; Williams et al. 2018). In the present chapter we will present the evolving information on the epidemiology, natural history, prevention, and treatment of these selected liver diseases.

Hepatitis B

Aetiologic agent

Hepatitis B virus (HBV) is a double-stranded, enveloped virus of the *Hepadnaviridae* family.

The hepatitis B virion consists of a surface and a core, which contains a DNA polymerase and the e antigen. The ~3,200 bp DNA structure is double-stranded and circular with four major genes: the S (surface), the C (core), the

P (polymerase), and the X (transcriptional transactivating)genes. The S gene consists of three regions: S, pre-S2, and pre-S1—that encode the small, middle, and large envelope proteins called in total the hepatitis B surface antigen (HBsAg). HBsAg is a lipoprotein of the viral envelope that circulates in the blood as spherical and tubular particles. The C gene is divided into two regions, the pre-core and the core, and codes for two different proteins, the core antigen (HBcAg) and the e antigen (HBeAg).

HBV strains are classified into ten genotypes designated A to J, based on divergence in nucleotide sequences of ≥8%, most of which have a characteristic geographical distribution (Cao 2009; Tanwar and Dusheiko 2012; Sunbul 2014). E.g. genotype A is highly prevalent in sub-Saharan Africa, Northern Europe, and Western Africa. Genotypes B and C are the major variants in South and South East Asia and the Pacific region. Genotype D is prevalent in Africa, Europe, the Mediterranean region, and India. Genotype E is restricted to West Africa. Genotype F is found in Central and South America. Genotype G has been reported in France, Germany, and the United States and genotype H is found in Central America (Kurbanov et al. 2010; Lin and Kao 2011). Genotypes A, B, C, D, and F have further been subdivided in up to four major subgenotypes and several minor genotypes, now identified by Arabic numerals. In recent years, massive immigration waves generate an impact on the geographic distribution of genotypes worldwide.

HBV infection is controlled by cellular and humoral immune responses (Tan et al. 2015; Corti et al. 2018). It can be tracked through serological detection of the viral antigens or the antibodies raised by the immune system to target the virus. The presence of antibodies to the hepatitis B surface and/or hepatitis B core antigens (anti-HBs and anti-HBc) in the absence of HBsAg is generally taken (with few exceptions) to indicate resolution of infection and provides evidence of previous HBV infection. Persistence of HBV infection (chronic carrier state) is diagnosed by the detection of HBsAg in the blood on two occasions, six months apart or through continuous detection of HBV DNA even in the absence of detectable HBsAg in patients with occult HBV infection (OBI) (Raimondo et al. 2013). HBeAg is a secretory protein processed from the pre-core gene and considered as a marker for high viral replication and infectivity. HBV DNA is used as an indicator for viral replication expressed as IU/ml. There is a clear association between serum hepatitis B viral load and prognosis. Thus, the cumulative

annual incidence of cirrhosis or HCC being 4.5 and 1.3%, respectively, in persons with DNA levels <50 IU/ml, while it is 36.2 and 14.9%, respectively, in persons with DNA levels of ≥ to 2 × 10^5 IU/ml). This relationship supports the rationale for treating patients with high levels of HBV DNA with antiviral agents (Chen et al. 2006).

Epidemiology and transmission

Globally, hepatitis B is one of the most common infectious diseases (MacLachlan et al. 2015; Schweitzer et al. 2015). There are however wide variations in published global estimates of persistent HBV infection ranging between 240 and 350 million HBsAg carriers (Basnayake and Easterbrook 2016), (up to ~3.5% of the world population). Estimates indicate that about 2 billion people (i.e. about 30% of the world population) have ever been infected with HBV worldwide (WHO 2012) residing mainly in Africa, Asia, and the Western Pacific regions. One recent burden estimate in 2015 proposed that 257 million people are living with chronic HBV infection while 2.7 million persons are co-infected with HBV and human immunodeficiency virus (HIV).

On the basis of sero-epidemiological surveys, the WHO classified countries into four levels of endemicity according to the prevalence of chronic HBsAg carriage. Thus, HBV prevalence can be defined as low (<2.0%), lower intermediate (2.0–4.99%), higher intermediate (5.0–7.99%), and high ≥8.0%) (Fig. 8.16.1).

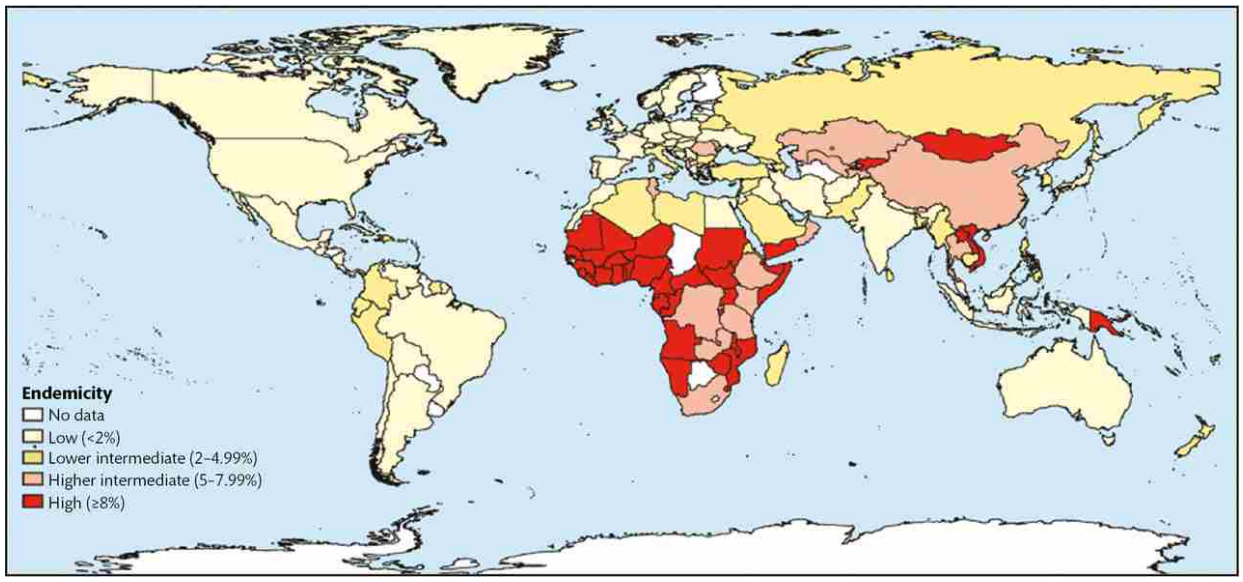


Fig. 8.16.1 Global HBsAg endemicity (1957–2013).
Reproduced from Schweitzer, A., Horn, J., Mikolajczyk, R.T., Krause, G., and Ott, J.J. (2015). Estimations of worldwide prevalence of chronic hepatitis B virus infection: a systematic review of data published between 1965 and 2013. *Lancet* 386, 1546–55, with permission from Elsevier.

HBV is transmitted by either percutaneous or mucous membrane contact with infected blood or other body fluid. The virus is found in highest concentrations in blood and serous exudates. The primary routes of transmission are vertical and perinatal, early childhood exposure (often called horizontal transmission), sexual contact, and percutaneous exposure to blood or infectious body fluids (i.e. injections, needle stick, blood transfusion, tattoos).

Most perinatal infections occur among infants of pregnant women with chronic-persistent HBV infection. The likelihood of an infant developing chronic HBV infection following vertical or horizontal transmission in the first year of life is 70–90% for those born to HBeAg-positive mothers (corresponding to high viral load) and less than 15% for those born to HBeAg-negative mothers. Most early childhood infections occur in households of persons with chronic HBV infection. The most probable mechanism involves unapparent percutaneous or permucosal contact with infectious body fluids (e.g. bites, breaks in the skin, dermatological lesions, skin ulcers). Sexual transmission has been estimated to account for 50% of new infections among adults in industrialized countries. The most common risk factors include multiple sex partners, men who have sex with men (MSM), history of a sexually transmitted infection and illicit drug injection.

Finally, unsafe injections and other unsafe percutaneous or permucosal procedures (such as cocaine snorting) are a major source of blood-borne pathogen transmission of HBV, hepatitis C virus (HCV), and HIV in many countries. The risk of HBV infection from needle stick exposure to HBsAg-positive blood (in patients and healthcare workers alike), is approximately 30% and worldwide unsafe injection practices account for approximately 8–16 million HBV infections each year.

In areas of high endemicity, the lifetime risk of HBV infection is more than 60%, and most infections occur during the perinatal period or during early childhood. In areas of intermediate endemicity, the lifetime risk of HBV infection varies between 20 and 60%, and infections occur in all age groups through the four modes of transmission, but primarily in infants and children. In areas of low endemicity, infection occurs primarily in adult life by sexual or parenteral transmission (e.g. through drug use). Acute infection is more often clinically overt in adults, while asymptomatic acute infection often remains unnoticed in young and older children. Infected infants and pre-school age children are at greatest risk of becoming chronic HBsAg carriers, thereby increasing the risk of cirrhosis and primary HCC later in life. The precise mechanism by which carrier rates are influenced by age of infection is unknown but probably relates to the effect of age on the immune system's ability to clear and eliminate the infection due to evolving immune tolerance to the virus.

Approximately 75% of the world's chronic HBV carriers live in Asian countries. In 2018, China ranks highest, with at least 87 million hepatitis B carriers, and India the second highest, with a carrier pool of 35 million (Tandon 1997; WHO Hepatitis Data and Statistics https://www.wpro.who.int/hepatitis/data/hepatitis_data_statistics/en/). Importantly, chronic carriers of HBV are not only at risk of developing long-term progression of the infection but also represent a significant source and reservoir of infection to others.

A model developed in 2005 estimated that for the year 2000, 620,000 persons died worldwide from HBV-related causes: 580,000 (94%) from chronic infection-related cirrhosis and HCC and 40,000 (6%) from acute HBV infection. Infections acquired during the perinatal period, in early childhood (<5 years old), and at 5 years of age and older accounted for 21%, 48%, and 31% of deaths, respectively (Goldstein et al. 2005).

Besides the acquired age of infection, some of the variation in outcome of HBV infection may be related to the genetic heterogeneity of the virus. Progression to chronic infection appears to occur more frequently following acute infection with genotypes A and D than with the other genotypes (Tanwar and Dusheiko 2012). Chronic infection with genotype A and B appears to have a better prognosis than genotype C and D which are often caused by pre-core mutations of HBV, most common in Asia, Mediterranean countries, and Southern Europe.

Clinical manifestations and immuno-pathogenesis of HBV infection

Acute hepatitis B has a long incubation period ranging from 15 to 180 days (90 days on average). Individual responses to the infection vary greatly. One-third of the individuals have subclinical infection; one-third experience a mild 'flu-like' illness without jaundice; and the remaining one-third develop jaundice with dark urine, extreme fatigue, anorexia, and abdominal pain. Jaundice usually peaks within 1–2 weeks and then gradually subsides. About 90–95% of infected adults and only 10–50% of infected newborns and babies recover spontaneously while the remaining subjects develop persistent HBV infection. Clinical recovery from acute infection is defined by normalization of ALT and AST associated with anti-HBs antibody seroconversion. Latent-occult HBV infection, also called OBI, manifested as intranuclear covalently closed circular (ccc) HBV DNA in the host hepatocytes, may persist for decades post so-called clinical recovery and anti-HBs antibody seroconversion (Rehermann et al. 1996). Such patients who may enter a dormant phase of persistent HBV infection are at risk for HBV reactivation when treated with immune-suppressive agents, including chemotherapy for malignant diseases or anti-inflammatory biologic agents, e.g. for inflammatory bowel disease, psoriasis, or rheumatoid arthritis (Shouval and Shibolet 2013). HBV reactivation rarely leads to immune clearance of HBV and more often to progression to cirrhosis and even fulminant liver failure. A small proportion (0.1–0.5%) of acutely infected adults develop fulminant hepatitis B, an exceptionally severe form of the disease which can be fatal requiring in two-thirds of cases an emergency liver transplant unless promptly treated with nucleos(t)ide analogue (Roche and Samuel 2014). About 5–10% of acutely infected adults and 50–90% of

newborns will become chronically infected and remain infectious (Kumar et al. 2007).

The immuno-pathogenesis of acute and chronic hepatitis B in infected newborns and babies as compared to adults may lead to different outcomes. Following acute infection, babies and young adults infected during the early years of life, often develop immune tolerance to HBV, associated with a high viral load and minimal liver injury. When immune tolerance to HBV is lost, the host immune system causes hepatocyte injury manifested by necro-inflammation and fibrosis, followed by an attempt at immune clearance of HBV, at a time when ALT levels often rise. A long-term, ongoing immune clearance phase causes more fibrosis and may lead to cirrhosis and compromised liver function. Once the host immune system is successful in inducing immune control against the virus, viral replication drops, and histologic evidence for liver injury may stabilize or resolve. However, resolution of the inflammatory process in the liver does not necessarily signifies cure since HBV DNA may still be present in the hepatocyte nucleus as cccHBV-DNA or undergo integration into the host hepatocyte genome. Thus persistent HBV infection is characterized by either overt viremia with detectable HBV DNA and HBsAg in serum or occult infection (OBI) extending to an inactive phase of the infection.

Extrahepatic manifestations of HBV infection

Extrahepatic manifestations following acute and chronic HBV infection occur in 10–20% of patients. These include serum sickness like syndrome, rash, arthralgia, glomerulopathies, and very rarely aplastic anaemia.

The natural history of chronic HBV infection can vary between individuals. Risk factors, which affect progression to chronic hepatitis, cirrhosis, and HCC include male gender, viral load, elevated alanine aminotransferase (ALT), genotype, and degree of fibrosis on liver biopsy. Patients with acute HBV infection are at risk for progression to a chronic carrier state. These patients, who may be still potentially infectious, may be asymptomatic with normal ALT and AST levels. A subgroup of HBsAg carriers may have clinically insignificant or minimal liver disease and never develop complications. Others will have clinically apparent chronic hepatitis with transition to cirrhosis and portal hypertension. Chronic infection with HBV can be either ‘replicative’ or ‘non-replicative’. The ‘replicative’ phase

or immune tolerance phase, characterized by positive HBeAg and high viral load (≥ 2000 IU/ml and often >10^6 IU/ml), is often present in newborns and children of HBsAg-positive mothers. In 'non-replicative' infection, the rate of viral replication in the liver is low, serum HBV DNA concentration is generally low, and HBeAg is not detected. In these inactive HBsAg carriers, reactivation of HBV and liver injury can occur either spontaneously or following reduction in dose of immune-suppressive agents (e.g. chemotherapy for malignant conditions) (Lalazar et al. 2007). Patients with chronic HBV and replicative infection generally have a worse prognosis and a greater chance of developing cirrhosis and/or HCC (Chen et al. 2006). In patients with HBV pre-core mutations, replicative infection can occur in the absence of detectable serum HBeAg. Finally, super or co-infection with the hepatitis delta virus (HDV) or HCV remain a constant threat for HBsAg-positive patients irrespective if they have an active inflammatory liver disease at time of exposure or not. (See later).

Treatment

Success in HBV treatment in CLD depends on the phase of chronic HBV infection. Viral replication by itself is not cythopathic and liver injury occurs only after the activation of host cellular and adoptive immune responses by the replicating virus, causing hepatocellular damage. Guidelines for treatment of chronic and acute HBV infection were developed by the American and European Associations for the Study of the Liver (European Association for the Study of the Liver 2017; Terrault et al. 2018). Treatment is indicated if ALT is elevated >X2 ULN, HBV DNA is ≥2,000 IU/ml or, if there are signs of moderate or severe liver fibrosis, or in case of treatment with chemotherapeutic or immune-suppressive agents.

The short-term goals of therapy for chronic HBV infection are suppression of viral replication and HBeAg seroconversion in HBeAg-positive patients or only viral suppression in anti-HBe positive patients. In some patients, HBsAg titres will drop or become undetectable by conventional assays but anti-HBs seroconversion, and loss of HBsAg which is an ultimate goal for therapy, is relatively rare. The long-term goals include prevention of cirrhosis and HCC. However, although control of inflammation and progression to cirrhosis are often attainable, complete resolution of infection with elimination of cccHBV-DNA is rarely achieved

in patients with persistent HBV infection. At present, treatment usually consists of a well-tolerated nucleos(t)ide analogue (p.o.) for an indefinite period and possibly lifelong. There are nowadays several licensed antiviral agents which are administered p.o., including lamivudine, telbivudine, entecavir, and tenofovir. Among these, entecavir and tenofovir are usually preferred, because of the higher genetic barrier (low probability of resistance) and higher potency. Treatment with the less tolerated interferon alpha by s.c. injections for about 48 weeks, is effective in inducing anti-HBe seroconversion in about a quarter of HBeAg-positive patients with chronic HBV; has the advantage of leading to prolonged suppression of viral replication when successful but is associated with significant adverse effects and still rare permanent resolution of infection. In summary, treatment of chronic HBV infections with oral antiviral agents is highly effective in suppression of viral replication and liver inflammation but rarely leads to anti-HBs seroconversion and complete resolution of infection. The recent detection of the sodium taurocholate co-transporting polypeptide (NTPC) which is an HBV receptor on the cell surface of hepatocytes is expected to trigger the development of new and more efficacious antiviral agents against HBV (Ni et al. 2014; Li 2015; Lin and Kao 2016).

Prevention

All major health authorities agree that the most effective approach to reducing the burden of HBV is primary prevention through universal vaccination as close to birth as possible and control of disease transmission. Interrupting the chain of infection requires knowledge of the mode of disease transmission and modification of behaviour through individual education to practice safe sex and good personal hygiene. Screening of all donated blood and maintenance of strict aseptic techniques with invasive health treatments using universal precaution measures, have reduced the likelihood of contracting HBV.

Safe and effective HBV vaccines have been available since the 1980s, and immunization with HBV vaccine remains the most effective means of preventing HBV disease and its consequences worldwide (Zanetti et al. 2008). Although the vaccine will not cure chronic hepatitis B, it is ~95% effective in preventing chronic infections, and is the first vaccine against a major human cancer.

After the development and licensure of plasma-derived vaccines in the 1980s, recombinant DNA technology has allowed the expression of HBsAg in bacteria, yeasts, mammalian cells and even plants (1986 and onwards). As a result, different manufacturers have successfully developed recombinant DNA vaccines against HBV, which are currently the backbone for most vaccination programmes worldwide (Shouval 2003).

Moreover, apart from monovalent vaccines against hepatitis B, a combination vaccine against hepatitis B and A as well as a broad range of combination penta and hexavalent vaccines that include an HBV component are available especially for vaccination during infancy and early childhood. Such vaccines, shown to be safe and immunogenic simultaneously, provide protection against a number of aetiologic agents including, diphtheria, tetanus, polio, pertussis (with either a whole-cell or an acellular component) and haemophilus influenza b.

More recently, the so-called third-generation hepatitis B vaccines—based on the S-, pre-S1-, and pre-S2-antigens, or using new adjuvants have been developed (Barry and Cooper 2007; Leroux-Roels 2015; Shouval et al. 2015). These vaccines generate a more robust humoral immune response against HBsAg as compared to conventional recombinant HBV vaccines leading to an enhanced immune response, especially in immune-compromised persons and non-responders to vaccination.

Immunization against hepatitis B requires the intramuscular administration of three doses of vaccine given at 0, 1, and 6 months. More rapid protection (i.e. for healthcare workers exposed to HBV or the susceptible sexual partner of a patient with acute hepatitis B) can be achieved through the adoption of an alternative schedule using three doses of vaccine administered at 0, 1, and 2 months followed by a booster dose given at 12 months. The extensive and long-term use of both plasma-derived and recombinant HBV vaccines has confirmed their excellent safety and tolerability. In 1998, several case reports from France raised concern that hepatitis B vaccination may lead to new cases or relapse of multiple sclerosis (MS) or other demyelinating diseases, including Guillain–Barré syndrome; however, no causal relation has been established (Duclos 2003). Hepatitis B vaccination is not contraindicated in pregnant or lactating women when absolute necessary.

Seroprotection against HBV infection is defined as having an anti-HBs level $\geq$10 IU/L after complete immunization (CDC 1987). Reviews on the

use of HBV vaccine in neonates and infants report seroprotective levels of anti-HBs antibodies at 1 month after the last vaccine dose for all schedules in 98–100% of vaccinees (Venters et al. 2004). While HBV vaccines generally induce an adequate immune response in over 95% of fully vaccinated healthy persons, a large interpersonal variability has been demonstrated in the anti-HBs immune response. The antibody response to hepatitis B vaccine has been shown to depend on the type, dosage, and schedule of vaccination used, as well as on the age, the gender, genetic factors, comorbidity, and the status of the immune system of the vaccinee (Hadler and Margolis 1992; Hollinger 1989). Immune-deficient patients i.e. with HIV, or those undergoing haemodialysis or immunosuppressant therapy, may require higher doses of vaccine and more injections (i.e. at months 0, 1, 2 and 6) to achieve an adequate and sustained immune response.

Follow-up studies have shown that vaccine-induced anti-HBs antibodies persist over periods of several decades and that duration of anti-HBs positivity is related to the antibody peak level achieved after primary vaccination (Jilg et al. 1988; Leuridan and Van Damme 2011). Follow-up of successfully vaccinated people has shown that the antibody concentrations usually decline over time, but clinically significant breakthrough infections are rare. Those who have lost antibody over time after a successful vaccination usually show a rapid anamnestic response when boosted with an additional dose of vaccine given several years after the primary course of vaccination or when exposed to the HBV. This means that the immunological memory for HBsAg can outlast the anti-HBs antibody detection, providing long-term protection against acute disease and the development of the HBsAg carrier state (Banatvala and Van Damme 2003). Hence, for immune-competent children and adults, the routine administration of booster doses of vaccine does not appear necessary to sustain long-term protection (Consensus-Group 2000). Such conclusions are based on data collected during the first 10–20 years of vaccination in countries of both high and low endemicity (Kao and Chen 2005; Zanetti et al. 2005).

Since the availability of hepatitis B vaccines in industrialized countries, early strategies for HBV control have stressed immunization of high-risk groups (e.g. MSM, healthcare workers, patients in sexually transmitted infection clinics, sex workers, IV drug users, people with multiple sex

partners, household contacts of chronically infected persons) and babies born to HBsAg-positive mothers. However, such vaccination strategies, directed at specific risk groups, had little impact on the overall burden of HBV in endemic countries and gradually were replaced by effective universal vaccination programmes of newborns.

In 1991, the WHO called for all children to receive the HBV vaccine. Substantial progress has been made in implementing this WHO recommendation: by the end of 2012, 179 countries had implemented or were planning to implement a universal HBV immunization programme for newborns, infants, and/or adolescents. In 2015, the global coverage with the third dose of hepatitis B vaccine reached 84%, which is not far from 90% goal, declared for the year 2020 by the WHO (WHO 2017). This high coverage leads to a major reduction in the incidence of chronic HBV infection in children below the age of 5 years worldwide as well as HCC. However, there are regional differences in coverage, especially in early administration of the birth dose of HBV vaccines in endemic countries. Thus, African, Eastern Mediterranean, and some European regions remain below the global average. Early vaccination within hours from birth, and high coverage with the primary vaccine series among infants had the greatest overall impact on the prevalence of chronic HBV infection in children (WHO 2004). According to model-based predictions, universal HBV infant immunization, (even without administration of a birth dose of vaccine to prevent perinatal HBV infection), would prevent up to 75% of global deaths from HBV-related causes, depending on the vaccination coverage for the complete series. Adding the birth dose would increase the proportion of deaths prevented up to 84% (Goldstein et al. 2005).

In countries with high or intermediate disease endemicity, the most effective strategy is to start immunization at birth (< 24 hours). In such regions a birth dose of a monovalent HBV vaccine is recommended even if vaccination protocols include later administration of combination vaccines, which also contain an HBV component. Countries with low HBV prevalence may consider immunization of children or adolescents as an alternative to newborn immunization (WHO 2004, 2006).

Indeed, the effectiveness of hepatitis B newborn and infant immunization programme has already been demonstrated in a variety of countries and settings (Andre and Zuckerman 1994; Lee 1997; WHO 2001). The results of effective implementation of universal hepatitis B programmes have become

apparent in terms of reduction not only in the incidence of acute hepatitis B infections, but also in the carrier rate in immunized cohorts and in hepatitis-B-related mortality from HCC and cirrhosis (Coursaget et al. 1994; Chang et al. 1997; Liaw 2013). In Taiwan, the HBsAg prevalence in children under 15 years of age decreased from 9.8% in 1984 to 0.7% in 1999 (Chan et al. 2004). In the Gambia, childhood HBsAg prevalence decreased from 10% to 0.6% since the introduction of the universal infant immunization programme (Viviani et al. 1999). Data in Hawaii show a 97% reduction in the prevalence of HBsAg since the start of the infant hepatitis B vaccination programme in 1991. The incidence of new acute hepatitis B infections in children and adults was reduced from 4.5/100,000 in 1990 to 0 in the period 2002–2004 (Perz et al. 2006b). In Bristol Bay, Alaska, 3.2% of children were HBsAg-positive before universal hepatitis B immunization; 10 years later, no child under 10 years of age was HBsAg-positive (Wainwright et al. 1997). Finally, surveillance data from Italy, where a universal programme was started in 1991 in infants as well as in adolescents, have shown that the incidence of acute hepatitis B progressively decreased from 1991 to 2005, with a 24-fold and 50-fold decrease in the 15–24-year and 0–14-year age groups, respectively (Mele et al. 2008).

Public health impact

HBV infection is a serious global health problem, which is under continuous surveillance by the WHO (WHO 2017). Prevention and elimination efforts of chronic HBV infection have utilized various estimates to assess the global burden of chronic HBV infection (Wiktor and Hutin 2016). Of the approximately 2 billion people who have been infected worldwide by 2012, more than 250 million are chronic carriers of HBV according to one estimate (WHO 2012). Approximately 15–40% of infected patients will develop cirrhosis, liver failure, or HCC. According to various estimates, HBV infection accounts for 600,000–900,000 deaths each year, mainly due to the consequences of chronic hepatitis, such as cirrhosis and liver cancer (Goldstein et al. 2005; Lavanchy 2004; Perz et al. 2006a; Schweitzer et al. 2015). Because these complications mainly occur in adults infected with HBV during childhood, most of the benefits of vaccination initiated 20–30 years ago have yet to be realized. Another consequence of persistent HBV

infection, which is often underestimated, is the stigma that HBV carries to the individual, in the family setting, as well as in social and professional life.

Hepatitis delta virus

HDV is a transmissible RNA pathogen that generally requires the help of a hepadnavirus like HBV for its own replication—similar to viroids or plant virus satellite RNAs (Hughes et al. 2011; Rizzetto 2015). Thus, in a natural setting, HDV is only found in patients who are also infected with persistent HBV since the HDV requires simultaneous presence of HBV for complete virion assembly and secretion. However, under rare circumstances, HDV may also replicate in immune suppressed, transplanted livers in the absence of overt HBV.

Aetiologic agent

The HDV genome is a small single-stranded RNA composed of approximately 1,680 bases with a unique circular conformation that is replicated using a host RNA polymerase. The HDV genome contains an open reading frame translated to HDV antigens—the small (S-HD) and large (L-HD) proteins which differ by 19 amino acids. The L-HD amino acid sequence is identical to S-HD with the addition of a carboxy-terminal extension of 19–20 amino acids following the editing of the S-HD stop codon during the viral RNA replication cycle. The S-HD is required for viral replication and might promote RNA polymerase II elongation of nascent HDV RNA, while L-HD is essential for HDV particle assembly (Tseng and Lai 2009).

Epidemiology and transmission

There are three types of HDV infection: simultaneous co-infection of HDV where HBV is transmitted by the same inoculum, super-infection of an HBV carrier with HDV (Rizzetto and Verme 1985) and isolated HDV infection without overt HBV re-infection (e.g. in immunosuppressed patients). A co-infection usually evolves into complete recovery, but the risk for severe acute hepatitis, even fulminant hepatitis, is higher than in the case of acute HBV mono-infection. A super-infection evolves towards chronicity in about

90% of cases. In isolated HDV infection, liver function tests are usually within normal range without significant histologic inflammatory activity.

According to the WHO, at least 5% of people with chronic HBV infection are co-infected with HDV, resulting in a total of 15–20 million persons infected with HDV worldwide (http://www.who.int/news-room/fact-sheets/detail/hepatitis-d) (Fig. 8.16.2).

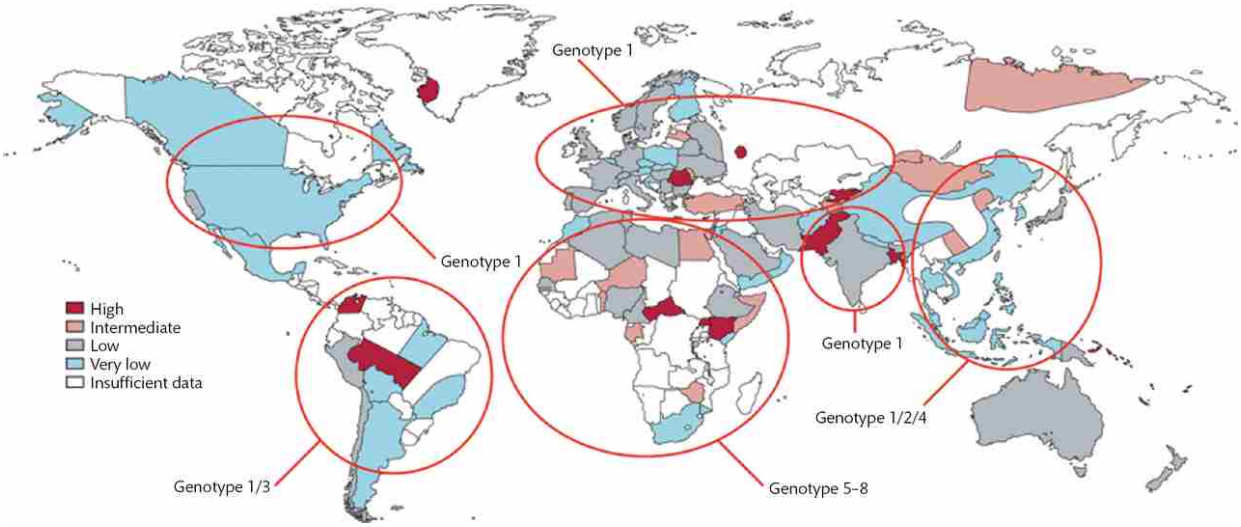


Fig. 8.16.2 Worldwide prevalence of HDV and the geographic distribution of its genotypes. Reproduced from Hughes, S.A., Wedemeyer, H., and Harrison, P.M. (2011). Hepatitis delta virus. *Lancet* 378, 73–85, with permission from Elsevier. Idilman, I.S., Keskin, O., Celik, A., Savas, B., Elhan, A.H., Idilman, R., and Karcaaltincaba, M. (2016).

However, this is a broad global estimate since many countries do not report the prevalence of HDV. Worldwide, the overall number of HDV infection has been declining since 1980s. This trend is mainly due to a successful control of transmission and more importantly, global HBV vaccination programmes. Eight HDV genotypes have been characterized to date on the basis of a small number of complete genome sequences with 19–38% divergence at the nucleotide level of complete genomes (Hughes et al. 2011). The most frequent genotypes are: genotype I (includes the European, North American, African, and some Asian HDV isolates), genotype II (found in Japan, Taiwan, and Eastern Europe), genotype III (found exclusively in South America), genotype IV (found mainly in West Africa) and genotype V (found in Central Africa). The global distribution of the

different genotypes of HBV and HDV is rather well known, although less is known on the types circulating among populations in remote areas. Little is also known on the viral transmission mechanisms driving the circulation of the endemic strains and epidemic outbreaks. Yet, on an individual basis, it appears that new and cell-to-cell transmission between hepatocytes is facilitated through the NTCP receptor.

In contrast to the US and Western Europe, the prevalence of HDV infection is relatively high in the equatorial subtropical and tropical zones, concentrating in certain population groups (Torres 1996). A putatively more severe form of HDV has been reported in Brazil and is, endemic in the Western Amazon states (Parana et al. 2008; Viana et al. 2005). HDV infection is also endemic in several Eastern European and Balkan countries as well as in Turkey. In Western Europe, HDV is becoming rare, affecting around 5% of chronic carriers (Rizzetto and Ciancio 2012). HDV infection is associated with an increased risk of developing advanced CLD, early cirrhosis development, and HCC (Farci 2003; Smedile and Bugianesi 2005).

In endemic areas, all HBV carriers should be tested for HDV. The diagnosis of HDV infection is based on serologic identification of anti-HDV (IgG) as evidence for previous or current exposure and anti-HDV(IgM) for identification of active /recent super or co-infection of an HBsAg carrier. Confirmation for delta virus viremia is obtained through HDV RNA testing by polymerase chain reaction (PCR), which may also be utilized quantitatively for evaluation of HDV treatment protocols.

Prevention and treatment

There is currently no vaccine for prevention of HDV. Consequently, it is imperative to prevent exposure of HBsAg carriers to contacts with HDV patients when applicable. Furthermore, no effective means are so far available for treatment of HDV infection. Long-term interferon alpha treatment may suppress HDV RNA in some patients and slow down progression of fibrosis but usually does not lead to resolution of HDV infection. The most effective means for controlling HDV infection is prevention of HBV infection through immunization. It is too early to judge if recently developed **e**xperimental pharmacologic agents against HDV i.e. blocking the NTCP receptor (Yurdaydin 2019), will reach licensure (Hughes et al. 2011; Yurdaydin 2019). Finally, for those patients with HDV co- or

super-infection who develop decompensation and liver failure, survival after liver transplantation is superior as compared to patients with HBV mono-infection (Roche and Samuel 2012).

Hepatitis C

Hepatitis C is a liver disease caused by the hepatitis C virus: the virus can cause both acute and chronic hepatitis infection, ranging in severity from an initially mild illness, lasting a few weeks to a lifelong illness with variable severity. Following acute HCV infection, 50–80% of patients will develop persistent infection and chronic hepatitis of variable severity, of whom ~20% will progress to cirrhosis with an annual risk for developing HCC of 4–5%.

Aetiologic agent

HCV is classified in the family *Flaviviridae*. Like other flaviviruses, HCV is an enveloped RNA virus with an inner nucleoprotein core. Its envelope contains two glycoproteins, E1 and E2, which form heterodimers (to form a functional subunit) at the surface of the virion (Dubuisson and Cosset 2014; Panel 2015). Efforts to identify the HCV succeeded by direct cloning and sequencing of its genome. Although the virus was identified almost three decades ago (Choo et al. 1989), its replication cycle is still only partially understood. An important feature of HCV is that the viral genome displays extensive genetic heterogeneity at an individual as well as the global level. Thus, even within the same host, the HCV genome population circulates as ‘quasispecies’ of closely related sequences. Worldwide, a high degree of genetic variation exists, resulting in at least six major genotypes and more than 100 distantly related subtypes (Forns and Bukh 1999). It has been reported that until recently, virus pathogenicity and sensitivity to current standards of treatment appear to vary with different subtypes (i.e. genotypes 2 and 3, responding better to interferon alpha treatment than genotype 1 and 4). This relative resistance to treatment with anti-viral drugs has practically been overcome through the development of new direct anti-viral agents.

Epidemiology and transmission

Estimates on the global prevalence of HCV and its consequences vary significantly (Alter 2007; WHO 2012; Basnayake and Easterbrook 2016; Polaris Observatory 2017). An extensive literature survey reports a wide range in estimates of the global prevalence of HCV carriers, ranging between 71 and 200 million of whom at least 2.28 million are co-infected with HIV (http://www.who.int/news-room/fact-sheets/detail/hepatitis). Yet reliable data from several countries remain unavailable. A review of the epidemiology and genotype distribution of HCV infection revealed a mean global prevalence of anti-HCV antibodies of 1.6% (range 1.3–2.1%) corresponding to 115 million past or present viraemic infections. Globally, genotype 1 was the most common variant (46%) followed by genotype 3 (22%). Genotype 4 is the most common genotype in African HCV patients. A significant number of those who are chronically infected will develop cirrhosis or liver cancer (Gower et al. 2014).

Figure 8.16.3 summarizes one estimate of the global prevalence of HCV. The highest prevalence of HCV has been reported from Egypt with a seroprevalence in the Nile delta ranging from 19% in the 10–19-year-old age group to 60% in the 30-year-old age group, and also associated with a high prevalence of liver cirrhosis. The high prevalence in the Nile delta is reported to be linked to parenteral anti-schistosomiasis therapy, which was carried out with inadequately sterilized injection material (Frank et al. 2000). Estimates in the United States reveal that 3.9 million Americans are chronically infected with HCV, with prevalence rates as high as 8–10% in African Americans (Smith et al. 2012). Haemodialysis patients, haemophiliacs, drug addicts, and people transfused with blood before 1990 are particularly affected by the disease. In Europe, 0.1–3.3% of the population has been infected, with the highest prevalence observed in Southern Europe (Italy) and Romania (Blachier et al. 2013).

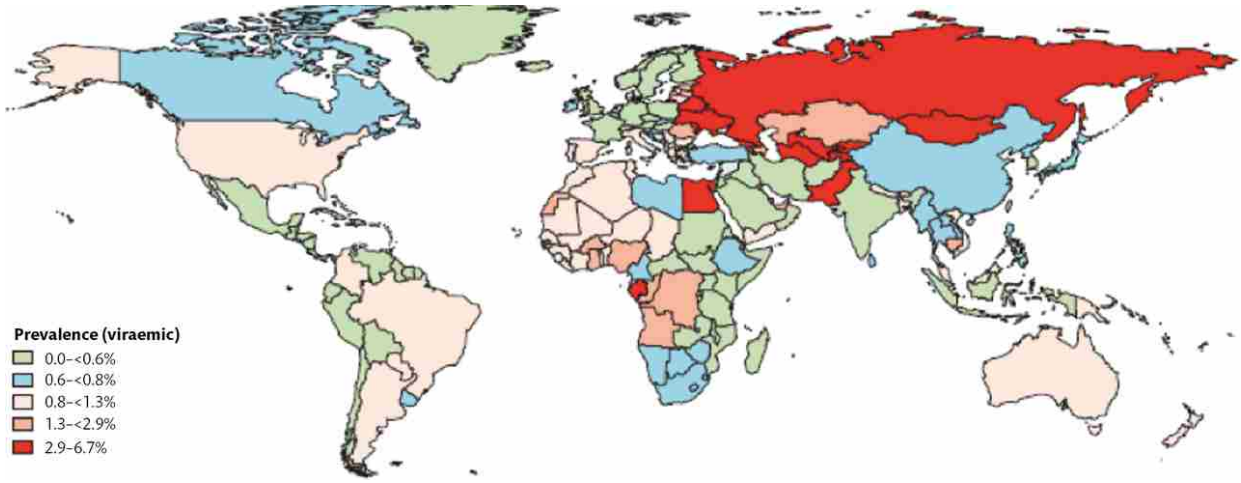


Fig. 8.16.3 Estimated prevalence of HCV.
Reproduced from Polaris Observatory, H.C.V.C. (2017). Global prevalence and genotype distribution of hepatitis C virus infection in 2015: a modelling study. *Lancet Gastroenterol Hepatol* 2, 161–76, with permission from Elsevier.

Despite infection control precautions, healthcare workers remain at risk for acquiring blood-borne HCV infections due to accidental exposure. Therapeutic injections are reported as accounting for 2 million new HCV infections each year. Many of these injections are performed in less than ideal conditions, often with reuse of needles or multidose vials and mainly, but not exclusively, in low- and middle-income countries. The residual risk of transmitting HCV through screened blood transfusions is very low in industrialized countries but safety of blood supply remains a major source of public concern in low- and middle-income countries.

Up to 60–70% of intravenous drug users living in urban areas are seropositive for HCV antibodies. The rate of infection depends on the length of drug use, with 25% of infections occurring during the first year of addiction, 50% after 5 years, and up to 90% after more than 5 years of intravenous drug use.

The global epidemic of HCV infection emerged in the second half of the twentieth century and has been attributed, at least in part, to the increasing use of parenteral therapies and blood transfusion during that period. In high-income countries, the rapid improvement of healthcare conditions and the introduction of anti-HCV screening for blood donors have led to a sharp decrease in the incidence of iatrogenic HCV (Prati 2006). Injectable drug use remains the main route of transmission, accounting for nearly 90% of new HCV infections. Mother-to-child transmission has been documented but is

lower as compared to HBV. The risk of perinatal infection in children from HCV-infected mothers ranges from 3% to 10% in different populations. Transmission is believed to occur *in utero*, as a consequence of a high viral load in the mother; in particular, from co-infected HIV positive mothers (Kato et al. 1994). There is no contraindication for infected mothers to breastfeed the newborn. Sexual transmission is thought to be relatively infrequent (up to 2–3%); as such, it is not recommended in monogamic and stable couples to use condoms. However, the large reservoir of HCV carriers provides multiple opportunities for exposure to potentially infected partners. Individuals with multiple sexual partners, MSM, prostitutes and their clients, patients with common sexually transmitted infections, and partners of HCV and HIV co-infected persons are at the highest risk of acquiring HCV sexually. In many cases of HCV infection, no recognizable transmission factor or route is identified (Memon and Memon 2002).

Clinical manifestations

The incubation period for hepatitis C before the onset of clinical symptoms ranges from 6 to 7 weeks on average. In acute infections, the most common symptoms are fatigue and jaundice; however, the majority of cases (between 60 and 70%), even those who develop chronic infection, are asymptomatic for years. Fulminant hepatitis C is very rarely observed. While most patients with acute HCV infection have mild symptoms or no symptoms, 50–85% of those infected develop chronic disease. Chronic disease is difficult to recognize because symptoms are mild and infection passes silently and insidiously from the acute to the chronic phase. Serological diagnosis of acute HCV infection is mainly based on recent HCV antibodies (anti-HCV) seroconversion. Persistence of HCV infection is diagnosed by the presence of HCV RNA in the blood for at least 6 months. The mechanisms leading to HCV persistence are currently unknown, although it is known that HCV chronicity develops despite humoral and cellular responses to HCV proteins (Park and Rehermann 2014). Factors associated with development of chronic disease appear to include older age at the time of infection, male gender, and an immunosuppressed state such as HIV infection (Lauer and Walker 2001).

Extrahepatic manifestations of hepatitis C

Hepatitis C virus infection provokes B-lymphocytes dysfunction with extrahepatic manifestations of autoimmunity. Several autoimmune and metabolic conditions and diseases have been associated with HCV infection: lichen planus, granuloma annulare, porphyria cutanea tarda, and psoriasis (Andrade et al. 2012). One important expression of autoimmunity in HCV carriers is HCV-related cryoglobulinaemia, which can cause purpura, vasculitis, glomerulonephritis, and peripheral neuropathy (Atta et al. 2010). Cryoglobulinaemia has been found in HCV carriers with lympho proliferative disorders, mainly B-cell non-Hodgkin lymphoma (Yu and Lin 2013). Furthermore, HCV infection has been associated with higher risk of insulin resistance and diabetes (Naing et al. 2013). On the other side, diabetes is a confirmed risk factor for the evolution to hepatocellular carcinoma.

The presence of non-organ specific autoantibodies is another expression of autoimmunity found in HCV carriers, which has been associated in some studies with advanced fibrosis and failure in the treatment of hepatitis C with the combined therapy of interferon-α plus ribavirin. Antibodies to self-antigens such as nucleoproteins (ANA), smooth muscle (SMA), liver–kidney microsomal type-1 antigen (LKM-1), immunoglobulin G (RF), neutrophil cytoplasm (ANCA), and phospholipids (APL), which are routinely used as biomarkers of autoimmune diseases, can be found in HCV carriers with varying prevalence. Their induction seems mainly to involve immune cross-reactions caused by molecular mimicry between HCV polyprotein and human auto antigens (Atta et al. 2010).

Treatment

The primary goals for treatment of HCV infection include eradication of infection as well as reduction of morbidity and mortality from CLD and HCC. Specifically, such effects are believed to occur through normalization of liver function tests such as ALT/AST, complete clearance of HCV, decline in disease progression leading to improved quality of life, and reducing the reservoir of chronic carriers, thereby controlling further transmission. In contrast to hepatitis B, chronic HCV infection is potentially completely curable. Virological cure or sustained viral response (SVR) is currently defined by undetectable HCV RNA for at least 24 weeks (SVR 24) after ending therapy (Gurusamy et al. 2013).

Treatment is recommended for HCV-infected patients who are at risk of developing cirrhosis or with potential harmful extrahepatic manifestations of HCV as well as for liver transplant recipients with graft re-infection of HCV.

In patients with cirrhosis, the benefits of therapy consist of reduction in of the risk of dying from a liver-related death including the risk of decompensation, liver failure, and development of HCC (Veldt et al. 2007).

Historically, and until 2011 effective SVR has been obtained in up to 50% of HCV patients with genotype 1 and 80% of patients with genotypes 2 or 3 who had received combined weekly pegylated interferon-based treatment with daily ribavirin for 24–48 weeks (Chevaliez and Pawlotsky 2007; Tan and Lok 2007). Following the development and licensure in 2011 of the first generation of direct antiviral protease inhibitors Telaprevir and Boceprevir, SVR rates in response to various combination protocols rose to 70–80%, in treatment naïve patients, albeit associated with significant adverse events (Kwo et al. 2010; Jacobson et al. 2011). Since 2014, several new, efficacious, safe, and well-tolerated direct antiviral agents (DAAs) have been developed and licensed worldwide for interferon free treatment of chronic HCV infection (i.e. various combinations of sofosbuvir, ledispasvir, simeprevir, daclatasvir, ritonavir-boosted paritaprevir, ombitasvir as well as dasabuvir, elbasvir and grazoprevir), leading to SVR rates >90% for the various genotypes treated for 8–12 weeks in treatment naïve patients (European Association for Study of the Liver 2015; Panel 2015; AASLD-IDSA Hepatitis C Guidelines, 2020).

The remarkable breakthrough in identification of HCV which lead to the development of efficacious anti-viral agents was recently acknowledged by the Swedish Nobel Prize committee. Thus the 2020 Nobel Prize in Physiology and Medicine was awarded to three scientists who have made the decisive contributions to the fight against blood-borne hepatitis, Harvey J. Alter, Michael Houghton and Charles M. Rice. Yet, the exceedingly high cost charged by the pharma industry for antiviral drugs has mellowed somewhat the excitement generated by the development of the new generation of DAAs. In spite of proven cost-effectiveness, low- and middle-income countries and even health insurers in the US and Europe face difficult challenges in coverage of the very high cost of curative treatment for the millions of patients around the globe (Chhatwal et al. 2015), (http://www.who.int/hepatitis/publications/hep-c-access-report/en/). The WHO is currently promoting an effort to simplify curative treatment

protocols for developing countries (http://www.who.int/hepatitis/news-events/hep-c-access-report-2018/en/).

Prevention

There is currently no licensed vaccine against HCV. The high mutation rates of the HCV genome complicate vaccine development and there is no consensus yet regarding the potential candidates for receiving such a vaccine once proven to be efficacious (Baumert et al. 2014; Mehta et al. 2014). Although 20–35% of patients with acute HCV infection clear the virus spontaneously, lack of knowledge of any effective protective immune response following HCV infection impedes vaccine research. Although some studies have shown the presence of virus-neutralizing antibodies, it is not fully clear whether and how the immune system is able to eliminate the virus. From a global perspective, an important impact on HCV disease burden will likely be achieved by a combination of efforts including public health measures as well as treatment of large cohorts of HCV carriers with the highly effective (but also very expensive) DAAs with a special emphasis on specific risk groups. It is imperative to continue and focus efforts on reducing the risk of HCV transmission from nosocomial exposures (e.g. screening of blood, rigorous implementation of infection control, reducing unsafe injection practices) and high-risk behaviours (e.g. injection drug use). Adherence to fundamental infection control principles, including safe injection practices and appropriate aseptic techniques, is essential to prevent transmission of blood-borne viruses in healthcare settings. Educational programmes aimed at the prevention of drug use and, for those already addicted, aimed at the prevention of shared needles and other equipment can decrease this source of infection. Some countries have established needle exchange programmes that provide easy access to sterile needles and syringes, accompanied by counselling and health education and instructions on the safe disposal of used syringes.

Public health impact

The WHO estimates that ~400,000 die each year from HCV infection and its complications; (http://www.who.int/news-room/fact-sheets/detail/hepatitis-c). It is estimated that 3–4 million persons are newly infected each year and

that ~20% of those infected with HCV progress to cirrhosis within the first 10 years after infection (Alter 2007; Gerberding and Henderson 1992). Although the prevalence of chronic HVC reached a peak in 2001 (according to a multiple cohort model) in the United States, the prevalence of hepatitis C cirrhosis and HCC is expected to continue to increase through the next decades (Davis et al. 2010). Furthermore, chronic HCV disease is the primary indication for liver transplantation in industrialized countries (Adam et al. 2012). In Europe the overall prevalence of chronic hepatitis C in the last decade was 0.13–3.26% and ~1% in countries of the European Union and Economic area (https://ecdc.europa.eu/en/hepatitis-c). It is of great concern that the majority of subjects infected with HCV and HBV globally are unaware of their status (Blachier et al. 2013).

Alcoholic liver disease

Alcoholic beverages have been used in human societies since the beginning of recorded history. In 2005, a study estimated that just over 40% of the world's adult population consumes alcohol and the average consumption per drinker is 17.1 L per year (Shield et al. 2013). The level of alcohol consumption worldwide in 2010 was estimated at 6.2 litres of pure alcohol per person aged 15 years and over (equivalent to 13.5 g of pure alcohol per day) (WHO 2014).

It has long been known that alcohol consumption is responsible for increased illness and death. Worldwide, alcohol causes ~3.3 million deaths each year (5.9% of all deaths), while liver cirrhosis and its complication accounts for 16.2% of the cases (WHO 2014). Furthermore, 139 million disability-adjusted life years (5.1% of the global burden of disease and injury), were attributable to alcohol consumption (WHO 2014). The burden is not equally distributed among countries. The highest disease load attributable to alcohol is found in the heavy-drinking countries of Eastern Europe and in Latin America (Figure 8.16.4). Of note, alcohol consumption over the last 20 years in the UK and Finland has increased significantly, while other countries such as France, Spain and Portugal were able to reduce the number of liver-related deaths (Sheron 2016; Williams et al. 2018).

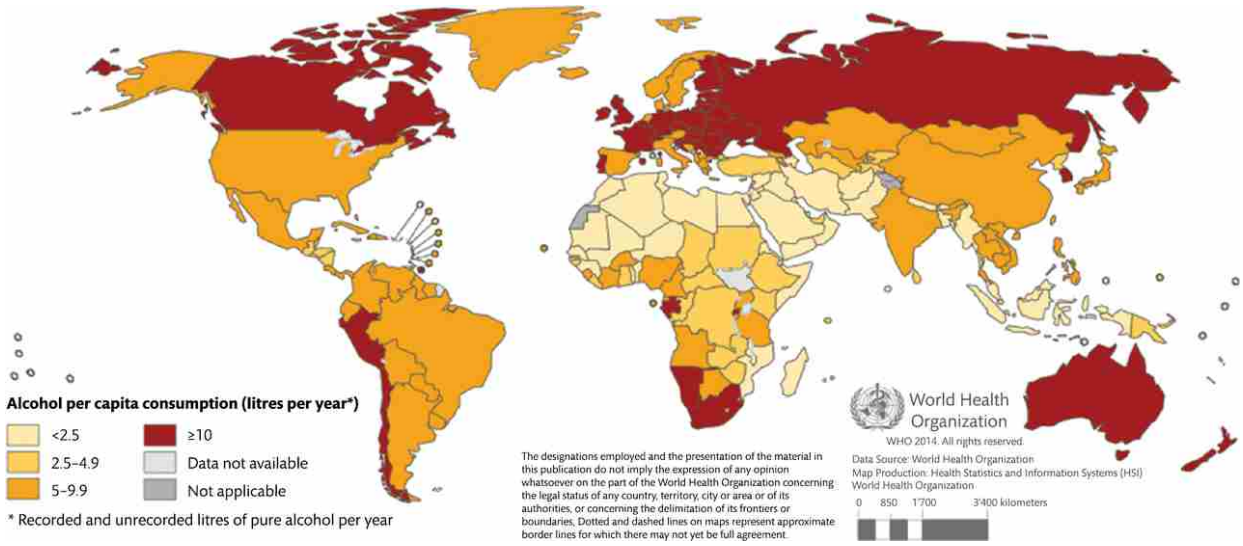


Fig. 8.16.4 Total (recorded and unrecorded) alcohol consumption per capita (aged 15 years and over), in litres of pure alcohol within a calendar year, by WHO region, projected estimates for 2012. Reproduced from WHO (2014) 'Global status report on noncommunicable diseases 2014' with permission from the World Health Organization.

Alcoholic liver disease constitutes the commonest reason for admitting patients with liver disease to a hospital and is responsible for at least 41% of the liver deaths in Europe (Sheron 2016). Between 2004 and 2013, the number of new liver transplants registrations for this aetiology in the US increased by 45% (Wong et al. 2015). In Europe, the proportion of patients transplanted for alcoholic liver disease increased by 8.3% between the time periods 1988–1995 and 1996–2000 (Burra et al. 2010). Alcohol (38%), viral hepatitis (43%), or both (6%) are the leading indications for liver transplantation following cirrhosis.

The costs to society from alcohol abuse cannot be overemphasized. In 2006, overall costs in the United States reached US$223.5 billion, out of which healthcare expenses accounted for 11%. Almost three-quarters of these costs were due to binge drinking (Bouchery et al. 2011). In Europe in 2003, alcohol consumption resulted in estimated costs of €125 billion, equivalent to 1.3% of gross domestic product, while the reported contribution of the alcohol industry to the European economy was about €9 billion (Rabinovich et al. 2009).

In 2010, 5.5% of the global burden of disease was attributable to alcohol, almost similar to the burden of tobacco-associated illness (6.3%) (Lim et al. 2012). Alcohol use is the leading risk factor for global disease burden in Eastern Europe, most of Latin America, and southern sub-Saharan Africa.

Morphology and natural history of alcoholic liver disease

Fatty liver (steatosis)

The first and most predictable hepatic change attributable to alcohol is the development of large droplet (macrovesicular) steatosis. This disorder usually resolves within 2 weeks if alcohol consumption is discontinued (Diehl 1997). Of note, about 5–15% of patients with alcohol-induced steatosis will develop cirrhosis during a 10-year follow-up period (Sorensen et al. 1984).

Alcoholic steatohepatitis

The spectrum of alcoholic steatohepatitis includes fatty infiltration of hepatocytes associated with hepatocellular injury including ballooning degeneration, Mallory bodies, inflammation with neutrophils and/or lymphocytes, and fibrosis with a perivenular, perisinusoidal, and pericellular disposition. These changes are present in 10–35% of all alcoholics. It is estimated that the probability of developing cirrhosis in patients with alcoholic liver disease is 10–20% per year and overall, 70% of patients with alcoholic hepatitis will eventually develop cirrhosis (Diehl 1997) and some will proceed to fatal liver decompensation.

Cirrhosis

In 2010, alcoholic cirrhosis accounted for 48% of cirrhosis-related deaths globally (Rehm et al. 2013). The long-term prognosis of alcoholic cirrhosis improves with abstinence. The 5-year survival in compensated cirrhosis patients who continue to drink is 70%, but can rise to 90% if patients abstain from further alcohol intake. In patients with decompensated cirrhosis, the 5-year survival drops to 30% in individuals who continue to drink, but is 60% in those who stay abstinent (Alexander et al. 1971; Diehl 1997).

Hepatocellular carcinoma

Alcohol is considered both as a primary cause of HCC as well as a co-factor for the malignant transformation. As hepatitis C is relatively frequent in alcoholics, the proportional contribution of alcohol in the reported HCC incidence rates in earlier studies is likely to be overestimated. Although the exact annual incidence rate of HCC in alcoholic cirrhosis is unknown, it is

clear that alcoholic cirrhosis is a significant risk factor for HCC, sufficient to warrant surveillance for HCC (Bruix et al. 2005).

Factors influencing the risk of alcoholic liver disease

Most authors agree that persons who drink heavily (50–60 g of ethanol daily) are at increased risk of developing liver disease (Becker et al. 2002). However, the absolute risk of acquiring alcoholic hepatitis or cirrhosis is relatively low (6.9% in the mentioned study) which suggests that genetic factors and/or environment may play a role in the pathogenesis.

Amount of alcohol

Excessive alcohol consumption is associated with an increased risk of cirrhosis. However, a specific dose–response relationship for cirrhosis has not been agreed upon. Evidence suggests that there is an increased risk for alcoholic liver disease with the ingestion of 60–80 g/day of alcohol in men and 20 g/day in women (Day 2000). 'Safe' limits of alcohol consumption for the liver are up to two drinks per day for women, and up to three drinks per day for men, with at least three alcohol-free days per week (Rehm et al. 2010; European Association for the Study of the Liver 2018).

Drinking behaviour

A survey of 30,630 persons from Denmark revealed that beer or spirits are more likely to promote liver disease than wine (Becker et al. 2002). At present, it is uncertain whether wine *per se* is responsible for this reduced risk of liver disease compared to the other alcoholic beverages, or whether it represents a surrogate for other healthy behaviours such as increased consumption of fruits/vegetables (Everhart 2003). Binge drinking, a mode of social behaviour nowadays, is the exaggerated form of non-mealtime drinking. It has been reported to increase the risk of alcoholic hepatitis fivefold (Barrio et al. 2004) and to increase the risk of all-cause mortality in men and women (Tolstrup et al. 2004). Importantly, a report from the OECD countries revealed that 30% of men and 12% of women binge-drink at least once per month (OECD 2017). Also drinking multiple types of alcohol has been shown to be related to the risk of cirrhosis and non-cirrhotic liver disease (Naveau et al. 1997). Nevertheless, there are still controversies

regarding the risks of binge drinking. A Danish survey of 55,000 participants showed that daily drinking was associated with the highest risk of liver cirrhosis (Askgaard et al. 2015), while a Finnish population-based health study of 6,366 subjects without baseline liver disease revealed that binge drinking was associated with an increased risk of liver disease independently of average alcohol intake (Aberg et al. 2017).

Gender

The proportion of global deaths attributable to alcohol differs based on gender, with 7.6% of deaths among males and 4.0% of deaths among females in 2012 (WHO 2014). Nevertheless, it is well recognized that women are more susceptible to alcohol-induced health disorders than men. Men and women have similar sized livers and similar metabolic rates when the rate of alcohol metabolism is normalized to liver mass. However, blood alcohol levels after comparable doses of alcohol will usually be higher in women than in men because of their lower body volume and the higher percentage of their body mass consisting of fat. Evidence from animal models has suggested that oestrogen increases the gut permeability to endotoxin and accordingly upregulates endotoxin receptors of Kupffer cells leading to an increased production of tumour necrosis factor alpha in response to endotoxin (Enomoto et al. 1999). It is also known that the gastric alcohol dehydrogenase activity can be different in men and women (Parlesak et al. 2002).

Comorbid conditions

Individuals with comorbid conditions affecting the liver exhibit a greater tendency to develop liver disease in response to alcohol consumption than persons being otherwise healthy (Fattovich et al. 2004). This was clearly demonstrated in the case of hepatitis C (Corrao and Arico 1998), hepatitis B, hereditary haemochromatosis (Fletcher et al. 2002), and obesity (Naveau et al. 1997), and most probably applies to other causes of chronic hepatitis.

Genetic polymorphisms

Strong epidemiological evidence suggests the existence of heritable susceptibility to alcoholic liver disease. This appears related to several gene

polymorphisms, some of which impact alcohol metabolism and others that influence hepatic immune responses.

Prevention

The overall management of alcoholism and of alcoholic liver disease and its complications is beyond the scope of this chapter. Nevertheless, general measures include alcohol withdrawal and providing nutritional support. A number of public health policies to reduce alcohol consumption have been shown to be helpful and cost-effective (Wagenaar et al. 2009; Wagenaar et al. 2010). Effective strategies include price based policies via taxation or by setting a minimum unit pricing, limitation of alcohol availability, restriction of marketing and advertising. Children are particularly sensitive to alcohol marketing shown to impact their drinking behaviour (Anderson et al. 2009). Setting a floor price for a unit of alcohol has been shown to be highly successful, reducing alcohol-related mortality by 32% in British Columbia (Stockwell et al. 2013). General practitioners have a pivotal role in screening for excessive alcohol ingestion and referral to a professional multidisciplinary team for withdrawal (European Association for the Study of the Liver 2018). Notably, there is good clinical evidence that cessation of drinking at any point in the natural history of the disease reduces the risks of disease progression and incidence of complications from cirrhosis (Masson et al. 2014).

Non-alcoholic fatty liver disease and non-alcoholic steatohepatitis

Definitions

Steatosis is defined as the accumulation of fat in the liver parenchymal cells or hepatocytes. A distinction is made between macrovesicular and microvesicular steatosis. Macrovesicular steatosis implies the presence of large fat vacuoles, containing predominantly triglycerides, and occupying a large part of the cell cytoplasm, displacing the nucleus towards the cell border. The hepatocytes may be enlarged by the presence of these fat vacuoles. Macrovesicular steatosis is graded according to the percentage of hepatocytes containing fat vacuoles: less than 5% is minimal or no steatosis;

5–30% is mild steatosis; over 30–60% is moderate; and greater than 60% is considered to be severe macrovesicular steatosis (D'Alessandro et al. 1991). In microvesicular steatosis, bipolar lipids are forming micelles, which are spread over the cytoplasm, and which do not displace the nucleus. The cells usually have normal dimensions. In many patients, both types of steatosis are present, called mixed type steatosis. In those cases, macrovesicular steatosis is usually predominant.

Individuals with fatty liver may present with a wide range of histological disease severity starting from pure steatosis without other abnormalities to steatosis accompanied by inflammation, tissue damage, liver fibrosis, and finally cirrhosis. The term *non-alcoholic fatty liver disease* (NAFLD) encompasses the entire spectrum of fatty liver diseases in individuals that lack secondary causes of hepatic fat accumulation such as significant alcohol consumption and use of steatogenic medications. Of note, NAFLD also implies the exclusion of other CLDs, including chronic viral hepatitis, toxic hepatitis (due to industrial toxins or solvents or to pharmacological agents), autoimmune liver disease, haemochromatosis, Wilson's disease, and some rare metabolic disorders. Hepatitis C, especially genotype 3, and Wilson's disease are two classical examples of liver diseases accompanied by steatosis, but they are not NAFLD.

NAFLD can be classified histologically into *non-alcoholic fatty liver (*NAFL) and *non-alcoholic steatoheatitis* (NASH). NAFL is defined as the presence of liver steatosis of more than 5% without evidence of hepatocellular injury or fibrosis. The risk of progression to liver cirrhosis in NAFL is considered minimal. NASH is defined as the presence of liver steatosis of more than 5% with inflammation and hepatocyte injury (especially ballooning of hepatocytes), with or without fibrosis. NASH might progress to cirrhosis, liver failure and hepatocellular carcinoma (Chalasani et al. 2018). The maximum daily alcohol consumption allowed for the definition of NAFLD is uncertain and inconsistent in the literature. As for eligibility for clinical trial, a consensus meeting recommended a limit of 21 and 14 standard drinks per week over a 2-year period in males and females, respectively (Sanyal et al. 2011). A standard alcoholic drink contains about 14 g of pure alcohol according to the National Institute of Alcohol Abuse and Alcoholism (NIAAA 2016). The diagnosis of alcohol consumption relies on thorough anamnesis and hetero-anamnesis, with a detailed 7-day diary of alcohol use. The diagnosis of the aspect of 'non-

alcoholic' therefore constitutes a first problem in the interpretation of any data on the prevalence and natural history of NAFLD.

Diagnosis

A first problem is the diagnosis of the aspect 'non-alcoholic' fatty liver. Patients may not accurately report the quantity of alcohol they consume. Laboratory tests, including elevation of AST (aspartate transaminase) more than ALT (alanine transaminase), and elevation of γ-GT (gamma-glutamyl transpeptidase) measurement may be helpful, but are inaccurate. Thorough anamnesis and hetero-anamnesis is the cornerstone of the diagnosis, which nevertheless may remain questionable.

A second problem is the distinction between steatosis and steatohepatitis. Abdominal ultrasound has a sensitivity of 70–75% and a specificity of 60–70% in diagnosing moderate to severe steatosis (Bellentani et al. 2000). Computed tomography scanning and magnetic resonance imaging are equally specific (100%) and sensitive (75%) in making the same distinction (Rinella et al. 2001). These non-invasive tools are thus not very sensitive, not able to accurately grade the steatosis, and not able to diagnose the presence of inflammation or fibrosis, and hence do not distinguish between NAFL and NASH.

Magnetic resonance imaging by spectroscopy or proton density fat fraction can accurately quantify the fat content of the liver, but the need for specific software and practical considerations limits its use to specific research centres (Heba et al. 2016; Idilman et al. 2016). Transient elastography based measurement of continuous attenuation parameters is a novel promising tool for quantifying hepatic fat in an ambulatory setting (de Ledinghen et al. 2016; Kwok et al. 2016). Yet, the utility of non-invasively quantifying steatosis in patients with NAFLD in routine clinical care is limited.

The diagnosis of steatohepatitis is even more complex. Laboratory tests, especially the elevation of aminotransferase levels, are inaccurate, although frequently regarded as a sign of liver cell damage and hence inflammation. Patients with abnormal liver tests may have pure steatosis without inflammation on liver biopsy, and 50% of the patients with biopsy-proven steatohepatitis have normal transaminases (Prati et al. 2002). The cut-off values for normal aminotransferase levels have been questioned, and

lowering the upper limit of normal ALT to 30 U/L in males and 19 U/L in females increases the sensitivity for the diagnosis of NASH from 42% to 74%, but specificity decreases from 80% to 42% (Kunde et al. 2005). Circulating levels of cytokeratin-18 fragments have been investigated as novel biomarkers for the presence of inflammation in patients with NAFLD (Chen et al. 2014; Cusi et al. 2014). However, this assay is currently not available in a clinical care setting. The gold standard for the diagnosis of steatohepatitis is still a liver biopsy. However, the invasive character of this procedure, limits its use on a larger scale.

Several well established non-invasive tool for assessment of advanced fibrosis in patients with NAFLD are currently available for clinicians. The NAFLD fibrosis score (NFS) is based on six readily available parameters (age, body mass index (BMI), hyperglycaemia, platelet count, albumin, and AST/ALT ratio). In a meta-analysis the NFS had an AUROC of 0.85 for predicting advanced fibrosis (Musso et al. 2011). FIB-4 index is an algorithm based on platelet count, age, AST, and ALT that offers dual cut-off values (patients with score <1.45 are unlikely, whereas patients with score >3.25 are likely to have advanced fibrosis) (Kaswala et al. 2016). The enhanced liver fibrosis (ELF) panel, that was recently approved for commercial use in Europe comprises of plasma levels of 3 matrix turnover proteins (hyaluronic acid, tissue inhibitor of metalloproteinase 1, and N-terminal procollagen III-peptide) had an AUROC of 0.87 with 89% sensitivity and 96% specificity for detecting advanced fibrosis (Rosenberg et al. 2004).

Vibration-controlled transient elastography (VCTE, FibroScan) is an ultrasound-based method to measures liver stiffness non-invasively, recently approved by the US FDA. Several studies examined the performance of VCTE in patients with suspected NAFLD. A study from the United States in 164 patients with biopsy-proven NAFLD (median BMI, 32.2 kg/m^2) showed sensitivity of 95% and specificity of 77% with AUROC of 93% for detection of advanced fibrosis. Interestingly, in 27% of the participants the VCTE yielded unreliable results (Tapper et al. 2016). In a similar Japanese study that included 142 patients with biopsy-proven NAFLD (mean BMI, 28.1 kg/m^2) the AUROC for identifying advanced fibrosis was 0.88 (Imajo et al. 2016). MR elastography is excellent for identifying varying degrees of fibrosis in patients with NAFLD (Kim et al. 2013; Loomba et al. 2014;

Imajo et al. 2016), however the availability of this tool beyond clinical trials and excellence research centres is currently limited.

Incidence and prevalence NAFLD

As already mentioned, the difficulty in diagnosing non-alcoholic steatosis, and the lack of accurate tools for the diagnosis of steatosis, constitute two major limitations in the acquisition of precise epidemiological data. Tissue sample selection (namely a liver biopsy) constitutes a third obstacle in reaching a diagnosis. Only a few studies evaluated the incidence of NAFLD over time. Two studies from Asia that followed subjects for 3–5 years with ultrasound or MRI reported an overall fatty liver incidence of 12–13.5% during the follow-up period (Sung et al. 2014; Wong et al. 2016), while a cohort study that followed 77,425 subjects free of NAFLD for an average of 4.5 years showed incidence rate of NAFLD documented by ultrasound of 29.7 per 1,000 person-years (Chang et al. 2016). A study from Israel stated an incidence rate of 28 per 1,000 person-years (Zelber-Sagi et al. 2012), and a similar incidence rate has been reported by an English study based on administrative ICD-10 coding (Whalley et al. 2007). A meta-analysis from 2016 approximated the pooled incidence of NAFLD to be 52 and 28 per 1,000 person-year in Asia and the West, respectively (Younossi et al. 2016).

The prevalence of NAFLD has been evaluated in several studies summarized in a 2016 meta-analysis that estimated the overall global prevalence of NAFLD diagnosed by imaging to be as high as 25% (Younossi et al. 2016). Regional variation exists with the highest prevalence of about 30–32% reported in South America and the Middle East and the lowest prevalence rate of about 13% reported from Africa (Younossi et al. 2016). Of note, the prevalence and the stage of liver disease increase with age (Frith et al. 2009; Koehler et al. 2012), and the prevalence in men is about two times higher than in women (Zelber-Sagi et al. 2006). Ethnic differences for NAFLD have been described (Masuoka and Chalasani 2013). These differences may be partially explained by the genetic variation related to the paratin-like phospholipase domain-containing protein 3 (PNPLA-3) gene (Dongiovanni et al. 2013).

Given that liver biopsy is the gold standard for the diagnosis of NASH and is not feasible in the general population, a direct assessment of NASH prevalence cannot be done. Yet, efforts to assess NASH prevalence indirectly were done. Fifty-nine per cent (59%) of NAFLD patients who

underwent a liver biopsy for a clinical indication were diagnosed with NASH. However, in NAFLD patients who had a liver biopsy without a specific clinical indication, NASH prevalence was estimated from 6.67 to 29.85%. Taking into account these numbers, the estimation of NASH in the general population ranges between 1.5 and 6.45% (Younossi et al. 2016).

NAFLD and NASH and the metabolic syndrome

The metabolic syndrome is defined by visceral overweight, dyslipidaemia, hyperinsulinemia or diabetes mellitus, and arterial hypertension, as proposed in the Third Report of the National Cholesterol Education Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel-ATP III) (Expert Panel on Detection and Treatment of High Blood Cholesterol 2001). This entity seems to be closely related with NAFLD and NASH. Some authors consider NAFLD and NASH as the hepatic manifestation of the metabolic syndrome. A number of epidemiological studies support a close relationship between the two entities. In patients with NAFLD, the metabolic syndrome, according to the criteria of the ATP III, is fully present in 30% of males and 60% of females. Visceral adiposity is present in 40% and 65% of males and females, respectively, and diabetes in 10% and 30%, respectively. These prevalence rates were significantly higher than in the control population. The metabolic syndrome is significantly more prevalent in patients with NASH compared to patients with simple steatosis (38% vs. 14%, $p = 0.004$) (Marchesini et al. 2003).

In patients with obesity, steatosis is present in 60–95%, according to the selection of patients and the procedure used for diagnosis (e.g. ultrasound or histology in a series of patients undergoing bariatric surgery). The BMI is an independent predictive factor for the accumulation of fat in the liver (Marchesini et al. 2003).

Globally, the prevalence of overweight and obesity has increased since 1980, and the increase has accelerated. The global age-standardized prevalence of obesity nearly doubled from 6.4% in 1980 to 12.0% in 2008 (Stevens et al. 2012). This increase of the prevalence of overweight in children and adolescents is of particular concern. The prevalence of diabetes is also increasing, and was estimated at around 9% worldwide in 2008 (Danaei et al. 2011). In the United States, 22% of the adult population fulfils the criteria of the metabolic syndrome (Lin and Pi-Sunyer 2007).

The natural history and outcome of NAFLD

Data on the natural history of NAFLD and NASH have the same three limitations as outlined for the prevalence data. In patients with NASH, 45% will exhibit fibrosis progression and 19% will ultimately develop cirrhosis (Fassio et al. 2004).

On the other hand, in a study that followed 109 NAFL patients (pure fatty liver without inflammation per biopsy) for a median follow-up time of 16.7 years, only one patient developed cirrhosis during the follow-up period, supporting the impression of a benign clinical course of this condition (Dam-Larsen et al. 2004). A long-term follow-up study (mean follow-up of 13.7 years) showed no increase in mortality in patients with elevated liver enzymes and pure steatosis on an initial biopsy. Patients with biopsy-proven NASH, on the other hand, had a higher risk of dying from cardiovascular disease (15.5% vs. 7.5%, $p = 0.04$) and from liver-related causes (2.8% vs. 0.2%, $p = 0.04$). Disease progression was, however, noted: 41% had fibrosis progression and 5.4% of patients developed cirrhosis, and this did not depend on features of inflammation on the initial biopsy (Ekstedt et al. 2006).

Patients with NAFLD have an increased overall mortality compared to matched control populations without NAFLD, with the most common cause of death being cardiovascular disease (Adams et al. 2005; Sayiner et al. 2017). Liver-related mortality is the second or third cause of death among patients with NAFLD while in the general population is only the twelfth leading cause (Younossi and Henry 2016).

In a recent meta-analysis, liver-specific mortality rates among NAFLD and NASH patients were determined to be 0.77 and 11.77 per 1,000 person-years respectively (Younossi et al. 2016). The most important histological factor of NAFLD associated with long-term mortality found to be the fibrosis stage (Younossi et al. 2011; Angulo et al. 2015).

Importantly, NASH is the second-most common cause for liver transplantation and will likely overtake viral hepatitis as the number one cause in the future, as more HCV patients are treated with effective antiviral regimens (Wong et al. 2015).

In patients with cryptogenic cirrhosis, more than 60% have features that might have been associated with NASH and in these patients, cirrhosis is believed to be an end stage of NASH (Ekstedt et al. 2006). Essentially,

cryptogenic cirrhosis accounts for 4% of the total indications for liver transplantation in Europe (Adam et al. 2012). Of note, NASH may recur after liver transplantation, further enforcing the concept of NASH as aetiology of cryptogenic cirrhosis (Maheshwari and Thuluvath 2006).

NAFLD is currently the third-most common cause of hepatocellular carcinoma (HCC) in the United States (Mohamad et al. 2016). NAFLD-related HCC incidence is growing by 9% annually, probably due to the increasing obesity and the metabolic syndrome (Younossi et al. 2015). In the Ekstedt series (Ekstedt et al. 2006), 2.3% developed HCC or 43% of those with documented cirrhosis. It is thus not clear whether the risk is comparable to the 10% cumulative risk usually reported in cirrhosis of any aetiology, but it might be higher (Smedile and Bugianesi 2005). Patients with NAFLD-related HCC are older, have a shorter survival time and are more likely to die from their primary liver cancer as compared to HCC patients of other aetiologies (Mohamad et al. 2016). Importantly, a Veteran Administration study reported that about 13% of HCC patients did not have cirrhosis, and having NAFLD was independently associated with HCC in the absence of cirrhosis. This study strengthens previous small studies reporting of HCC in NAFLD patients without cirrhosis (Mittal et al. 2016) and raise questions regarding the need to screen non-cirrhotic NASH patients for HCC development, although it is estimated that the risk for developing HCC in non-cirrhotic NAFLD patients is very low given the enormously large number of patients with NAFLD without cirrhosis within the general population (Younossi et al. 2015).

Risk factors reported to be associated with an increased risk of fibrosis are: age (40 or 50 years of age), the presence of diabetes, BMI 25 or 28 or 30, hypertriglyceridaemia, elevated transaminases two times the upper limit of normal, and AST/ALT ratio of 1 (Angulo et al. 1999; Adams et al. 2005). Patients with NAFLD and diabetes have a higher probability of cirrhosis and liver-related death, compared to NAFLD patients without diabetes (Abrams et al. 2004). In the Ekstedt series (Ekstedt et al. 2006), a 41% progression of fibrosis was associated with higher levels of ALT, a higher weight gain during follow-up, more severe insulin resistance, and more pronounced fatty infiltration. Of note, patients with NASH more frequently meet the criteria of the metabolic syndrome and are more likely to have visceral obesity compared to patients with simple steatosis. As it is believed that NASH is a subgroup of NAFLD at risk for progressive fibrosis, the metabolic syndrome

and its components clearly constitute a risk factor for fibrosis and cirrhosis, which will be a major burden of disease in view of the epidemic of obesity and diabetes and their related conditions.

Prevention and treatment

No specific treatment is clearly defined for NAFLD/NASH. There is however a consensus concerning implementation of behavioural and lifestyle change measures, as well as treating cases of advanced disease (Sourianarayanane et al. 2013). NASH treatment is driven by the stage of the disease and associated metabolic abnormalities such as overweight, hyperglycaemia, and hyperlipidaemia. In more advanced stages, drugs must be added to classical behavioural measures. Up to now, it is clear that patients with NASH benefit from physical exercise and balanced well oriented diets, probably through weight loss and decreased insulin resistance (Carulli et al. 2013).

In a meta-analysis of eight randomized controlled trials (RCT), patients who were able to lose at least 5% of body weight had improvement in hepatic steatosis, whereas 7% body weight reduction was associated also with improvement of the NAFLD activity score that includes steatosis, lobular inflammation, and ballooning (Musso et al. 2012). A recent prospective trial (including paired liver biopsies) in 261 patients, combining caloric restriction and physical activity, demonstrated a dose–response between weight loss and improvement in histopathology in NASH (Vilar-Gomez et al. 2015). Importantly, patients losing more than 5% body weight stabilized or improved fibrosis in 94% of the cases, however, only 50% of patients were able to achieve at least a 7% weight loss at 12 months in this trial (Vilar-Gomez et al. 2015).

Bariatric surgery is recommended in patients with a BMI greater than 40 kg/m^2 or BMI greater than 35 kg/m^2 and comorbidities. This procedure is highly effective and may be indicated even in patients with well-compensated cirrhosis, but it has high risk of morbidity if the patients have portal hypertension.

For patients who evolve towards advanced liver disease or HCC, liver transplantation can be indicated, although NASH relapse post transplantation is quite frequent.

Specific pharmacotherapy includes drugs used to treat the metabolic syndrome (e.g. statins, bezafibrate) and diabetes, as well as drugs with putative antioxidants effects. Two meta-analyses concluded that Metformin did not improve liver histology in patient with NAFLD and hence it not recommended for the treatment of NASH (Li et al. 2013). Pioglitazone, a ligand for the nuclear transcription factor peroxisome proliferator-activated receptor gamma, has been evaluated in several trial for the treatment of NASH and overall shown to improve liver histology. Of note, pioglitazone has been associated with potential harmful adverse events, such as weight gain and possibly bladder cancer (Aithal et al. 2008; Sanyal et al. 2010; Yau et al. 2013; Cusi et al. 2016; Tuccori et al. 2016). The American Association for the Study of Liver Disease (AASLD) practice guidelines recommend that this drug may be used to treat patients with biopsy-proven NASH, while the risks and benefits should be discussed with each patient before starting therapy (Chalasani et al. 2018).

Glucagon-like peptide-1 (GLP-1) analogues are used for the treatment of the metabolic syndrome and may be attractive option for NASH. A recent RCT evaluated treatment with liraglutide for 48 weeks in NASH patients and found improvement in resolution of steatohepatitis and decreased progression of fibrosis that was associated with weight loss (Armstrong et al. 2016).

Antioxidants are considered promising drugs for NASH, but the studies are considered preliminary. Among the putative antioxidant drugs, vitamin E has been evaluated as having the most consistent results, even with documented amelioration in liver pathology (Sanyal et al. 2010). On the other hand, prolonged use of vitamin E has been associated with adverse reactions such as risk of cerebral vascular haemorrhage, coagulation disturbance, and prostatic disorders.

Despite benefits observed in the aforementioned drug therapies, the long-term use of these drugs in NASH patients in terms of efficacy and safety remains unknown.

Overall conclusion

In spite of the availability of safe and effective vaccines and their proven effectiveness in reducing the chronic consequences of HBV infections, the current burden of disease associated with hepatitis B remains substantial. To

finally achieve the WHO goal of HBV elimination, continuous efforts will be required to keep prevention of hepatitis B on the agenda of public health officers worldwide, and to continue to improve treatment options for those already suffering chronic hepatitis B.

Even though the present burden of disease caused by HCV is somewhat lower as compared to HBV, the lack of an effective vaccine makes control of this global epidemic very challenging. Thus, the increased burden of CLD resulting from the high rates of HCV infection 20–30 years ago (the baby boomer generation, born 1946–1964), pose a substantial challenge to public health authorities. However, the recent introduction of highly efficient DAAs should contribute to the WHO goal of HCV elimination within the next decade on a global scale, provided the cost of treatment can be cut significantly.

Alcoholic liver disease remains a major cause of morbidity and mortality worldwide. Indeed, alcohol-related cirrhosis mortality rates have decreased in some countries during the past 30 years. However, rates are no longer declining in several other countries and are even increasing in low-and middle-income regions. Thus there is concern that, worldwide, the rate of alcoholic liver disease may increase in the next decades, especially in view of recent data which indicate that alcohol consumption is increasing in such low- and middle-income nations. In addition, rates of excessive alcohol intake appear to be rising in women and binge drinking has become a common pattern of excessive alcohol use.

Although data on the prevalence and natural history of NAFLD/NASH are scarce and suffer from multiple methodological problems, it is clear that, because of their association with the metabolic syndrome and its components, which are increasing to epidemic proportions in the Western population, NAFLD, NASH and HCC will constitute a major health problem in the near future. Moreover, recent data suggest that NAFLD may play a detrimental role in the outcome of COVID-19 infection (Portincasa et al. 2020). To cope with this challenge the pharma industry is involved in a major effort for development of new and effective anti-NASH agent(s) (Issa et al. 2018).

Key points

Liver cirrhosis and primary liver cancer are important public health problems worldwide, killing millions of people each year.
Viral hepatitis B and C, and alcoholic as well as non-alcoholic fatty liver disease, represent the major causes for CLDs worldwide, all of which have oncogenic potential.
Despite the availability and widespread use of effective and low cost hepatitis B vaccines, efforts will be required to keep the momentum of universal immunization programmes against HBV with special emphasis on administration of the birth dose of such a vaccine.
The new DAAs against HCV are highly efficacious in clearing chronic HCV infection. Equal access to the new antivirals is of the utmost importance.
As the development of a hepatitis C vaccine has not yet resulted in success, prevention and control measures form a major challenge to all those involved in public health.
Alcohol consumption remains a major health problem comprising ~6% of all deaths globally.
Fatty liver disease, steatohepatitis and CLDs associated with the metabolic syndrome, are rising to epidemic proportions globally.
New and efficient pharmacologic agents are urgently needed to cope with the rising number of patients with fatty liver disease.

ACKNOWLEDGEMENT

The authors of the present extensively revised edition of this chapter are indebted to the authors of the previous edition who developed the original template for this review.

REFERENCES

AASLD-IDSA Hepatitis C guidelines. (2020). *Hepatology*, **71**, 686–721.
Aberg, F., Helenius-Hietala, J., Puukka, P., and Jula, A. (2017). Binge drinking and the risk of liver events: a population-based cohort study. *Liver International*, **37**, 1373–81.
Abrams, G.A., Kunde, S.S., Lazenby, A.J., and Clements, R.H. (2004). Portal fibrosis and hepatic steatosis in morbidly obese subjects: a spectrum of nonalcoholic fatty liver disease. *Hepatology*, **40**, 475–83.
Adam, R., et al. (2012). Evolution of indications and results of liver transplantation in Europe. A report from the European Liver Transplant Registry (ELTR). *Journal of Hepatology*, **57**, 675–88.
Adams, L.A., et al. (2005). The natural history of nonalcoholic fatty liver disease: a population-based cohort study. *Gastroenterology*, **129**, 113–21.

Aithal, G.P., et al. (2008). Randomized, placebo-controlled trial of pioglitazone in nondiabetic subjects with nonalcoholic steatohepatitis. *Gastroenterology*, **135**, 1176–84.

Alexander, J.F., Lischner, M.W., and Galambos, J.T. (1971). Natural history of alcoholic hepatitis. II. The long-term prognosis. *American Journal of Gastroenterology*, **56**, 515–25.

Alter, M.J. (2007). Epidemiology of hepatitis C virus infection. *World Journal of Gastroenterology*, **13**, 2436–41.

Anderson, P., de Bruijn, A., Angus, K., Gordon, R., and Hastings, G. (2009). Impact of alcohol advertising and media exposure on adolescent alcohol use: a systematic review of longitudinal studies. *Alcohol & Alcoholism*, **44**, 229–43.

Andrade, D.L., et al. (2012). [A study about hepatitis C virus infection in patients with psoriasis in a Brazilian reference center]. *Acta Gastroenterologica Latinoamerica*, **42**, 285–90.

Andre, F.E., and Zuckerman, A.J. (1994). Review: protective efficacy of hepatitis B vaccines in neonates. *Journal of Medical Virology*, **44**, 144–51.

Angulo, P., Keach, J.C., Batts, K.P., and Lindor, K.D. (1999). Independent predictors of liver fibrosis in patients with nonalcoholic steatohepatitis. *Hepatology*, **30**, 1356–62.

Angulo, P., et al. (2015). Liver fibrosis, but no other histologic features, is associated with long-term outcomes of patients with nonalcoholic fatty liver disease. *Gastroenterology*, **149**, 389–97.e310.

Armstrong, M.J., et al. (2016). Liraglutide safety and efficacy in patients with non-alcoholic steatohepatitis (LEAN): a multicentre, double-blind, randomised, placebo-controlled phase 2 study. *The Lancet*, **387**, 679–90.

Askgaard, G., Gronbaek, M., Kjaer, M.S., Tjonneland, A., and Tolstrup, J.S. (2015). Alcohol drinking pattern and risk of alcoholic liver cirrhosis: a prospective cohort study. *Journal of Hepatology*, **62**, 1061–7.

Atta, A.M., Oliveira, I.S., Sousa, G.M., Parana, R., and Atta, M.L. (2010). Serum cytokine profile in hepatitis C virus carriers presenting cryoglobulinaemia and non-organ-specific autoantibodies. *Microbial Pathogenesis*, **48**, 53–6.

Banatvala, J.E., and Van Damme, P. (2003). Hepatitis B vaccine—do we need boosters? *Journal of Viral Hepatitis*, **10**, 1–6.

Barrio, E., et al. (2004). Liver disease in heavy drinkers with and without alcohol withdrawal syndrome. *Alcoholism: Clinical and Experimental Research*, **28**, 131–6.

Barry, M., and Cooper, C. (2007). Review of hepatitis B surface antigen-1018 ISS adjuvant-containing vaccine safety and efficacy. *Expert Opinion on Biological Therapy*, **7**, 1731–7.

Basnayake, S.K., and Easterbrook, P.J. (2016). Wide variation in estimates of global prevalence and burden of chronic hepatitis B and C infection cited in published literature. *Journal of Viral Hepatitis*, **23**, 545–59.

Baumert, T.F., Fauvelle, C., Chen, D.Y., and Lauer, G.M. (2014). A prophylactic hepatitis C virus vaccine: a distant peak still worth climbing. *Journal of Hepatology*, 61, S34–44.

Becker, U., Gronbaek, M., Johansen, D., and Sorensen, T.I. (2002). Lower risk for alcohol-induced cirrhosis in wine drinkers. *Hepatology*, **35**, 868–75.

Bellentani, S., et al. (2000). Prevalence of and risk factors for hepatic steatosis in Northern Italy. *Annals of Internal Medicine*, **132**, 112–17.

Beste, L.A., Leipertz, S.L., Green, P.K., Dominitz, J.A., Ross, D., and Ioannou, G.N. (2015). Trends in burden of cirrhosis and hepatocellular carcinoma by underlying liver disease in US veterans, 2001–2013. *Gastroenterology* 149, 1471–82.e1475; quiz e1417–78.

Blachier, M., Leleu, H., Peck-Radosavljevic, M., Valla, D.C., and Roudot-Thoraval, F. (2013). The burden of liver disease in Europe: a review of available epidemiological data. *Journal of Hepatology*, **58**, 593–608.

Bouchery, E.E., Harwood, H.J., Sacks, J.J., Simon, C.J., and Brewer, R.D. (2011). Economic costs of excessive alcohol consumption in the U.S., 2006. *American Journal of Preventive Medicine*, **41**, 516–24.

Bruix, J., Sherman, M., and Practice Guidelines Committee (2005). Management of hepatocellular carcinoma. *Hepatology*, **42**, 1208–36.

Burra, P., et al. (2010). Liver transplantation for alcoholic liver disease in Europe: a study from the ELTR (European Liver Transplant Registry). *American Journal of Transplantation*, **10**, 138–48.

Cao, G.W. (2009). Clinical relevance and public health significance of hepatitis B virus genomic variations. *World Journal of Gastroenterology*, **15**, 5761–9.

Carulli, L., Maurantonio, M., Hebbard, L., Baldelli, E., Loria, P., and George, J. (2013). Classical and innovative insulin sensitizing drugs for the prevention and treatment of NAFLD. *Current Pharmaceutical Design*, **19**, 5280–96.

Centers for Disease Control and Prevention (CDC) (1987). Recommendations of the Immunization Practices Advisory Committee. Update on hepatitis B prevention. *Morbidity and Mortality Weekly Report*, **36**, 353–60.

Chalasani, N., et al. (2018). The diagnosis and management of nonalcoholic fatty liver disease: practice guidance from the American Association for the Study of Liver Diseases. *Hepatology*, **67**, 328–57.

Chan, C.Y., Lee, S.D., and Lo, K.J. (2004). Legend of hepatitis B vaccination: the Taiwan experience. *Journal of Gastroenterology & Hepatology*, **19**, 121–6.

Chang, M.H., et al. (1997). Universal hepatitis B vaccination in Taiwan and the incidence of hepatocellular carcinoma in children. Taiwan Childhood Hepatoma Study Group. *New England Journal of Medicine*, **336**, 1855–9.

Chang, Y., et al. (2016). Metabolically healthy obesity and the development of nonalcoholic fatty liver disease. *American Journal of Gastroenterology*, **111**, 1133–40.

Chen, C.J., et al. (2006). Risk of hepatocellular carcinoma across a biological gradient of serum hepatitis B virus DNA level. *Journal of the American Medical Association*, **295**, 65–73.

Chen, J., Zhu, Y., Zheng, Q., and Jiang, J. (2014). Serum cytokeratin-18 in the diagnosis of non-alcoholic steatohepatitis: a meta-analysis. *Hepatology Research*, **44**, 854–62.

Chevaliez, S., and Pawlotsky, J.M. (2007). Hepatitis C virus: virology, diagnosis and management of antiviral therapy. *World Journal of Gastroenterology*, **13**, 2461–6.

Chhatwal, J., Kanwal, F., Roberts, M.S., and Dunn, M.A. (2015). Cost-effectiveness and budget impact of hepatitis C virus treatment with sofosbuvir and ledipasvir in the United States. *Annals of Internal Medicine*, **162**, 397–406.

Choo, Q.L., Kuo, G., Weiner, A.J., Overby, L.R., Bradley, D.W., and Houghton, M. (1989). Isolation of a cDNA clone derived from a blood-borne non-A, non-B viral hepatitis genome. *Science*, **244**, 359–62.

Consensus-Group (2000). Are booster immunisations needed for lifelong hepatitis B immunity? European Consensus Group on Hepatitis B Immunity. *The Lancet*, **355**, 561–5.

Corrao, G., and Arico, S. (1998). Independent and combined action of hepatitis C virus infection and alcohol consumption on the risk of symptomatic liver cirrhosis. *Hepatology*, **27**, 914–19.

Corti, D., Benigni, F., and Shouval, D. (2018). Viral envelope-specific antibodies in chronic hepatitis B virus infection. *Current Opinion in Virology*, **30**, 48–57.

Coursaget, P., et al. (1994). Twelve-year follow-up study of hepatitis B immunization of Senegalese infants. *Journal of Hepatology*, **21**, 250–4.

Cusi, K., et al. (2014). Limited value of plasma cytokeratin-18 as a biomarker for NASH and fibrosis in patients with non-alcoholic fatty liver disease. *Journal of Hepatology*, **60**, 167–74.

Cusi, K., et al. (2016). Long-term pioglitazone treatment for patients with nonalcoholic steatohepatitis and prediabetes or type 2 diabetes mellitus: a randomized Trial. *Annals of Internal Medicine*, **165**, 305–15.

D'Alessandro, A.M., et al. (1991). The predictive value of donor liver biopsies for the development of primary nonfunction after orthotopic liver transplantation. *Transplantation* 51, 157–63.

Dam-Larsen, S., et al. (2004). Long term prognosis of fatty liver: risk of chronic liver disease and death. *Gut*, **53**, 750–5.

Danaei, G., et al. (2011). National, regional, and global trends in fasting plasma glucose and diabetes prevalence since 1980: systematic analysis of health examination surveys and epidemiological studies with 370 country-years and 2.7 million participants. *The Lancet*, **378**, 31–40.

Davis, G.L., Alter, M.J., El-Serag, H., Poynard, T., and Jennings, L.W. (2010). Aging of hepatitis C virus (HCV)-infected persons in the United States: a multiple cohort model of HCV prevalence and disease progression. *Gastroenterology*, **138**, 513–21, 521, e511–16.

Day, C.P. (2000). Who gets alcoholic liver disease: nature or nurture? *Journal of the Royal College of Physicians London*, **34**, 557–62.

de Ledinghen, V., et al. (2016). Controlled attenuation parameter for the diagnosis of steatosis in non-alcoholic fatty liver disease. *Journal of Gastroenterology & Hepatology*, **31**, 848–55.

Diehl, A.M. (1997). Alcoholic liver disease: natural history. *Liver Transplant Surgery*, **3**, 206–11.

Dongiovanni, P., Anstee, Q.M., and Valenti, L. (2013). Genetic predisposition in NAFLD and NASH: impact on severity of liver disease and response to treatment. *Current Pharmaceutical Design*, **19**, 5219–38.

Dubuisson, J., and Cosset, F.L. (2014). Virology and cell biology of the hepatitis C virus life cycle: an update. *Journal of Hepatology*, 61, S3–S13.

Duclos, P. (2003). Safety of immunization and adverse events following vaccination against hepatitis B. *Journal of Hepatology*, 39 Suppl 1, S83–8.

Ekstedt, M., et al. (2006). Long-term follow-up of patients with NAFLD and elevated liver enzymes. *Hepatology*, **44**, 865–873.

El-Serag, H.B., and Kanwal, F. (2014). Epidemiology of hepatocellular carcinoma in the United States: where are we? Where do we go? *Hepatology*, **60**, 1767–75.

Enomoto, N., et al. (1999). Estriol sensitizes rat Kupffer cells via gut-derived endotoxin. *Am Journal of Physiology*, 277, G671–677.

European Association for Study of the Liver (2015). EASL recommendations on treatment of hepatitis C 2015. *Journal of Hepatology*, **63**, 199–236.

European Association for the Study of the Liver (2017). EASL 2017 clinical practice guidelines on the management of hepatitis B virus infection. *Journal of Hepatology*, **67**, 370–98.

European Association for the Study of the Liver (2018). EASL clinical practice guidelines: management of alcohol-related liver disease. *Journal of Hepatology*, **69**, 154–181.

Everhart, J.E. (2003). In vino veritas? *Journal of Hepatology*, **38**, 245–6.

Expert Panel on Detection and Treatment of High Blood Cholesterol (2001). Executive summary of the third report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III). *Journal of the American Medical Association*, **285**, 2486–97.

Farci, P. (2003). Delta hepatitis: an update. *Journal of Hepatology*, 39 Suppl 1, S212–19.

Fassio, E., Alvarez, E., Dominguez, N., Landeira, G., and Longo, C. (2004). Natural history of nonalcoholic steatohepatitis: a longitudinal study of repeat liver biopsies. *Hepatology*, **40**, 820–6.

Fattovich, G., Stroffolini, T., Zagni, I., and Donato, F. (2004). Hepatocellular carcinoma in cirrhosis: incidence and risk factors. *Gastroenterology*, 127, S35–50.

Fletcher, L.M., Dixon, J.L., Purdie, D.M., Powell, L.W., and Crawford, D.H. (2002). Excess alcohol greatly increases the prevalence of cirrhosis in hereditary hemochromatosis. *Gastroenterology*, **122**, 281–9.

Forns, X., and Bukh, J. (1999). The molecular biology of hepatitis C virus. Genotypes and quasispecies. *Clinical Liver Disease*, **3**, 693–716, vii.

Frank, C., et al. (2000). The role of parenteral antischistosomal therapy in the spread of hepatitis C virus in Egypt. *The Lancet* 355, 887–91.

Frith, J., Day, C.P., Henderson, E., Burt, A.D., and Newton, J.L. (2009). Non-alcoholic fatty liver disease in older people. *Gerontology*, **55**, 607–13.

Gerberding, J.L., and Henderson, D.K. (1992). Management of occupational exposures to bloodborne pathogens: hepatitis B virus, hepatitis C virus, and human immunodeficiency virus. *Clinical Infectious Diseases*, **14**, 1179–85.

Goldstein, S.T., Zhou, F., Hadler, S.C., Bell, B.P., Mast, E.E., and Margolis, H.S. (2005). A mathematical model to estimate global hepatitis B disease burden and vaccination impact. *International Journal of Epidemiology*, **34**, 1329–39.

Gower, E., Estes, C., Blach, S., Razavi-Shearer, K., and Razavi, H. (2014). Global epidemiology and genotype distribution of the hepatitis C virus infection. *Journal of Hepatology*, 61, S45–57.

Gurusamy, K.S., et al. (2013). Is sustained virological response a marker of treatment efficacy in patients with chronic hepatitis C viral infection with no response or relapse to previous antiviral intervention? *PLoS One*, 8, e83313.

Hadler, S.C., and Margolis, H.S. (1992). Hepatitis B immunization: vaccine types, efficacy, and indications for immunization. *Current Clinical Topics in Infectious Diseases*, **12**, 282–308.

Heba, E.R., et al. (2016). Accuracy and the effect of possible subject-based confounders of magnitude-based MRI for estimating hepatic proton density fat fraction in adults, using MR spectroscopy as reference. *Journal of Magnetic Resonance Imaging*, **43**, 398–406.

Hollinger, F.B. (1989). Factors influencing the immune response to hepatitis B vaccine, booster dose guidelines, and vaccine protocol recommendations. *American Journal of Medicine*, **7**, 36S–40S.

Hughes, S.A., Wedemeyer, H., and Harrison, P.M. (2011). Hepatitis delta virus. *The Lancet* 378, 73–85.

Idilman, I.S., et al. (2016). A comparison of liver fat content as determined by magnetic resonance imaging-proton density fat fraction and MRS versus liver histology in non-alcoholic fatty liver disease. *Acta Radiology*, **57**, 271–8.

Imajo, K., et al. (2016). Magnetic resonance imaging more accurately classifies steatosis and fibrosis in patients with nonalcoholic fatty liver disease than transient elastography. *Gastroenterology*, **150**, 626–37.e627.

Issa, D., Patel, V., Sanyal, A. J. (2018). Future therapy for non-alcoholic fatty liver disease. *Liver International*. **38** (Suppl 1):56–63.

Jacobson, I.M., et al. (2011). Telaprevir for previously untreated chronic hepatitis C virus infection. *New England Journal of Medicine*, **364**, 2405–16.

Jilg, W., Schmidt, M., and Deinhardt, F. (1988). Persistence of specific antibodies after hepatitis B vaccination. *Journal of Hepatology*, **6**, 201–7.

Kao, J.H., and Chen, D.S. (2005). Hepatitis B vaccination: to boost or not to boost? *The Lancet*, **366**, 1337–8.

Kaswala, D.H., Lai, M., and Afdhal, N.H. (2016). Fibrosis assessment in nonalcoholic fatty liver disease (NAFLD) in 2016. *Digestive Diseases & Sciences*, **61**, 1356–64.

Kato, N., et al. (1994). Genetic drift in hypervariable region 1 of the viral genome in persistent hepatitis C virus infection. *Journal of Virology*, **68**, 4776–84.

Kim, D., Kim, W.R., Talwalkar, J.A., Kim, H.J., and Ehman, R.L. (2013). Advanced fibrosis in nonalcoholic fatty liver disease: noninvasive assessment with MR elastography. *Radiology*, **268**, 411–19.

Koehler, E.M., et al. (2012). Prevalence and risk factors of non-alcoholic fatty liver disease in the elderly: results from the Rotterdam study. *Journal of Hepatology*, **57**, 1305–1311.

Kumar, M., et al. (2007). A randomized controlled trial of lamivudine to treat acute hepatitis B. *Hepatology*, **45**, 97–101.

Kunde, S.S., Lazenby, A.J., Clements, R.H., and Abrams, G.A. (2005). Spectrum of NAFLD and diagnostic implications of the proposed new normal range for serum ALT in obese women. *Hepatology*, **42**, 650–6.

Kurbanov, F., Tanaka, Y., and Mizokami, M. (2010). Geographical and genetic diversity of the human hepatitis B virus. *Hepatology Research*, **40**, 14–30.

Kwo, P.Y., et al. (2010). Efficacy of boceprevir, an NS3 protease inhibitor, in combination with peginterferon alfa-2b and ribavirin in treatment-naïve patients with genotype 1 hepatitis C infection (SPRINT-1): an open-label, randomised, multicentre phase 2 trial. *The Lancet* 376, 705–16.

Kwok, R., et al. (2016). Screening diabetic patients for non-alcoholic fatty liver disease with controlled attenuation parameter and liver stiffness measurements: a prospective cohort study. *Gut*, **65**, 1359–68.

Lalazar, G., Rund, D., and Shouval, D. (2007). Screening, prevention and treatment of viral hepatitis B reactivation in patients with haematological malignancies. *British Journal of Haematology*, **136**, 699–712.

Lauer, G.M., and Walker, B.D. (2001). Hepatitis C virus infection. *New England Journal of Medicine*, **345**, 41–52.

Lavanchy, D. (2004). Hepatitis B virus epidemiology, disease burden, treatment, and current and emerging prevention and control measures. *Journal of Viral Hepatology*, **11**, 97–107.

Lee, W.M. (1997). Hepatitis B virus infection. *New England Journal of Medicine*, **337**, 1733–45.

Leroux-Roels, G. (2015). Old and new adjuvants for hepatitis B vaccines. *Medical Microbiology & Immunology*, **204**, 69–78.

Leuridan, E., and Van Damme, P. (2011). Hepatitis B and the need for a booster dose. *Clinical & Infectious Diseases*, **53**, 68–75.

Li, W. (2015). The hepatitis B virus receptor. *Annual Review of Cellular & Developmental Biology*, **31**, 125–47.

Li, Y., Liu, L., Wang, B., Wang, J., and Chen, D. (2013). Metformin in non-alcoholic fatty liver disease: a systematic review and meta-analysis. *Biomedical Reports*, **1**, 57–64.

Liaw, Y.F. (2013). Reduction of cirrhosis and hepatocellular carcinoma with antiviral therapy in chronic hepatitis B. *Hepatology*, **58**, 1856.

Lim, S.S., et al. (2012). A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010. *The Lancet*, **380**, 2224–60.

Lin, C.L., and Kao, J.H. (2011). The clinical implications of hepatitis B virus genotype: recent advances. *Journal of Gastroenterology & Hepatology*, 26 Suppl 1, 123–30.

Lin, C.L., and Kao, J.H. (2016). Review article: novel therapies for hepatitis B virus cure—advances and perspectives. *Alimentary Pharmacology & Therapeutics*, **44**, 213–22.

Lin, S.X., and Pi-Sunyer, E.X. (2007). Prevalence of the metabolic syndrome among US middle-aged and older adults with and without diabetes—a preliminary analysis of the NHANES 1999–2002 data. *Ethnicity & Disease*, **17**, 35–9.

Loomba, R., et al. (2014). Magnetic resonance elastography predicts advanced fibrosis in patients with nonalcoholic fatty liver disease: a prospective study. *Hepatology*, **60**, 1920–8.

MacLachlan, J.H., Locarnini, S., and Cowie, B.C. (2015). Estimating the global prevalence of hepatitis B. *The Lancet* 386, 1515–17.

Maheshwari, A., and Thuluvath, P.J. (2006). Cryptogenic cirrhosis and NAFLD: are they related? *American Journal of Gastroenterology*, **101**, 664–8.

Marchesini, G., et al. (2003). Nonalcoholic fatty liver, steatohepatitis, and the metabolic syndrome. *Hepatology*, **37**, 917–23.

Masson, S., et al. (2014). Clinical but not histological factors predict long-term prognosis in patients with histologically advanced non-decompensated alcoholic liver disease. *Liver International*, **34**, 235–42.

Masuoka, H.C., and Chalasani, N. (2013). Nonalcoholic fatty liver disease: an emerging threat to obese and diabetic individuals. *Annals of New York Academy of Sciences*, **1281**, 106–22.

Mehta, B., Kumar, V., Chawla, S., Jindal, H., and Bhatt, B. (2014). Hepatitis C: is a vaccine the solution? *Human Vaccines & Immunotherapy*, **10**, 417–19.

Mele, A., et al. (2008). Acute hepatitis B 14 years after the implementation of universal vaccination in Italy: areas of improvement and emerging challenges. *Clinical Infectious Diseases*, **46**, 868–75.
Memon, M.I., and Memon, M.A. (2002). Hepatitis C: an epidemiological review. *Journal of Viral Hepatitis*, **9**, 84–100.
Mittal, S., et al. (2016). Hepatocellular carcinoma in the absence of cirrhosis in United States veterans is associated with nonalcoholic fatty liver disease. *Clinical Gastroenterology & Hepatology*, **14**, 124–31.e121.
Mohamad, B., et al. (2016). Characterization of hepatocellular carcinoma (HCC) in non-alcoholic fatty liver disease (NAFLD) patients without cirrhosis. *Hepatology International*, **10**, 632–9.
Musso, G., Cassader, M., Rosina, F., and Gambino, R. (2012). Impact of current treatments on liver disease, glucose metabolism and cardiovascular risk in non-alcoholic fatty liver disease (NAFLD): a systematic review and meta-analysis of randomised trials. *Diabetologia*, **55**, 885–904.
Musso, G., Gambino, R., Cassader, M., and Pagano, G. (2011). Meta-analysis: natural history of non-alcoholic fatty liver disease (NAFLD) and diagnostic accuracy of non-invasive tests for liver disease severity. *Annals of Medicine*, **43**, 617–49.
Naing, C., Mak, J.W., Wai, N., and Maung, M. (2013). Diabetes and infections-hepatitis C: is there type 2 diabetes excess in hepatitis C infection? *Current Diabetes Reports*, **13**, 428–34.
Naveau, S., Giraud, V., Borotto, E., Aubert, A., Capron, F., and Chaput, J.C. (1997). Excess weight risk factor for alcoholic liver disease. *Hepatology*, **25**, 108–11.
Ni, Y., et al. (2014). Hepatitis B and D viruses exploit sodium taurocholate co-transporting polypeptide for species-specific entry into hepatocytes. *Gastroenterology*, **146**, 1070–83.
NIAAA (2016). *What is a Standard Drink?* [Online] Available at: https://pubsniaaanih gov/publications/practitioner/pocketguide/pocket_guide2htm
OECD (2017). Alcohol consumption among adults. In: *Health at a Glance*, pp. 72–3. [Online] Available at: https://www.oecd-ilibrary.org/docserver/health_glance-2017-17-en.pdf?expires=1600277418&id=id&accname=guest&checksum=2A74BD93A67555F6DF7299B70037222F
Panel, A.I.H.G. (2015). Hepatitis C guidance: AASLD-IDSA recommendations for testing, managing, and treating adults infected with hepatitis C virus. *Hepatology*, **62**, 932–54.
Parana, R., Vitvitski, L., and Pereira, J.E. (2008). Hepatotropic viruses in the Brazilian Amazon: a health threat. *Brazilian Journal of Infectious Diseases*, **12**, 253–6.
Park, S.H., and Rehermann, B. (2014). Immune responses to HCV and other hepatitis viruses. *Immunity*, **40**, 13–24.
Parlesak, A., Billinger, M.H., Bode, C., and Bode, J.C. (2002). Gastric alcohol dehydrogenase activity in man: influence of gender, age, alcohol consumption and smoking in a caucasian population. *Alcohol & Alcoholism*, **37**, 388–93.
Perz, J.F., Armstrong, G.L., Farrington, L.A., Hutin, Y.J., and Bell, B.P. (2006a). The contributions of hepatitis B virus and hepatitis C virus infections to cirrhosis and primary liver cancer worldwide. *Journal of Hepatology*, **45**, 529–38.
Perz, J.F., Elm, J.L., Jr., Fiore, A.E., Huggler, J.I., Kuhnert, W.L., and Effler, P.V. (2006b). Near elimination of hepatitis B virus infections among Hawaii elementary school children after universal infant hepatitis B vaccination. *Pediatrics*, **118**, 1403–8.
Pimpin, L., et al. (2018). Burden of liver disease in Europe: epidemiology and analysis of risk factors to identify prevention policies. *Journal of Hepatology*, 69(3), 718–35.
Polaris Observatory, H.C.V.C. (2017). Global prevalence and genotype distribution of hepatitis C virus infection in 2015: a modelling study. *Lancet Gastroenterology & Hepatology*, **2**, 161–76.
Portincasa, P., Krawczyk, M., Smyk, W. et al. (2020). COVID-19 and non-alcoholic fatty liver disease: Two intersecting pandemics. *European Journal of Clinical Investigation*, 50, e13338.

Prati, D. (2006). Transmission of hepatitis C virus by blood transfusions and other medical procedures: a global review. *Journal of Hepatology*, **45**, 607–16.
Prati, D., et al. (2002). Updated definitions of healthy ranges for serum alanine aminotransferase levels. *Annals of Internal Medicine*, **137**, 1–10.
Rabinovich L, B.P.-B., de Vries H, Tiessen J, Clift J, and Reding A. (2009). *The Affordability of Alcoholic Beverages in the European Union: Understanding the Link Between Alcohol Affordability, Consumption and Harms*. Rand Report. [Online] Available at: http://bc.ias.org.uk/pdfs/advocacy-course/rand-technical-reports-2009-tr689.pdf
Raimondo, G., Caccamo, G., Filomia, R., and Pollicino, T. (2013). Occult HBV infection. *Seminars in Immunopathology*, **35**, 39–52.
Rehermann, B., Ferrari, C., Pasquinelli, C., and Chisari, F.V. (1996). The hepatitis B virus persists for decades after patients' recovery from acute viral hepatitis despite active maintenance of a cytotoxic T-lymphocyte response. *Nature Medicine*, **2**, 1104–8.
Rehm, J., Samokhvalov, A.V., and Shield, K.D. (2013). Global burden of alcoholic liver diseases. *Journal of Hepatology*, **59**, 160–8.
Rehm, J., et al. (2010). Alcohol as a risk factor for liver cirrhosis: a systematic review and meta-analysis. *Drug & Alcohol Review*, **29**, 437–45.
Rinella, M.E., et al. (2001). Body mass index as a predictor of hepatic steatosis in living liver donors. *Liver Transplantation*, **7**, 409–14.
Rizzetto, M. (2015). Hepatitis D virus: introduction and epidemiology. *Cold Spring Harbor Perspectives in Medicine*, 5, a021576.
Rizzetto, M., and Ciancio, A. (2012). Epidemiology of hepatitis D. *Seminars in Liver Disease*, **32**, 211–19.
Rizzetto, M., and Verme, G. (1985). Delta hepatitis--present status. *Journal of Hepatology*, **1**, 187–93.
Roche, B., and Samuel, D. (2012). Liver transplantation in delta virus infection. *Seminars in Liver Disease*, **32**, 245–55.
Roche, B., and Samuel, D. (2014). Prevention of hepatitis B virus reinfection in liver transplant recipients. *Intervirology*, **57**, 196–201.
Rosenberg, W.M., et al. (2004). Serum markers detect the presence of liver fibrosis: a cohort study. *Gastroenterology*, **127**, 1704–13.
Sanyal, A.J., et al. (2011). Endpoints and clinical trial design for nonalcoholic steatohepatitis. *Hepatology*, **54**, 344–53.
Sanyal, A.J., et al. (2010). Pioglitazone, vitamin E, or placebo for nonalcoholic steatohepatitis. *New England Journal of Medicine*, **362**, 1675–85.
Sayiner, M., et al. (2017). Variables associated with inpatient and outpatient resource utilization among medicare beneficiaries with nonalcoholic fatty liver disease with or without cirrhosis. *Journal of Clinical Gastroenterology*, **51**, 254–60.
Schweitzer, A., Horn, J., Mikolajczyk, R.T., Krause, G., and Ott, J.J. (2015). Estimations of worldwide prevalence of chronic hepatitis B virus infection: a systematic review of data published between 1965 and 2013. *The Lancet*, **386**, 1546–55.
Sheron, N. (2016). Alcohol and liver disease in Europe--Simple measures have the potential to prevent tens of thousands of premature deaths. *Journal of Hepatology*, **64**, 957–67.
Shield, K.D., Rylett, M., Gmel, G., Gmel, G., Kehoe-Chan, T.A., and Rehm, J. (2013). Global alcohol exposure estimates by country, territory and region for 2005--a contribution to the Comparative Risk Assessment for the 2010 Global Burden of Disease Study. *Addiction*, **108**, 912–22.
Shouval, D. (2003). Hepatitis B vaccines. *Journal of Hepatology*, 39 Suppl 1, S70–6.
Shouval, D., Roggendorf, H., and Roggendorf, M. (2015). Enhanced immune response to hepatitis B vaccination through immunization with a Pre-S1/Pre-S2/S vaccine. *Medical Microbiology & Immunology*, **204**, 57–68.

Shouval, D., and Shibolet, O. (2013). Immunosuppression and HBV reactivation. *Seminars in Liver Disease*, **33**, 167–77.
Smedile, A., and Bugianesi, E. (2005). Steatosis and hepatocellular carcinoma risk. *Europeans Review for Medical & Pharmacological Sciences*, **9**, 291–3.
Smith, B.D., Jorgensen, C., Zibbell, J.E., and Beckett, G.A. (2012). Centers for Disease Control and Prevention initiatives to prevent hepatitis C virus infection: a selective update. *Clinical Infectious Diseases*, 55 Suppl 1, S49–53.
Sorensen, T.I., Orholm, M., Bentsen, K.D., Hoybye, G., Eghoje, K., and Christoffersen, P. (1984). Prospective evaluation of alcohol abuse and alcoholic liver injury in men as predictors of development of cirrhosis. *The Lancet*, **2**, 241–4.
Sourianarayanane, A., Pagadala, M.R., and Kirwan, J.P. (2013). Management of non-alcoholic fatty liver disease. *Minerva Gastroenterologica e Dietologica*, **59**, 69–87.
Stevens, G.A., et al. (2012). National, regional, and global trends in adult overweight and obesity prevalences. *Popular Health Metrics*, **10**, 22.
Stockwell, T., et al. (2013). Minimum alcohol prices and outlet densities in British Columbia, Canada: estimated impacts on alcohol-attributable hospital admissions. *American Journal of Public Health*, **103**, 2014–20.
Sunbul, M. (2014). Hepatitis B virus genotypes: global distribution and clinical importance. *World Journal of Gastroenterology*, **20**, 5427–34.
Sung, K.C., Wild, S.H., and Byrne, C.D. (2014). Development of new fatty liver, or resolution of existing fatty liver, over five years of follow-up, and risk of incident hypertension. *Journal of Hepatology*, **60**, 1040–5.
Tan, A., Koh, S., and Bertoletti, A. (2015). Immune response in hepatitis b virus infection. *Cold Spring Harbor Perspectives in Medicine*, 5, a021428.
Tan, J., and Lok, A.S. (2007). Update on viral hepatitis: 2006. *Curr Opin Gastroenterol* 23, 263–267.
Tandon, B.N. (1997). Epidemiological trends of viral hepatitis in Asia. In: Rizzetto, M., Purcell, R.H., Gerin, J.L., and Verme, G. (eds.) *Viral Hepatitis and Liver Disease*, pp. 559–61. Turin: Edizioni Minerva Medica.
Tanwar, S., and Dusheiko, G. (2012). Is there any value to hepatitis B virus genotype analysis? *Current Gastroenterology Reports*, **14**, 37–46.
Tapper, E.B., Challies, T., Nasser, I., Afdhal, N.H., and Lai, M. (2016). The performance of vibration controlled transient elastography in a US cohort of patients with nonalcoholic fatty liver disease. *American Journal of Gastroenterology*, **111**, 677–84.
Terrault, N.A., et al. (2018). Update on prevention, diagnosis, and treatment of chronic hepatitis B: AASLD 2018 hepatitis B guidance. *Hepatology*, **67**, 1560–99.
Tolstrup, J.S., Jensen, M.K., Tjonneland, A., Overvad, K., and Gronbaek, M. (2004). Drinking pattern and mortality in middle-aged men and women. *Addiction*, **99**, 323–30.
Torres, J.R. (1996). Hepatitis B and hepatitis delta virus infection in South America. *Gut*, 38 Suppl 2, S48–55.
Tseng, C.H., and Lai, M.M. (2009). Hepatitis delta virus RNA replication. *Viruses*, **1**, 818–31.
Tuccori, M., Filion, K.B., Yin, H., Yu, O.H., Platt, R.W., and Azoulay, L. (2016). Pioglitazone use and risk of bladder cancer: population based cohort study. *British Medical Journal*, 352, i1541.
Veldt, B.J., et al. (2007). Sustained virologic response and clinical outcomes in patients with chronic hepatitis C and advanced fibrosis. *Annals of Internal Medicine*, **147**, 677–84.
Venters, C., Graham, W., and Cassidy, W. (2004). Recombivax-HB: perspectives past, present and future. *Expert Review of Vaccines*, **3**, 119–29.
Viana, S., Parana, R., Moreira, R.C., Compri, A.P., and Macedo, V. (2005). High prevalence of hepatitis B virus and hepatitis D virus in the western Brazilian Amazon. *American Journal of Tropical Medicine & Hygiene*, **73**, 808–14.

Vilar-Gomez, E., et al. (2015). Weight loss through lifestyle modification significantly reduces features of nonalcoholic steatohepatitis. *Gastroenterology*, **149**, 367–78.e365; quiz e314–65.

Viviani, S., Jack, A., Hall, A.J., Maine, N., Mendy, M., Montesano, R., and Whittle, H.C. (1999). Hepatitis B vaccination in infancy in the Gambia: protection against carriage at 9 years of age. *Vaccine*, **17**, 2946–50.

Wagenaar, A.C., Salois, M.J., and Komro, K.A. (2009). Effects of beverage alcohol price and tax levels on drinking: a meta-analysis of 1003 estimates from 112 studies. *Addiction*, **104**, 179–90.

Wagenaar, A.C., Tobler, A.L., and Komro, K.A. (2010). Effects of alcohol tax and price policies on morbidity and mortality: a systematic review. *American Journal of Public Health*, **100**, 2270–8.

Wainwright, R.B., Bulkow, L.R., Parkinson, A.J., Zanis, C., and McMahon, B.J. (1997). Protection provided by hepatitis B vaccine in a Yupik Eskimo population--results of a 10-year study. *Journal of Infectious Diseases*, **175**, 674–7.

Whalley, S., Puvanachandra, P., Desai, A., and Kennedy, H. (2007). Hepatology outpatient service provision in secondary care: a study of liver disease incidence and resource costs. *Clinical Medicine (London)*, **7**, 119–24.

World Health Organization (WHO) (2001). *Expanded Programme on Immunization. Introduction of Hepatitis B Vaccination into Childhood Immunization Services: Management Guidelines, Including Information for Health Workers and Parents (WHO/V&B/01.31)*. Geneva: WHO. [Online] Available at: http://www.who.int/vaccines-documents/DocsPDF01/www613.pdf

World Health Organization (WHO) (2004). Hepatitis B vaccines (World Health Organization position paper). *Weekly Epidemiological Record*, **79**, 255–63.

World Health Organization (WHO) (2006). *Vaccines and Biologicals. WHO Vaccine Preventable Disease Monitoring System. Global Summary 2006 (Data up to 2005)*. [Online] Available at: http://www.who.int/vaccines-documents/GlobalSummary.pdf

World Health Organization (WHO) (2012). *Prevention and Control of Viral Hepatitis Infection: Framework for Global Action*. Geneva: WHO.

World Health Organization (WHO) (2014). *Global Status Report on Noncommunicable Diseases 2014*. Geneva: WHO Library Cataloguing-in-Publication Data.

World Health Organization (WHO) (2017). *Global Hepatitis Report, 2017*. [Online] Available at: http://www.whoint/hepatitis/publications/global-hepatitis-report2017/en/.

Wiktor, S.Z., and Hutin, Y.J. (2016). The global burden of viral hepatitis: better estimates to guide hepatitis elimination efforts. *The Lancet* 388, 1030–1.

Williams, R., et al. (2018). Disease burden and costs from excess alcohol consumption, obesity, and viral hepatitis: fourth report of the Lancet Standing Commission on Liver Disease in the UK. *The Lancet*, 391(10125), 1097–107.

Wong, R.J., et al. (2015). Nonalcoholic steatohepatitis is the second leading etiology of liver disease among adults awaiting liver transplantation in the United States. *Gastroenterology*, **148**, 547–55.

Wong, V.W., et al. (2016). Long-term clinical outcomes after fatty liver screening in patients undergoing coronary angiogram: a prospective cohort study. *Hepatology*, **63**, 754–63.

Yau, H., Rivera, K., Lomonaco, R., and Cusi, K. (2013). The future of thiazolidinedione therapy in the management of type 2 diabetes mellitus. *Current Diabetes Reports*, **13**, 329–41.

Younossi, Z., and Henry, L. (2016). Contribution of alcoholic and nonalcoholic fatty liver disease to the burden of liver-related morbidity and mortality. *Gastroenterology*, **150**, 1778–85.

Younossi, Z.M., Koenig, A.B., Abdelatif, D., Fazel, Y., Henry, L., and Wymer, M. (2016). Global epidemiology of nonalcoholic fatty liver disease-meta-analytic assessment of prevalence, incidence, and outcomes. *Hepatology*, **64**, 73–84.

Younossi, Z.M., et al. (2015). Association of nonalcoholic fatty liver disease (NAFLD) with hepatocellular carcinoma (HCC) in the United States from 2004 to 2009. *Hepatology*, **62**, 1723–30.

Younossi, Z.M., et al. (2011). Pathologic criteria for nonalcoholic steatohepatitis: interprotocol agreement and ability to predict liver-related mortality. *Hepatology*, **53**, 1874–82.

Yu, S.C., and Lin, C.W. (2013). Early-stage splenic diffuse large B-cell lymphoma is highly associated with hepatitis C virus infection. *Kaohsiung Journal of Medical Science*, **29**, 150–6.

Yurdaydin, C. (2019). New treatment options for delta virus: Is a cure in sight? *Viral Hepatitis*, **26**(6), 618–26.

Zanetti, A.R., et al. (2005). Long-term immunogenicity of hepatitis B vaccination and policy for booster: an Italian multicentre study. *The Lancet*, **366**, 1379–84.

Zanetti, A.R., Van Damme, P., and Shouval, D. (2008). The global impact of vaccination against hepatitis B: a historical overview. *Vaccine*, **26**, 6266–73.

Zelber-Sagi, S., et al. (2012). Predictors for incidence and remission of NAFLD in the general population during a seven-year prospective follow-up. *Journal of Hepatology*, **56**, 1145–51.

Zelber-Sagi, S., Nitzan-Kaluski, D., Halpern, Z., and Oren, R. (2006). Prevalence of primary non-alcoholic fatty liver disease in a population-based study and its association with biochemical and anthropometric measures. *Liver International*, **26**, 856–63.

8.17

Emerging and re-emerging infections

David L. Heymann and Vernon J.M. Lee

Introduction to emerging and re-emerging infections

The microbial world is complex, dynamic, and constantly evolving. Infectious organisms reproduce rapidly, mutate frequently, cross the species barrier between animal hosts and humans, and adapt with relative ease to their new environments. Because of these characteristics, many infectious organisms are able to alter their epidemiology, their virulence, and their susceptibility to anti-infective drugs.

When disease is caused by an organism that is newly identified and not known previously to infect humans, or has changed in susceptibility to an anti-infectious drug, it is commonly called an emerging infectious disease, or simply an emerging infection. When a disease is caused by and infectious organism that has re-entered human populations or changed in epidemiology, it is called a re-emerging infection. A report published by the United States Institute of Medicine in 1992 first called attention to emerging and re-emerging infectious diseases as evidence that the fight against infectious diseases was far from won, despite great advances in the development of anti-infective drugs and vaccines (Lederberg et al. 1992).

All forms of infectious organisms—bacteria, viruses, parasites, and prions—are able to emerge or re-emerge in human populations, and it is estimated that up to 70% or more of all emerging infections have a source in animals (zoonotic infections). Infected humans may become clinically ill or remain asymptomatic; human-to-human transmission may or may not occur; and virulence may remain the same, or increase or decrease. Changes in clinical expression, transmissibility and virulence can occur gradually through a process of adaptive mutation, or abruptly as the result of a sudden genetic exchange (reassortment) during reproduction and/or replication.

Epidemiology of emerging and re-emerging infections

In general, when an infectious organism emerges or re-emerges in human populations there are three potential outcomes. Some do not transmit from humans to other humans; some have limited human-to-human transmission, with the potential to re-emerge when the conditions are favourable; and others emerge, spread widely, and become endemic.

There are many factors that influence the transmission of an infectious organism to and among humans including the type of organism, the incubation period (time from exposure to the infectious organism to clinical disease and infectiousness), clinical outcomes, mutations, and human genetics. The following subsections describe some examples of common infections based on their ability to transmit among humans.

Infections that do not transmit from humans to other humans

Rabies is a zoonotic infection that commonly occurs after the bite of a rabid animal, and variant Creutzfeldt–Jakob disease (vCJD) is a zoonotic infection associated with beef and bovine spongiform encephalopathy (BSE) or ‘mad cow disease’. These infections cause illness in humans but cannot transmit from one human to another unless there is iatrogenic transmission through non-sterile medical procedures, blood transfusion, or organ transplant. In several instances, corneal transplantation from a person who died with undiagnosed rabies infection has caused rabies in transplant recipients. The identification of several humans with vCJD associated with blood transfusion demonstrates its potential to spread iatrogenically within the human population.

Infections with limited human-to-human transmission

Some zoonotic infections in humans are able to initially transmit from person to person, but then are unable to sustain transmissibility. Thought to have a rodent reservoir in the sub-Saharan rain forest, the monkeypox virus infects humans who come in contact with an infected animal. Transmission among humans is through direct contact with infected droplets, and it is thought that smallpox vaccination (vaccinia virus vaccine) protects against infection with the monkeypox virus. Most outbreaks of human monkeypox

cease spontaneously. As smallpox immunity continues to wane since the cessation of vaccination after the certification of smallpox eradication in 1980, the epidemiology of human monkeypox appears to be changing with longer chains of human-to-human transmission after re-emergence before outbreaks spontaneously cease (Pru Rimoin et al. 2010). The case fatality rate, that can approach 10% when infection becomes generalized, appears to remain the same as transmissibility declines.

The monkeypox virus caused a limited, multistate outbreak of non-generalized cutaneous infection in the United States, linked to Gambian giant rats imported from West Africa and distributed to pet shops as exotic pets. In the pet shops, prairie dogs, thought to have been the source of the outbreak, appear to have been infected with the monkeypox virus by the Gambian giant rats, and they then transmitted the virus onwards to humans (Guarner et al. 2004).

Another zoonotic infectious organism with limited human-to-human transmission is the Ebola virus. With a relatively short incubation period, infection leads to rapid progression to disease with a severe and often fatal outcome that limits the potential for contact of infected humans with others. Because of its high virulence and short incubation period, most epidemiologists believe that the Ebola virus cannot sustain prolonged transmission, and will therefore not become endemic in humans in its current form. The 2014 Ebola outbreak in West Africa spread more widely than previous outbreaks in part due to the lack of a robust emergency response when it was first identified. The same time-proven response measures together with adequate capacity to implement these measures can stop Ebola outbreaks: hospital infection control and protection of health workers; contact tracing with contacts placed under daily fever surveillance and isolation if Ebola symptoms develop; and community engagement with health education on transmission, safe burial practices, and prevention. The 2014 outbreak spread in an area where health systems were being rebuilt after years of civil strife and there was thus inadequate infection control, low access to healthcare, and great movement of populations. This led to the World Health Organization (WHO) declaration of a Public Health Emergency of International Concern (PHEIC), and the outbreak was finally contained after the input of substantial international aid resources over the prolonged outbreak period of more than 2 years (http://apps.who.int/ebola/ebola-situation-reports).

One common infectious organism that is able to sustain human-to-human transmission for varying periods of time—the influenza virus—demonstrates the great diversity of infectious organisms. The RNA viruses that cause seasonal human epidemics of influenza are highly unstable genetically and mutate frequently during replication. Mutation, or genetic drift, results in susceptibility of individuals regardless of their exposure to the viruses before drift, requiring annual antigenic modifications in seasonal influenza vaccines to ensure protection. The short incubation period, and the mild illness caused by the majority of infections, especially among pre-school and school-aged children, result in rapid transmission potential that leads to seasonal influenza epidemics.

Different subtypes of influenza viruses are found across a variety of avian species, and these occasionally infect humans directly from birds, or from intermediary animal populations that have been infected by birds. The H5N1 type A avian influenza virus that was first identified in Hong Kong as the cause of human infections in 1997 continues to cause occasional severe infections in humans. However, H5N1 remains a zoonotic disease and does not transmit easily from person to person. This virus, like many other avian influenza viruses, has the potential to mutate and gain the epidemiological characteristics that would permit it to spread easily among humans to cause an epidemic or a pandemic.

A novel A(H1N1) virus however, emerged in 2009 and unlike the H5N1 virus, spread easily among humans with sustained transmission around the world, resulting in a pandemic (http://www.who.int/csr/disease/swineflu/en/).

Pandemics are, by definition, infectious disease epidemics that spread to affect susceptible populations over much of the world, and they can in some cases lead to endemicity. The highly fatal 1918 influenza pandemic is thought to have originated from an avian influenza virus that may have undergone adaptive mutation in non-human mammals over time, attaining human transmissibility characteristics before crossing the animal/human species (Taubenberger and Morens 2006).

Other influenza pandemics in 1957, 1968, and 2009 are thought to have been caused by genetic reassortments during the intracellular replication process in an animal dually infected with viruses containing human and avian influenza genetic components (Guan et al. 2010; York and Donis 2012. The 2009 H1N1 pandemic influenza virus, for example, showed genes from

swine, avian, and human influenza. Although the 2009 pandemic virus is thought to have its human origins in the Americas, risk factors for emergence of influenza viruses in humans are thought to be highest in areas such as East and Southeast Asia, shown by the many other type A influenza viruses that have recently emerged there, where there are large populations of aquatic birds (the hosts of many different types of avian influenza viruses) and where humans live in close proximity to animals such as pigs and poultry that may be easily infected by these aquatic birds.

Finally, the Middle East Respiratory Syndrome Coronavirus (MERS Co-V) and the Nipah virus periodically cause zoonotic infections in humans that can lead to outbreaks of varying size, generally caused by iatrogenic transmission in health facilities with poor infection control (Cho et al. 2016; Hassan et al. 2018).

Infections with sustained transmission

Infection with the human immunodeficiency virus (HIV), thought to have emerged from a non-human primate sometime in the late nineteenth or early twentieth century, has a long incubation (latent) period during which it is difficult to detect clinically. Though amplification of transmission is thought to have begun during the 1970s through risky sexual behaviour, HIV escaped detection until the early 1980s by which time it had already spread throughout the world and caused a series of epidemics that led to its endemicity. During 2016, an estimated 36.7 million persons were living with HIV infections, and an estimated 1 million of them died from AIDS-related illness worldwide, placing HIV infection high on the list of causes of human morbidity and mortality (http://www.aidsdatahub.org/sites/default/files/publication/UNAIDS_Global_AIDS_Update_2017_Data_book_2017_en.pdf).

It is generally believed by most medical epidemiologists that most other endemic infectious diseases likewise had their origins as emerging infections at some time in the past when the conditions for emergence lined up in such a way that emergence across the species barrier occurred.

Geographic distribution of emerging and re-emerging infections

Emerging infections have the potential to occur in every country and on every continent (Fig. 8.17.1). Though the term emerging infections was introduced in the early 1990s, the previous 40 years had seen a panoply of newly identified infections in humans on every continent. The year 1976 was especially illustrative of this phenomenon with the identification of the 1976 swine flu virus (H1N1), thought to be a direct descendant of the virus that caused the pandemic of 1918, at a military base in Fort Dix (United States) (Sencer 2011); the identification of *Legionella pneumophila* as the cause of an outbreak of severe respiratory illness among a group of veterans staying at a hotel in downtown Philadelphia (United States), initially feared to be a human outbreak of swine influenza (H1N1) (Centers for Disease Control and Prevention (CDC) 2012a); and the identification of the Ebola virus as the cause of simultaneous outbreaks of haemorrhagic fever in Sudan and the Democratic Republic of Congo (then called Zaire) (CDC 2012b).

Nine years earlier, in 1967, the Marburg virus was identified for the first time in an outbreak in Germany that caused 25 primary infections and seven deaths among laboratory workers who were infected by handling monkeys from Uganda; and six secondary cases in health workers who took care of primary cases, with subsequent spread to family members (Slenczka and Klenk 2007). A member of the same filovirus family as Ebola, the Marburg virus has caused sporadic outbreaks in Africa since the 1970s, followed by the major West African outbreak from 2013 to 2015 (http://www.who.int/news-room/fact-sheets/detail/ebola-virus-disease).

Since the Marburg virus was first identified in 1967 there have been over 40 other newly identified infectious organisms in humans, an average of one per year. Some of the most recent emerging/re-emerging infections are MERS Co-V and Zika virus infections, the latter associated with microcephaly in children born to women who were infect during pregnancy (Heymann et al 2016).

Economic impact of emerging and re-emerging infections

Outbreaks caused by emerging and re-emerging infections are costly (Fig. 8.17.2). They consume healthcare resources and divert them from endemic disease problems, result in productivity loss, and decrease trade and tourism revenue. At times they economically devastate entire sectors. This has occurred after major outbreaks of emerging or re-emerging infections during

the past 20 years, with economic losses ranging from an estimated US$36 million after the re-emergence of cholera in Tanzania in 1998 (WHO 2007a, 2013), to approximately US$39 billion after the emergence of BSE in the United Kingdom during the period 1990–1998 (WHO 2007a). During 2009, after the pandemic of H1N1 had been initially identified in Mexico, the Mexican economy lost an estimated US$2 billion, mainly due to decreased trade and tourism, and the nosocomial outbreak of MERS Co-V in South Koreas has been unofficially estimated by some to have cost the economy a loss of revenue of up to US$ 10 billon (https://www.adb.org/news/features/middle-east-respiratory-syndrome-testing-asia-s-ability-address-emerging-diseases).

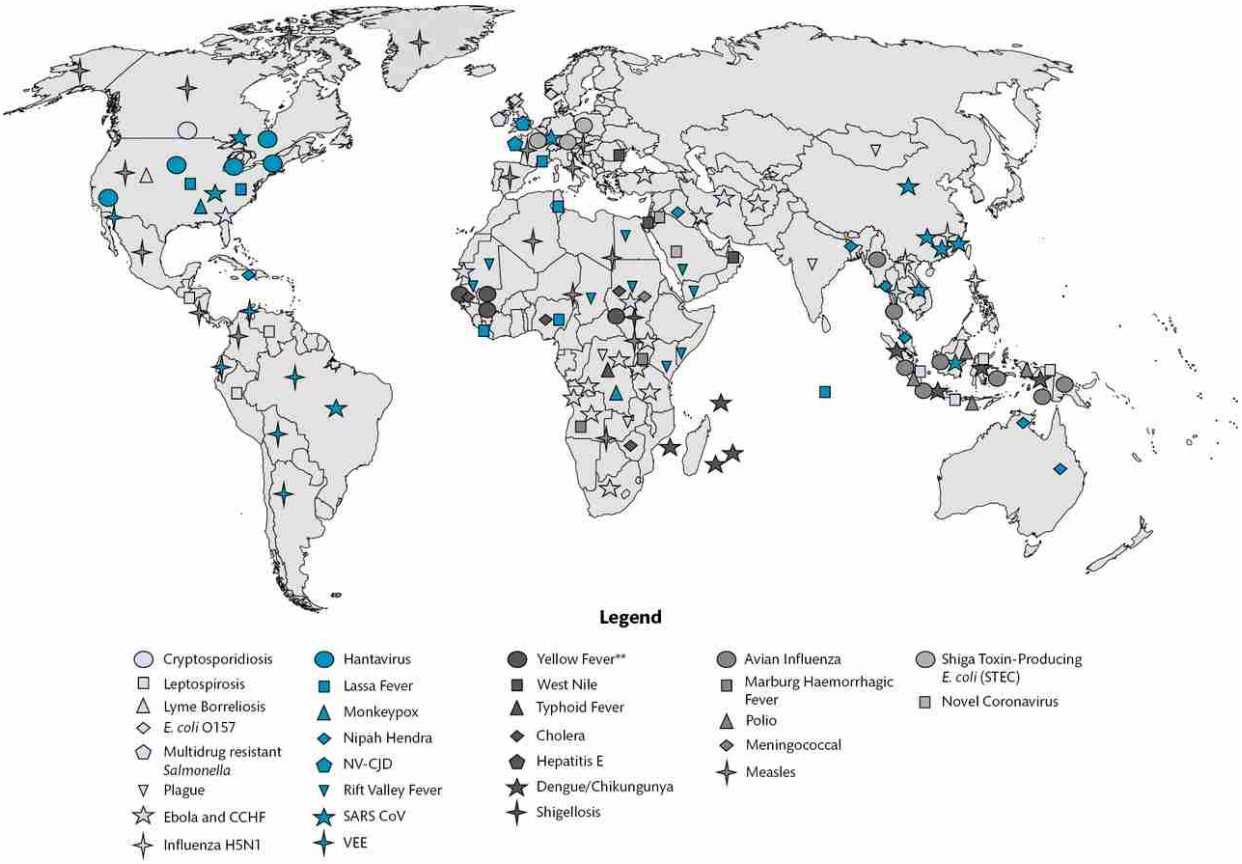


Fig. 8.17.1 Selected emerging and re-emerging infectious diseases, 1996–2012.
Source: data from World Health Organization (WHO), Global Alert and Response (GAR), Disease Outbreaks by Year, Copyright © WHO 2014, available from http://www.who.int/csr/don/archive/year/en/.

Costs from outbreaks can be direct, indirect, or intangible. The most apparent direct costs are those associated with medical care of patients. Indirect costs arise at the individual level from loss of work and caregiving by relatives, and through interventions aimed at reducing the impact of the outbreak such as quarantining of contacts or closure of schools. Indirect costs also arise at the macro-economic level because of their impact on travel, tourism, trade, and consumer confidence. Fear of transmission causes international tourists to choose alternative holiday locations, and local populations to avoid any perceived source of infection such as foods from animals that may have been implicated in the outbreak, restaurants, and other public leisure venues—sectors of the economy that are significant contributors to the gross domestic product (GDP) of many countries. Misinformed consumers avoided purchasing pork during the 2009 pandemic of H1N1 influenza even though it was not a continued source of human infection as clearly demonstrated by a 10% decrease in pork trade on the Chicago Mercantile Exchange during the month after news of the pandemic was first announced; and avoided fresh vegetables during the 2011 outbreak of *Escherichia coli* O104 in Europe when speculation occurred as to its cause. Finally, there are intangible costs such as the loss of life. Economists have attempted to estimate the economic cost of death, but fail to capture other non-economic and sociocultural changes that can be a consequence of a highly lethal outbreak.

The severe acute respiratory syndrome (SARS) was a good example of an emerging infection with severe indirect costs, as it was responsible for sizeable economic losses and insecurity in financial markets across Asia and worldwide. With fewer than 9,000 reported cases, the outbreak was estimated by the Asian Development Bank to have cost Asian countries an estimated US$18 billion in GDP terms for 2003, and up to US$60 billion of gross expenditure and business losses (Asian Development Bank 2003).

In 2008, the World Bank estimated that the cost of a severe influenza pandemic (which would apply to other pandemics of similar severity) could be up to US$2–3 trillion taking into account healthcare costs and impact on the economy (Burns et al. 2008). It is therefore important to prepare for and respond to infectious disease outbreaks in order to minimize their healthcare and economic impact. Though the economic impact of the 2009 pandemic of H1N1 did not reach these proportions, the avian influenza virus and other

novel flu viruses continues to circulate among poultry in Asia with potential to cause a highly lethal and costly pandemic in poultry and/or humans.

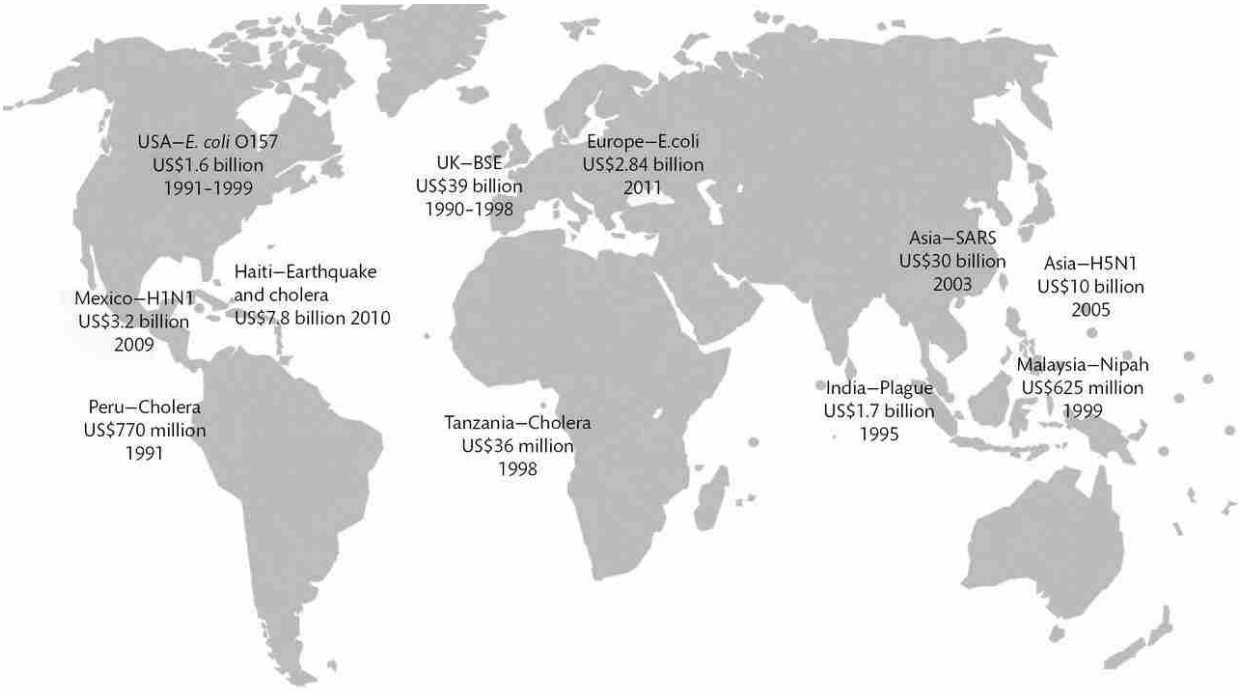


Fig. 8.17.2 Direct economic impact of selected infectious disease outbreaks, 1990–2017.
Source: data from Food Poison Journal, German *E. coli* O104:H4 Outbreak—$2.84 Billion in Human Damage, June 2011, Copyright © 2011 Marler Clark, LLP, available from http://www.foodpoisonjournal.com/foodborne-illness-outbreaks/german-e-coli-o104h4-outbreak—284-billion-in-human-damage/#.UPi_lB2R89U; European Commission, Factsheet Haiti Two Years on: European Commission's Actions to Help Rebuild the Country, 2012, Copyright © European Union, 1995–2014, available from http://ec.europa.eu/europeaid/news/documents/factsheet_haiti_en.pdf; Marc P. Girarda et al., The 2009 A (H1N1) influenza virus pandemic: A review, *Vaccine*, Volume 28, Issue 31, pp. 4895–4902, Copyright © 2010 Published by Elsevier Ltd; and WHO (2006, 2007a).

Factors influencing emergence and re-emergence

Many external factors may align in such a manner as to provide opportunities for the emergence or re-emergence of infectious diseases. They range from weakened public health infrastructure and failure of safety procedures/regulations to increases in population, anthropogenic activities,

or climate change, civil disturbance/human displacement, and human behaviour that varies from misperceptions about the seriousness of the risks that cause emerging infections and the use of anti-infective drugs, to the safety of public health interventions and the desire to deliberately cause terror and harm.

Weakened public health infrastructure

Weakening of public health infrastructure resulted in part from decreased investment in public health during the second half of the twentieth century. This is compounded by poverty in lower-resourced countries where there are competing priorities for development, and suboptimal access to healthcare and sanitation. Deficient infrastructure also leads to a lack of the robust public health practices that could prevent the emergence of infectious diseases, or prepare communities to respond to them should they emerge.

Aedes aegypti has now become well established in many large cities worldwide following the deterioration of mosquito control campaigns during the 1970s. The resurgence of the *Aedes* species has been confounded by the adoption of modern consumer habits in urban areas where discarded household appliances, tyres, plastic food containers, and jars create abundant mosquito breeding sites. With the increase in *Aedes* species there has been an increased risk of dengue outbreaks. Prior to 1970, nine countries, mainly in Latin America, reported outbreaks of dengue. In 1983, 13 countries in Latin America and Asia reported dengue outbreaks, and by 1998, 1.2 million cases were reported from 56 countries. During 2001, 69 countries reported dengue outbreaks, and it is now endemic in more than 100 countries in Africa, the Americas, the Eastern Mediterranean, South East Asia, and the Western Pacific. By 2010, reported dengue infections to WHO exceeded 2.2 million annually with more than 20,000 deaths, and estimates placed the actual number of infections worldwide at between 50–100 million (WHO 2012a). By 2015 reported cases of dengue had increased to 3.2 million, with major outbreaks in Brazil, Indonesia, Thailand, Vietnam, Bangladesh, India, Pakistan, and Cape Verde. During 2012, an unexpected outbreak of dengue occurred in Portugal, with 191 probable cases reported (WHO 2012b). Currently WHO estimates that there are approximately 400 million dengue infections each year, while reporting of cases by countries remains much lower (http://www.who.int/denguecontrol/epidemiology/en/).

In 2005, the Chikungunya virus, likewise transmitted by *Aedes* species, emerged and spread throughout several Indian Ocean islands. A total of 3,100 human infections were reported by a sentinel network on La Réunion within the first 6 months of the outbreak, leading to an estimate of over 204,000 human infections by March 2006. In 2007, the Chikungunya virus spread north to Europe, where it caused an outbreak with 292 suspected human infections in northern Italy, and during 2010–2011 Chikungunya epidemics occurred throughout southern India (WHO 2012c). During 2017 Chikungunya continued to cause outbreaks in Asia, Latin America, and Africa, and in 2017 Italy again reported an outbreak of Chikungunya with 183 reported cases (http://www.who.int/csr/don/29-september-2017-chikungunya-italy/en/).

During the twentieth century the Zika virus spread from Africa to Asia, then in 2013 was first identified in island nations in the Pacific, and subsequently spread to the Americas. A major outbreak of Zika virus infection in South and Central America in 2015 caused concern because of the association between Zika virus infection and microcephaly and other neurologic defects among newborns to mothers who were infected with Zika, and Guillain Barre syndrome in adults (Heymann et al 2016). As there is no vaccine for Zika, vector control and personal protection against mosquito bites were recommended, but the outbreak subsided in Brazil and other Latin American and Caribbean countries by the end of 2016 (http://www.who.int/emergencies/zika-virus/situation-report/en/).

A major outbreak of yellow fever was first reported by Angola in late 2015, and yellow fever then spread to neighbouring urban areas in the Democratic Republic of Congo, and by unvaccinated Chinese guest workers to China who returned while still in the incubation period. Major vaccination campaigns were held, resulting in a worldwide shortage of yellow fever vaccine and the successful use of fractional doses of existing vaccine supplies (http://www.who.int/emergencies/yellow-fever/mediacentre/qa-fractional-dosing/en/). In 2017, Brazil reported major yellow fever outbreaks throughout the country, and as in Africa, major vaccination campaigns were conducted using fractional doses of yellow fever vaccine because of the continued worldwide shortage of vaccine.

Lapses in routine childhood immunization coverage due to weakened childhood immunization programmes in Russia resulted in the re-emergence of diphtheria, with major epidemics in the early 1990s. Reported cases of

diphtheria in the Russian Federation increased from just over 1,200 in 1990 to over 5,000 in 1993. In 2017 there were major diphtheria outbreaks in Bangladesh and Yemen among the children of displaced persons who had poor access to vaccines (http://www.who.int/csr/don/22-december-2017-diphtheria-yemen/en/; http://www.who.int/csr/don/13-december-2017-diphtheria-bangladesh/en/).

Lapses in routine yellow fever vaccination programmes in sub-Saharan Africa since the 1950s had likewise left large susceptible populations in both rural and urban areas of the region, with sporadic urban outbreaks in cities in Côte d'Ivoire (2001), Senegal and Guinea (2002), and Burkina Faso (2004) (Jamison et al. 2006), prior to the major outbreak in Angola in 2015. This is in contrast to the successful smallpox eradication programme that achieved sustained investment and close links between policy and political support, research, and field programmes that contributed to its success. Progress in the eradication of polio that has decreased the annual number of children paralysed from an estimated 1 000 children per day in 1988 to fewer than 100 in 2017 worldwide (http://polioeradication.org/polio-today); and the decrease in measles mortality by 74% (2007) after intensified efforts at measles mortality reduction were begun during 2000 (WHO 2008a). By 2016, there were 89,780 measles deaths reported globally—marking the first year that measles deaths had fallen below 100,000 per year (http://www.who.int/immunization/diseases/measles/en/).

Surveillance systems are critical in the early detection of outbreaks and during their containment, and to monitor progress. Setting up of surveillance systems requires substantial investment in infrastructure and manpower, and the commitment to sustain these systems in the long term. Most epidemiologists recognize that it was in part because of substandard surveillance systems in developing countries that HIV rapidly spread during the late 1970s, and was not detected until it was first identified in the United States in the early 1980s.

Failure of safety procedures/regulations

Substandard universal precautions and hospital regulations during the 1980s led to breaches in sterile injection practices and nosocomial infections of HIV in the former USSR and Romania, together infecting over 250 children, accompanied by high levels of hepatitis B in both patients and health

workers (Matic et al. 2006). Likewise substandard universal precautions led to nosocomial outbreaks of Ebola haemorrhagic fever in the Democratic Republic of Congo in 1976 and 1995, where syringes and/or failed barrier nursing amplified the transmission to patients, health workers, and the community (CDC 2012b). Continued lapses in universal precautions led to the amplification of recent outbreaks of Ebola infection in the Democratic Republic of Congo, Uganda, Sierra Leone, Liberia, and Guinea; and contributed to nosocomial transmission of SARS in hospitals in China, Hong Kong, Taiwan, Singapore, Vietnam, and Canada. Outbreaks then spread to family members of infected health workers, and then into the communities in which they lived. In 2015 the MERS coronavirus, thought to be transmitted to humans directly from camels, was first identified in humans in the Kingdom of Saudi Arabia in 2012, and has since caused nosocomial outbreaks in hospitals in the Middle East, and a major nosocomial outbreak in South Korea after a 68 year old male developed severe respiratory symptoms after returning from the Middle East. Nosocomial spread occurred to at least 15 hospitals in South Korea before transmission was finally interrupted (Oh et al. 2018).

Blood transfusions are another potential source of infection, especially for emerging infections that are not yet known and therefore impossible to detect by screening. A substantial proportion of HIV cases in the early phases of the epidemic were transmitted by blood transfusion (Curran et al. 1984). Although most countries now test banked blood for HIV and many other known infectious organisms, many resource-poor countries struggle to bank blood and ensure that transfusions are safe (Goodnough et al. 2003).

Changes in the process of rendering the carcasses of ruminant animals for the preparation of bone meal, which was fed to other ruminant animals, are thought to have been the cause of the outbreak of BSE in cattle that also led, in May 1995, to the death of a 19-year-male in the United Kingdom, the first human death from what is now known to be variant vCJD or human bovine spongiform encephalopathy (hBSE) (Trevitt and Singh 2003). The BSE and hBSE outbreaks demonstrate the health consequences of regulations for rendering that had changed over a 10-year period prior to 1995, inadvertently permitting rendered parts of cattle infected with the BSE-causing prion to contaminate bone meal used for livestock feed. The most likely source of human infection is thought to be through the preparation or consumption of contaminated meat and/or beef products. The BSE and

hBSE outbreaks led to the recognition of the need for stronger government intervention along the entire ‘feed to food’ continuum to better ensure the safety of foodstuffs for human consumption.

Major food borne outbreaks have also been related to contamination of the food chain with organisms such as *E coli* and *Salmonella* species. A major outbreak of bloody diarrhoea, with a high frequency of complications and death occurred in Germany in 2011, for example, linked to *E coli O104:H4* —contaminated fenugreek seeds imported from Egypt (Buchholz et al. 2011); and major outbreaks in North America in 2017 have been caused by *E.coli*, *Salmonella* species, listeria, and norovirus (CDC 2012a–d).

Population shifts including rapid increase and uncontrolled urbanization

Population shifts can lead to the emergence, re-emergence, and transmission of infectious diseases. Classic examples are the plague epidemics in Europe in the fourteenth century thought to have originated in Asia and spread first along the Silk Road, then by trade on merchant ships sailing the Mediterranean; and smallpox that is thought to have arrived in the Americas from Europe during the years of exploration and colonization that began in the fifteenth century. These epidemics incited fear and killed many who became infected, threatening to bring down entire civilizations. The world is much more easily travelled and interconnected since then, with large numbers of people moving from continent to continent every day. This has facilitated the international spread of emerging diseases as illustrated by the SARS epidemic that spread from a medical doctor who had treated patients with a severe atypical pneumonia in Guangdong Province, China. He became ill, but travelled to Hong Kong for a wedding and spent a night in a hotel where he passed the infection to tourists and other travellers who then seeded outbreaks in Vietnam, Singapore, Canada, and other parts of the world within days.

Population growth can result in increased disease transmission, especially where the infrastructure is unable to cope. The world’s population more than doubled in the second half of the twentieth century, accelerating most rapidly in the developing countries of the tropics and subtropics. Rural–urban migration has resulted in crowded living conditions, and inadequacy of water and sanitation systems, and other basic infrastructure. In 1950, there

were two urban areas in the world with populations greater than 7 million; by 1990, this number had risen to 23. As of 2017, as urban migration continues, there are 47 megacities, each with more than 10 million people, some of which have not been able to keep up with the development of infrastructure that will prevent major outbreaks related to water and sanitation.

Population increases in Latin America resulted in breakdowns in sanitation and water systems in large coastal cities. In 1991, when cholera re-emerged in Peru after having been quiescent for approximately 100 years, it rapidly spread throughout other urban areas in Latin America. Thought to have originated from contaminated seafood on the Peruvian coast, the disease caused nearly 400,000 reported cases and over 4,000 deaths in 16 South American countries that year. By 1995, there were more than 700,000 reported cases and just over 6,000 deaths in the Americas, and in 2012, cholera remains endemic in many areas (Delgado et al. 2011). A cholera outbreak in Haiti that began in 2010 after a major earthquake that destroyed Haitian infrastructure, caused by a strain of *V cholera* imported from Nepal, had resulted in approximately 10,000 deaths by 2017, and is now endemic in Haiti causing occasional outbreaks; while in 2017 a major outbreak of cholera began in Yemen among displaced populations, and has continued during the first half of 2018 (http://www.who.int/gho/epidemic_diseases/cholera/cases/en/).

Uncontrolled urbanization, with overcrowding and substandard living conditions in slum areas of Asia and Africa, has likewise contributed to the re-emergence of tuberculosis (TB) and plague (WHO 2011). HIV infection, also concentrated in urban areas, has facilitated the re-emergence of TB, showing the sometimes dynamic interaction among infectious agents. An outbreak of plague occurred in five states in India in 1994, where almost 700 suspected bubonic or pneumonic plague cases and 56 deaths were reported; a smaller outbreak of 16 persons with pneumonic plague was reported from northern India in 2002 (Gupta and Sharma 2007). In 2016 an outbreak of pulmonic and bubonic plague was reported from rural Madagascar with over 2,000 cases, and it rapidly spread to urban areas in just under 50% of all administrative areas in the country (http://www.who.int/news-room/fact-sheets/detail/plague/).

Anthropogenic activities or climate change

Deforestation that disrupts natural habitats of animals, and forces animals, searching for food, into closer contact with humans has been linked to the emergence and re-emergence of Lassa Fever in West Africa, and sine nombre virus in North America. First identified in 1969 when two nurses died with a haemorrhagic fever syndrome in Nigeria, the Lassa Fever virus is now known to be transmitted to humans from human food supplies and/or the household environment contaminated by urine and/or other excreta of infected rodents (Fichet-Calvet and Rogers 2009; Senior 2009). A major Lassa Fever outbreak began in 2018 in Nigeria, with approximately 1,800 suspected cases during the first 3 months of the year, occurred in areas where low levels of rainfall are thought to have forced rodents to move closer to humans in search of food (http://www.who.int/csr/don/01-march-2018-lassa-fever-nigeria/en/).

In many instances, rats invade human living spaces in search of food because rainforests, a natural habitat, have been destroyed. Sin nombre virus is a hantavirus, first identified in an outbreak in the southwestern United States in 1993. It is now known to spread from infected rodents to humans through aerosolized excreta found in dust of homes that have been invaded by rodents (CDC 2012c).

The Nipah virus outbreaks in Malaysia and Singapore in 1998–1999 are another example where evidence suggests that anthropogenic activities, including clearing of forests for agricultural purposes, encroached on the natural reservoir for the Nipah virus, the fruit bat. The virus, excreted in bat guano, is thought to have infected pigs that were farmed in areas that had penetrated the forests. The virus subsequently spread to humans in close contact with the infected pigs, and resulted in outbreaks with approximately 250 human infections and 100 deaths from Nipah-caused viral encephalitis. Additional outbreaks of the Nipah virus from 2001 to 2009 in Bangladesh suggested indirect transmission from bats to humans through palm sap collected from palm trees where fruit bats are known to feed, and nosocomial human-to-human transmission in hospitals (Epstein et al. 2006, Ref Islam NS); and an outbreak occurred among health workers in southern India in 2018 (http://www.who.int/csr/disease/nipah/en/).

Another virus thought to be carried by bats, the Ebola Reston virus, infected pigs in the Philippines in 2009, and the virus spread from pigs to pig farmers and workers in slaughterhouses causing asymptomatic infection (Miranda and Miranda 2011).

In Latin America, Chagas disease re-emerged as an important human disease after deforestation caused triatomine populations to move from their wild natural hosts to involve humans and domestic animals in the transmission cycle, eventually transforming the disease into an urban infection that can also be transmitted by blood transfusion. Other emerging infections influenced by changing habitats of animals include Lyme borreliosis in Europe and North America, transmitted to humans who come into contact with ticks that normally feed on rodents and deer, the natural reservoir of *Borrelia burgdorferi*.

Climatic changes, both anthropogenic and from natural variations, also influence the development of epidemics. The narrow band of desert in sub-Saharan Arica, in which epidemic *Neisseria meningitides* infections traditionally occur, has enlarged as drought spread south to involve Uganda and Tanzania. Climate extremes, whether involving excessive rainfall or drought, can likewise displace animal species and bring them into closer contact with human settlements, or increase vector breeding sites. A 1998 outbreak of Japanese encephalitis in Papua New Guinea has been linked to extensive drought, which led to increased breeding sites for the *Culex* mosquito as rivers dried into stagnant pools (Erlanger et al. 2009). It is thought that mosquitoes then transmitted the Japanese encephalitis virus from infected pigs and/or wild birds to humans. The Japanese encephalitis virus is now widespread in Southern Asia from India and Thailand to Malaysia, and as far north as Korea and Japan.

Above-normal rainfall associated with the occurrence of the warm phase of the El Niño Southern Oscillation phenomenon is thought to have been the cause of extensive flooding in East Africa from December 1997 to March 1998, increasing the breeding sites for *Aedes* mosquitoes. Mosquitoes then facilitated the transfer of the Rift Valley fever (RVF) virus from infected cattle, sheep, and/or goats to humans who had been forced to live in close proximity to animals on islands of dry land surrounded by flood water. During this period, the largest RVF outbreak ever reported in East Africa occurred in Kenya, Somalia, and Tanzania. The total number of human infections in northern Kenya and southern Somalia alone was estimated at 89,000 with an estimated 478 deaths. During 2006 and 2007, RVF caused outbreaks with approximately 250 human infections and 130 deaths in Kenya, Somalia, and Sudan; and an outbreak of 172 cases and 15 deaths occurred in 2010 in South Africa (Hightower et al. 2012). Outbreaks have

also occurred in Mauritania (2012) and Niger (2016), demonstrating the board geographic range of this infection first identified in East Africa in 1931 (http://www.who.int/emergencies/diseases/rift-valley-fever/en/).

Public health consequences of civil disturbance, human displacement, and natural disasters

Human population movements on a large scale as a result of war, conflict, or natural catastrophe often result in crowded, unhygienic, and impoverished living conditions. This in turn heightens the risk of emergence and re-emergence of infectious diseases. In the aftermath of civil disturbance in Rwanda in 1994, over 48,000 cases of cholera and 23,800 deaths were reported within 1 month among Rwandans who had been displaced to refugee camps in Goma, Democratic Republic of Congo (Goma Epidemiology Group, 1995).

A collateral impact of war, conflict, or natural catastrophe such as earthquakes is the destruction or weakening of health systems with diminished capacity to detect, prevent, and respond to infectious disease outbreaks. One consequence of the 27-year civil war in Angola was the outbreak of Marburg haemorrhagic fever in 2004 that spread to more than 200 humans, 90% of whom died. Emergence of the Marburg virus was detected late and transmission was amplified in overcrowded and understaffed health facilities where lack of investment during the war had resulted in substandard infection control. Another large outbreak of Marburg virus infection was identified in late 1998 in the Democratic Republic of Congo, also a conflict-ravaged country. This emergence resulted in sporadic cases with small chains of transmission over a 2-year period in a remote area where civil war had interrupted supply lines and communication to health facilities in the region (Bausch et al. 2006).

In addition to cholera outbreaks among those displaced by war and conflict in Yemen, polio outbreaks, caused by the wild polio virus, thought to have been imported from Pakistan, occurred among children w Syria in 2013 and 2014 who had been unvaccinated, the first reported polio since 1999. Another polio outbreak occurred in in Syria in 2017, caused by a circulating vaccine derived polio virus type 2, again among unvaccinated children displaced by the civil war (http://polioeradication.org/where-we-work/syrian-arab-republic/).

Natural disasters can also create conditions that facilitate the direct or indirect spread of infectious disease. Flooding commonly disrupts clean water supplies, and in 2009 in the Philippines, where leptospirosis is endemic, more than 2000 reported human infections, with an estimated case fatality rate of 10%, were associated with flooding caused by a major typhoon (Amilasan et al. 2012). The deterioration of clean water supplies and health services in Haiti after the earthquake in 2010, discussed previously, served as fertile ground for the cholera importation that has led to the continued outbreaks occurring throughout the country (CDC 2012d; Piarroux and Faucher 2012).

Human behaviour

Occupation

Throughout history, some human occupations have been associated with infectious disease. Anthrax has been called wool-sorters' disease because of transmission of its spores from infected animals to humans who shear sheep and other wool-producing animals, or whose occupation leads to exposure to their hides. It has also been associated with butchers who come into contact with infected animals at the time of slaughter or during preparation of meat for markets. Anthrax spores infect humans either intradermally, causing cutaneous anthrax, or by ingestion or inhalation, causing enteric or pulmonary/inhalation anthrax.

Though intensive research has failed to confirm the origins of Ebola haemorrhagic fever outbreaks, infection is thought to occur as humans encounter animal sources, possibly infected bats and/or non-human primates infected by bats, somewhere in the transmission cycle. An outbreak of Ebola haemorrhagic fever in humans in 1995 was linked to a woodsman, who worked deep within the tropical rainforest making charcoal, where he was believed to have become infected with the Ebola virus that he then carried back to his home village. An outbreak of Ebola in Gabon in 1996 is thought to have had a point source in a chimpanzee that was killed and butchered by a group of young hunters; and in 1994 a Swiss researcher infected with the Ebola virus in a forest reserve in West Africa was thought to have become infected while conducting chimpanzee autopsies in search of the cause of a major die-out of the animals (Daszak 2006). More recent outbreaks of Ebola in Uganda (2008) and Guinea (2013) have been associated with exposure to

bat guano in gold mines and in the rainforest, respectively (Food and Agricultural Organisation 2011, http://www.who.int/csr/disease/ebola/one-year-report/virus-origin/en/).

In 2003, a veterinarian in the Netherlands became infected with the influenza A (H7N7) virus during an investigation of influenza outbreaks in poultry and later died from acute respiratory failure. A total of 89 humans, including the veterinarian, were confirmed to have H7N7 influenza virus infection linked to this poultry outbreak and no further deaths occurred. The majority of human infections are thought to have occurred as a result of direct contact with infected poultry; but there were three possible instances of transmission of infection from poultry workers to family members (Jong et al. 2009). The H5N1 influenza virus has caused major lethal epidemics of influenza among poultry in many Asian countries, and exposure of children or adults to infected poultry being raised in backyards or industrial poultry farms has resulted in periodic human infections. By August 2012, a total of 608 sporadic infections had been reported to the WHO with an estimated case fatality rate of approximately 59% (WHO 2012e), and sporadic human infections have continued to be reported every year since then. Several countries have reported small clusters of persons infected with H5N1 infection that do not sustain transmission, usually occurring in family members or others who have close contact with an infected person. Other avian influenza viruses including H7N9 and other subtypes, continue to cause human infection among persons who are in close contact with poultry in China (http://www.who.int/influenza/en/),

demonstrating the diversity of this family of RNA viruses.

Healthcare workers are exposed to emerging and re-emerging infections on two fronts. Laboratory and health workers are at especially high risk of emerging and re-emerging infections, and can themselves be conduits of transmission. In the 1995 outbreak of Ebola haemorrhagic fever in Kikwit (Democratic Republic of Congo), almost one-third of those infected were health workers, and in the Ebola outbreak in West Africa from 2013 to 2015 up to 40% of all deaths have been estimated to have occurred among health workers (https://www.cdc.gov/vhf/ebola/history/2014-2016-outbreak/index.html), while in the 2003 SARS outbreak in Singapore, ten health workers were thought to have been infected while treating an infected health worker colleague who is also thought to have infected her husband, three other patients, and seven visitors to the hospital (Wilder-Smith et al.

2005). These examples show the potential for health workers to become infected, and in some instances to sustain and amplify transmission in hospitals, and through their patients and family members, to the community. Laboratory workers are also at risk of infection: the last human case of smallpox was caused by a laboratory accident in the United Kingdom, and the last-known human cases of SARS occurred in laboratory accidents in Singapore and China.

Mistrust and misinformation

During 2003, unsubstantiated rumours circulated in northern Nigeria that the oral polio vaccine (OPV) was unsafe and vaccination of young children could cause infertility. Mistrust and misinformation that followed led to the government-ordered suspension of polio immunization in two northern states and substantial reductions in polio immunization coverage in those states, and a large number of others (Larson and Heymann 2010). The result was a polio outbreak across northern Nigeria that then spread to, and re-emerged in, previously polio-free areas in sub-Saharan Africa, the Middle East, and as far as Indonesia. Over 70% of all children worldwide who were paralysed by polio during the following year, 2004, were living in Nigeria—or in other parts of sub-Saharan Africa that had been re-infected by polio virus genetically linked to viruses that had a Nigerian origin (United Nations Children's Fund 2006; Modlin 2010).

Misinformation about the safety of vaccines against pertussis, measles, and hepatitis B has likewise led to decreases in vaccine uptake among children, and in some instances industrialized country outbreaks of pertussis and measles. An example is a decrease in uptake of the measles, mumps and rubella (MMR) vaccine in the United Kingdom from a coverage of 92% at the time of a falsified publication in 1998 that suggested a link between MMR and autism, to a low of 79% during 2003 (Pearce et al. 2008). Persistent low coverage has led to measles outbreaks in England and Wales during the following years, with the outbreak still ongoing during the first six months of 2018 (https://www.gov.uk/government/news/measles-outbreaks-confirmed-in-leeds-liverpool-and-birmingham). To ensure vaccine uptake is maximized, public confidence in the benefit, safety, and effectiveness of vaccines must be upheld, which in turn, relies on a comprehensive evidence base underpinning vaccination strategies (Larson et al. 2011).

Deliberate use to cause terror and harm

Infectious organisms have been deliberately used to cause terror and harm throughout history, with accounts of the use of highly infectious and lethal organisms such as smallpox to kill or weaken the opposition during wars in the eighteenth century (Riedel 2005), and this has continued into the twenty-first century. It was graphically illustrated in 1979 in an outbreak of inhalation anthrax in Sverdlovsk, 1,400 km east of Moscow, in the then Soviet Union. Attributed at first by government officials to the consumption of contaminated meat, it was later shown to have been caused by the unintentional release of anthrax spores from a Soviet military microbiology facility. It is estimated that up to 358 humans were infected and that between 45 and 199 died (Meselson et al. 1994) (see Chapter 8.18).

In the United States in 1984, salad bars in restaurants in Oregon were deliberately contaminated with *Salmonella enterica* by a group of followers of a spiritual leader in order to cause illness at the same time as county elections. The result was food poisoning of over 750 persons (https://www.cdc.gov/phlp/docs/forensic_epidemiology/Additional%20Materials/Articles/Torok%20et%20al.pdf).

In late September 2001, the deliberate dissemination of anthrax spores in four known letters sent through the United States Postal Service caused massive disruption of postal services in the United States, and in many other countries around the world. The anthrax letters—dated 11 September 2001, and postmarked 7 days later—caused huge public alarm and prompted a massive public health response. A total of 22 persons are thought to have been infected by anthrax spores sent through the postal system; 11 developed cutaneous anthrax and the remaining 11 developed inhalation anthrax, of whom five died. Twenty of the 22 patients were exposed to work sites that were found to be contaminated with anthrax spores. Nine of them had worked in mail processing facilities through which the anthrax letters had passed (Bouzianas 2009).

Other bacteria, viruses, mycotic organisms, and biological toxins are also considered to have the potential for deliberate use to cause harm to humans. Great concern has been expressed by many countries about the potential health consequences that could be caused by the deliberate introduction of infectious organisms such as the variola virus into a human population where smallpox vaccination is no longer practised, or the plague bacillus that could potentially cause an outbreak of pneumonic plague. Stockpiles for

vaccines against infections such as smallpox and anthrax are currently being maintained, at great expense, as part of emergency contingency planning in some countries.

In addition to the deliberate use of existing infectious agents, new laboratory research can result in the creation of novel agents that have the potential for human spread. One example is the laboratory research on H5N1 influenza viruses to determine the mutations that will facilitate human spread. While the research was performed to understand the determinants of transmission in an academic setting, if these more transmissible viruses were to fall into the wrong hands, it could result in a pandemic of unprecedented proportions. This has resulted in much debate among the scientific and regulatory communities on the merits of research with potential dual uses.

Antimicrobial drug resistance—susceptibility of infectious organisms to anti-infective drugs

Bacteria, viruses, and parasites all have the potential to develop resistance to anti-infective drugs by spontaneous mutation and natural selection, or through the exchange of genetic material between strains and species. These resistant organisms can then transmit from human to human, replacing more susceptible organisms with resistant strains. Soon after development of the first antibiotics, warning signs of microbial resilience began to appear. By the end of the 1940s, resistance of hospital strains of *Staphylococcus aureus* to penicillin emerged in the United Kingdom with levels as high as 14%, and by the end of the 1990s, levels had risen to of 95% or greater. Resistance has also developed to semi-synthetic penicillin derivatives as *S. aureus* has acquired genes encoding resistance to methicillin and other narrow-spectrum beta-lactamase-resistant antibiotics. Methicillin-resistant *S. aureus* (MRSA), first identified in the United Kingdom in 1961 within a year of the introduction of methicillin, is now widespread in hospital-acquired infections in the United Kingdom and throughout the world (Johnson 2011).

The bacterial and viral infections that contribute most to human disease are also those in which antimicrobial resistance is rapidly emerging: diarrhoeal diseases such as dysentery; respiratory tract infections, including pneumococcal pneumonia and TB; sexually transmitted infections such as gonorrhoea and HIV; and infectious organisms that have now accumulated

resistance genes to virtually all currently available anti-infective drugs such as extremely drug-resistant TB (XDR-TB).

Since 2008, Gram-negative Enterobacteriaceae have been shown to be resistant to a broad range of beta-lactam antibiotics including carbapenems that are widely used for their treatment. Carbapenem-resistant bacterial infections have been identified in many countries across the world with emergence thought to have occurred somewhere in Asia (Kumarasamy et al. 2010).

The causes of antimicrobial drug resistance (AMR) are complex, and involve both the human and animal health sectors, and others such as plant and fish agriculture. AMR is a major global public health issue because it threatens the effectiveness of antimicrobials in combating infection, which has been one of the most important advances in clinical medicine. A lack of effective antimicrobials could lead the world back to the pre-antimicrobial era where simple infections often led to severe outcomes, and where effective measures to control the spread of disease, especially those for which there are no vaccines, would be lost. AMR also increases healthcare and other indirect costs as more expensive treatments are required, and illnesses may be prolonged.

Use of anti-infective drugs in human health

Behaviours such as over- or underprescribing of antibiotics by health workers, excessive demand for antibiotics by the general population, or the use of substandard drugs with inferior anti-infective drug content, have had a remarkable impact on the selection and survival of resistant microbes, rapidly increasing levels of microbial resistance.

In Thailand, among 307 hospitalized patients in the late 1980s, 36% of persons who were treated with anti-infective drugs did not have an infectious disease (Aswapokee et al. 1990). Overprescribing of anti-infective drugs occurs in most other countries as well. In Canada, it has been estimated of the more than 26 million people treated with anti-infective drugs, up to 50% were treated inappropriately (Kondro 1997). This has also led to community-based resistance where otherwise healthy individuals may be carrying resistant microbes which can result in transmission or infection. Findings from community surveys of *Escherichia coli* in the stool samples of healthy children in China, Venezuela, and the United States suggest that although

multiresistant strains were present in each country, they were more widespread in Venezuela and China, countries where less control is maintained over antibiotic prescribing and sales (Lester et al. 1990).

The availability of multiple drugs to treat TB since the 1940s has reduced death rates from TB substantially. Resistant organisms have emerged, however, facilitated by inappropriate prescribing and poor adherence to treatment regimes. Multidrug-resistant TB (MDR-TB) has now become widespread, especially in prison settings in Eastern Europe, and resistant organisms are now being transmitted to healthy populations throughout the world. During 2016 WHO estimates that there were a total of 490,000, over 50% of them in China, India, and the Russian Federation (http://www.who.int/tb/publications/global_report/Exec_Summary_13Nov2017.pdf?ua=1).

In 2006, an outbreak of extremely drug-resistant TB (XDR-TB), against which almost all TB drugs are ineffective, occurred in South Africa (WHO 2012f), and all XDR-TB cases that were tested for HIV were found to be HIV-positive as well. WHO has estimated that up to 9% of the 490,000 MDR MDR-TB cases that occurred in 2016 were XDR-TB, and with HIV infection endemic, the world provides continued fertile ground for the transmission of all forms of TB, including MDR- and XDR-TB (http://apps.who.int/iris/bitstream/handle/10665/259366/9789241565516-eng.pdf?sequence=1).

Carbapenem-resistant Enterobacteriaceae (CRE) are carbapenemase-producing Enterobacteriaceae (CPE) that have become resistant to the carbapenem class of antibiotics. Resistance has spread from country to country in persons infected during medical procedures or in the community, and treatment of resistant infections requires prolonged hospitalization and attempts to cure with various antibiotic combinations (Tängdén 2015)

Malaria is another major infection with high mortality that has been reduced substantially with the introduction of antimalarial drugs. By 1976, however, chloroquine-resistant *Plasmodium falciparum* malaria had developed and was highly prevalent in South East Asia. Ten years later, chloroquine-resistant malaria was found worldwide, as was high-level resistance to two second-line drugs, sulphadoxine/pyrimethamine and mefloquine. Subsequently, use of artemisinin combination therapy is now recommended by the WHO to ensure effective treatment, but since 2009 there are reports of artemisinin resistance in South East Asia, threatening to

derail efforts to control malaria in Asia, and worldwide (Nayyar et al. 2012). Multidrug-resistant malaria has now emerged and spread in Cambodia, Laos, Thailand, and Vietnam (Takala-Harrison and Jacob 2015). Surveillance and monitoring for resistance continues both in Asia where efforts are being made to contain its spread to other continents including Africa.

Prevention and control of resistant infections such as MRSA and *Clostridium difficile* in healthcare facilities is an important key to reducing the spread of AMR, and control of such hospital-acquired infections must be a priority.

Use of anti-infective drugs in animal husbandry and agriculture

Large quantities of anti-infective drugs are indiscriminately added to animal feed in some countries to promote animal growth, resulting in the selection of resistant bacterial strains in animals. Antibiotics are also used in agriculture for spraying of fruit trees, rice paddies, and flowers to avoid bacterial blights.

Some of the organisms that infect animals freely circulate between animals and humans, providing opportunities for zoonotic infection with resistant organisms, and for swapping or exchanging resistant genes, thus increasing the speed with which anti-infective resistance evolves in both agriculture and human populations. Alternative approaches can reduce such antibiotic use as in Norway, where the introduction of effective vaccines in farmed salmon and trout, and improved health management, reduced the annual anti-infective use in farmed fish by 98% from 1987 to 2004 (WHO 2012g). Similar outcomes have occurred in North America as well, where vaccines are also used intensively in salmon farming in the Pacific Northwest. During the 1990s, Norway and other countries throughout Europe banned the non-discriminate use of antibiotics for growth promotion in animal feed; and in the United States in 2017 a voluntary ban on use for growth promotion of those antibiotics used in humans was put into place by the Food and Drug Administration.

The weakness of surveillance and monitoring systems for antimicrobial resistance, partially the result of the lack of easy-to-use surveillance tools, makes it difficult to understand the extent and evolution of the AMR

problem globally. The lack of financial investment for research into new antimicrobials and vaccines is also a major global problem. Many of the anti-infective drugs currently in use were developed decades ago and the lack of drugs that target non-similar areas on microbes limits treatment modalities available to combat infections. Mechanisms are beginning to be put in place in an attempt to increase the development of new antibiotics and vaccines with additional funding or research and development, and de-linkage of marketing from research and development is seen as a potential means of prolonging new antibiotic life by decreasing the marketing pressure that inevitably occurs as companies attempt to rapidly recuperate their investment.

The WHO proposed a five-pronged containment strategy for AMR (Fig. 8.17.3) through activities set up around a World Health Day in 2011(WHO 2012h), and a report on AMR in 2012. In 2015 a Global Action Plan on Antimicrobial Resistance was launched. Underpinning this strategy is the need for national commitments for sufficient and sustainable resources to be made available to combat AMR, and for countries to developing national action plans (http://www.who.int/antimicrobial-resistance/global-action-plan/en/).

In 2016, at the United National General Assembly, Member States recognized AMR as a global problem and adopted a declaration of the high-level meeting on AMR. The United Nations also established an Interagency Coordination Group on AMR, to provide guidance on approaches to ensure sustained effective global action on AMR.

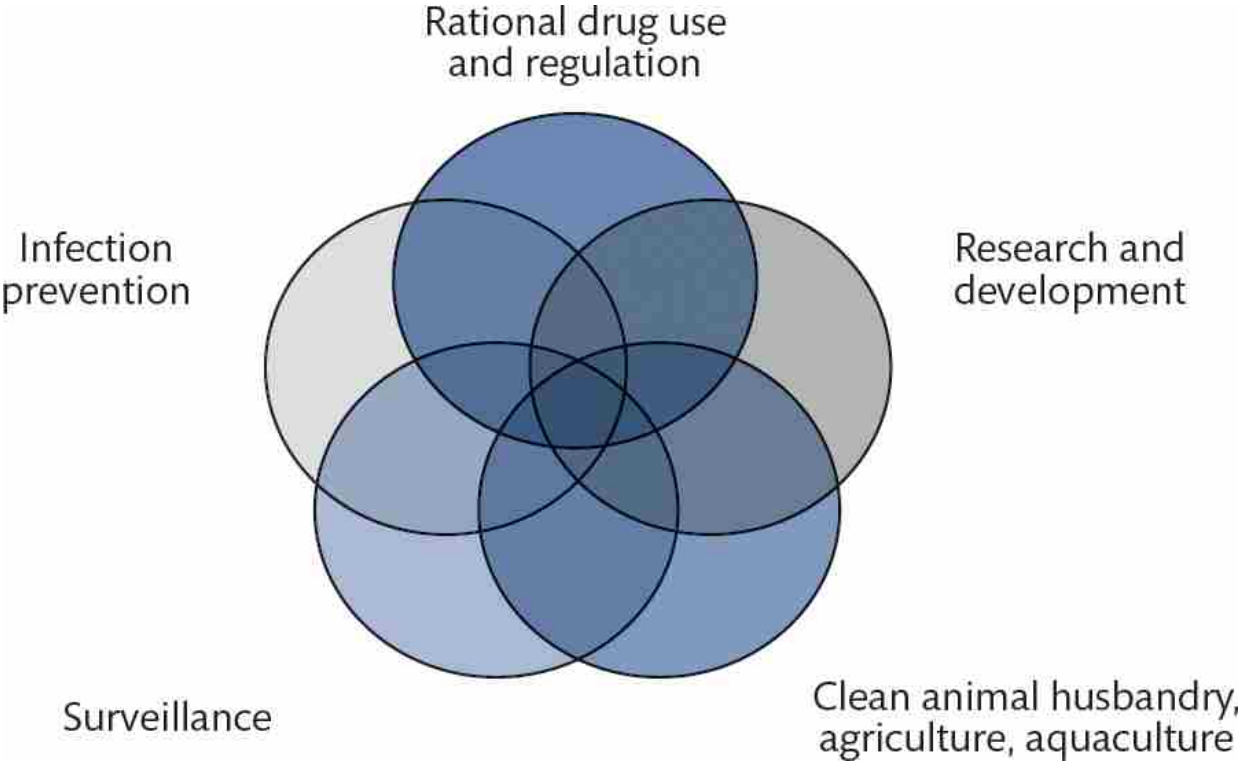


Fig. 8.17.3 The five key areas for containment of antimicrobial resistance.
Source: data from World Health Organization (WHO), Model based on the WHO Global Strategy for Containment of Antimicrobial Resistance, Copyright © WHO 2001, available from http://www.who.int/csr/resources/publications/drugresist/en/EGlobal_Strat.pdf.

Public health security: globalization and emerging/re-emerging infectious disease organisms

Our world of increased human mobility and interdependence facilitates the transfer of infectious organisms from country to country, and across continents. Infectious organisms efficiently travel in humans, insects, food, and animals, and can spread around the globe and emerge in new geographic areas with ease and speed. Some are transported by migratory birds. Others, such as disease-carrying mosquitoes, travel in the passenger cabin or luggage hold of jets, to cause tropical infections in temperate countries when they feed on airport workers or those who live nearby. They thus threaten public health security—our collective vulnerability to acute infectious disease outbreaks across international borders (Heymann 2003).

In 2000, among 312 athletes participating in an international triathlon held in Malaysia, 33 became infected with leptospirosis and returned to their

home countries during the incubation period. While leptospirosis lacks human transmissibility and therefore did not set up local foci or transmission, the outbreak of SARS in 2003 clearly did, and demonstrates the full potential of emerging infectious organisms to spread internationally. From a medical doctor infected in Guangdong Province of China, who unknowingly carried the new infectious organism to a Hong Kong hotel, SARS spread in individual chains of transmission from infected hotel guests to 8,422 reported cases in North and South America, the Middle East, Europe, and Asia with a case fatality rate of approximately 11% (WHO 2003).

During the years 1969–2003, 18 instances of airport malaria were reported to the WHO—malaria infections in workers at airports or in persons who live nearby who had not travelled to malaria-endemic countries (WHO 2007b). Their infection originated from malaria-infected *Anopheles* mosquitoes that had travelled from countries with endemic malaria and took a blood meal from airport workers or other persons upon landing. In 2003, a seven-person outbreak of malaria among persons in Florida who had not previously travelled to malarious areas demonstrated the potential for imported malaria to set up local foci of transmission if the *Anopheles* mosquito is present (Morbidity and Mortality Weekly Report 2004). In 2013, 4 cases of malaria were reported in persons living within a 2 km radius of the Tunis International Airport who had no history of international travel, the first malaria reported in Tunisia since 1980 (Siala and Gamara 2015), and ‘airport malaria’ continues to regularly reported from industrialized countries, demonstrating that malaria remains a risk in countries where there is no endemic malaria transmission.

Livestock, animal products, and food can also carry infectious organisms that emerge or re-emerge in non-endemic countries. VCJD, identified in persons in 11 countries, most of whom had not previously travelled to the United Kingdom (Diack et al. 2012), clearly demonstrates the ability of infections to travel in animals or animal products. RVF emerged in humans in Yemen and Saudi Arabia in 2000, 2 years after a major outbreak of RVF in East Africa. Infection has since become endemic in livestock in the Arabian Peninsula, and is thought to have been imported from East Africa in livestock traded across the Red Sea (Balkhy and Memish 2003). In 2000, imported raspberries contaminated with *Cyclospora* caused an outbreak in the United States. It is hypothesized that the raspberries imported from

Guatemala were contaminated when surface water was used to spray them with fungicide before harvest (Ho et al. 2002). In 2011, the German outbreak caused by *Escherichia coli* O104 linked to sprouts grown from imported beans resulted in over 3,000 infections and 31 deaths from haemolytic uramic syndrome (WHO 2006; European Food Safety Authority 2012).

Concerns about the international spread of emerging and re-emerging infections and the need for strong public health security are not new. By the fourteenth century, governments had recognized the capacity for international disease transmission and legislated preventive measures, as reflected in the establishment of quarantine in the city-state of Venice. Arriving ships were not permitted to dock for 40 days, and people at land borders were held in isolation for 40 days in the hope of keeping plague from entering.

Many European leaders of the mid-nineteenth century, worried by the cholera pandemic of the time, threats of plague, and the weakness of quarantine measures, began to recognize that controlling the spread of infectious diseases from one nation to another required cooperation between those nations. International conventions were organized and draft covenants signed, almost all of which related to some type of quarantine regulations.

From 1851 to 1897, ten international sanitary conferences were held among a group of 12 European countries, focusing exclusively on the containment of epidemics in their territories. The inaugural 1851 conference in Paris lasted 6 months and was followed in 1892 by the first International Sanitary Convention that dealt with cholera. Five years later, at the 10th International Sanitary Conference, a similar convention that focused on plague was signed.

New policies then emerged in the late nineteenth century, such as the obligatory telegraphic notification of first cases of cholera and plague, a model that a small group of South American nations followed when they signed the first set of international public health agreements in the Americas during the 1880s. In addition to cholera and plague, often carried by immigrants arriving from Europe, the agreements in the Americas covered yellow fever that was endemic in much of the American region at that time, and that from time to time caused major urban epidemics.

During the following decade, 12 countries attended the First International Sanitary Convention of the American Republics in Washington, DC, leading to the creation of the Pan American Sanitary Bureau (now called the Pan

American Health Organization) in 1902. Its counterpart in Europe was the Office International d'Hygiène Publique (OIHP), established in 1907, and based in Paris.

In 1951, 3 years after its founding, the WHO adopted a revised version of the International Sanitary Regulations (1892) that remained focused on the control of cholera, plague, and yellow fever and was rooted firmly in the preceding agreements of the nineteenth and twentieth centuries.

The International Health Regulations

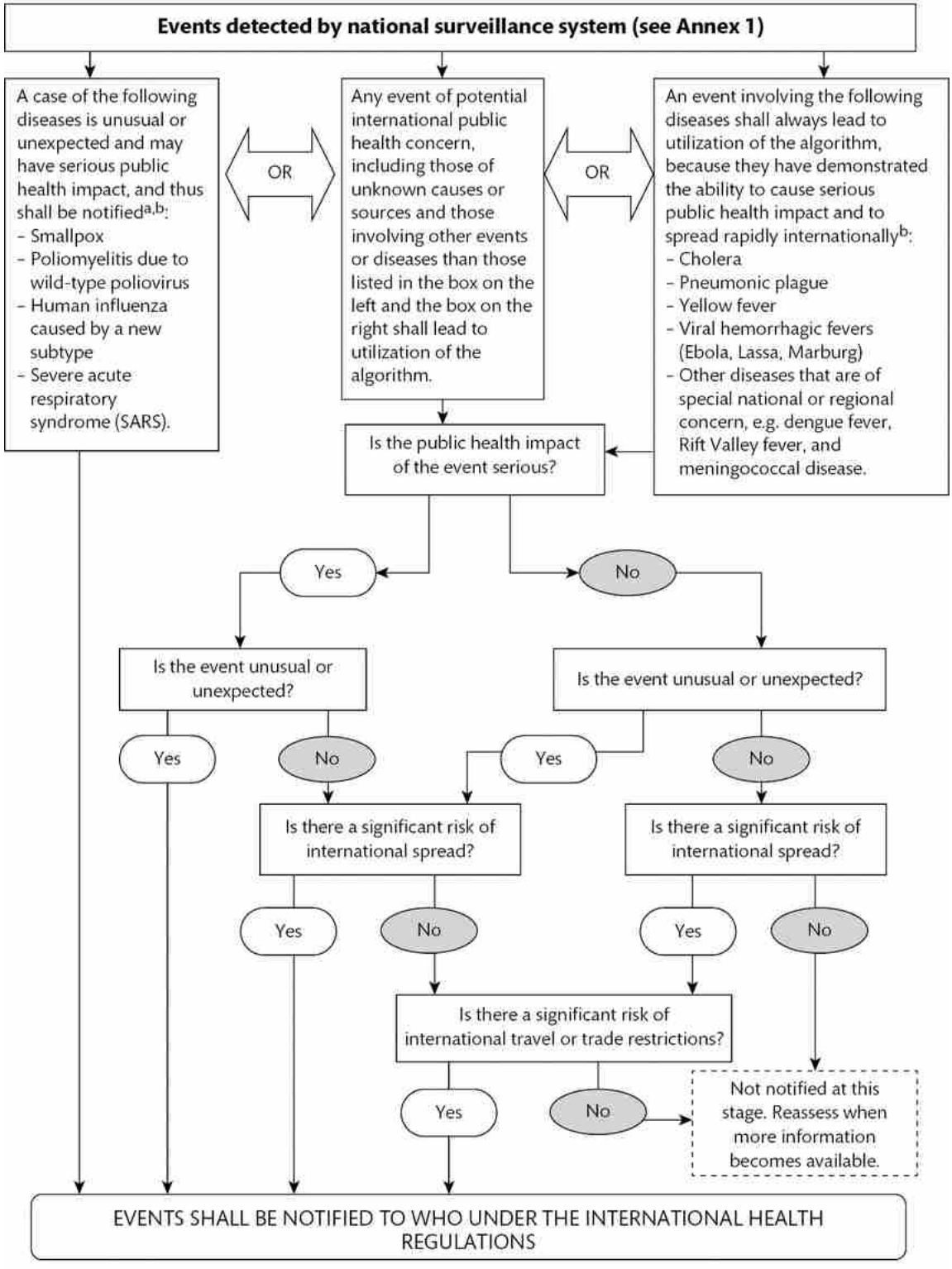
In 1969, the Member States of the WHO agreed to a new set of international sanitary regulations—the International Health Regulations (IHR)—aimed at better ensuring public health security with minimal interruption in travel and trade. In addition to requiring reporting of four infectious diseases—cholera, plague, yellow fever, and smallpox—the IHR were aimed at stopping the spread of disease by pre-established control measures at international borders. They included reporting requirements for these four diseases, and described appropriate responses such as yellow fever and smallpox vaccination for passengers arriving from countries where yellow fever or smallpox outbreaks had been reported. The IHR provided a legal framework for global surveillance and control of spread of infectious diseases.

By 1996, it had become clear, however, that the IHR were not able to ensure public health security as had been envisioned. Over the previous 30 years, border posts proved inadequate to stopping the international spread of infectious diseases which often crossed borders while still in the incubation period in humans, or silently in non-human hosts—insects, animals, food, and agriculture goods. At the same time, countries reported the occurrence of cholera, plague, and yellow fever late or not at all because of fear of stigmatization and economic repercussions (smallpox had been removed from the list when it was certified eradicated in 1980). The IHR also did not meet the challenges caused by other emerging and re-emerging infectious disease and their rapid transit round the globe.

From 1996 until 2005, WHO Member States therefore undertook a process to examine and revise the IHR. The result—the IHR (2005) (WHO 2008b)—provided a more up-to-date legal framework requiring reporting of any PHEIC, and the use of real-time evidence to recommend measures to stop their international spread. A PHEIC is defined as an extraordinary event

that could spread internationally or might require a coordinated international response (WHO 2012i). Under the IHR (2005) an event is evaluated for its potential to become a PHEIC by the country in which it is occurring even though reporting might legitimately come from elsewhere, using a decision tree instrument developed for this purpose (Fig. 8.17.4). If the criteria for a PHEIC are met, an official notification must be provided to the WHO. Notification is also required for even a single occurrence of a disease that would always threaten global public health security—smallpox, poliomyelitis caused by a wild-type poliovirus, human influenza caused by a new virus subtype, and SARS. In addition, there is a second list that includes diseases of documented—but not inevitable—international impact. An event involving a disease on this second list, which includes cholera, pneumonic plague, yellow fever, Ebola, and the other haemorrhagic fevers, still requires the use of the decision tree instrument to determine if it is a PHEIC. Thus, two safeguards create a baseline of public health security by requiring countries to respond, in designated ways, to well-known threats.

ANNEX 2
DECISION INSTRUMENT FOR THE ASSESSMENT AND NOTIFICATION OF EVENTS THAT MAY CONSTITUTE A PUBLIC HEALTH EMERGENCY OF INTERNATIONAL CONCERN



[a] As per WHO case definitions.
[b] The disease list shall be used only for the purpose of these Regulations.

Fig. 8.17.4 Decision tree for assessment of a potential public health event of international concern. Reproduced with permission from World Health Organization (WHO), International Health Regulations, Second edition, Copyright © WHO 2005, available from: http://www.who.int/ihr/publications/9789241596664/en/.

In contrast to previous regulations, the IHR (2005) also introduced a set of core capacity requirements for surveillance and response. All countries were to meet these requirements during the first years of implementation of the IHR (2005) in order to rapidly detect, assess, notify, report, and contain the events covered by the regulations so that their potential for international spread and negative economic impact would be minimized. The IHR (2005) likewise require collective action by all WHO Member States in the event that an emerging or re-emerging infectious disease threatens to spread internationally, and the free sharing of information pertaining to this threat. They thus provide a safety net against the international spread of emerging or re-emerging infections, requiring collaboration between all states to ensure the timely availability of surveillance information and technical resources that better guarantee international public health security.

Other international frameworks have also been developed to contain and curtail the international spread of emerging infections. Among them is the WHO Global Action Plan on Antimicrobial Resistance (WHO 2015) described earlier in this chapter (see ‘Use of anti-infective drugs in animal husbandry and agriculture’). Though not legally binding, this framework calls on countries to work across the human health, animal health, agricultural, and trade sectors to ensure more rational use of anti-infective drugs in order to limit the factors that accelerate the selection and proliferation of anti-infective drug-resistant microbes.

Recommendations for the future

National and international systems for addressing the threat of infectious diseases

To enable the world to deal effectively with the threat of emerging and re-emerging diseases, there is the need for countries to follow the international frameworks laid out by such agreements as the IHR (2005) and the WHO Action Plan on Amtimicrobial Resistance (2015); and also by the United Nations Framework Convention on Climate Change, still under negotiation.

These international agreements call for action at the global, national, and subnational levels that will decrease the risks factors that influence emergence and re-emergence of infectious diseases. At the same time, if followed they will result in better detection and response where and when emerging and re-emerging public health events occur, be they at the animal/human interface or in other situations where a series of risk factors align in such a way as to permit their occurrence. Effectively preventing, or rapidly detecting, assessing, and responding to infectious disease outbreaks will save lives and protect economies.

The link between health and overall human development is also important, as many factors such as poverty, lack of education, and basic community infrastructure play a role in emergence and re-emergence. As concluded by the WHO Commission on Macroeconomics and Health (2001b), health considerations must be placed squarely at the centre of economic development, and all sectors of society must work together to achieve this goal, as it is not the sole purview of the health sector alone.

A better understanding of our ecosystem and causative factors will enable preventive actions that reduce the likelihood of disease emergence, especially infections at the animal/human interface. One example is safe and healthy market programmes to encourage proper handling and distribution of livestock and maintenance of strict sanitation in markets where animals or animal products are sold that could be a source of infection. This fits with the one health agenda that calls for joint collaborative activities between the animal and human health sectors in such areas as surveillance and response, and risk assessment.

For emerging and re-emerging infections that are able to transmit from person to person, the window of opportunity for effective intervention often closes quickly. The most important defence against their international spread is robust surveillance and preparedness, response protocols, and mitigation systems. Infectious disease outbreaks must be detected rapidly and the response must be timely and effective. This includes highly sensitive national surveillance systems, public health laboratories that can rapidly detect such outbreaks, and mechanisms for timely outbreak containment and mitigation. It also necessitates good communications systems, and risk communication and health education programmes, to enable coordinated preparedness and response by all stakeholders. These are the core capacities required by the IHR (2005), and it is clear that the same systems that detect

and contain naturally occurring outbreaks also permit detection and initial response to deliberately caused outbreaks of infectious disease.

The requirement of the IHR for core capacity development is key to effective prevention and control of emerging and re-emerging infections—it will equip countries with the ability to detect and respond when and where outbreaks occur, preventing their spread across borders. The collaborative action required by IHR (2005) when emerging or re-emerging infections threaten to spread internationally provides a safety net of global surveillance and response that can be called into action if and when outbreaks begin to spread internationally. The IHR (2005) explicitly acknowledge that non-state sources of information about outbreaks often precede official notification, especially when countries may be reluctant to reveal an event in their territories. Collaboration among countries for public health security in the face of emerging and re-emerging infections is vital. If an outbreak of emerging or re-emerging infectious disease occurs in a country that requires international assistance to help in containment activities, as agreed by the WHO and the affected country after confidential and pro-active consultation, other countries are obligated to provide the technical expertise required to the country through WHO or bilaterally. To this end, global databases of professionals with expertise in specific diseases or epidemiological techniques are maintained as part of the requirement of the IHR (2005), together with non-governmental organizations present in countries and in a position to reach remote areas. Such mechanisms, which are further supported by the WHO network of Collaborating Centres (national laboratories and institutes throughout the world serving as international reference centres), help optimize the use of global expertise and resources. The IHR (2005) were instrumental in the coordination of the international response to the H1N1 influenza pandemic in 2009. Upon the recommendation of the IHR Emergency Committee set up by the Director General of the WHO, the pandemic was declared as a PHEIC within weeks of its onset, and triggered a series of obligations for collaboration among countries under the IHR (2005). Since then PHEICs have been declared by the Director General for polio, Ebola, and Zika virus infections.

At the same time operational protocols that had been developed as part of pandemic planning and preparedness activities in countries around the world were implemented, together with coordinated actions under the IHR (2005).

Another platform that supports disease response is the Global Outbreak Alert and Response Network (GOARN). GOARN was formed in 2000 by WHO together with partner institutions, organizations, and networks with interests in global epidemic surveillance and response. GOARN contributes to outbreaks response, to providing technical assistance to affected states rapidly, and to long-term preparedness and capacity building. GOARN brings together resources for rapid identification, confirmation, and response to outbreaks of international importance. During outbreaks, WHO coordinates international outbreak response using resources from GOARN partners that volunteer their services, to ensure that the right expertise and provided where and when they are needed most.

The 2020 SARS-CoV2 pandemic

In 2020 the world is experiencing the worst rapidly-spreading pandemic since the 1917–18 H1N1 Influenza pandemic. As of this writing, April 2020, almost 200 million cases and 120,000 deaths have been reported from countries on every continent, and the pandemic continues to spread.

On 21 December 2019, it is understood that an ophthalmologist and several other clinicians in Wuhan, Hubei Province, China began reporting to their colleagues cases of an unusual respiratory disease with pneumonia. By 27 December local public health officials reported the occurrence of this unusual disease to the Jianghan Centers for Disease Control. It was thought that the Huanan fish and wild animal market was the source of the first cases of the new disease, and it was closed on 31 December. On this same date an epidemiologic alert was sent to the China Centers for Disease Control in Beijing which then forwarded the alert to the WHO and to other countries. On 5 January the WHO published in its Disease Outbreak News a description of the clusters of pneumonia, an assessment of its seriousness, and provided advice based on the reports coming from Beijing. By 12 January the causative agent had been identified with a genetic sequence closely related to two other mutated coronaviruses which had caused epidemics of SARS and Middle Eastern Respiratory Syndrome, and the genetic sequence was shared publicly. The new agent was designated as Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) and the disease labelled as COVID-19. A genetic similarity to bat coronaviruses was noted suggesting that a bat virus had been transmitted to an animal

intermediary from which it emerged to cause human infection and disease. As of this writing an animal intermediary host has not been confirmed, and based on further analysis of the genetic sequence, it is hypothesized that the virus emerged several weeks earlier than originally thought.

The magnitude of the epidemic in China, as confirmed by polymerase chain reaction (PCR) testing, and the infectious nature of the virus prompted a strict lockdown of Wuhan City 23 January and subsequently further lockdown of an additional 15 other cities from 24 to 27 January. The tight lockdown of the major cities in China ultimately achieved the goal of slowing the epidemic in China and decreasing its international spread, and by mid-April Wuhan and other Chinese cities had pulled back on the intensity of the lockdown. By the time of this writing, China was reporting more imported cases than were occurring domestically (Wu et al. 2020a, 2020b, 2020c).

On 30 January 2020, the WHO declared the epidemic a PHEIC, and in March declared SARS-CoV2 a global pandemic with cases reported from other Asian countries, and countries around the world including reports of severe epidemics by Italy, Spain, and other parts of Europe where outbreaks were overwhelming demand on hospital systems. The United States, as of the time of writing, has reported more cases and more deaths than any other country in the world including China. Singapore, Germany, Taiwan, South Korea, and Hong Kong, which have kept the reproductive number less than 1 by implementing strong outbreak containment measures including patient identification and isolation, and contact tracing with rigorous daily follow-up.

Other countries have implemented these measures less consistently, and all countries are attempting to limit population exposure by issuing personal self-protective policies that include frequent hand washing and physical distancing, and social distancing measures such as closing schools and non-essential businesses, and the cancelling large gatherings of people for concerts, sports, and other events. The wearing of a face mask or other means of protecting the spread of droplets from potentially infected persons has become policy for the general public in some countries, and all are recommending the use of face mask for persons who have cough, fever, or other symptoms, or who are unable to physically distance such as workers in homes for the elderly. At the same time hospitals in all countries are

increasing their infection prevention and control measures to prevent infection of health facility workers and nosocomial transmission.

Despite early examples of outbreak control in Asian countries, many other countries did not heed the warning, and the different experiences in dealing with the pandemic underscore the importance of stronger preparedness of basic public health strategies such as alert systems, containment strategies including patient isolation and rigorous contact tracing, surge capacity in health systems, and application of quarantine measures when required based on continuing risk assessment. It is also generally acknowledged that in the future there must be stronger surveillance and risk assessment at the human-animal interface.

In the majority of persons infected with SARS-CoV2 disease manifestations are mainly those of mild respiratory illness, but in the elderly, and in those with underlying medical conditions such as diabetes, hypertension, cardiac disease, and chronic pulmonary disease, there is a risk of severe disease and death. Severe disease and death have also been observed in other age groups and in persons with no known underling medical conditions, but at the time of writing they have not been clearly quantified, nor is the percentage of infections that remain asymptomatic fully understood. The case fatality ratio has ranged from under 1% to as high as 13–14% depending on the case definition used, which has usually been for severe disease and therefore does not include persons with less severe illness in the calculation; and the capacity of hospitals to support those who require intensive care is also a factor that contributes to mortality. Common symptoms of illness include low-grade fever, dry cough which can become continuous, shortness of breath which can progress to respiratory failure, and extreme fatigue. The incubation period from exposure to development of symptoms is thought to range from 1 to 14 days, and transmission has been observed from individuals who subsequently developed symptoms one or two days after they had been exposed to persons who later developed infection. The role of earlier transmission from asymptomatic persons is currently not completely understood, nor is the immune response. Antibody testing facilitates determination of the prevalence of disease and may be helpful in understanding the proportion of infections that are asymptomatic, but at the time of writing it is not known whether the antibodies detected are protective, and if so for how long; and the sensitivity and specificity of these tests is not yet understood, with validation continuing. It is known, however,

that immunity to some of the common coronaviruses in humans is not long-lived.

The impact of the on the global economy of forced physical distancing through the lockdown of certain economic sectors has been severe. Stock markets have fallen, jobless rates have increased as restaurants and other small businesses have closed, and the socioeconomic gap between the wealthy and poorer population groups is feared to be widening in all countries. It is also feared that an excess of deaths will occur in persons not infected with SARS-CoV2 because of restricted access to routine medical care. In some countries forced physical distancing measures have resulted in tension between public health leaders and business leaders about how long to continue the strict lockdown policies, and recovery, once the pandemic has been controlled, is anticipated to take several years or longer. Though the destiny of this virus is not yet fully understood and cannot be predicted with accuracy, some epidemiological modellers suggest that it could continue to cause waves of transmission for years to come, and others are predicting a decrease in transmission at certain times of the year.

The future

The vision embraced by the IHR (2005) is of a world alert and ready to respond collectively to the threat of emerging and re-emerging infections that represent an acute threat to public health security. This requires an unbroken line of defence within countries and globally, using the most up-to-date technologies of surveillance and response.

The IHR (2005) require adherence to new norms and standards for reporting and responding to emerging and re-emerging infections despite the economic consequences that may result. Their full achievement will provide the highest level of public health security possible. National core capacities as described in the IHR (2005) must be put in place within the national public health system that can detect, investigate, communicate, and contain events that threaten public health security as soon as they appear. An interconnected system must be operating at the national and international levels, engaged in specific threat and risk assessment and management activities that minimize collective vulnerability to public health events.

The newly established and voluntary WHO joint external evaluation (JEE) process provides countries with the platform to help them assess their core

capacities in public health, followed by an action plan and budget that permits them to mobilize resources (https://www.jeealliance.org/). At the same time, development agencies are gearing up to provide resources for such capacity strengthening through their global health security efforts.

Countries must sustain and/or increase their resources to strengthen core capacity in public health. In today's mobile, interdependent, and interconnected world, threats arising from emerging and epidemic-prone diseases affect all countries. Such threats reinforce the need for shared responsibility and collective action in the face of universal vulnerability, in sectors that go well beyond health.

REFERENCES

Amilasan, A.-S., et al. (2012). Outbreak of leptospirosis after flood, the Philippines, 2009. *Emerging Infectious Diseases*, 18(1), 91–4.

Asian Development Bank (2003). SARS: economic impacts and implications. [Online] Available at: http://www.adb.org/publications/sars-economic-impacts-and-implications

Aswapokee, N., Vaithayapichet, S., and Heller, R.F. (1990). Pattern of antibiotic use in medical wards of a university hospital, Bangkok, Thailand. *Reviews of Infectious Diseases*, 12(1), 136–41.

Balkhy, H.H. and Memish, Z.A. (2003). Rift Valley fever: an uninvited zoonosis in the Arabian Peninsula. *International Journal of Antimicrobial Agents*, 21(2), 153–7.

Bausch, D.G., et al. (2006). Marburg hemorrhagic fever associated with multiple genetic lineages of virus. *The New England Journal of Medicine*, 355(9), 909–19.

Bouzianas, D.G. (2009). Medical countermeasures to protect humans from anthrax bioterrorism. *Trends in Microbiology*, 17(11), 522–8.

Buchholz, U., et al (2011). German outbreak of *Escherichia coli* 104:H4 associated with sprouts. *The New England Journal of Medicine*, **365**, 1763–70.

Burns, A., Mensbrugghe van der, D., and Timmer, H. (2008). Evaluating the economic consequences of avian influenza. [Online] Available at: http://siteresources.worldbank.org/EXTAVIANFLU/Resources/EvaluatingAHIeconomics_2008.pdf

Centers for Disease Control and Prevention (CDC) (2012a). *Legionella Topic Home*. [Online] Available at: http://www.cdc.gov/legionella/index.htm

Centers for Disease Control and Prevention (CDC) (2012b). *Questions and Answers about Ebola Hemorrhagic Fever*. [Online] Available at: http://www.cdc.gov/ncidod/dvrd/spb/mnpages/dispages/ebola/qa.htm

Centers for Disease Control and Prevention (CDC) (2012c). *Tracking a Mystery Disease: The Detailed Story of Hantavirus Pulmonary Syndrome (HPS)*. [Online] Available at: http://www.cdc.gov/hantavirus/hps/history.html

Centers for Disease Control and Prevention (CDC) (2012d). *Haiti Cholera Outbreak*. [Online] Available at: http://www.cdc.gov/haiticholera/

Cho, S.Y. et al. (2016). MERS-CoV outbreak following a single patient exposure in an emergency room in South Korea: an epidemiological outbreak study. *The Lancet*, 388(10048), 994–1001.

Curran, J.W., et al. (1984). Acquired immunodeficiency syndrome (AIDS) associated with transfusions. *The New England Journal of Medicine*, 310(2), 69–75.

Daszak, P. (2006). Risky behavior in the Ebola zone. *Animal Conservation*, 9(4), 366–7.
Delgado, G., Morales, R., Mendez, J.L., and Cravioto, A. (2011). The re-emergence of cholera in the Americas. In: Ramamurthy, T. and Bhattacharya, S.K. (eds.) *Epidemiological and Molecular Aspects on Cholera*. New York: Springer, pp. 79–95.
Diack, A.B., et al. (2012). Constant transmission properties of variant Creutzfeldt–Jakob disease in 5 countries. *Emerging Infectious Diseases*, 18(10), 1574–9.
Epstein, J., Field, H., and Luby, S. (2006). Nipah virus: impact, origins, and causes of emergence. *Current Infectious Disease Reports*, 8(1), 59–65.
Erlanger, T.E., Weiss, S., Keiser, J., Utzinger, J., and Wiedenmayer, K. (2009). Past, present, and future of Japanese encephalitis. *Emerging Infectious Diseases*, 15(1), 1–7.
European Food Safety Authority (2012). *Shiga Toxin-Producing E. coli Outbreak(s)*. [Online] Available at: http://www.efsa.europa.eu/en/topics/topic/ecolioutbreak2011.htm.
Fichet-Calvet, E. and Rogers, D.J. (2009). Risk maps of Lassa fever in West Africa. *PLoS Neglected Tropical Diseases*, 3(3), e388.
Food and Agricultural Organisation (2011). *Investigating the Role of Bats on Emerging Zoonoses*. Rome: FAO. [Online] Available at: http://aitoolkit.org/site/DefaultSite/filesystem/documents/BATS ZOONOSES.pdf
Goma Epidemiology Group (1995). Public health impact of Rwandan refugee crisis: what happened in Goma, Zaire, in July, 1994? *The Lancet*, 345(8946), 339–44.
Goodnough, L., Shander, A., and Brecher, M.E. (2003). Transfusion medicine: looking to the future. *The Lancet*, 361(9352), 161–9.
Guan, Y., Vijaykrishna, D., Bahl, J., Zhu, H., Wang, J., and Smith, G.J.D. (2010). The emergence of pandemic influenza viruses. *Protein & Cell*, 1(1), 9–13.
Guarner, J., et al. (2004). Monkeypox transmission and pathogenesis in prairie dogs. *Emerging Infectious Diseases*, 10(3), 426–31.
Gupta, M.L. and Sharma, A. (2007). Pneumonic plague, northern India, 2002. *Emerging Infectious Diseases*, 13(4), 664–6.
Hassan, M., et al. (2018). Nipah virus contamination of hospital surfaces during outbreaks, Bangladesh, 2013–2014. *Emerging Infectious Diseases*, 24(1), 15–21.
Heymann, D.L. (2003). The evolving infectious disease threat: implications for national and global security. Harvard University/Commission on Human Security for Global Health. *Journal of Human Development*, **4**, 191–207.
Heymann, D.L. et al (2016). Zika virus and microcephaly: why is this situation a PHEIC? *The Lancet*, 387(10020), 719–21.
Hightower, A., et al. (2012). Relationship of climate, geography, and geology to the incidence of Rift Valley fever in Kenya during the 2006–2007 outbreak. *American Journal of Tropical Medicine and Hygiene*, 86(2), 373–80.
Ho, A.Y., et al. (2002). Outbreak of cyclosporiasis associated with imported raspberries, Philadelphia, Pennsylvania, 2000. *Emerging Infectious Diseases*, 8(8), 783–8.
Jamison, D., et al. (2006). *Vaccine-Preventable Diseases*. Washington, DC: World Bank. [Online] Available at: http://www.ncbi.nlm.nih.gov/books/NBK2284/
Johnson, A.P. (2011). Methicillin-resistant *Staphylococcus aureus*: the European landscape. *The Journal of Antimicrobial Chemotherapy*, 66(Suppl. 4), iv43–8.
Jong, M.C.M.D., Stegeman, A., Goot, J.V.D., and Koch, G. (2009). Intra- and interspecies transmission of H7N7 highly pathogenic avian influenza virus during the avian influenza epidemic in the Netherlands in 2003. Interspecies transmission of H7N7 low pathogenic avian influenza virus to poultry. *Revue Scientifique et Technique*, 28(1), 333–40.
Kondro, W. (1997). Study reveals extent of Canada's overprescribing. *The Lancet*, 349(9059), 1156.
Kumarasamy, K., et al. (2010). Emergence of a new antibiotic resistance mechanism in India, Pakistan and the UK: a molecular, biological and epidemiological study. *The Lancet Infectious*

Diseases, 10(9), 597–602.
Larson, H.J., Cooper, L.Z., Eskola, J., Katz, S.L., and Ratzan, S. (2011). Addressing the vaccine confidence gap. *The Lancet*, 378(9790), 526–35.
Larson, H.J. and Heymann, D.L. (2010). Public health response to influenza A(H1N1) as an opportunity to build public trust. *Journal of the American Medical Association*, 303(3), 271–2.
Lederberg, J., Shope, R., and Oaks, S.C.J. (1992). *Emerging Infections: Microbial Threats to Health in the United States*. Washington, DC: National Academy Press.
Lester, S., del Pilar, P.M., Wang, F., Perez Schael, I., Jiang, H., and O'Brien, T. (1990). The carriage of *Escherichia coli* resistant to antimicrobial agents by healthy children in Boston, in Caracas, Venezuela, and in Qin Pu, China. *The New England Journal of Medicine*, 323(5), 285–9.
Matic, S., Lazarus, J.V., and Donoghoe, M.C. (2006). *HIV/AIDS in Europe: Moving from Death Sentence to Chronic Disease Management*. Geneva: WHO.
Meselson, M., et al. (1994). The Sverdlovsk anthrax outbreak of 1979. *Science*, 266(5188), 1202–8.
Miranda, M.E.G. and Miranda, N.L.J. (2011). Reston ebolavirus in humans and animals in the Philippines: a review. *The Journal of Infectious Diseases*, 204(Suppl. 3), S757–60.
Modlin, J.F. (2010). The bumpy road to polio eradication. *The New England Journal of Medicine*, 362(25), 2346–9.
Morbidity and Mortality Weekly Report (2004). Multifocal autochthonous transmission of malaria —Florida, 2003. *Journal of American Medical Association*, 292(3), 324–5.
Nayyar, G.M.L., Breman, J.G., Newton, P.N., and Herrington, J. (2012). Poor-quality antimalarial drugs in southeast Asia and sub-Saharan Africa. *The Lancet Infectious Diseases*, 12(6), 488–96.
Oh, M.D., et al. (2018). Middle East respiratory syndrome: what we learned from the 2015 outbreak in the Republic of Korea. *Korean Journal of International Medicine*, 33(2), 233–46.
Onder, G., Rezza, G., and Brusaferro, S. (2020). Case-fatality rate and characteristics of patients dying in relation to COVID-19 in Italy. *Journal of the American Medical Association*, 323(18), 1775–6.
Pearce, A., Law, C., Elliman, D., Cole, T.J., and Bedford, H. (2008). Factors associated with uptake of measles, mumps, and rubella vaccine (MMR) and use of single antigen vaccines in a contemporary UK cohort: prospective cohort study. *British Medical Journal (Clinical Research ed.)*, 336(7647), 754–7.
Pedersen, S.F. and Ho, Y.C. (2020). SARS-CoV-2: a storm is raging. *Journal of Clinical Investment*, 130(5), 2202–5.
Piarroux, R. and Faucher, B. (2012). Cholera epidemics in 2010: respective roles of environment, strain changes, and human-driven dissemination. *Clinical Microbiology and Infection*, 18(3), 231–8.
Pru Rimoin, A.W., et al. (2010). Major increase in human monkeypox incidence 30 years after smallpox vaccination campaigns cease in the Democratic Republic of Congo. *Proceedings of the National Academy of Sciences U S A*, 107(37), 16262–7.
Riedel, S. (2005). Edward Jenner and the history of smallpox and vaccination. *Proceedings (Baylor University, Medical Center)*, 18(1), 21–5.
Sencer, D.J. (2011). Perspective: swine-origin influenza: 1976 and 2009. *Clinical Infectious Diseases*, 52(Suppl. 1), S4–7.
Senior, K. (2009). Lassa fever: current and future control options. *The Lancet Infectious Diseases*, 9(9), 532.
Siala, E. and Gamara, D. (2015). Airport malaria: report of four cases in Tunisia (2015). *Malaria Journal*, **14**, 42.
Slenczka, W. and Klenk, H.D. (2007). Forty years of marburg virus. *Journal of Infectious Diseases*, 196(Suppl. 2), S131–5.

Takala-Harrison, S., et al. (2015). Independent emergence of artemisinin resistance mutations among Plasmodium falciparum in Southeast Asia. *Journal of Infectious Diseases*, 211(5), 670–9.
Tängdén, T. and Giske, C.G. (2015) Global dissemination of extensively drug-resistant carbapenemase-producing Enterobacteriaceae: clinical perspectives on detection, treatment and infection control. *Journal of Internal Medicine*, 277(5), 501–12.
Taubenberger, J.K. and Morens, D.M. (2006). 1918 Influenza: the mother of all pandemics. *Emerging Infectious Diseases*, 12(1), 15–22.
Trevitt, C.R. and Singh, P.N. (2003). Variant Creutzfeldt–Jakob disease: pathology, epidemiology, and public health implications. *American Journal of Clinical Nutrition*, 78(3), 651S–6.
United Nations Children's Fund (2006). Japanese support for polio eradication and malaria prevention in Nigeria. [Online] Available at: http://www.unicef.org/media/media_35425.html.
Wilder-Smith, A., Teleman, M.D., Heng, B.H., Earnest, A., Ling, A.E., and Leo, Y.S. (2005). Asymptomatic SARS coronavirus infection among healthcare workers, Singapore. *Emerging Infectious Diseases*, 11(7), 9–12.
World Health Organization (WHO) (2001a). *WHO Global Strategy for Containment of Antimicrobial Resistance*. Geneva: WHO. Available at: http://whqlibdoc.who.int/hq/2001/WHO_CDS_CSR_DRS_2001.2.pdf
World Health Organization (WHO) (2001b). *Macroeconomics and Health: Investing in Health for Economic Development (Vol. 8)*. Geneva: WHO.
World Health Organization (WHO) (2003). Summary of probable SARS cases with onset of illness from 1 November 2002 to 31 July 2003. [Online] Available at: http://www.who.int/csr/sars/country/table2004_04_21/en/index.html
World Health Organization (WHO) (2006). International Food Safety Authorities Network (INFOSAN), successful strategies in controlling avian influenza. [Online] Available at: http://www.who.int/foodsafety/fsmanagement/No-04-Avianinfluenza-Aug06-en.pdf
World Health Organization (WHO) (2007a). *The World Health Report 2007. A Safer Future: Global Public Health Security in the 21st Century*. Geneva: WHO.
World Health Organization (WHO) (2007b). Infectious disease threats to our public health security: from SARS to avian influenza. [Online] Available at: http://www.ama-assn.org/resources/doc/cphpdr/david_heymann.pdf
World Health Organization (WHO) (2008a). Global measles deaths drop by 74%. [Online] Available at: http://www.who.int/mediacentre/news/releases/2008/pr47/en/index.html
World Health Organization (WHO) (2008b). *International Health Regulations* (2005) (2nd ed.). Geneva: WHO.
World Health Organization (WHO) (2011). Tuberculosis: MDR-TB & XDR-TB 2011 progress report. [Online] Available at: http://www.who.int/tb/challenges/mdr/factsheet_mdr_progress_march2011.pdf
World Health Organization (WHO) (2012a). Impact of dengue. [Online] Available at: http://www.who.int/csr/disease/dengue/impact/en/
World Health Organization (WHO) (2012b). Dengue fever in Madeira, Portugal. [Online] Available at: http://www.who.int/csr/don/2012_10_17/en/index.html
World Health Organization (WHO) (2012c). *Chikungunya*. [Online] Available at: http://www.who.int/mediacentre/factsheets/fs327/en/
World Health Organization (WHO) (2012d). *Poliomyelitis*. [Online] Available at: http://www.who.int/mediacentre/factsheets/fs114/en/index.html
World Health Organization (WHO) (2012e). cumulative number of confirmed human cases of avian influenza A(H5N1) reported to WHO. [Online] Available at: http://www.who.int/influenza/human_animal_interface/H5N1_cumulative_table_archives/en/index.html
World Health Organization (WHO) (2012f). *Drug-Resistant Tuberculosis*. [Online] Available at: http://www.who.int/tb/challenges/mdr/tdrfaqs/en/index.html

World Health Organization (WHO) (2012g). New WHO book showcases ways to safeguard medications: the evolving threat of antimicrobial resistance—options for action. [Online] Available at: http://www.who.int/mediacentre/news/notes/2012/amr_20120308/en/index.html.

World Health Organization (WHO) (2012h). The evolving threat of antimicrobial resistance options for action. [Online] Available at: http://whqlibdoc.who.int/publications/2012/9789241503181_eng.pdf?ua=1

World Health Organization (WHO) (2012i). *International Health Regulations (2005)*. [Online] Available at: http://www.who.int/ihr/en/

World Health Organization (WHO) (2013). *Global Alert and Response (GAR), Disease Outbreaks by Year*. [Online] Available at: http://www.who.int/csr/don/archive/year/en/

World Health Organization (WHO) (2020). *Coronavirus disease 2019 (COVID-19) Situation Report*. [Online] Available at: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports

Wu, D., Wu, T., Liu, Q., and Yang, Z. (2020a). The SARS-CoV2 outbreak; what we know. *Journal of Infectious Diseases*, **94**, 44–8.

Wu, Z. and McGoogan, J.M. (2020b). Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China: summary of a report of 72 314 cases from the Chinese Center for Disease Control and Prevention. *Journal of the American Medical Association*, 323(13), 1239–42.

Wu, Z. (2020c). *The COVID19 Epidemic in China*. *Conference on Retroviruses and Opportunistic Infections*. Boston MA. [Online] Available at: https://www.croiconference.org/abstract/outbreak-of-covid-19-in-china-and-the-countrys-responses/

Wu, Z. and McGoogan, J. (2020). Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China. *Journal of American Medical Association*, 323(13), 1239–42.

York, I. and Donis, R.O. (2012). The 2009 pandemic influenza virus: where did it come from, where is it now, and where is it going? *Current Topics in Microbiology and Immunology*, **370**, 241–57.

8.18

Bioterrorism

Peter Katona

Introduction

Whether from a natural pandemic, antibiotic resistance, a biological accident, or an act of terrorism, deadly disease can be a threat to all of us and to our national security as the fine line between natural and man-made infection is getting thinner. With earth's 7.4 billion people, 20 billion chickens, and 400 million pigs, we have the ideal scenario for creating and spreading deadly microbes between humans, animals, and even to and from plants. Trade and travel connect points of the globe in a matter of hours not days or weeks as in the past. Tens of millions of people are living in the microbe rich and sanitation poor mega-city slums or markets of the developing world such as Lagos or Mexico City or Wuhan where a small outbreak can spread into a pandemic with or without a terrorist's help. Add to this a sophisticated bioterrorist using a microbe on an unprecedented scale with today's advancing weaponization technology and one can further understand our need for preparedness.

The evolution, risk, and relevance of biological weapons

Biological agents are living organisms that may or may not have disease or transmissibility properties. Biological weapons are living biological agents or their toxins that are intentionally used to cause harm to humans, plants, or animals. More specifically, the draft Model State Emergency Health Powers Act of 2001 defined bioterrorism as 'the intentional use of any microorganism, virus, infectious substance, or biological product that may be engineered as a result of biotechnology, or any naturally occurring or

bioengineered component of any such microorganism, virus, infectious substance, or biological product, to cause death, disease, or other biological malfunction in a human, an animal, a plant, or another living organism in order to influence the conduct of government or to intimidate or coerce a civilian population'. These organisms may be easily obtained, and, with the right equipment and technical skills, grown and disseminated relatively easily. If weaponized, they can infect a large number of hosts with a small quantity of product. They can be contagious or non-contagious from person to person.

If disseminated efficiently, biological weapons can cause high morbidity and mortality. Whether weaponized or naturally occurring, they can be seen as part of a larger group of evolving emerging and re-emerging biological agents. This evolution is caused or aggravated by a number of factors including climate change, antibiotic resistance, the overuse or misuse of antibiotics, new or changing disease vectors, new habitats, advancing, mostly genetic, technology, or a terrorist act. Microbial agents may be harmful or beneficial to society as shown in Table 8.18.1.

Table 8.18.1 Microbial agents

Dr Jekyll	Mr. Hyde
• Antibiotics and vaccines • Fungi that eat plastic, provide clean fuels, and clean up radiation polluted areas • Better understanding of the genome, proteome, and microbiome	• Isolated infections • Outbreaks, epidemics, and pandemics • Biological control byproducts • Emerging and re-emerging agents • Biological weapons • Increasing resistance

Biological weaponization is key to intentionally causing significant morbidity and mortality to large numbers of people, animals, or plants. The process of making a biologic agent weaponized first involves achieving a particle size of 1–5 microns. Also important is the chemical coating, the absence of an electrostatic charge, the non-hydrophobic non-self-adhering or 'flowability' properties, virulence, the viability of the agent within each

particle, antibiotic resistance, and ease of milling to remove clumps and to stabilize the particle.

There are several types of serious intentional biological events. They can be biological crimes targeted against individuals, acts of bioterrorism making a violent political statement against a target audience between or by nations (see Fig. 8.18.1), or agroterrorism directed against animals or crops. There have also been accidental laboratory releases. A bioattack may be a high impact low probability event such as Nissan Talib's low probability high impact 'black swan', or Donald Rumsfeld's unanticipated 'unknown unknown'. All may constitute public health emergencies.

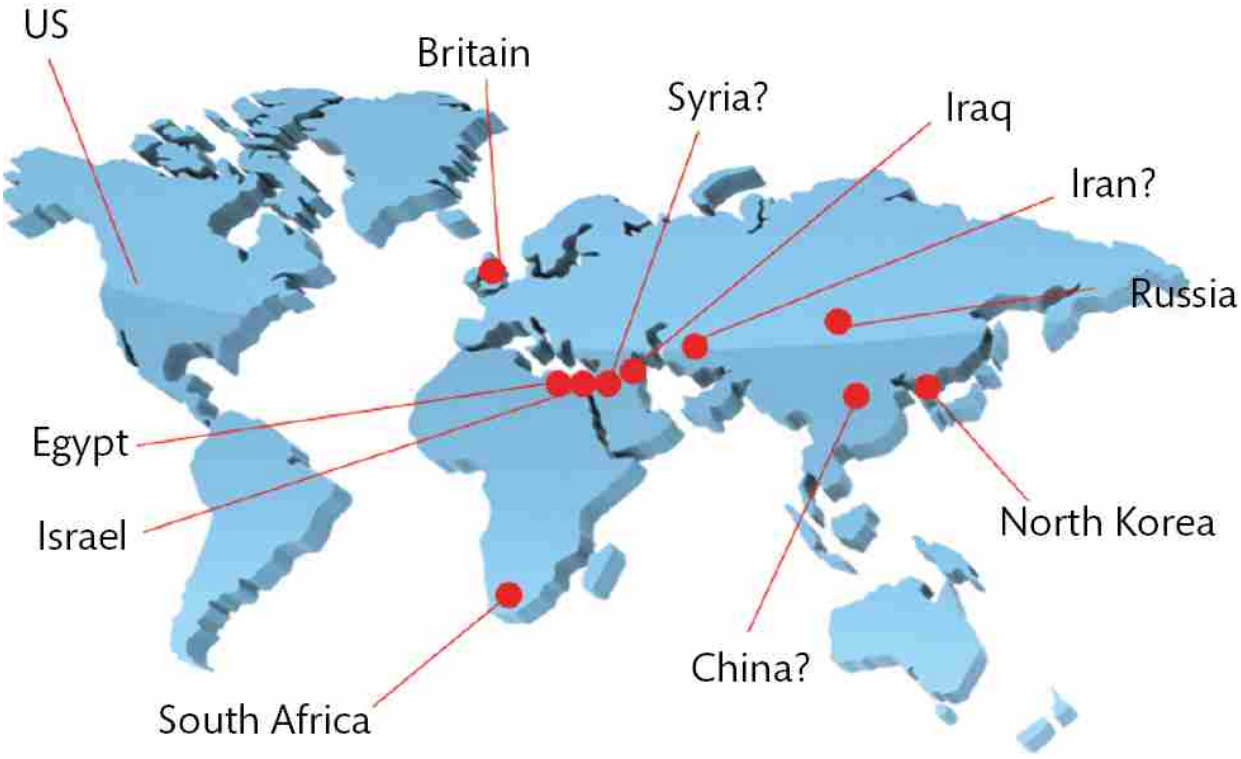


Fig. 8.18.1 Biowarfare programmes.

The risk of a biological attack is loosely defined as its probability times its consequences. This takes into account prevention, mitigation, resilience, preparedness, avoidance, post-event emergency response with detection, containment, treatment, and recovery. The term 'dual use' is important in that care is needed in not letting a terrorist get the upper hand with peacetime technology that could later be converted into a weapon (dual use). Other

factors promoting bioweapons are the low cost, the compensation for military inferiority, the attribution difficulty, and the fact that it enables hyperempowered individual(s) to achieve a nefarious and more significant goal. Ultimately, for a bioterror attack to be successful there needs to be a critical system breach, a widespread mechanism, and a prevention or mitigation failure.

Treaties, initiatives, directives, and legislation

Bill Gates warned world leaders that terrorist use of a genetically engineered virus could kill more people than a nuclear weapon when he delivered a chilling address at the 2017 Munich Security Conference. His concern was supported by US and British intelligence agencies, which felt that 'scientific terrorists' have access to the necessary tools to design biological weapons of mass destruction capability. In the face of this knowledge, he said that governments around the world remain complacent and neither prepared nor equipped to respond. Gates suggested preparing for epidemics 'the way the military prepares for war' with more exercises and training (Gates 2017).

Gates specifically mentioned organizations like the Coalition for Epidemic Preparedness Innovations, a public-private coalition that aims to derail epidemics by speeding up the development of vaccines, much like what the federal government is doing with programmes like the Biomedical Advanced Research and Development Authority (BARDA), but with less paperwork. Research was needed, he said, to dramatically reduce the time to develop vaccines. He added that there needs to be adequate laboratory resources to identify and monitor suspect pathogens, as well as trained medical personnel, germ games, and other preparedness exercises.

Post 9/11 US presidents have also shown an interest. In 2002 President George W. Bush said: 'Bioterrorism is a real threat to our country. It's a threat to every nation that loves freedom. Terrorist groups seek biological weapons; we know some rogue states already have them … It's important that we confront these real threats to our country and prepare for future emergencies.' Many programmes addressing this were started under his administration.

The Public Health Security and Bioterrorism Preparedness and Response Act of 2002 was critical. It established preparation procedures for bioterrorism and public health emergencies. It created the National Disaster

Medical System, and security risk assessment of all individuals who have access to select agents and toxins, especially from the CDC Category A list mentioned later.

National Security and Presidential Directives HSPD 10 in 2001 and NSPD 33 in 2004 addressed ‘Biodefense for the 21st Century’ for threat awareness, prevention, and protection, surveillance and protection, as well as response and recovery in an attempted comprehensive way. The National Security Preparedness Memorandem of 2018 (NCPM-14) emphasized that biothreats are persistent, originate from multiple sources, and don’t respect borders. Biothreats require multisectoral cooperation for prevention and response, and need science and technology to advance globally.

There were many developments regarding biopreparedness following the events of 9/11. Congress established a Senate Subcommittee on Bioterrorism and Public Health Preparedness (SSBPHP). The House established the House Subcommittee on Emergency Preparedness, Response and Communications (HSEPRC). There was even a position created for a Special Assistant to the President for Biodefense (ASPR).

President Obama spoke of the need to ‘come together to prevent, detect and fight every kind of biological danger, whether it’s a pandemic like H1N1, or a terrorist threat’. His mention of bioterrorism in the 2010 State of the Union address, and his issuance of several executive orders increased congressional attention (Obama Administration Gets an ‘F’ for Bioterrorism Defense 2010). The absence of a comprehensive US capability to rapidly recognize, respond to, and recover from a disease-based attack was, he claimed, the most ‘significant failure’ identified by the Commission on the Prevention of Weapons of Mass Destruction Proliferation and Terrorism.

In 2018 President Trump spoke of establishing a Biodefense Steering Committee (BSC), led by the Health and Human Services (HHS) Secretary, to coordinate biodefense efforts across the federal government. He would also task the National Security Adviser with reviewing threats and updating preparations on an annual basis. He claimed that his administration was focused on strengthening our nation’s defences against the full range of threats to our health and security, including those of a biological nature. The planning, he said, was meant to counter both man-made biological weapons and naturally occurring diseases like Ebola.

With threats to food, water, agriculture as well as to humans there have been many federal initiatives addressing weapons of mass destruction

(WMD) in general, and bioterrorism specifically. USDA, HHS, and the Department of Homeland Security (DHS) are tasked with initiatives inside the US, while the Department of Defense (DoD), the Department of State (DoS) and the intelligence community are tasked to the foreign side. Prevention was placed under DoS, the Department of Justice (DoJ), DHS, HHS, and the intelligence services.

Many HHS efforts are housed within the Centers for Disease Control and Prevention (CDC), which supervises the Bio-Intelligence Center, the Strategic National Stockpile (SNS) of countermeasures available on short notice from distribution centres strategically placed around the country. CDC also manages the Cities Readiness Initiative (CRI) programme designed to enhance biopreparedness in the nation's largest population centres, where nearly 60% of the population resides. The purpose was to effectively respond to large-scale public health emergencies needing life-saving medicines and medical supplies. HHS, like DHS, also funds regional Centers of Excellence to lead and coordinate research on defences against both bioterrorism and emerging infectious diseases. CDC has a unit dedicated to emergency preparedness and response and has also published numerous guidelines (e.g. Emergency Preparedness and Response 2018).

The Pandemic and All-Hazards Preparedness Act (PAHPA) of 2006, reaffirmed by the 2013 Pandemic and All-Hazards Preparedness Reauthorization Act (PAHPRA) established the HHS Assistant Secretary for Preparedness and Response (ASPR) as the principal adviser to the HHS Secretary. They were responsible for providing integrated policy coordination and strategic direction for all matters related to public health, medical preparedness, and deployment of the federal response for public health emergencies and incidents. Most of these initiatives are housed under HHS' Science and Technology Directorate.

PAHPA transferred the National Bioterrorism Hospital Preparedness Program (NBHPP) from the Health Resources and Services Administration (HRSA) to ASPR. The focus was now all-hazard preparedness from the original focus on bioterrorism. The purpose of the now renamed Hospital Preparedness Program (HPP) was to enhance the ability of hospitals and healthcare systems to prepare for and respond to bioterrorism and other public health emergencies. Programme priority areas included interoperable communication systems, National Incident Medical System compliance and training, bed tracking, personnel management, mass fatality management

planning, and hospital evacuation and shelter in place planning. During the past 5 years, HPP funds have also improved medical surge capacity and resources, decontamination capabilities, isolation capacity, pharmaceutical supplies, training, and education drills and exercises. Much of this responsibility has more recently been transferred to regional healthcare coalitions across the country.

Louisiana State University's National Center for Biomedical Research and Training (NCBRT) is a member of the National Domestic Preparedness Consortium (NDPC), a partnership of public and private organizations, the majority academic, and is committed to serving emergency first responders by providing quality, cost-effective counterterrorism training. The goal is to provide a focused threat responsive, long-term national capability to execute and sustain comprehensive and coordinated domestic WMD emergency responder education courses, and is specifically tasked with educating first responders about biological terrorism and community disaster preparedness.

Established in 2016, The National Biodefense Analysis and Countermeasures Center (NBACC) is a national biodefense research laboratory created by DHS. It supports intelligence assessments, preparedness planning, response, emerging threat characterization and bioforensic analyses. It was the first national laboratory created by DHS.

NBACC is comprised of two sections: The National Bioforensic Analysis Center (NBFAC) and the National Biological Threat Characterization Center, the lead federal facility to conduct and facilitate the technical forensic analysis and interpretation of materials recovered following a biological attack in support of the appropriate lead federal agencies like the FBI. It supports federal law enforcement conducting experiments to better understand biological vulnerabilities and hazards. The NBFAC provides the FBI with bioforensic analysis and testing of suspected biothreat samples and has supported over 100 federal cases. The Center is partnered in the National Interagency Confederation for Biological Research at Fort Detrick and has laboratories with biosafety levels 1 2,3, and 4. Having the highest level of safety and experimental capabilities this enables research on pathogens that do not have a treatment or vaccine. Currently, there are only 13 BSL-4 laboratories in the US. NBACC works with the DoD and National Institutes of Health to organize and conduct research with the goal of addressing the gaps and needs of biodefense.

DHS's National Biosurveillance Integration Center (NBIC) looks at the biothreat awareness, detection, and information sharing needed to respond to incidents that may cascade into a nationally significant event. This includes bioterrorism attacks, disease outbreaks, and environmental catastrophes as well as data on medical intervention, food safety, travel, and wildlife health. By tapping into all of these databases, their centralized system would be able to track, map, and predict national patterns and outbreaks. With enough data across so many sources and sectors, they hope to eventually generate predictive algorithms using artificial intelligence.

DHS surveillance and communication systems such as BioWatch are used to detect the release of biopathogens into the air as part of a terrorist attack on major American cities. It was created in 2001 in response to the increased threat of bioterrorism on postal facilities sparked by the 2001 anthrax attacks, and was announced in President George W. Bush's State of the Union Address of 2003. The programme, described as 'the nation's first early warning network of sensors to detect biological attack', operates via a system of filters located within existing Environmental Protection Agency (EPA) air filters that monitor air quality. The CDC analyses results from these filters and passes any significant results to the FBI. Unfortunately, the system also had generated more than 50 false alarms between 2003 and 2008. State and local health officials have never ordered evacuations or distributed emergency medicines in response to a false positive reading from the system.

BioSense, also operated by the CDC, has near real-time electronic connectivity with 32 hospital emergency departments in 10 cities. It is designed to give federal and local health officials an immediate indication of a bioterrorism or deadly infectious disease outbreak via syndromic surveillance of symptoms.

An overview of these agencies and laws is highlighted in Table 8.18.2.

Table 8.18.2 Bio acronyms mentioned in the text

DHHS	DHS	DOD	Laws	Directives
CDC	RCE	USAMRIID	PAHPA	BSC
ASPR	NBACC	DTRA	PAHPRA	HSPD-10
HRSA	NDPC		SSBPHP	NSPD-33
BioSense	NACC		HSEPRC	NCPM-14
	NBIC			
	BioWatch			

The CDC-defined agents

There are many factors that make biological terrorism appealing. Biothreat agents can be cheap and—with proper expertise—relatively easy to handle as production recipes have been available on the Internet. There is an incubation period between contact with the agent and the development of symptoms, which can range from a day for plague to a few weeks for smallpox. Perpetrators may be halfway around the world before any symptoms are recognized. Detection is difficult as they are tasteless and odourless. Most important is that they can be efficiently disseminated by aerosol especially in large closed spaces.

Since there have been very few large-scale bioterrorist incidents, CDC had to rely on a panel of experts to decide which agents were most important to prepare for. They classified agents into categories A, B, or C based on their ease of transmission, severity of morbidity and mortality, and likelihood of use. Category A agents can be easily disseminated or transmitted from person to person, result in high mortality rates and have the potential for major public health impact. They might cause public panic and social disruption and require special action from public health authorities. Most Category A agents are considered especially dangerous due to the potential for airborne transmission. They are shown on Box 8.18.1

Box 8.18.1 CDC Category A agents

Variola major (smallpox)*
Bacillus anthracis (anthrax)
Yersinia pestis (plague)*
Clostridium botulinum toxin (botulism)
Francisella tularensis (tularemia)
Filoviruses Ebola and Marburg*
Arenaviruses Lassa (Lassa fever), Junin (Argentine), and related viruses*

* contagious from person to person including only the pneumonic form of plague.

CDC Category B agents are moderately easy to disseminate, result in moderate morbidity and low mortality, and require CDC's diagnostic capacity and enhanced disease surveillance. They include Coxiella burnetti (Q fever); Brucella species (brucellosis); Burkholderia mallei (glanders); Rickettsia prowazekii (Typhus fever); alphaviruses Venezuelan, Eastern, and Western equine encephalitis; Ricinus communis (ricin toxin from castor beans); epsilon toxin of Clostridium perfringens; and Staphylococcus enterotoxin B. The CDC list also contains a group of Category B agents loosely defined as 'food- or waterborne pathogens'. These include Salmonella species, Shigella dysenteriae, *Escherichia coli* O157:H7, Vibrio cholera, Cryptosporidium, and Noroviruses.

Category C agents are emerging pathogens that could be engineered for mass dissemination because of their availability, ease of production and dissemination, potential for morbidity and mortality, and major health impact being less significant than Category A and B agents. This includes Nipah virus, hantaviruses, tick-borne haemorrhagic fever viruses, tick-borne encephalitis viruses, yellow fever, and multidrug-resistant tuberculosis.

Although clinical symptoms of these agents may vary from diarrhoea to a skin rash, a commonality of virtually all of these biothreats agents is a 'flu-like illness' consisting of fever, chills, and body aches. Contagiousness is important but not necessary. Pandemic influenza and coronaviruses were not even on this list.

There are many possible routes of exposure to a biologic agent. There is ingestion, where the agent enters when the host eats, drinks, or puts contaminated fingers or other objects into his or her mouth. An example is the Salmonella salad bar attack by the Rajneesh. There is absorption, where

the agent enters the host through mucosal surfaces such as eyes, nose, or throat. Intact skin is hard to breech except by T2 mycotoxin, a fungal byproduct. There is injection, where the organism enters the host through the skin via a vector such as animal bites, hypodermic needles, any sharp object, or defects in the skin surface as in the Markov ricin incident. The most effective route is inhalation, where the agent enters when the host takes a breath. This occurred with the anthrax letters. The optimal particle size, as mentioned, is 1–5 microns, so the agent can easily establish itself in the lungs, cause local disease, or even enter the bloodstream. Inhalation exposure can occur with slowly descending aerial bombs or bomblets, spray tanks, carefully designed ballistic missile warheads, or, as we now know, letters.

Clinical recognition indicators of an intentional bioattack may include unusual time of year or locale for a particular set of common flu-like symptoms. There may be an unusual or unexpected number of sick or dying people or animals. There may be an unscheduled or unusual dissemination of sprayed material such as mosquito sprayers or crop dusters. We may even find abandoned sprayer dispersion devices.

Initial detection may be from a ‘sentinel’ case where an astute physician has diagnosed the disease in a patient. There may be arrival in large numbers overwhelming emergency rooms. There may be gradual recognition by the medical community, an indicator from a public health or animal or syndromic surveillance system, a lab report, or the opening of a suspicious letter. The best-case scenario is prevention by infiltration into a terrorist organization from good intelligence gathering. The worst-case scenario is after mass chaos has occurred.

Like nuclear weapons, to optimize the dissemination of a weaponized biothreat agent ideally one needs to do experiments, which the US military originally did in the 1950s. Both liquid agents, preferred by the Soviets, and dry powder, preferred by the US, were tested. Testing was done in Alaska to assess the influence of extreme cold weather (it was not a factor). In Texas, an experiment was done utilizing flying aircraft to assess if there was sufficient energy to deliver a respiratory sized particle aerosol using liquid agent (there was). In New York City, experiments were done to see if subway cars generated sufficient wind currents to disseminate a dry powder product (they did). In San Francisco Bay experiments assessed the influence

of weather conditions. They found that outdoor dissemination of an agent is very dependent on inversions, air temperature, wind speed, and humidity.

More recently there is an EPA and DHS project designed to help develop the tools and techniques necessary for the rapid return of service to a subway system after it had been contaminated with a biological organism. In one experiment they wrapped a subway car in cellophane to assess how quickly they could decontaminate it. All told here were three field studies done since 2013.

Indoor release has fewer potential liabilities. Researchers concluded that a well-planned weaponized aerosol agent disseminated through the air can have far-reaching consequences, especially in large enclosed spaces such as closed stadiums, subway systems, and convention halls. All this information has been available to terrorists.

There were also functional WMD management exercises, the most prominent being TOPOFF and Dark Winter. TOPOFF consisted of four WMD scenarios played out every other year by ‘top officials’. The first TOPOFF exercise was held in 2000 in Denver, Colorado, and Portsmouth, New Hampshire where participants faced a simulated attack with plague. TOPOFF 3, held in 2005, involved a biological attack in New Jersey, also with plague. Operation Dark Winter was a senior-level bioterrorist attack simulation conducted in 2001. It was designed to carry out a covert and widespread smallpox attack on Oklahoma City with additional attacks in Georgia and Pennsylvania. Organized by DoJ and FEMA, it was specifically designed to spiral out of control. The purpose was to assess the nation’s crisis and consequence management capability by local, state, and federal responses to geographically dispersed terrorist threats and acts.

History: from nations to cults to lone wolves to unknown perpetrators

There have been large-scale national programmes, cults, individuals, and unknown perpetrators involved in intentional bioattacks. There were also accidental releases. While early biowarfare started by contaminating a city’s reservoir with a dead horse going back thousands of years, we will start with an incident in the 14th century.

During the 1346 battle of Caffa the Tartar army hurled the corpses of plague victims over the walls of the city, a seaport on the Crimean coast. Gabriel de Mussis, a notary, saw the Tatar attack on this well-fortified Genoese-controlled seaport in modern Ukraine. He described how the plague-weakened aggressors catapulted dead victims of plague over the walls into the town. He wrote: '[The Tatars], fatigued by such a plague and pestiferous disease, stupefied and amazed, observing themselves dying without hope of health, ordered cadavers placed on their hurling machines (catapults) and thrown into the city of Caffa, so that by means of these intolerable passengers the defenders died widely.' An epidemic of plague within the walls of the city followed, forcing a retreat of the Genoese forces. The exported disease continued to spread in Europe. Some speculate that this may have set off the 2nd European Black Death plague pandemic of the Middle Ages that resulted in 30–40 million deaths, half the worldwide population. This was repeated in 1710, when the Russians, besieging Swedish forces at Reval in Estonia, also catapulted bodies of people who had died from plague.

The French and Indian Wars were not spared from acts of biological terrorism. British forces used smallpox-tainted clothing as a weapon in the 1785 siege of La Calle. Earlier, British officers such as Sir Jeffrey Amherst had plans to intentionally transmit smallpox to Native Americans during Pontiac's Rebellion near Fort Pitt (present-day Pittsburgh) in 1763. It is not clear whether they actually carried out these plans. But, from whatever its source, smallpox did spread rapidly among susceptible Native Americans in the area during and after that rebellion, killing off as many as 50% of the Indian population.

Japan's infamous Unit 731 operated between 1937 and 1945. Unit 731 was a biowarfare unit masquerading as a water purification facility with a 150-building complex just outside of Harbin, Manchuria, with another related site at Changchun named Unit 100. In one experiment plague-infected fleas were dropped with grain from airplanes using uniquely designed fragmentation bomblets over Manchuria. The grain attracted rats, and the rats became infected from the fleas. This brought the disease deep into the human population. All told they field-tested biological weapons on the Chinese, on prisoners of war and on other civilians. Tens of thousands died as a result of these plague, cholera, and anthrax experiments.

In their facilities the Soviet Union produced biological agents in unconscionable quantities., including greater than 1,000 metric tonnes of anthrax and plague, and 100 metric tonnes of smallpox. Produced in Biopreparat and other facilities, this was enough bioagent to eradicate every living thing on the entire galaxy. But there were also problems. In 1979 the Soviet Union had an anthrax accident at Sverdlovsk, a Soviet military compound. The accident resulted in 77 anthrax cases with 66 deaths, all downwind from the facility that had failed to replace a defective filter, and all from less than 1 gm of weaponized anthrax spores. Cases occurred within the known incubation period of 4 to 45 days. The Russians, originally stating that this was naturally occurring intestinal anthrax, did not admit this was from a military accident until many decades later. Biopreparat employed 60,000 people in the production of offensive biological weapons. Russia may well still be experimenting with bioweapons in other facilities.

During World War I, the German Army developed anthrax, glanders, cholera, and a wheat fungus, specifically for use as biological weapons. They allegedly spread plague in St. Petersburg, Russia, infected mules with glanders in Mesopotamia, and attempted to do the same with the horses of the French Cavalry.

The US also had a programme housed at the US Army Medical Research Institute for Infectious Disease (USAMRIID), which stopped offensive bioweapons experiments in 1969, but has continued studying defensive technology, studying agents that had a vaccine or therapy. National biowarfare programmes are shown on the graphic.

With the defection of Iraqi General Hussein Kamal Hassan in 1995, evidence was found that the Iraqi biological warfare programme was more advanced than previously believed. Iraq had both biological and chemical weapons in the 1990s. Inspectors found that they produced 19,000 litres of botulinum toxin, 2,200 litres of aflatoxin, and 8,500 litres of anthrax. They filled 25 missiles and 166 bombs with anthrax and botulinum toxin. They also had delivery systems using SCUD missiles, bombs, and even a remote-controlled MIG-21 equipped with a 500-gallon sprayer. The reason these agents were not used in Gulf War I remains unclear.

In 1978 Bulgarian exile Georgi Markov, while in London, was injected in the leg with a steel ball impregnated with ricin via a specially constructed umbrella. He felt immediate pain at the injection site and within 5 hours became weak and dizzy. Fifteen to 24 hours later, Markov developed a fever,

became nauseated, and was admitted into a hospital. He eventually died. The reason for the area of induration and redness on his leg was unknown until the autopsy. Markov represents, in recent history, the first example of state-supported bioterrorism. The assassination was carried out by the communist Bulgarian government with technology supplied by the Soviet Union. The platinum-iridium pellet was the size of the head of a pin and cross-drilled with 0.016-inch holes to contain the toxin.

Only 10 days earlier, a similar assassination attempt was made against another dissident, Vladimir Kostov, in Paris. In this case only heavy clothing prevented the steel ball from entering any farther than Kostov's subcutaneous tissue. After he learned of his comrade's death, he went in for an examination and the pellet was found before any of the toxin was absorbed. Forty years later a similar presumed Russian attack using the chemical agent novichok occurred just outside of London.

Aum Shinrikyo was a religious cult based primarily in Japan. During the morning of March 12, 1995 they unleashed the chemical agent sarin in the Tokyo subway. The five simultaneous attacks resulted in 5,000 injuries and 12 deaths in a very poorly executed attack consisting of punching holes into plastic bags containing liquid sarin with an umbrella. Less well-known were eight unsuccessful anthrax attacks the group tried. They had purchased a 48,000-acre farm in Australia to test biological agents on livestock. They went to Africa for Ebola samples, had two biologic research centres, and had $300 million, $80 million of which was earmarked for biological weapons. Fortunately they lacked the expertise to successfully carry out an attack.

The 2001 letters laden with weaponized anthrax resulted in 22 cases, including five deaths. Initially, the letters were sent to NBC News, the New York Post, and AMI, a newspaper publishing firm. Three weeks later two more letters were sent to Senators Tom Daschle and Patrick Leahy, each bearing the handwritten return address of a non-existent 'Greendale School' in Franklin Park, New Jersey. Dr Steven Hatfill was the FBI's first 'person of interest', but eventually USAMRID researcher Bruce Ivins became the focus of the FBI investigation until his suicide just days before his arrest. The anthrax letters had many consequences: the largest investigation in law enforcement history, very high cost, numerous reports, commissions, meetings, books, articles, a building frenzy of BSL-4 labs, many dead ends, and many new federal programmes.

What's the possible role of ISIS and other current terrorist actors in bioterrorism? In 2014 an ISIS laptop was recovered containing a 19-page document on how to develop biological weapons, and in 2017 Kenyan authorities disrupted an anthrax plot by a medical student and associates affiliated with ISIS. In 2018 South Korea raised concerns that North Korea possessed biological weapons and could use drones to carry out attacks. A North Korean defector also showed evidence of anthrax exposure by antibody studies.

Although the nature of the 'next 9/11' was factored well into preparedness, worry over the unpredictable domestic lone wolf seems to dominate. With minimal resources the harm they inflict may be limited, but the inspiration from groups such as ISIS and the Taliban will continue to be problematic. Using airplanes as weapons was unpredictable as was using cars and trucks to run over pedestrians. Would drones or something else be next?

Connecting the agent, the perpetrator, the victim, and the target

For a successful bioattack you need an effective terrorist group or perpetrator(s) and a target population—which may not be the victim. The selected biological agent has to be available in large enough quantities, weaponized, with an effective dissemination method. The target should be 'soft' or vulnerable.

A biological attack will have three main phases (Fig. 8.18.2). First, there's the pre-exposure planning phase where drug prophylaxis, immunization, and training are done. Second is the incubation phase, which may range from a few hours for toxins, or days to several weeks for bacteria or viruses. This is the period between acquisition of infection and manifestation of disease. Finally, there is the period of overt disease where diagnosis, treatment, and multidisciplinary communication become increasingly important. The 'window period' between the onset of illness in the first few cases and confirmation of the agent is very important. Vital decisions will have to be made during this brief period of time (hours to a few days) with limited data on hand. The longer-term implications of these decisions are going to be

important regarding resource allocation and public statements. Afterwards there is the post-recovery or resilience phase.

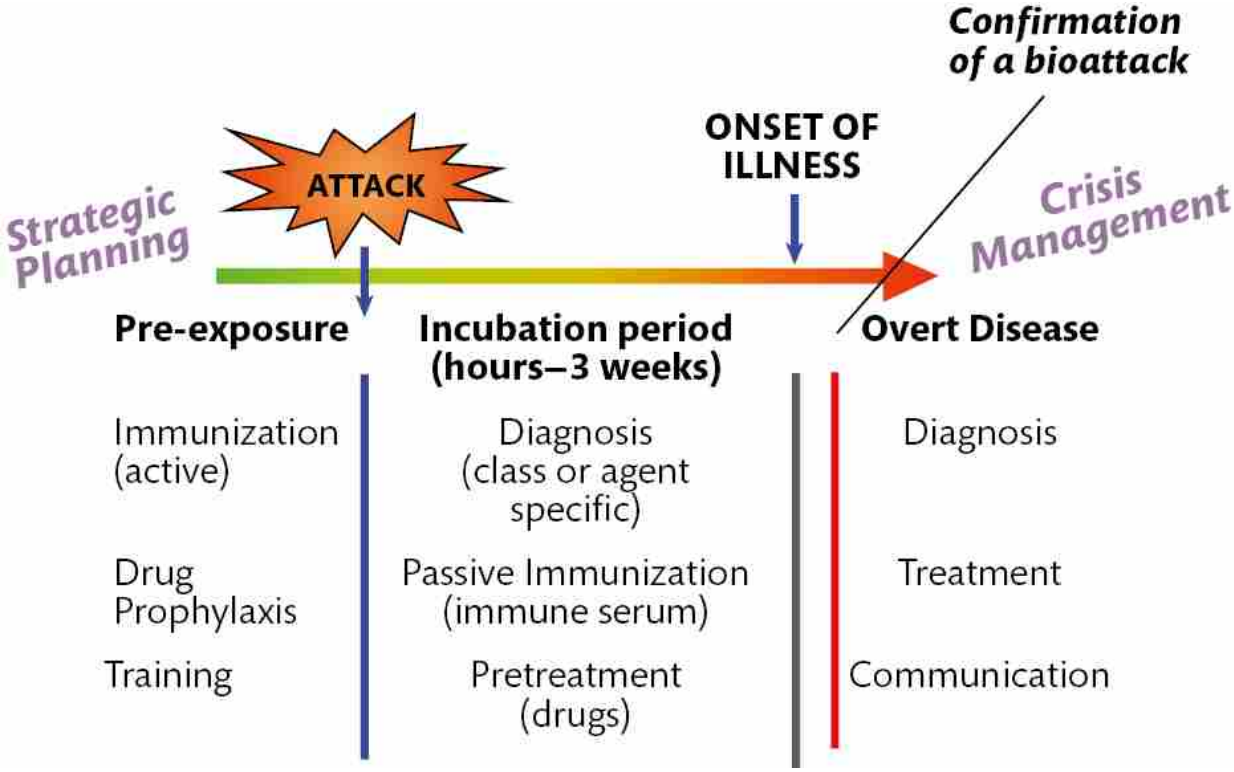


Fig. 8.18.2 Sequence of events.

Preparedness, disease management, and resilience

A few basic questions: How prepared are we for a biological attack? How prepared do we need to be? In a perfect world, how prepared can we be? What type and distribution of training should be made available to optimize preparedness and response? Irwin Redlener, Director of the Earth Institute at Columbia University, said that preparedness is the right place on the continuum between mindless complacency and all-consuming paranoia. In normal times we look for medical affordability, quality, and accessibility (Redlener 2006). We plan for dual use and maintain stockpiles or caches. Disasters, especially biologic ones, require a different approach, including functional resilience, the ability to surge, having access to stockpiles,

avoiding panic, and understanding that, with limited resources, we're dealing with a new type of mass casualty ethics.

At all levels there are four components of disaster preparedness and response that come into play with a biodisaster. They are public health, delivery of healthcare, public safety, and emergency management.

Public health deals with the distribution of countermeasures, leading the epidemiologic investigation in conjunction with the criminal investigation, public relations messaging to the public, and setting up quarantine, isolation and 'cordon sanitaire' procedures. CDC, with broad resources and jurisdiction, does, on a national level, what health departments do on local and state levels. Public health will have the best biological expertise.

In the US healthcare is a complex mosaic including very large health systems, single physician practices, public and private payers, research institutions, medical device developers, software companies, and a diverse patient population. Today's hospital system is the key ingredient, and in some ways is like independent fire stations without agreements for multiple alarm fires. There is no broad incident command (ICS) structure. The hospital ICS, or HICS, is a good start for facilities, but it is internal to each facility usually without shared resources with other hospitals. Preparedness exercises are not practised enough. Electronic medical records, if used wisely, can help. There will be patient confidentiality considerations. Large hospital systems devote more resources and thus seem to be better prepared. A contagious agent might present more complex problems. Healthcare providers have the needed medical expertise.

Public safety includes law enforcement and fire. Law enforcement has many jurisdictions including local city police, county Sheriff, state Highway Patrol, federal Marshals, and FBI, all with different roles in a biodisaster. They deal with the criminal investigation as well as enforcing quarantines. The FBI will ultimately be the lead agency. Public safety provides security expertise and emergency medical services.

All disasters are local, and emergency management is the public entity that leads coordination in a unified command structure at the local level. The emergency manager leads mitigation, planning, response, and recovery efforts for natural, technological, and man-made disasters, including acts of bioterrorism. Is their access to data and the ability to quickly analyse and coordinate it the new source of authority? Emergency management provides the crisis management component. See Figure 8.13.3.

Fig. 8.18.3 Healthcare planning and recovery.

Surveillance is paramount to preparedness, and falls into several categories: general routine, syndromic, non-specific (i.e. drug sales, absenteeism, animal disease), and targeted surveillance that is intensified once we know that an outbreak has started. But surveillance methodologies are changing. There are now prediction models, social networking, Google Trends, integrated diverse human and non-human systems, as well as big data mining and complex analytics.

Global concerns and interactions

The World Health Organization (WHO) lacks adequate appropriations and was slow to recognize the 2014 Ebola as a 'Disease of International Concern'. CDC, the lead US domestic public health agency, has assets and capabilities that reach well into both domestic and international communities. A review of WHO's Blueprint list of priority diseases occurred 6–7 February 2018. Given their potential to cause a public health emergency

and the absence of efficacious medical countermeasures, there was an urgent need for accelerated research and development for agents shown in Box 8.18.2.

Box 8.18.2 WHO blueprint agents

Crimean-Congo haemorrhagic fever (CCHF)
Ebola virus disease and Marburg virus disease
Lassa fever
Middle East respiratory syndrome coronavirus (MERS-CoV)
Severe acute respiratory syndrome (SARS)
Nipah and Henipa viruses
Rift Valley fever (RVF)
Zika
Disease X

Disease X represents the knowledge that a serious international epidemic could be caused by a pathogen currently unknown to cause human disease, and so the R&D Blueprint explicitly seeks to enable crosscutting R&D preparedness that is also relevant for an unknown 'Disease X' as far as possible. Note that only Ebola from the CDC A List was mentioned, and influenza was not mentioned at all.

There have been a number of international bioweapons agreements. The Geneva Protocol of 1925 was signed by 108 nations. This was the first multilateral agreement that extended prohibition of chemical agents to biological agents. In 1972 the US became a signatory to the United Nations Convention on the Prohibition of the Development, Production and Stockpiling of Bacteriological and Toxin Weapons and or Their Destruction (usually called the Biological and Toxin Weapons Convention or BTWC) By November 2001, there were 162 signatories with 144 countries ratifying the convention. Unfortunately, without established verification procedures in place, dual use, and lack of inspections, many nations such as the North Korea, Iran, Soviet Union, and Iraq didn't abide by its terms. This preceded President Ronald Reagan's 'Trust but verify' comment. The UN Security Council has also held sessions devoted to preventing the proliferation of chemical, biological, and nuclear weapons, although not much has come out of these meetings.

The future

Will offensive technology outgrow defensive technology? Will risk change significantly over time as we feel (often falsely) complacent? From the sociological perspective there will always be enemies. And if these enemies have the capability and desire to attack with a biological weapon they will do so.

The historian Yuval Harari noted that we are uneasily poised at the conjuncture of standard history and science fiction, of sober analysis and mad prophecy, of nightmare and Utopia. With different backgrounds terrorists are prone to do different things. Osama bin Laden studied economics and used airplanes as weapons. He was in the construction business.

Should we be preparing for a conventional weaponized agent, or an exotic new threat such as mixes of spores, bacteria, toxins, and viruses; altered antigenicity agents for vaccine evasion; antibiotic resistant organisms; chimera or fusion agents? Probably not, but there is worrisome technology out there regarding altering genetics with technologies such as CRISPR. Gene drives and CRISPR technology have enormous implications for genetic manipulation. Cheap and simple CRISPR kits are now available on the Internet, allowing almost anyone to edit the genes of bacteria and viruses making them more virulent and/or transmissible. Eventually, there is the nightmarish prospect of engineered targeted diseases to targeted individuals. There's also the combination of a civilian, military, and/or cyber targeted attack in conjunction with a biological attack. Wirth this in mind, can we actually prepare for the 'unknown unknown'?

There are examples of genetic manipulation. The highly pathogenic H5N1 influenza virus kills most of the people it infects, and has artificially acquired the ability to transmit easily. This wasn't the naturally occurring strain, but one created in research laboratories in Holland and the US with the best of scientific intentions. Scientists had created a deadly strain of avian influenza virus that transmits easily in ferrets, which have a remarkably similar immune response to humans. Pressured not to publish the entire genetic blueprint, this heated up the debate about where you draw the publishing line. CDC and WHO committees disagreed about harm versus benefit of publishing this research. The argument was that while no one wants bioterrorists to get hold of such a recipe, if researchers cannot pool

their knowledge then we hamper their ability to give us the best chance of averting or surviving a flu pandemic with a such high kill rate. The results were eventually published.

Another example: Australian and American researchers inserted into a mousepox virus a gene that creates large quantities of interleukin-4. Shockingly, this totally suppressed the 'cell-mediated response'—the arm of the immune system that combats viral infections. The engineered virus also appeared unnaturally resistant to attempts to vaccinate the mice. Could superviruses created in laboratories inspire terrorists to create their own deadly diseases? Where is all this leading us? Both optimistic and pessimistic scenarios for the future are shown on Table 8.18.3.

Table 8.18.3 Optimism vs. pessimism

An optimistic scenario	A pessimistic scenario
• Defensive technology dominates • Violent extremism recedes • Weaponization becomes harder to achieve • Antibiotic resistance recedes • We don't let our guard down psychologically • US appropriations and global health pandemic preparedness spending and interest increases	• Offensive technology dominates • Violent extremism expands as social media and slick advertising keeps pushing people to extremes • Weaponization technology gets easier to access • Antibiotic resistance grows as we run out of effective antibiotics • We don't remain vigilant • Appropriations diminish for both global and domestic health pandemic preparedness and management

Walter Laquer wrote in 'No End to War' that a disproportionate number of doctors are terrorist leaders. Ayman al-Zawahiri is the leader of al Qaeda, George Habash founded the Popular Front for the Liberation of Palestine; Dr Fathi Shikaki was a founder of Palestinian Islamic Jihad, Dr Kefah Basheer Hussein was both an ISIS Minister and a rheumatologist. Seven of the eight suspects in an attempted car bombing in London and Glasgow were doctors. With so many doctors involved in terrorist leadership, could use of a biological agent be more appealing?

Finding and stopping disease outbreaks at the earliest possible moment no matter where they emerge is important to reduce illness and death, increase

national security, and maintain the economic gains made over the previous decades. Disease threats, after all, require only the smallest opening to take root and spread. In today's tightly connected world a disease can be transported from an isolated, rural village to any major city in as little as 36 hours.

How much damage can a weaponized biological agent actually do? Can it wipe out humanity or wipe out much of our infrastructure? Can it cause massive direct and indirect damage to civil society with rampant disease, civil unrest, anarchy, and a greater need to trade freedom for security? Is biological terrorism an existential risk, a catastrophic threat, a weapon of mass disruption, or a WMD? Is it a 'black swan' low-probability high-consequence event, or merely a weapon that temporarily overwhelms local and regional resources and responses, but then quickly recovers? It's hard to know since this has never happened on a large enough scale in modern history.

In today's complex world we need to minimize biological risk. There needs to be continued vigilant multisource pre-event intelligence on terrorist countries, groups, and individuals that have potential biocapabilities. Effective disease recognition surveillance systems, and use of big data and artificial intelligence are needed. Intelligence driven analytical threat mitigation, as well as good predictive modelling is essential.

We need to keep countermeasures such as vaccines, antitoxins, and antibiotics in the pipeline. Development of better rapid on the scene (point of care) diagnostics needs to be encouraged. Continued multidisciplinary exercises, training, cooperation, and communication are important. We need more effective well thought out laws and treaties with appropriations that back them up. Public–private partnerships in biomedical and interdisciplinary research and collaboration should be encouraged. In the event of an attack, public health needs to work well with clinical medicine, public safety, and disaster management. Programmes like SNS, BARDA, and BioWatch need to continue.

Expert opinions differ on the plausibility or risk of a large-scale bioattack. The US Office of the Director of National Intelligence stated in 2008 that bioterrorism is a more likely threat than nuclear terrorism. That same year, Director Mike McConnell stated that of all WMD, biological weapons were his personal greatest worry. Other defence experts and scientists insist that the possibility of any attack, especially a large-scale one, is small, given the

immense challenges to obtaining, weaponizing, and deploying biological agents. Regardless, most biosecurity experts acknowledge that the potential of an attack should not be ignored. Moreover, specific preparations will likely benefit the response to other types of public health emergencies.

REFERENCES

Alibek, K. and Handelman, S. (2000). *Biohazard*. New York: Dell Pub.

Bacon, J. (2017, 28 December). Anthrax antibodies in defector raises fears North Korea is developing chemical weapons. [Online] Available at: https://www.usatoday.com/story/news/world/2017/12/27/anthrax-defector-raises-new-fears-north-korea-developing-biological-weapons/983810001/

Barlas, S. (2007). CDC's BioSense Network identifying possible bioterrorism, but skeptics abound. *Emergency Medicine News*, 29(9). doi:10.1097/01.eem.0000296435.09253.ff

Baumgaertner, E. and Broad, W. (2019, 15 January). North Korea's less-known military threat: biological weapons. [Online] Available at: https://www.nytimes.com/2019/01/15/science/north-korea-biological-weapons.html

Biodefense for the 21st Century (2004, 28 April). [Online] Available at: https://fas.org/irp/offdocs/nspd/hspd-10.html

Biological Warfare (2007, 23 April). [Online] Available at: https://www.cia.gov/library/reports/general-reports-1/iraq_wmd_2004/chap6.html

Biological Warfare (2019, 22 January). [Online] Available at: https://en.wikipedia.org/wiki/Biological_warfare

Biological Weapons, Bioterrorism, and Vaccines (2018). [Online] Available at: https://www.historyofvaccines.org/content/articles/biological-weapons-bioterrorism-and-vaccines

Bioterrorism. (n.d.). [Online] Available at: https://www.ready.gov/Bioterrorism

Broad, W. (2003, November 01). Bioterror researchers build a more lethal mousepox. [Online] Available at: http://online.sfsu.edu/rone/GEessays/BioterrorMousepox.htm

Charles, D. (2018, November 07). Is the Pentagon modifying viruses to save crops—or to wage biological warfare? [Online] Available at: https://www.npr.org/sections/thesalt/2018/11/07/658471432/is-the-pentagon-modifying-viruses-to-save-crops-or-to-wage-biological-warfare

Chemical and Biological Defense (2018, October 10). [Online] Available at: https://www.dhs.gov/science-and-technology/chemical-and-biological-defense

Derbes, V. J. (1966). De Mussis and the Great Plague of 1348. *Journal of the American Medical Association*, 196(1). doi:10.1001/jama.1966.03100140113030

Emergency Preparedness and Response (2018, 26 October). [Online] Available at: https://emergency.cdc.gov/bioterrorism/index.asp

Forgotten Horrors: The Human Experiments of Unit 731 (2013). [Online] Available at: https://knowledgenuts.com/2013/07/23/forgotten-horrors-the-human-experiments-of-unit-731/

Gates, B. (2017, 18 February). Bill Gates: a new kind of terrorism could wipe out 30 million people in less than a year. [Online] Available at: https://www.businessinsider.com/bill-gates-op-ed-bio-terrorism-epidemic-world-threat-2017-2

Green, M.S., Leduc, J., Cohen, D., and Franz, D.R. (2019). Confronting the threat of bioterrorism: realities, challenges, and defensive strategies. *The Lancet Infectious Diseases*, 19(1), e2–e13.

Holm, S. (2017). The bioethicist who cried 'synthetic biology': an analysis of the function of bioterrorism predictions in bioethics. *Cambridge Quarterly of Healthcare Ethics*, 26(02), 230–8.

Hooker, E. (n.d.). Biological warfare facts & history of biological agents. [Online] Available at: https://www.emedicinehealth.com/biological_warfare/article_em.htm

Katona, P. (2018, 12 March). Biological terrorism: an existential threat or merely a weapon of mass disruption? [Online] Available at: https://www.risksciences.ucla.edu/videos/2018/3/12/biological-terrorism-an-existential-threat-or-merely-a-weapon-of-mass-disruption

Katona, P. (2012). Botulinum toxin: therapeutic agent to cosmetic enhancement to lethal biothreat. *Anaerobe*, 18(2), 240–3.

Katona, P., Sullivan, J.P., and Intriligator, M.D. (2011). *Global Biosecurity: Threats and Responses*. London: Routledge.

LeClaire R.D. and Pitt M.L.M. (2005) Biological weapons defense. In: Lindler L.E., Lebeda F.J., and Korch G.W. (Eds.) *Biological Weapons Defense*. *Infectious Disease.* New York: Humana Press.

Martin, A. (2018, 28 March). DHS, EPA recap four-year project to improve response to bioterrorism subway attacks. [Online] Available at: https://homelandprepnews.com/stories/27566-dhs-epa-recap-four-year-project-improve-response-bioterrorism-subway-attacks/

National Biodefense Analysis and Countermeasures Center (2018, 12 October). [Online] Available at: https://www.dhs.gov/science-and-technology/national-biodefense-analysis-and-countermeasures-center#

Obama Administration Gets an 'F' for Bioterrorism Defense (2010, 26 January). [Online] Available at: https://www.nti.org/gsn/article/obama-administration-gets-an-f-for-bioterrorism-defense/

Paddock, C. (2011, 21 December). US govt asks scientists to keep lab-bred bird flu blueprint secret. [Online] Available at: http://www.medicalnewstoday.com/articles/239510.php

Pruitt, S. (2018, 4 October). When anthrax-laced letters terrorized the nation. [Online] Available at: https://www.history.com/news/anthrax-attacks-terrorism-letters

Redlener, I. (2006). *Americans at Risk: Why We Are Not Prepared for Megadisasters and What We Can Do Now*. New York: Knopf.

Reeves, R.G., Voeneky, S., Caetano-Anollés, D., Beck, F., and Boëte, C. (2018). Agricultural research, or a new bioweapon system? *Science*, 362(6410), 35–7.

Schnirring, L. (2018, 18 September). White House shines high-level spotlight on biodefense. [Online] Available at: http://www.cidrap.umn.edu/news-perspective/2018/09/white-house-shines-high-level-spotlight-biodefense

Sibley, R. (2007). When healers become killers: the doctor as terrorist. *Canadian Medical Association Journal*, 177(6), 688.

Sullivan, P. (2018, 18 September). Trump directs government to step up defenses against bioterrorism. [Online] Available at: https://thehill.com/policy/healthcare/407244-trump-directs-government-to-step-up-defenses-against-bioterrorism

U.S. Government's Biological Warfare Response Team May be Leaking its Secrets (2018, 18 December). [Online] Available at: https://sputniknews.com/us/201812181070775373-US-Biological-Warfare-Leaking-Secrets/

U.S. Office of the Press Secretary (Producer) (2012, 31 July). *National Strategy for Biosurveillance.* [Online] Available at: https://obamawhitehouse.archives.gov/the-press-office/2012/07/31/national-strategy-biosurveillance

Wampler, R.A. and Benton T.S. (eds.) (2001). The September 11th sourcebooks, volume v: anthrax at Sverdlovsk, 1979. [Online] Available at: https://nsarchive2.gwu.edu/NSAEBB/NSAEBB61/

8.19

Genetic epidemiology

Elizabeth H. Young and Manjinder S. Sandhu

Introduction to genetic epidemiology

What is genetic epidemiology?

Classical epidemiology asks whether, across a study population, a measured exposure is consistently associated with an observed disease or trait. Similarly, *genetic* epidemiology in its simplest paradigm asks whether a *genetic* exposure is associated with a disease or health outcome in a given population. To expand on this, genetic epidemiology can be defined as population-based research that focuses on the aetiological role of genes in the occurrence of disease, in order to provide insights into disease biology and mechanism by unpicking the causal pathway from genes to disease.

Over the last decade there has been an extraordinary pace of change in genetic epidemiology driven by rapid advances in technology. These changes have revolutionized the scope and scale of what can be done, ranging from studies of single gene variants measured in small samples through to the emergence of high-throughput genotyping, genome-wide association studies (GWAS), and next-generation whole genome sequencing technologies which produce information on millions of genetic variants in thousands of individuals. In parallel, as these genomic technologies provide new opportunities to better understand disease and develop targeted therapeutics for patients, it is becoming increasingly apparent that a knowledge of genomic medicine will be essential for public health practitioners in meeting the public's future healthcare needs.

Genetic epidemiology and public health practice

Ultimately, the goal of genetic epidemiology lies in the application of research findings to clinical care and improved health for the population. The discovery and confirmation of genetic variants associated with diseases and traits may have translational implications—such as advances in treatment and genetic risk prediction. First, the identification of genetic determinants of traits, together with a better understanding of the underlying physiology and regulation, should help inform drug development and improve therapeutic effectiveness. Secondly, in relation to risk prediction, as more disease loci are identified it may be useful to assess whether genetic testing could improve risk prediction of disease beyond traditional risk factors. Additionally, use of genetic data makes risk stratification possible at a young age, when signs and symptoms of disease are not yet present—the clinical and ethical implications of this must be carefully considered. Thus, in this twenty-first-century era of genetic epidemiology and genomic science, public health practitioners cannot overlook the impact of genetic epidemiology on our understanding of models of disease, individualized medicine, and the relationship between individuals and populations, treatment, and prevention strategies.

Heredity, genetics, and disease

DNA, RNA, and proteins: the structure and function of genes

The human genome carries the chemical information that allows the exact transmission of genetic information from one cell to its daughter cells and from one generation to the next. Our genome is made up of deoxyribonucleic acid (DNA), which consists of a long sequence of nucleotide bases of four types: adenosine (A), cytosine (C), guanine (G), and thymine (T). Strong covalent bonds bind the bases together along a single strand, and weaker hydrogen bonds pair A with T, and C with G, between the two strands in their double-helix form. The double-stranded DNA acts as a template for its own replication: separation of the two strands, and construction of a new complementary strand for each, results in two identical copies of the original.

DNA also acts as a template for its transcription and translation into proteins. The precise order (the ‘sequence’) in which nucleotide bases are arranged along the DNA strand determines the amino acid sequences of the

proteins. First, a single strand of DNA can act as a template for a complementary strand of ribonucleic acid (RNA), through the process of *transcription*. Importantly, in certain regions of DNA, which can be called genes, the transcribed RNA encodes instructions that tell the cell how to assemble amino acids to make proteins through a process known as *translation*. Thus, just as DNA directs the sequence of RNA, so RNA directs the sequence of amino acids and proteins. Only 1–2% of these RNA transcripts encode for protein; the vast majority are transcribed as non-coding RNA which has important biological functions that influence gene expression through a variety of mechanisms (Amaral and Mattick 2008; Kaikkonen et al. 2011).

At its simplest, a *gene* can be thought of as a length of DNA containing the code for the amino acid sequence of a protein. However, few genes exist as continuous coding sequences; rather, most genes contain one or more non-coding regions, called introns, that interrupt the coding regions, or exons. It was previously assumed that non-coding regions of DNA had no biological function. However, we now know that many types of non-coding DNA sequences may have active roles in the regulation of protein-coding sequences and gene modulation (The ENCODE Project Consortium 2011).

The complete human DNA sequence is the human genome, and contains around 3 billion DNA base pairs. The human genome is distributed among 46 chromosomes, composed of 23 pairs: 22 homologous pairs of autosomes, and one pair of sex chromosomes (X and Y). In each pair of autosomes, one chromosome is inherited from the mother and the other from the father. Chromosomal DNA will contain many genes, with DNA segments of varying length between them. A given gene may exist in alternative forms, differing in DNA sequence, known as *alleles*. The *genotype* of an individual refers to the combination of two alleles inherited (one inherited maternally, the other paternally) at a specific locus—for example, AA, AG, TC, GG, etc. If the alleles are the same, the genotype is homozygous; if different, heterozygous.

Generating genetic diversity

The DNA of the human genome is subject to a variety of different types of genetic variation, defined as any change in the nucleotide sequence or arrangement of DNA. Some genetic variants may have no phenotypic effect

and, therefore, no clinical consequences. Other variants will profoundly alter disease susceptibility or, in extreme instances, be incompatible with life.

Genetic variants can be inherited from a parent, or acquired during a person's lifetime (somatic). They may be caused by radiation, viruses, and mutagenic chemicals, as well as errors that occur during meiosis (cell division) or DNA replication, or induced by cellular processes within an organism. Globally, human populations show structured genetic diversity as a result of geographical dispersion, selection and drift (Campbell and Tishkoff 2008; McCarthy et al. 2008; Novembre and Ramachandran 2011). Understanding this genomic variation can provide insights into our human origins and the evolutionary processes that shape variation in phenotypic traits and disease risk.

In the context of genetic epidemiology, we are concerned with understanding how disease susceptibility may be conferred by different kinds of genetic variation. Within an individual, genetic variation can occur at three levels:

Genome variation: changes that affect the number of intact chromosomes in the cell.
Chromosome variation: changes that alter the structure of individual chromosomes.
Genetic variation: changes that alter the sequence of nucleotide bases in a length of DNA.

Genetic variants

Gene variants are alterations in DNA sequence. Different versions of a particular DNA sequence at one specific chromosomal location (locus) are called alleles. When a particular allele is rare—conventionally, with a frequency of less than 1% in the general population—it is known as a *rare variant* or *mutation*. By contrast, where an allele is found in more than 1% of the population it is called a *genetic polymorphism* (literally, 'many forms') or *common variant*. There are many different types of DNA sequence variants, that can range from a change in a single nucleotide base to a change in many thousands of base pairs (Bochud 2012).

The most studied class of variant is the *single nucleotide polymorphism* (SNP, pronounced 'snip'; also known as a point mutation), where the DNA sequence varies in a single base—for example, A is present instead of T, C is

present instead of G. SNPs are the most common type of variant; as of March 2018, there are at least 113 million known SNPs in the human genome (NCBI dbSNP Build 151). SNPs lying in protein-coding regions of the gene can alter the code for an amino acid sequence and may (or may not) in turn affect the function or availability of a protein and cause disease. Non-coding SNPs, occurring in intronic (non-coding sequence of a gene) or intergenic (sequence of DNA between genes) regions of the genome can also cause disease, by affecting gene transcription and function. In addition to SNPs, there are also other types of structural variations in the human genome such as variable number tandem repeats (VNTRs), insertions and deletions of DNA bases, and copy number variants (CNVs), which have been used successfully to identify disease genes (Bochud 2012).

Patterns of inheritance, genotypes, and phenotypes

An individual's *genotype* can be understood as the combination of alleles present at a specific position (known as a locus) occupied by a gene on a chromosome. A *phenotype* is an observable, measurable characteristic or trait (biochemical, physiological, morphological, or behavioural) of an individual, other than its genotype. A phenotype could be a quantitative characteristic (such as systolic blood pressure) or a binary disease state (presence or absence of hypertension).

Many variations at the genotype level result in little or no observable variation in phenotype. If, however, a phenotype is present in individuals with just one copy of an abnormal allele, the allele is dominant. By contrast, if the phenotype is only expressed in the presence of two abnormal alleles, the allele is said to be recessive.

Reduced penetrance and variable expressivity are additional factors that influence the phenotypic expression of a particular genetic change. The *penetrance*, or positive predictive value, of a given genotype is the percentage of people with a disease-associated genotype who develop the disease. In cases where all individuals with a particular genotype develop the disease, the gene shows 100% penetrance. In contrast to this, some phenotypes are not expressed in all genotypically predisposed individuals; here, the gene is said to show *reduced penetrance*.

The concept of penetrance is often confused with *variable expressivity*, which is the variation in the severity of expression of a phenotype. Although

some genetic disorders exhibit little variation, most have signs and symptoms that differ among affected individuals, in some instances ranging from mild symptoms through to life-threatening complications. Variable expressivity refers to the range of severity of clinical features that occur in different people who have the same genotype.

Single-gene disorders

Single-gene, or monogenic, disorders run in families and are caused by a single mutated gene. Single-gene disorders are often called ‘Mendelian’ because they display obvious and characteristic patterns of inheritance as described by the nineteenth-century scientist Gregor Mendel, in his study of how traits were passed from one generation of garden peas to the next. Mendelian disorders are usually characterized by extreme phenotypes in rare patient subgroups that are linked to causal genes with rare mutations. These disorders, while individually rare and not contributing substantially to the global burden of disease, might collectively cause significant disease and death in a population. They can also provide invaluable insights into the underlying biology of disease and help to inform the search for susceptibility genes for more common diseases and traits.

Complex multifactorial diseases

Familial aggregation

In contrast to monogenic disorders, genetic disorders may also be complex (also known as multifactorial, or polygenic or non-Mendelian), meaning they are likely to be associated with the effects of multiple genes in combination with lifestyle and environmental factors, with no clear mode of inheritance. Multifactorial disorders include many common diseases such as cancers, heart disease, and diabetes, and quantitative traits such as height, weight, blood pressure, and blood chemistries. Patterns of inheritance for complex disorders do not fit simple patterns as with Mendelian diseases. This makes it difficult to determine a person’s risk of inheriting or passing on these disorders. Complex disorders are also difficult to study and treat because the specific factors that cause most of these disorders have not yet been identified.

However, this does not mean that the genes cannot eventually be located and studied. Because relatives share a greater proportion of their genes with one another than with unrelated individuals in the population, a primary feature of diseases or traits with a genetic basis is that affected individuals tend to cluster in families. Of course, familial aggregation does not necessarily mean there must be a genetic contribution; family members may develop a disease or trait simply by chance alone, particularly if it is common in the general population, or as a result of a shared environmental exposure. Given this, a helpful measure of familial aggregation is the *recurrence risk ratio* (λ_r), which takes into account the population prevalence of a disease. λ_r estimates the ratio of the risk of developing disease when a relative already has the disease compared with the risk in the general population (i.e. background prevalence).

The correlation of a trait or disease among relatives provides another useful estimate of genetic influence, on the assumption that the degree of similarity in the values of trait (e.g. body mass index (BMI)) measured among relatives is proportional to the number of alleles they share at the loci for that trait. Based on this premise, the more closely related the individuals are in a family the more likely they are to share alleles at loci and—in this instance—the more strongly correlated will be their BMI values. Thus a tendency for BMI to be more similar among relatives than among the general population is further evidence for a genetic aetiology. Familial correlations can also be used to calculate the *heritability* (h_2) of a trait, a measure that quantifies the percentage of phenotypic variation in a trait that is attributable specifically to genotypic differences. The higher the heritability, the greater the contribution of genetic variation among people in causing the variability of a trait for a given population.

Twin studies

To help disentangle the relative contributions of genes and environment in the aetiology of disease, twin and adoption studies are used (Boomsma et al. 2002). Comparison of disease concordance rates in identical (monozygotic (MZ)) and non-identical (dizygotic (DZ)) twins helps to assess the genetic influences on a disease. MZ twins are genetically identical, sharing the same alleles at every locus. DZ twins, on the other hand, share on average half the alleles at all loci, like all siblings. MZ twins provide an opportunity to study

disease occurrence when genetic factors are held constant, by measuring concordance in relatives with identical genotypes who may or may not be reared together in the same environment. Concordance in MZ twins in early life can be the result of genes or a shared environment. However, in later life most twins live apart and concordance at that stage may suggest a genetic component to the disease. Concordance less than 100% in MZ twins is strong evidence that non-genetic factors are involved in disease aetiology.

DZ twins reared together provide an opportunity to study disease occurrence when genetic differences are present but environmental factors are shared. Greater concordance in MZ versus DZ twins is strong evidence of a genetic component to the disease. Comparison of concordance rates in MZ versus DZ twins can be used to estimate the heritability of a disease.

As with family studies, twin studies are not without their limitations. First, despite being referred to as ‘identical’, MZ twins do not have identical gene expression of their identical genotypes; for example, somatic rearrangements will differ between MZ twins, which may give rise to phenotypic discordance and differences in disease susceptibility. Secondly, environmental exposures may not be the same for twins; even the intrauterine environment may differ within a twin pair (e.g. differential placental blood supply in MZ twins). Furthermore, studies comparing MZ to DZ twins are based on the assumption that they differ only by the number of genes they share and that within each pair the exposure to environment factors is identical; yet it is likely that MZ twins share more of their environment than DZ twins (particularly compared with opposite-sex DZ twins). Ascertainment bias poses an additional limitation, where twins who are very alike will tend to participate in studies more than twins who are very discordant for a phenotype. Studies of twins reared apart are easier to conduct than adoption studies, and similarly are an effective way of investigating the influence of shared genes and different environments. However, even with MZ twins separated at birth, intrauterine environmental influences cannot be distinguished from genetic influences.

Adoption studies

Adoption studies provide another valuable source of information about the genetic component of a disease. In this study design, a disease or trait is identified in adoptees and then studied in both their adoptive family (shared environmental factors only) and biological family (shared genetic factors

only). While undoubtedly providing the best way of separating out genetic and environmental influences, adoption studies carry obvious difficulties: information may be lacking about the biological family, compounded by their wish not to be identified and approached for details. In addition, relatively small sample sizes for which both adoptive and biological parents are available have limited the use of these studies.

The results of family studies must be interpreted with care. Importantly, familial aggregation of a continuous trait or discrete disease—however measured—does not indicate that genes are wholly responsible. Families share many factors besides their genes, such as common environmental exposures, cultural attitudes, and behaviours. This means that family studies are likely to overestimate the genetic component of a disease, and the strong environmental component to many complex, multifactorial diseases must not be overlooked.

Role of the environment and gene–environment interactions

Nearly all human diseases originate in large part from a complex interplay between genetic susceptibility factors and modifiable environmental exposures (including chemical, physical, infectious, nutritional, and behavioural factors). Therefore, in understanding the aetiology of disease and assessing the role of susceptibility alleles in disease risk, it is important to examine genetic factors *and* environmental factors and—importantly—the relation between the two: so-called gene–environment interactions.

Studies of gene–environment interactions present a number of methodological challenges (Hunter 2005). The assessment of the relation between genes and environment requires information on both components. Gene–environment interaction may be measured by the different effects of an environmental exposure on disease risk among individuals with different genotypes, or by the different effects of a genotype on disease risk among individuals with different environmental exposures (Yang and Khoury 1997). However, the power of studies to detect statistically relevant interactions is highly dependent on the precision with which genetic and non-genetic exposures are measured. As discussed elsewhere in this chapter, with the advent of next-generation genotyping and sequencing technologies, it is now possible and affordable to generate massive genomic datasets. By contrast, a challenge still remains to achieve accurate and objective measures

of environmental exposures such as dietary intake or physical activity that are feasible to collect in a large-scale epidemiological context.

In addition, the need for suitably large sample sizes is a major obstacle to overcome in the design of interaction studies (Hunter 2005). The sample size required to detect the joint effect of two variables is much larger than the size required to evaluate the main effect of each variable; given that many epidemiological studies are underpowered to detect main effects, these studies will be substantially underpowered to detect interactions. Careful consideration of study design, and the accurate and detailed assessment of genetic factors, environmental exposures, and disease phenotypes, is therefore crucial for planning interaction studies that will yield meaningful results.

Determining the details of gene–environment interactions will help us to understand the aetiology of disease and might lead to opportunities for targeted prevention strategies for some high-risk individuals (Hunter 2005; Khoury et al. 2005). Indeed, there may be significant public health benefits in using genetic information to improve existing approaches to identify and modify environmental risk factors, including stratifying the allocation of environmental interventions that prevent disease (Khoury et al. 2005). It may be, for example, that individuals with a certain genotype are particularly susceptible to the negative consequences of specific lifestyle behaviours, and that they would have most to gain from a targeted preventive intervention programme. Understanding how to detect these individuals and which environmental factors a programme should attempt to manipulate is a major goal of studies that attempt to unravel gene–environment interactions.

Finding disease genes: methods in genetic epidemiology

From monogenic to multifactorial disease

A number of strategies have been used to try and uncover genes involved in the aetiology of common, multifactorial disease (Table 8.19.1). Non-Mendelian disorders may depend on the interplay of common variants—polymorphisms—in several susceptibility genes together with a variable contribution from the environment. These genetic variants are presumed to have only modest effects accounting for a small percentage of variation in a trait, at least when examined individually; however, they may play a more

important role through interaction with other gene variants and be responsible for significant population attributable risk for disease.

Table 8.19.1 Common strategies for finding disease genes

Study design	Participants	Strengths	Limitations
Linkage studies	Relatives: large pedigrees, affected sib-pairs, or discordant sib-pairs	Can study multiple genetic markers simultaneously Good for rare traits	Only defines a locus or region, not specific genes
Family-based association studies	Parent–child trios, siblings	Less prone to population stratification Provides a rich context for evaluating shared genetic and environmental influences	Difficult to separate genetic from environmental influences Challenging for outcomes of older age: finding families with more than one affected generation will be difficult Findings may not be applicable to general population
Candidate gene association studies	Unrelated individuals	Powerful study design for common traits	Led by the state of current knowledge: success depends on selecting the right candidate gene
Genome-wide association studies	Unrelated individuals	Genotyping platforms developed to simultaneously examine hundreds of thousands of genetic variants	Lots of false positives; association does not equate to causation Replication of initial results is a challenge
Sequencing studies	Unrelated individuals	Good for identifying rare variants	Limited by costs and existing technological capabilities

Linkage studies, fine mapping through mutation screening, and case–control and cohort-based association studies of patients—combined with data from animal models—have uncovered the genes causing many Mendelian, or monogenic, disorders. In turn, a knowledge of the genetic basis of these Mendelian disorders has helped further our understanding of human physiology and played a vital part in establishing the causal role for disease risk factors and in the development of drug targets. In recent years

there has been considerable progress in mapping—that is, identifying and localizing—loci that influence susceptibility to common human diseases.

Genetic linkage scans

Linkage analysis is often the first stage in the genetic investigation of a trait, since it can be used to identify broad chromosomal regions of the genome (as opposed to specific alleles) that might contain a disease gene (Teare and Barrett 2005; Altshuler et al. 2008). Genetic linkage is the tendency of genes that are located close to each other on a chromosome to be transmitted together to offspring. Genes that are positioned nearer to each other are less likely to be separated onto different parts of the chromosome during *genetic recombination* (the breakage and rejoining of DNA strands to produce new combinations of alleles, encoding a novel set of genetic information) and are therefore said to be genetically linked. On this basis, susceptibility or causal genes can be localized by searching for genetic markers that co-segregate with the disease trait; if an allele of gene A causes disease, but is otherwise undetectable, whereas the allele of gene B can be easily detected and distinguished, gene B can be used as a marker of the inherited disease. Linkage studies use DNA from family members to identify areas of the genome that show evidence of containing the unknown disease-related genetic variant on the basis of correlation with DNA variation. Polymorphic markers positioned across the genome are genotyped in families, followed by calculating the degree of linkage of the marker to a disease trait. An advantage of linkage analysis is that this approach attempts to map genes purely by position and does not presume any a priori hypotheses about their location, biological function, or number.

Linkage analyses have proved highly successful for finding genes causing Mendelian disorders. However, linkage studies only work well if the phenotype of interest is caused by variation in the same gene region, and have generally been unsuccessful at identifying genes for common disorders where multiple genes are implicated (Hirschhorn and Gajdos 2011). Linkage analyses can only identify large regions and, while a very strong candidate gene might exist within the linkage region, such regions may contain hundreds of genes, many of which are biologically plausible candidates. To refine any association may require extensive genotyping—both time-

consuming and expensive—and potentially produce many false positive results.

Association studies

Genetic association studies provide an alternative approach to finding disease genes, which overcome some of the limitations of linkage studies. Genetic association studies aim to detect an association between an allele and trait (unlike linkage studies, which focus on a genetic locus as the unit of analysis regardless of the alleles present at that locus). Association differs from linkage in that the same allele (or alleles) is associated with the trait in a similar manner across the whole population, while linkage allows different alleles to be associated with the trait in different families. Whereas linkage studies can only be conducted using family data, association analyses can be done using unrelated and/or related people. Association studies can be conducted at the level of a candidate gene (candidate gene association study), or at the level of the entire genome (GWAS).

Candidate gene association studies

The candidate gene approach has many parallels with identifying and ranking risk factors in an epidemiological study (Tabor et al. 2002): from a large number of potential factors, researchers must pick those factors that are most likely to be associated with the outcome of interest. The selection of candidate genes focuses the search onto specific susceptibility genes that are identified a priori on the basis of their presumed biological relevance to the disease or trait of interest. Once a candidate has been selected, common variants—usually SNPs—in the gene are then identified (either *de novo* from sequencing/mutation screening, or based on previously published work). The SNPs are then genotyped and tested for association with the outcome of interest. In the case of a binary outcome, SNPs will be genotyped in cases and controls to assess whether the variants are associated with the phenotype of interest; where the outcome of interest is a quantitative trait, SNPs are typed in population cohorts. A strength of this gene-centric method is that it is a highly efficient way of identifying susceptibility genes if biological plausibility is high. Conversely, a drawback of the method is that it is, by definition, limited by the state of current knowledge. Thus countless genes may be overlooked because biological candidacy has not yet been

established. Issues of multiple testing and false positives remain limitations of the candidate gene approach.

Genome-wide association studies

GWAS have recently emerged as a powerful new tool for identifying the genetic variants related to common, complex diseases by effectively combining the association and linkage approaches to gene identification (McCarthy et al. 2008; Hirschhorn and Gajdos 2011). The basic paradigm of GWAS is to catalogue thousands of very common SNP markers located throughout the human genome, and then to genotype them in large numbers of unrelated humans who have already been carefully characterized for a wide variety of common diseases and traits. The GWAS approach affords the possibility of exploring the whole genome without any a priori knowledge of underlying disease biology. This allows for hypothesis-free testing of a substantial fraction of common genetic variation for a role in determining disease risk, and potentially enables the discovery of unexpected genetic factors or pathways for common diseases.

A central factor in the development of GWAS is having a comprehensive map of carefully selected SNPs (Box 8.19.1). Over the last two decades, a number of international efforts have successfully done this, creating denser and denser genetic maps. As these have emerged, it has become apparent that many SNPs are strongly correlated (Palmer and Cardon 2005). Consequently, if the pattern of allelic association can be described, direct assay of all existing polymorphisms is unnecessary, because genotypes of common SNPs may be inferred from knowledge of only a few representative SNPs. This is the principle underlying the tagging-SNP method: that it is possible to select the maximally informative set of polymorphisms (tagSNPs) to genotype in an association study such that all known SNPs are either directly genotyped or exceed a threshold level of association with a tagSNP. Given the practical limitations on genotyping, selection of tagging SNPs can usefully decrease the amount of genotyping required and thus improve the cost-effectiveness of studies. Inevitably, there is a trade-off between this improved efficiency and the loss of power that occurs when SNPs are discarded. However, by specifying various parameters (e.g. the level of allelic correlation, and the likelihood of genetic linkage) it is possible to try and maximize efficiency while maintaining statistical power.

Box 8.19.1 International scientific efforts to describe common human genetic diversity

The Human Genome Project, the SNP Consortium, the International HapMap Project, and the 1000 Genomes Project have collectively identified around 10 million common genetic variants in humans. Knowledge of these SNPs and their patterns of correlation have enabled GWAS which have successfully identified hundreds of novel disease genes (Donnelly 2008; International HapMap 3 Consortium 2010).

Completed in 2003, the Human Genome Project was an international collaborative research programme whose goal was to discover all the estimated 20,000–25,000 human genes and map the complete sequence of the 3 billion DNA bases (International Human Genome Mapping Consortium 2001; International Human Genome Sequencing Consortium 2001; International SNP Map Working Group 2001; Roberts et al. 2001; Collins et al. 2003). The official Human Genome Project was meant to have been a 15-year project spanning 1990–2005, but progress was faster than expected. Conducted on an industrial scale at a few large genome centres across the world, genetic maps were developed ahead of the original schedule, and the final stage of large-scale DNA sequencing was facilitated by developments in automated high-throughput sequencing techniques. The need to manage and store the huge amounts of sequencing data that were being produced led to the development of large electronic databases. Importantly, a key aim of this endeavour was to make these genomic data publicly available and accessible to the scientific community at large.

The SNP Consortium was established in 1999 as public–private partnership of several companies and institutions to produce a public resource of SNPs in the human genome (Kruglyak 2008). The initial goal was to discover 300,000 SNPs in 2 years, but the final results exceeded this, as 1.4 million SNPs had been released into the public domain at the end of 2001.

The SNP Consortium subsequently joined forces with the International HapMap Project. Launched in 2002, the HapMap Project is a global consortium of scientists and researchers with the aim of providing a public resource to inform and accelerate genetic medical research. Specifically, the project set out to describe the common patterns of human DNA sequence variation across the human genome in order to find genes affecting health, disease, and responses to drugs and environmental factors. Whereas the SNP Consortium was concerned with individual SNPs, the HapMap Project focused on examining combinations of SNPs that are inherited together, known as haplotypes. By 2007, the project had genotyped more than 3 million SNPs which captured the majority of all common variants with frequencies above 5% (International HapMap Consortium 2007). In its current phase, ‘HapMap 3’, the project is expanding its scope to identify rarer genetic variants that may be implicated in human disease and achieve a higher resolution map of human genetic variation (International HapMap 3 Consortium 2010).

The 1000 Genomes Project (1000 G) was launched in 2008 as an extension of the HapMap Project. Its aim is to create a public reference database containing an even more detailed characterization of the human genome by extending the catalogue of human genetic variants to include 95% of rarer variants, occurring at frequency of 1–5% (The 1000 Genomes Project Consortium 2010). As of November 2012, using a combination of whole genome sequencing and exome sequencing, the architecture and functional spectrum of human genetic variation has now been characterized in the genomes of 1,092 individuals from 14 populations (The 1000 Genomes Project Consortium 2012).

In 2007, the Wellcome Trust Case Control Consortium (WTCCC) set the benchmark for GWAS of common disease on an unparalleled scope and scale. Representing a UK-wide collaboration of more than 50 research groups and 200 scientists in the field of human genetics, the WTCCC genotyped over 500,000 SNPs in DNA samples from 17,000 people across the United Kingdom. Over 2 years, they analysed almost 10 billion pieces of genetic information, and substantially increased the number of genes known to play a role in the development of some of our most common diseases such as heart disease, hypertension, Crohn's disease, and diabetes. Many of these genes that have been found are in areas of the genome not previously thought to have been related to the diseases (Wellcome Trust Case Control Consortium 2007). More recently, advances from commercial companies such as Affymetrix and Illumina in the availability of dense genotyping chips with approximately 5 million SNPs are making even larger GWAS both cost-effective and feasible, and there has been considerable progress in the search for disease-related loci and genetic variants using the GWAS approach.

Searching for rare variants

GWAS have now been completed for many common human diseases and related traits, identifying a wealth of common gene variants. However, GWAS have a number of limitations: first, their scope is defined by the current knowledge of existing genetic variants that are included on a specific genotype chip. Second, most of the disease-associated SNPs found in GWAS have very small effect sizes, and explain only a modest fraction of the genetic component of human disease. This is to be expected if much of the missing heritability is due to gene variants that are less common or too rare to be detected by GWAS and have relatively modest effect sizes (Cirulli and Goldstein 2010). On this basis, gene discovery efforts have in recent years expanded to search for rare genetic variants.

To capture such rare variants, it will be necessary to use large samples sizes (thousands of individuals) and sequence entire genomes, rather than genotype a catalogue of variants. Until recently, this was technically unfeasible. However, the technological advances of the last couple of years (so-called next-generation sequencing technology) have made this possible (Metzker 2010; Zhang et al. 2011): now, billions of DNA bases can be

sequenced in a matter of days. The most complete study of rare disease variants will require *whole genome sequencing* in large sample sizes of human participants. Recent initiatives such as the 100,000 Genomes Project in the UK, which aims to sequence 100,000 genomes from around 70,000 individuals, show that with the rapid pace of technological development and the precipitant fall in the costs, such a shift to whole genome sequencing association studies on a massive scale is becoming more feasible.

In the meantime, *targeted sequencing* and *exome sequencing* offer more affordable strategies for finding rare genetic variants. Targeted sequencing is focused on genomic regions of specific interest—rather than the whole genome—and enables a systematic assessment of the full spectrum of genetic variation, including common and rare variants, for high-throughput sequencing at a lower cost per sample. Similarly, selectively sequencing only the exome constitutes a powerful and efficient alternative to whole genome sequencing (Bamshad et al. 2011). Approximately 85% of all disease-causing variants in humans can typically be found in the subset of the human genome that is protein coding—known as the exome. With the availability of next-generation sequencing technology, exome sequencing has become a cost-effective method for the identification of rare sequence variants (Choi et al. 2009; Ng et al. 2009; Pussegoda 2010). This strategy (also known as targeted exome capture) has been used successfully to identify disease-causing variants for a number of rare Mendelian disorders (Choi et al. 2009; Ng et al. 2010; Bamshad et al. 2011) and is an effective technique to detect causal variants associated with disease even when few individuals with disease are sequenced (Cirulli and Goldstein 2010). Although this method has been predominantly used in the study of Mendelian disorders, these studies provide the framework for applying this method to complex diseases (Cirulli and Goldstein 2010); for example, recent whole exome sequencing studies have identified genetic variants associated with low-density lipoprotein (LDL) cholesterol and risk of myocardial infarction and coronary heart disease (Bomba et al. 2017).

Generating genomic data: analytical issues

At present GWAS typically generate information on several million genetic variants, on many thousands of individuals. Many computational and analytical steps are required to translate massive raw genomic data into

usable, high-quality data, and may include techniques for genotype imputation. Imputation is the term used to describe the substitution of some value for missing data. In relation to GWAS data, genotype imputation is the process of predicting or imputing genotypes that are not directly assayed in a study sample of individuals, based on reference panels of SNPs (Li et al. 2009). These *in silico* genotypes can then be used to increase the total number of SNPs that can be analysed for association, thus improving the statistical power of the study. Handling such vast datasets, and identifying the few true association signals among the masses of false positive signals, requires extensive computing and storage infrastructure, and advances in computational resources and statistical techniques for processing, analysis, and interpretation (Hunter and Kraft 2007). In this context, cloud computing —with its advantages such as virtualized resources, parallel processing, and data service integration with scalable data storage—has become a powerful tool to perform massive-scale and complex computing with big data (Hashem et al. 2015; Langmead and Nellore 2018).

A key determinant of the quality in any association study—genetic or non-genetic—is its sample size (Ioannidis et al. 2003). This is particularly pertinent in genetic studies, which must be powered to detect variants that have only modest associations with the outcome of interest. Small initial studies rarely find the correct result and, even when they do, are likely to overestimate the true effect size (Ioannidis et al. 2003; Lohmueller et al. 2003; Wacholder et al. 2004). An implication of looking at modest effect sizes of genetic variants is the need for sufficiently large sample sizes to avoid underpowered studies with false negative results (Lohmueller et al. 2003; Ioannidis et al. 2006). In addition to ensuring adequate samples sizes, care must be taken to ensure optimal ascertainment of cases and controls, to improve study power—for example, by selecting cases with minimal phenotypic heterogeneity, or focusing on extreme phenotypes such as early-onset severe disease in order to enrich for disease-causing alleles (Hattersley and McCarthy 2005).

Multiple testing is also an issue for these large datasets: when testing so many SNP–disease associations simultaneously, the probability of a nominally significant result from within the entire experiment occurring by chance is higher than the individual *p*-value associated with that result, because of random variation in the data and the large number of tests that are performed (Thomas and Clayton 2004). In other words, there is an increased

chance of false positives, and one cannot be certain at the outset whether there are any true associations to be found at all. Consequently, the individual *p*-values of the tests may not be an appropriate guide to actual statistical significance.

There are several statistical techniques which may be used to correct for multiple testing, including a reduction in the number of tests, or the setting of a more stringent *p*-value, or permutation testing (Hirschhorn and Daly 2005). When adjusting for multiple testing, larger samples will be required so that if an initial association is statistically significant, it will remain significant after correction (Benjamini and Hochberg 1995; Hirschhorn and Daly 2005). Where GWAS findings indicate an association between a genetic variant and a trait or disease, validation and replication studies are essential to determine whether the result of the initial study represents a true reproducible association with genome-wide statistical significance (McCarthy et al. 2008; Ioannidis et al. 2009).

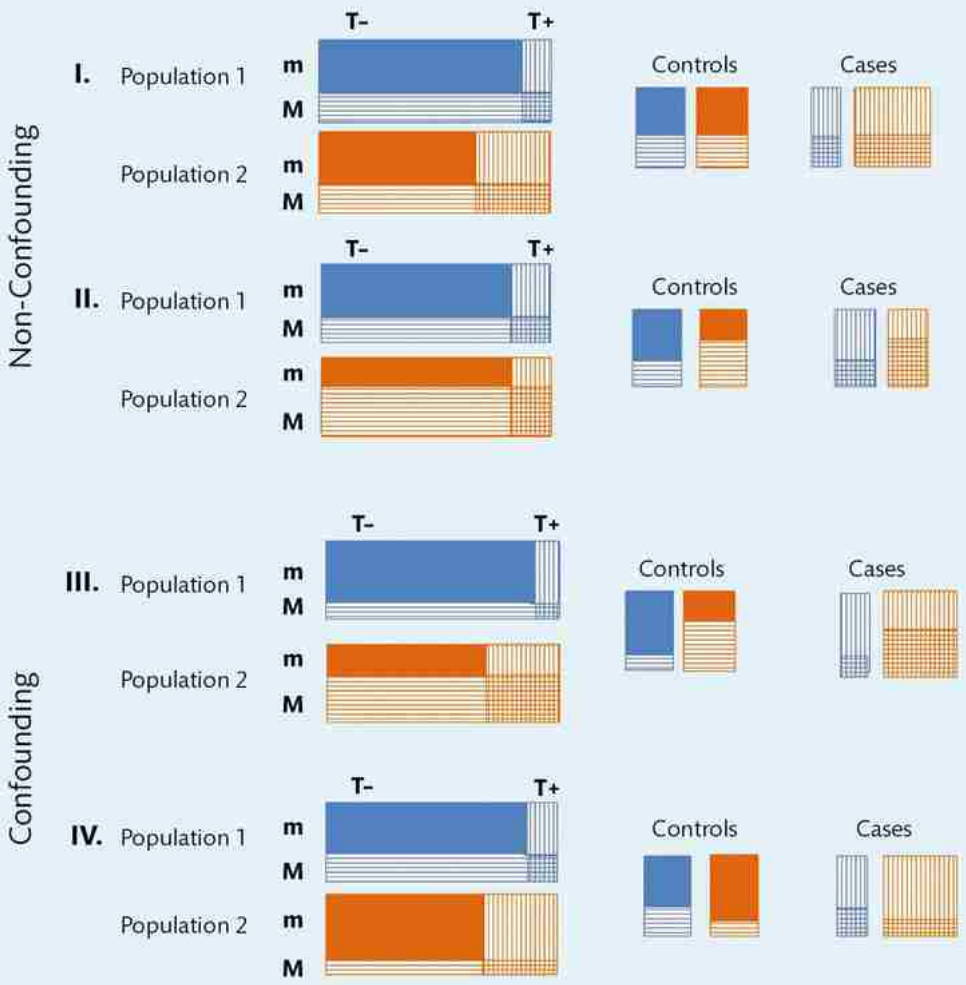
Causal inference

In observational epidemiological studies, causal inferences are potentially susceptible to problems of bias and confounding that, in addition to chance, provide alternative explanations for an observed association between an exposure and a disease. Likewise, the assessment of causality in gene–disease associations is vulnerable to possible other explanations (Geneletti et al. 2011). The role of *bias* must be considered in any observed gene–disease association. In contrast to studies of non-genetic exposures, which are potentially prone to recall bias, an individual's genotype is an exposure that is fixed at conception—thus the association between genetic variants and disease outcomes will not usually be biased in this way. However, a key potential bias in genetic association studies is the differential genotyping of cases and controls. Wherever possible this can be minimized by randomly allocating cases and controls in the genotyping process, so that systematic genotyping errors affecting solely cases or controls are unlikely to occur. Selection bias due to non-comparable cases and controls may also explain the observed associations. Bias, though, need not necessarily be a problem: it can also be used positively to enhance a study. For instance, one might deliberately select cases enriched for genetic variants (e.g. based on their

family history of a disease or trait), in order to improve the power of a study to detect a statistically significant association.

In the context of genetic association studies, classical *confounding* variables such as age and sex are not usually relevant. This is because in order to confound the association between genotype and phenotype, these characteristics would need to be associated with both the genotype and the specific outcome. In the case of genetic association studies this is not applicable, based on the assumption of random assortment—that is, during gamete formation alleles are allocated arbitrarily from parent to offspring independently of other traits. This results in population distributions of genetic variants that are generally independent of behavioural and environmental factors that typically confound epidemiological associations between risk factors and disease.

Nevertheless, we cannot ignore confounding as a result of *undetected population stratification* due to ethnic admixture, which can mimic the signal of association and give rise to false positives and produce associations of disease with genotype at loci that are not linked to any locus that affects disease susceptibility (Hoggart et al. 2003; Freedman et al. 2004; Marchini et al. 2004). Ethnic admixture refers to the presence of multiple subgroups within a larger population, caused by non-random mating, which have differing allele frequencies. This can have implications for studies of gene–disease association if the subgroups have different genotypic distributions (Fig. 8.19.1). If cases and controls are drawn differentially from two or more subpopulations in which marker allele frequencies *and* disease prevalence differ across subpopulations, this will result in an over-representation of one or more genetic subgroups among individuals chosen as disease cases in an association study. In this instance, any SNP with allele proportions that differ between the subgroup and the general population will be associated with disease status. Although the association is real from a statistical point of view, it is regarded as spurious since it does not yield any information about the biology of the phenotype or about the genomic location of the disease-modifying locus; rather, the association between the allele and the disease is confounded by the existence of genetic heterogeneous subpopulations (Wacholder et al. 2000; Ziv and Burchard 2003).



Rectangles on the left of each panel represent two subpopulations from which cases and controls are selected. The rectangles on the right represent the sample of cases and controls, sampled without regard to membership in genetic subpopulations. The horizontal lines represent the proportion of the population with the candidate allele, **M**. The vertical lines represent the proportion of the population with the trait of interest, **T+**. In each population on the left, there is no association between M and T+

I. The frequency of the trait is higher in population 2, but because the M allele frequency is the same in both populations, the frequency of M is the same in both cases and controls. Therefore, while cases are over-represented by population 2, there will be no spurious association as the trait is equally likely to be associated with the M allele in both populations.

II. The prevalence of the trait is the same in both populations, but the frequency of M is greater in population 2. Because the frequency of M is the same for both controls and cases (a control and a case are equally likely to have allele M), there is no spurious association.

III. Both the prevalence of the trait and the frequency of M is higher in population 2. This will result in a positive confounding association because as a larger number of cases will be sampled from population 2, cases will have a disproportionately high frequency of M compared with controls, resulting in a false-positive association between allele M and trait T+

IV. The trait prevalence is higher and the allele frequency of M is lower in population 2 compared with population 1. This will result in a negative confounding association because as a larger number of cases will be sampled from population 2, cases will have a disproportionately low frequency of M, resulting in a false-negative association between allele M and trait T+

Fig. 8.19.1 Spurious associations as result of population stratification.
Reproduced from Ziv E, and Burchard EG. (2003) Human population structure and genetic association studies, Pharmacogenomics, Volume 4, pp. 431–41, with permission from Future Medicine.

There are a number of different ways of addressing population stratification, such as matching cases and controls according to ethnicity and geographical origin. This matching can be difficult to achieve, and several different control populations might have to be used in any given study. To address this problem, a variety of statistical methods that infer genetic ancestry may be used to control for confounding by population stratification (Cordell and Clayton 2005; Balding 2006; Price et al. 2010). A commonly used statistical method for controlling population stratification is *genomic control*, in which the test statistic is corrected using an inflation factor (lambda, λ). To do this, associations between marker loci across the genome and traits or diseases are used to assess population structure and determine an average effect of population structure across the whole genome. The residual effects of stratification for all investigated markers are then adjusted by a uniform λ. Thus the further λ deviates from the value 1, the greater the correction, and the more evidence for presence of population stratification or other biases; conversely, the closer λ is to 1, the more unlikely it is that stratification is present.

An alternative method is to use *principal components analysis* (PCA) to correct for stratification, which can be carried out using software such as EIGENSTRAT. PCA explicitly models ancestry differences between cases and controls. The resulting genomic correction factor, λ, is specific to a candidate marker's variation in frequency across ancestral populations, minimizing spurious associations while maintaining power to detect true associations. Other approaches include family-based tests such as the transmission disequilibrium test, and mixed models—the latter having the additional advantage of simultaneously addressing population stratification along with other systematic differences, such as family structure or relatedness, that may be present in the data (Price et al. 2010).

Even after many convincing replications of a SNP–disease association, causality may not be inferred; the causal variant may still need to be identified. In this instance, the parallels with classic epidemiology are clear: an association between a SNP and trait might exist if: (1) the SNP with the observed association has a causal role; (2) the SNP has no causal role but is highly correlated (associated) with the true causal variant. It is also possible

that the currently identified SNPs might not fully describe genetic diversity, as a result of *incomplete capture of genetic variation*. One therefore cannot exclude the possibility that the unidentified and untyped variants might have shown statistically significant associations with disease and might include among them the true causal variant. Ultimately, to refine the association signal at a locus of interest requires knowledge of the SNP correlations in that region which, in turn, depends on the genetic variation in the region being fully characterized (Ioannidis et al. 2009).

The need for functional validation

For conclusive evidence of causality, the epidemiological approaches just described need to be complemented by efforts to establish which regions of the human genome sequence truly have a biological function and thus may be relevant to disease mechanism and human health. The recent discoveries from genome-wide studies represent a significant advance in our unravelling of the genetic basis of common human disease (Kruglyak 2008). Despite the large number of loci and the robustness of their associations with diseases or disease traits, these loci tend to have small individual effects on phenotype, and even collectively tend to explain only a small fraction of the heritable component. Moreover, it is not yet clear what most of the causal variants or even the causal genes are at these loci. Thus, whether the GWAS approach can yield biologically relevant insights into human physiology remains to be established.

A next step is to identify and validate functional variants through studies of gene expression, in order to elucidate the biological role of proteins or other gene products encoded by specific genomic regions and to understand the clinical and phenotypic consequences of genome sequence variation; an example of this is the ENCODE project (Box 8.19.2).

Problematically, some genes may only be expressed locally—for example, in specific brain nuclei—and so attaining samples for physiological studies may prove unfeasible. *In vitro* studies of gene activity, as well as *in vivo* overexpression and knockout studies in experimental animal models, will continue to play an important role in elucidating the specific molecular pathways that are relevant to human physiology and validating GWAS results. Promisingly, recent developments in cellular and functional genomics, including CRISPR-Cas9 gene-editing technology, are increasingly

being used to help clarify molecular and biological mechanisms at GWAS loci for disease traits (Klann et al. 2018; Roman and Mohlke 2018). If functional studies can confirm a causal role for genetic variants, these results may help identify potential new drug targets.

Using genetic epidemiology to identify new environmental exposures

A key challenge of classical epidemiology is establishing whether an association between an exposure and a disease outcome is causal or not. Given the observational nature of epidemiological research, an association may also be completely explained by confounding. In this scenario, the relationship is explained by a third factor that is associated with both the exposure and disease risk; the exposure is not independently or causally linked to disease risk. Epidemiological approaches generally use multivariable statistical analysis to reduce confounding. However, because of residual confounding (as a result of measurement error) and unknown confounders, standard observational epidemiology cannot resolve whether an observed association has a causal basis. The conceptual ambiguity of any disease mechanism (specifically, defining mediators and confounders) also limits statistical modelling to reduce confounding. Furthermore, the observed association may be due to reverse association/causation—that is, the relationship between biomarker and disease could be the result of undiagnosed or early disease rather than the risk factor, and thus a consequence of disease rather than a cause.

Box 8.19.2 The ENCODE project: identifying functional properties of the human genome sequence

ENCODE, the Encyclopedia of DNA Elements, is an international collective effort to interpret the human genome sequence and make this information available for medical and scientific communities to apply it to understand disease biology and improve human health (The ENCODE Project Consortium 2011). The project aims to define and annotate the functional elements encoded in the human genome sequence, such as those regions that regulate transcription from DNA into RNA, or regulate the expression of genes. In its initial pilot phase in 2003–2007, the ENCODE investigators set out to annotate functional elements in a defined 1% of the human genome. In 2007, the ENCODE project was expanded to study the entire human genome, to locate and describe discrete regions of the genome with a specific biological function. All data and derived results are readily available

through a freely accessible public database. In September 2012, the project released an extensive set of results, in 30 papers published simultaneously in several journals, including *Nature*, *Genome Biology*, and *Genome Research*. As part of ENCODE's expansion, a subproject named GENCODE was launched, run by an international team led by the Wellcome Trust Sanger Institute, specifically to annotate the genes within the human genome with even higher accuracy (Harrow et al. 2012).

Establishing the causal inference of environmental risk factors can be strengthened by the *Mendelian randomization* approach, in which the biology underpinning genetic epidemiology can be exploited to make causal inferences and establish aetiological associations (Davey Smith and Ebrahim 2003; Lawlor et al. 2008; Sandhu et al. 2008, 2010). The assignment of alleles from parents to offspring that occurs during gamete formation and conception tends to be random, and not correlated either with environmental exposures or with alleles at other loci. This means that the inheritance of one trait is independent of (that is, randomized with respect to) the inheritance of other traits, and characteristics that may confound any association between an exposure and disease are equally distributed among the relevant genetic variants. Thus the association between a disease and a genetic variant is generally much less susceptible to reverse causation or confounding by classical or environmental risk factors.

Given this, a comparison of groups of individuals defined by a genetic variant, based on a Mendelian randomization design, is equivalent to a randomized comparison, with only the relevant exposure differing across the relevant genetic variant. Examining the relationship between variants in the genome that show unequivocal associations with the relevant exposure and disease risk is therefore a potential method of assessing whether an exposure might be causally linked to disease (Fig. 8.19.2).

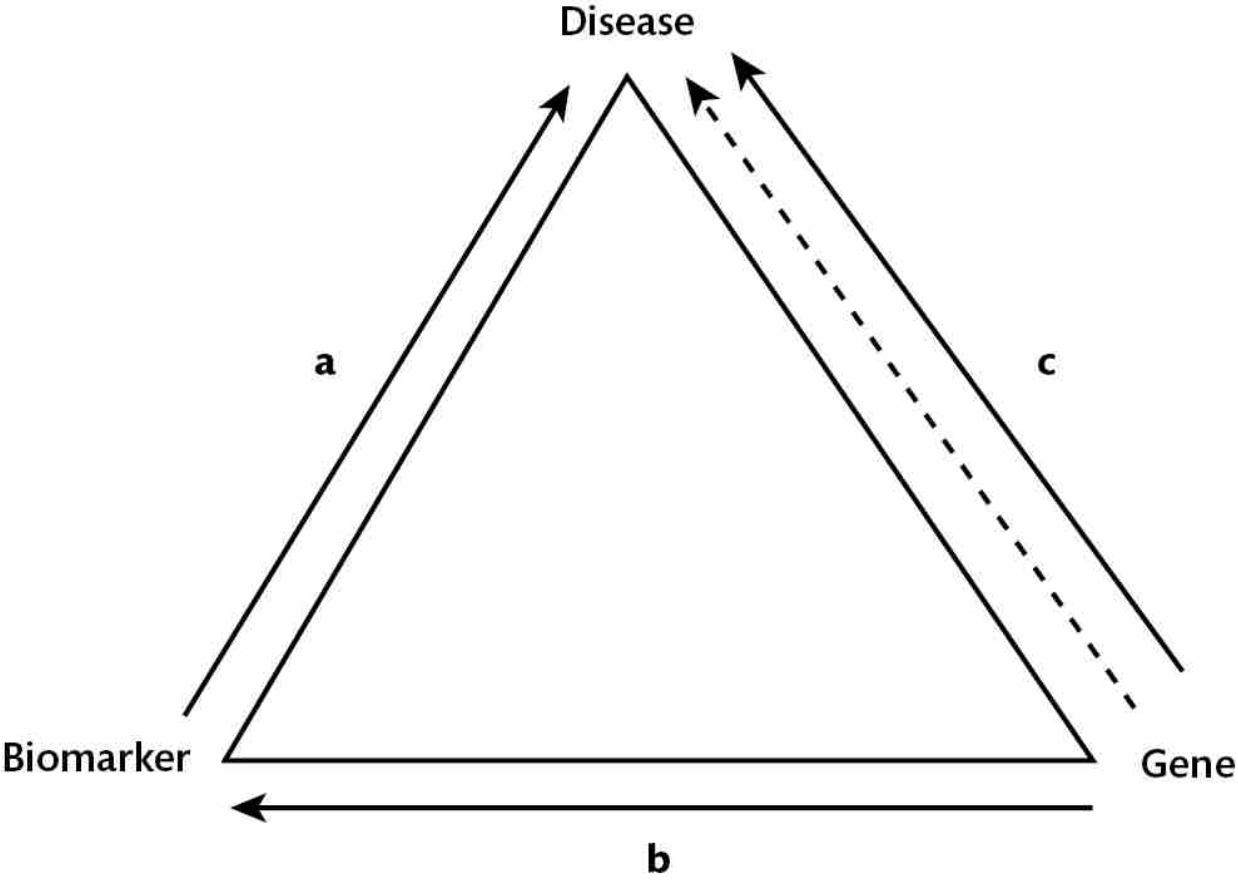


Fig. 8.19.2 Mendelian randomization. The Mendelian randomization design provides a strategy to help clarify whether there is a causal relation between an exposure and disease. Information on the magnitude of the association between the exposure and disease (a), combined with information on the magnitude of the association between the genetic variant and the exposure, (b), is used to estimate the expected magnitude of the association between the genetic variant and disease risk (dashed line in (c)). Based on several assumptions, including a linear association between the exposure and disease, the direct assessment of (c) (indicated by the solid line) provides an unconfounded assessment of the association between the exposure and disease risk, (a).

Reprinted from Springer, *Diabetologia*, Volume 51, Issue 2, 2008, pp. 211–13, Mendelian randomisation studies of type 2 diabetes: future prospects, Sandhu MS et al, Figure 1.

One example of the success of the Mendelian randomization approach is its role in providing evidence of a causal association between lipoprotein(a) and coronary heart disease. Lipoprotein(a) is a composite particle comprising LDL cholesterol linked to apolipoprotein(a) and has long been studied as a possible cardiovascular risk factor. Despite large-scale meta-analyses previously demonstrating an association between lipoprotein(a) and disease risk (Emerging Risk Factors Collaboration et al. 2009), it remained unclear whether the observed association reflected a causal relationship or simply an unexplained correlation. However, a Mendelian randomization

study of around 40,000 participants demonstrated that the magnitude of heart attack risk associated with genetically raised concentrations of lipoprotein(a) was consistent with that associated with an equivalent difference in measured lipoprotein(a) levels (Danesh and Erqou 2009; Kamstrup et al. 2009). These findings are consistent with a causal association between lipoprotein(a) and increased risk of heart disease, and highlight lipoprotein(a) as a potential therapeutic target for the management of cardiovascular disease risk. Mendelian randomization analyses have since been used to investigate a number of biomarkers of coronary heart disease—for example, to provide evidence that genetic variants underlying triglycerides are likely to be causal and, conversely, to refute a causal role for high-density lipoprotein (HDL) cholesterol in coronary heart disease (Triglyceride Coronary Disease Genetics Consortium and Emerging Risk Factors Collaboration et al. 2010; Voight et al. 2012).

As with all genetic epidemiology studies, there are problems associated with the need for large sample sizes (Ebrahim and Davey Smith 2008). Mendelian randomization studies, like other genetic epidemiological studies, require reliable identification of statistical associations between genetic variants and biomarkers and disease risk. Importantly, the principal requirement is the ability of Mendelian randomization studies to detect an association of equivalent magnitude to that predicted by a proportional change in the relevant trait or risk factor. Given that individual genetic variants are likely to explain only a very small proportion of the variation in a biomarker trait or risk factor, these assessments will require very large sample sizes. Mendelian randomization findings may also be limited by the effects of biological compensation (also known as canalization) and pleiotropy (Davey Smith and Ebrahim 2003; Lawlor et al. 2008). Canalization refers to the process by which potentially disruptive influences on normal development from genetic variations are buffered by compensatory developmental and physiological adaptations. This developmental compensation can lead to inconsistencies in the magnitude of the triangulated associations. Furthermore, the pleiotropic nature (multiple biological and phenotypic effects) of some genetic variants may produce confounded associations between genetic variants and phenotypes. Despite these limitations, Mendelian randomization provides a potential research framework using genetic variants to make causal inferences about modifiable (non-genetic) risk factors for disease and health-related

outcomes. These studies, when correctly performed, will provide insights into aetiological mechanisms and causality, informing potential therapeutic and preventive strategies.

Why is genetic epidemiology important for public health?

Genetics has increasingly become an essential part of the epidemiology paradigm. As we have seen, the integration of epidemiological methods with genome-wide technologies has provided unprecedented insights into the aetiology of complex traits and diseases in human populations and has made substantial and independent contributions to the identification of genetic loci associated with a broad range of complex traits and diseases, including diabetes, heart disease and cancer (Visscher et al. 2012). In parallel, there has been increasing speculation and expectation over how these findings may translate into practical applications for improving human health and lead to public health benefit.

The integration of human genomics discoveries into clinical and public health practice raises a number of questions. Public health is, by its very definition, concerned with improving health from a population perspective (Khoury et al. 2011a, 2011b). How does genomic discovery translate into population health benefit and provide new tools for achieving public health goals? After all, public health prevention and health promotion initiatives have, conventionally, focused on risk factors that are modifiable whereas genetic risk factors are unmodifiable. As we make the case in this section, the discovery and confirmation of genetic variants associated with diseases and traits may have translational implications and major impacts on both individual and population health: first, through elucidation of underlying disease biology and advances in therapeutics and, second, through risk prediction and personalized medicine. In addition, *genetic testing*—undertaken in order to diagnose an inherited disease that is caused by a specific genetic mutation—represents another application of genomic medicine, and is briefly mentioned here for completeness; for a full discussion of the topic, see Chapter 2.5.

Drug discovery and development

The aetiological mechanisms underpinning complex disorders almost certainly involve a combination of many factors: multiple susceptibility genes of modest effect, gene–gene interactions, gene–environment interactions, and heterogeneity of genetic and environmental exposures. The most beneficial application of genetic association studies is likely to be in unravelling the complex biological mechanisms of disease and providing novel insights into disease pathways. In turn, a better understanding of human biology may offer new possibilities for treatment and prevention at the population level (Kingsmore et al. 2008; McCarthy et al. 2008).

Clinical advances in novel therapeutic targets

First, the identification of susceptibility genes for common diseases and traits, together with a better understanding of their underlying functional properties and advances in knowledge of human physiology, should help inform drug development and help determine whether changes in gene activity have pharmacological potential. This is likely to have broad application at the population level (Table 8.19.2). For example, if overexpression of a variant gene product was protective against hypercholesterolaemia, then—even if the effect size was small among carriers—a drug which mimicked and inflated the biological effect of the causal variant might be an effective therapy for treatment of raised blood cholesterol. Such a drug could confer clinical benefit across the cholesterol range, and could be used in the general population irrespective of a person's genotype. Such translational opportunities present very real benefits for public health, although expectation must be tempered with realism: the time from gene discovery to an approved and marketable drug takes, on average, at least 10 years. Thus while the greatest impact of genomic medicine is likely to result from its application to novel biological pathways and new therapies, the full success of this will only be appreciable in decades to come (Hirschhorn and Gajdos 2011).

Table 8.19.2 Examples of matches between GWAS traits and drugs licensed for clinical use

Drug name or class	Gene	Drug indication	GWAS trait
Immunosuppressants and immunomodulators	*IL2RA*	Relapsing-remitting multiple sclerosis	Multiple sclerosis
Antidiabetic drugs	*KCNJ11*	Type 2 diabetes	Type 2 diabetes
Statins	*HMGCR*	Hypercholesterolaemia	LDL cholesterol
Lipid-regulating drugs	*LPL*	Hyperlipidaemia	HDL cholesterol
Gonadotrophins	*FSHR*	Female infertility	Polycystic ovary syndrome

Source: data from *Nature Biotechnology*, Volume 30, Issue 4, pp. 317–20, Sanseau P et al, Use of genome-wide association studies for drug repositioning.

Improved responses to existing drug treatments

A second, and more immediate, application of genomic data may be for drug repositioning—that is, the identification or refining of clinical indications for existing drugs (Sanseau et al. 2012). To do this, a first step would be to identify genes that are already being pursued as drug targets and the diseases that these drugs are currently being used to treat. Next, by utilizing data from GWAS, one can see whether those genes may be associated with additional traits or disease, which in turn could be explored as new potential drug indications. As an illustration, denosumab is a drug currently licensed for the treatment of osteoporosis, bone metastases, and rheumatoid arthritis. It acts by targeting the gene *TNFSF11*. In GWAS, the *TNFSF11* gene has been associated with Crohn's disease and thus it is reasonable to speculate that denosumab may have a therapeutic role in the management of Crohn's disease (Sanseau et al. 2012). Taking this approach thus affords translational opportunities for repositioning existing drugs for many common diseases with widespread benefit to public health.

Pharmacogenomics

Patients can vary in their response to prescribed drugs; some patients do not respond at all to certain drug treatments, and for other drugs the dose required to achieve a therapeutic effect can vary (Daly 2010). Personalized

treatment for diseases according to genotype might be feasible where risk alleles predict an individual's response to treatment or susceptibility to adverse drug reactions—for example, those genes that influence the effect of coumarin anticoagulant drugs such as warfarin. Warfarin is widely prescribed to reduce blood clotting and is used to protect high-risk patients from stroke, pulmonary embolism, and thrombosis. However, there is a large variation in the dose required to achieve safe yet effective anticoagulation; thus, reducing the uncertainty in establishing the therapeutic dose for individual patients would improve the safety and efficacy of anticoagulation therapy and potentially avoid the clinical complications associated with under- or overcoagulation. Warfarin is metabolized by the hepatic cytochrome P450 enzyme expressed by the *CYP2C9* gene. Its anticoagulant effect is mediated by the enzyme VKORC1, which is the target enzyme inhibited by warfarin. Among Caucasians, warfarin dosing is more strongly correlated with genetic variation in the *VKORC1* and *CYP2C9* genes than other patient-related factors (Cooper et al. 2008; Takeuchi et al. 2009; Teichert et al. 2009; Jorgensen et al. 2012). Common variants in *CYP2C9* are associated with slower metabolism of warfarin, higher warfarin concentrations, decreased dose requirements, and susceptibility to bleeding. Polymorphisms in *VKORC1* lead to hypocoagulable states, and necessitate lower dosage.

The translational applications of these observations are not yet fully established. In principle, an individual's genotype could, potentially, be used as a predictor of warfarin dosage. Indeed, several studies have suggested that combining genetic information on variants in *VKORC1* and *CYP2C9* together with clinical data of a patient may help better predict the warfarin dosage—both for initiation of therapy and for maintenance—required to achieve satisfactory anticoagulation (Hatch et al. 2008; International Warfarin Pharmacogenetics Consortium et al. 2009; Horne et al. 2012; Tatarunas et al. 2012). To date, though, genotype-guided dosing algorithms are not widely used, as they have not yet been shown to be cost-effective in preventing adverse events. However, a number of large clinical trials are now underway, designed specifically to assess whether using genetic data to inform warfarin dosing will improve outcomes for patients being treated with warfarin.

Another example of how genomic information has been successfully applied to understanding drug responses is in the treatment of chronic

hepatitis C infection. Standard therapy for the last decade has been a combination of pegylated interferon and ribavirin. It has been well documented that treatment response has only limited efficacy, even for more responsive viral genotypes. However, there is increasing evidence that marked variation in prognosis and response to hepatitis C treatment may be due to differences in human genome diversity; in particular, polymorphisms at the *IL28B* gene locus, encoding interferon-λ-3 (IFN-λ-3), predict response to interferon with genome-wide significance (Ge et al. 2009; Thomas et al. 2009).

A number of other GWAS on pharmacogenomics have identified several novel genetic variants that may affect drug responses or reactions, although many of these associations have failed to reach statistical significance probably due to insufficient sample sizes (Daly 2010). Nevertheless, these findings potentially could be used to develop genetic tests to determine the most appropriate drug for treatment or the optimum dose to prescribe. While it is too early to say if this is of clinical relevance in terms of predicting the trajectory of disease, the prospect of individually tailored treatment in the future raises exciting possibilities—see later for further discussion. Genotype-based randomized controlled clinical trials will be needed to formally evaluate the potential utility of genotype-specific treatment (Manolio et al. 2009; Marteau et al. 2010b).

Genomic medicine

The emergence of genomic medicine represents a shift from traditional medical genetics—with its focus on rare genetic diseases—towards improved risk prediction and, potentially, greater personalization in the practice of medicine and prevention of common diseases.

Risk prediction

Complex diseases and traits—including coronary heart disease, diabetes, hypertension, and most cancers—are caused by multiple genetic and environmental factors. For such conditions, individual risk assessment is an important component of clinical decision-making, potentially identifying people who might benefit from earlier or more intensive interventions. However, individual risk assessment is multifaceted, involving an evaluation

of risks and benefits to the individual and to society (British Medical Association Board of Science 2005; Grosse and Khoury 2006). It is also complicated by the debate over statistical methods to assess the utility of risk factors (including genetic variants) for disease risk prediction (Fig. 8.19.3) (Wood and Greenland 2009; Romanens et al. 2010; Talmud et al. 2010). Ultimately—and crucially—the utility of predictive models depends on their ability to alter clinical management which in turn relies on the availability of clinical interventions that are universally indicated but might be beneficial or cost-effective for higher-risk patient subgroups (Hirschhorn and Gajdos 2011).

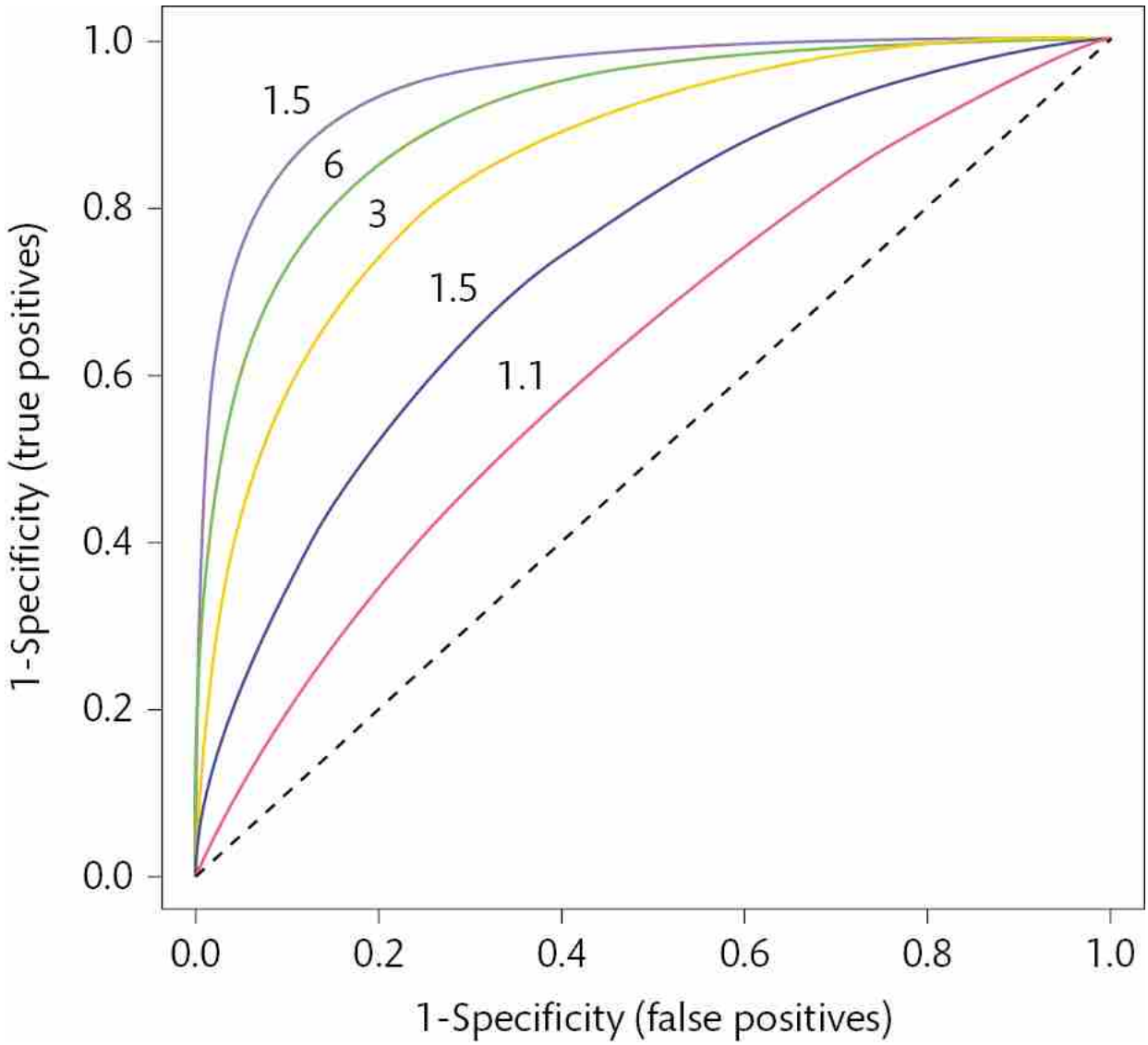


Fig. 8.19.3 Genomic risk prediction and personalized medicine. (a) Individual disease risk prediction from theoretical ROC curves for familial relative risk. Measures of risk discrimination include the sensitivity (the proportion of people who develop disease who were correctly categorized as having a high risk of disease) and specificity (the proportion of people who do not go on to develop disease who were correctly categorized as having a low risk of disease) of the test. These metrics are usually presented in a receiver operator characteristic (ROC) curve, and an overall assessment is given by the area under the ROC curve. This area is equal to the probability that a random person with disease has a higher value of predicted risk than a random person without disease; thus the larger the area, the better the test.

These metrics can be used to determine the positive predictive value—the proportion of people categorized as high risk that go on to develop disease. The figure shows ROC curves for various values of FRR. For a given disease prevalence, when the FRR is high (for example, FRR=15), the curve rises steeply, indicating the test has a relatively high sensitivity and specificity; the corresponding area under the ROC curve increases. As the FRR drops, sensitivity and specificity fall and the area under the curve decreases. For lower FRRs (for example, FRR = 1.1), the curve approaches the diagonal dotted line, indicating a test with only 50% probability of correctly predicting

risk of disease. (c) The spectrum of genomic medicine. In practice the three elements of risk prediction, stratified medicine, and personalized medicine, are not distinct but form a continuum of patient management. Genomic risk prediction involves identifying genetic variants that predict disease risk within a population. Further along the spectrum lies stratified medicine, which potentially allows targeting of treatments to specific disease pathways, identification of treatments effective for particular groups of patients, and development of diagnostics to ensure the right patient gets the right treatment at the right time. In the context of genomic medicine, individuals with disease are tested for a genetic marker which indicates whether a patient is likely to show a response to therapy. A further refinement of this approach takes us to personalized (individualized) medicine, where patients receive custom-made treatment.

Reproduced from *The Lancet*, Volume 376, Issue 9750, Sandhu M. et al., Genomic risk prediction, pp. 1366–67, with permission from Elsevier.

The identification of several genetic variants, each independently impacting on disease risk, has led to a reassessment of whether these variants improve risk prediction. To date, genetic variants identified for complex diseases explain only a small proportion of the total variation in disease risk. Accordingly, currently identified genetic susceptibility loci do not have the discriminatory power to substantially improve disease risk prediction (Kraft et al. 2009). However, with large-scale studies and new genomic sequencing technologies, there is the potential to identify new susceptibility loci. Could these variants improve clinical risk prediction? Data indicate that the variation in disease risk explained by many of these unidentified genetic variants may be very small; their detection may always be beyond the resolution of genomic research strategies and available sample sizes. Hence, a small proportion of the heritability of a disease may not be identified (Clayton 2009; Wray et al. 2010).

At the population level, the familial relative risk (FRR) represents the risk prediction achievable if all relevant loci were identified—assuming that the magnitude of the FRR is accurate and that it primarily reflects shared genetic variants rather than environments (Xu et al. 2009; Wray et al. 2010). With these caveats, for rare or common diseases with greater heritability the discovery of all disease susceptibility variants might provide opportunities to more reliably identify people at greater risk of disease (Fig. 8.19.3). However, for common diseases, such as type 2 diabetes, heart disease, and some cancers, with relatively small FRRs, it is unlikely that the discovery of all susceptibility loci will achieve the necessary predictive accuracy to provide clinically useful information (Wald et al. 1999; Ware 2006; Park et al. 2010).

The ultimate utility of genomic risk prediction in the context of complete genome sequencing will depend on the underlying genetic architecture, the

proportion of the variation in disease attributable to genetic factors, and the disease prevalence. Population-wide, it seems unlikely that genomic risk prediction alone will attain the discriminatory resolution to predict individual disease risk for many common diseases with only modest heritability. However, the use of genomic factors for individual disease risk prediction in conjunction with conventional risk factors, or for identifying subgroups for screening, remains unclear (Pharoah et al. 2008; Xu et al. 2009). While at present we lack the ability to provide clinically useful risk prediction, as more disease loci are identified it may be beneficial to re-assess whether genetic testing could improve risk prediction of common disease beyond traditional risk factors.

Stratified medicine and the emergence of personalized medicine

In addition to risk prediction and a better understanding of disease aetiology, genetic data may be increasingly used in the field of stratified medicine (Fig. 8.19.3). Patient response to drug treatments and therapeutic interventions varies markedly across the population as a result of differing underlying mechanisms of disease and patient responses to both disease and treatment. Stratified medicine can be described as identifying the different strata within a disease using a biomarker or genotype, in order to identify a patient subpopulation which is more likely to benefit or experience an adverse reaction in response to a given therapy. This approach—with patients better matched to therapies—is anticipated to have a major effect on both clinical practice and the development of new drugs and diagnostics (Trusheim et al. 2007). At the extreme of this patient matching are ‘personalized’ medicines, which by definition vary for each individual’s unique genetic make-up, such as customized stem-cell based therapies. However, the logistical, clinical, and ethical implications of this must be carefully considered.

Integrating genomics into electronic health records

As outlined earlier, personalized, or precision, medicine has the potential to transform medicine by tailoring treatment to individuals and subpopulations. In this context, large-scale integration of genetic data and prospective health data is vital for translating genetic discovery into direct clinical applications. If applied appropriately, such resources could revolutionize our

understanding of genetic risk of disease susceptibility and severity, as well as our understanding of drug response and adverse events in the clinical setting. Electronic health record (EHR) frameworks provide a unique platform to do this. EHR data can provide important information for the development of predictive algorithms, to help target screening and management interventions at specific individuals, and help target interventions required for certain diseases. Integrating genetics into predictive analytics has been shown to inform clinical interventions, and produce algorithms that can materially improve prediction of patient outcomes, response to drug, and the probability of adverse events with specific medication.

Importantly, the value—clinically and economically—of integrating genetic data into routine clinical care has not been evaluated; providing clear evidence for the utility of such initiatives will be a critical first step to realizing the full potential of genomics at population health scale.

Limitations and challenges of genetic epidemiology

Human genetics has advanced at an extraordinary pace during the last 20 years. The number of, and the ease with which, genetic markers of disease are being discovered has gone from dozens to millions. As we enter the era of next-generation high-throughput genotyping and sequencing, it is important to remember the limitations of this knowledge in order to have realistic expectations of what it can, and cannot, deliver and to apply it appropriately and responsibly.

Ethical, legal, and social implications

The generation of genetic information has contributed greatly to our understanding of the molecular basis of disease and, increasingly, to the development of effective treatments. In parallel, there are concerns about the ethical, legal, and social implications of this knowledge on the well-being of individuals and populations (Box 8.19.3).

Box 8.19.3 Examples of some of the ethical, legal, and social implications of genetic epidemiology

Genetic discrimination

A primary concern is that genetic information could be harmful and could lead to discrimination such as: (1) denial of health insurance or employment, based on genetic susceptibility; and (2) stigmatization of a specific ethnic group.

Genetic determinism

This stems from a misconception about the causal relationship between genetics and behaviour, and the erroneous belief that genetic information is somehow more predictive and certain than other types of health information because ‘you cannot change your genes’. This can lead to an unhelpful and inaccurate sense of inevitability and the use of genetic determinism to explain health disparities.

Data protection

The generation of vast amounts of genetic data, and its potential identifiability of individuals, raises significant issues around the storage of genetic data, data processing, data access, and confidentiality. Additional concerns exist around sharing sensitive data with other family members, such as in the context of reproductive choice.

Return of individual research findings (including incidental findings)

Disclosing study results and presenting clinical risk information can have important effects on health outcomes and health behaviours among study participants. Understanding probability information about disease risk and the benefits of intervention is a vital component of an individual’s decision-making capacity. The obligation of the health professional to return genetic findings—particularly when the clinical validity and utility is unclear—needs to be considered. Additionally, disclosing results for which treatments may be unavailable or prohibitively expensive raises particular ethical issues.

Informed consent

National and international research guidelines agree that valid consent to participate in research must be consent that is informed in a manner that is linguistically and culturally appropriate and is understood by the study participant. This can be particularly challenging in the context of genomics research, which is complex and often difficult to understand, and where the implications of research findings are largely unknown (including incidental findings which are not relevant to the original research question but may have clinical significance). An additional challenge is to resolve issues around individual consent versus family consent, including the right ‘not to know’ the findings of genetic studies.

Fair distribution of benefits from research

As with all fields of research, particular individuals, groups, or communities should neither bear an unfair share of the direct burdens of participating in research, nor should they be unfairly excluded from the potential benefits of research participation. This is particularly pertinent when genetic research is conducted in countries where genetic services are not available or may not be readily accessed. Inclusiveness in genetic research and fair distribution of benefits and burdens should be important considerations for researchers, research ethics boards, and sponsors. Issues of fair and equitable treatment arise in deciding whether and how to include individuals, groups or communities in research, and the basis for the exclusion of some.

Accounting for unexplained variation

Despite the success of GWAS in identifying disease genes, a substantial proportion of the heritability of disease traits remains unexplained (Manolio et al. 2009). So far, loci shown to be associated with disease only account for a small fraction of the total genetic contribution; the great majority of susceptibility loci remain to be identified. How do we account for this remaining variation? Unexplained heritability may be due to unknown common variants of small effect, or rare variants with larger effects, or influences by environmental factors. To better understand the missing heritability of disease traits, larger-scale studies are required, using denser SNP arrays. In addition, exome sequencing, whole genome sequencing, or whole genome association studies in large populations will be needed to identify all potential variants within a gene or locus. Sequencing candidate genes in individuals at the extremes of a quantitative trait (e.g. the extremes of lipid levels) may identify new variants, common or rare, and help further our understanding of disease biology and prevention.

Genetic epidemiology in low- and middle-income countries

To date, most human genomics research has been done in populations of European ancestry, and has successfully provided new insights into the biology of chronic diseases and their risk factors in these populations. However, the relevance of many recent genomic findings to populations in low- and middle-income countries (LMICs) is not known. It may be difficult to meaningfully apply the findings from Western populations to LMIC populations, where both genetic and non-genetic susceptibility may be different.

To fully appreciate the impact of a genetic variant on disease risk it will be important to establish whether the findings from one population apply to other populations of different ancestry. Globally, human populations show structured genetic diversity as a result of geographical dispersion, selection and drift (Campbell and Tishkoff 2008; Novembre and Ramachandran 2011). Accordingly, among many populations in LMICs there is marked genomic diversity—in particular, among populations of African descent, where there is more human genetic diversity than anywhere else (Tishkoff et al. 2009). Given this, understanding the genomic basis of common diseases and their risk factors in LMICs is likely to provide additional insights into

disease aetiology and potential therapeutic strategies (Hill 2006; Campbell and Tishkoff 2008; Novembre and Ramachandran 2011); an example of this is the African Genome Variation Project (AGVP). Representing the most comprehensive characterization of African diversity to date, the AGVP highlighted the utility of genomic information from African populations for identifying genes related to disease susceptibility and resistance and facilitating medical genetics in Africa (Gurdansani et al. 2015). Now, in addition to genotyping resources such as the AGVP, sequencing efforts are increasingly underway to significantly advance the comprehensive catalogue of human genetic variation in Africa, providing a mechanism to evaluate the clinical relevance of genetic diversity both in African populations and globally.

Thus the value of GWAS will be enhanced further by including populations with greater genetic diversity and less correlation between variants, and will benefit from more precise characterization of phenotypes and environmental exposures (Manolio et al. 2009). Ultimately, large-scale international collaborations will facilitate studies in different global populations and help to determine whether loci identified in initial GWAS (usually done in populations of European ancestry) are relevant in non-European individuals. As a result of such studies, our ability to treat disease and reduce risk will be much more complete.

In addition to studies of the human genome, techniques from genetic epidemiology can also be used to examine pathogen genomes, yielding important applications for human health—for example, in helping to understand the causes of resistance to malaria treatment. First-line treatment for uncomplicated falciparum malaria is artemisinin-based combination therapy (ACT), which has been highly effective in treating this disease. However, in recent years, resistance to artemisinin has become increasingly prevalent in Cambodia and Thailand, characterized by slow parasite clearance from the blood (Uhlemann and Fidock 2012). Antimalarial drug resistance hinders malaria control and is a major public health threat to global initiatives to control and eliminate the disease. Using a variety of techniques, including targeted association studies of candidate genes for ACT and genotyping of variants across the malaria genome, a number of studies have identified that resistance to artemisinins is, in part, genetically determined by variants in the malaria-parasite genome and conclude that parasite genetics have a central role in determining parasite clearance rates

(Cheeseman et al. 2012; Phyo et al. 2012). Furthermore, these studies report that the heritability (h_2)—that is, the contribution of parasite genetics that explains the variation in parasite clearance—has increased from 30% in 2001–2004 to 66% in 2007–2010 (Phyo et al. 2012). The mechanisms underlying the reported associations are not yet known, but understanding these biological processes could potentially inform new approaches for effective malaria treatment in the future.

The application of genetic epidemiology and genomic research to public health practice merits specific discussion in the context of LMICs—where the global burden of infectious and chronic, complex diseases is greatest. Additionally, in LMICs healthcare is often severely constrained by limited financial and human resources and a lack of appropriate infrastructure (Burke et al. 2010). It would be simplistic to suggest that because of the great health needs of LMICs—overwhelmingly determined by economic and social factors—genomics is an irrelevance and has no role to play in improving health. Rather, to omit genomic discoveries from the healthcare agenda of LMICs would serve only to promote existing inequities between rich and poor, between economically advantaged and disadvantaged. Public health professionals have an essential role to play in leading the response to genomic medicine in LMICs, ensuring that healthcare providers and policymakers are informed by a strong evidence base for the effectiveness of genomic approaches in improving population health.

Improving phenotypic characterization and classification of disease

Associations between genetic variants and phenotypes are limited by the accuracy with which each is measured. The ability to measure genotypes now far exceeds our ability to measure phenotypes (Altshuler et al. 2008). We need to match this with more detailed, and more accurate, assessments of clinical phenotypes. Indeed, carefully designed GWAS with thorough phenotypic characterization have the potential to redefine disease classification (Kingsmore et al. 2008). In some cases, it will become increasingly possible to unpick the phenotypic complexity of diseases that involve several pathological processes and allow us to establish whether genetic factors distinctly contribute to different phenotypes of the condition. An example of this is in the area of coronary artery disease (CAD), where

GWAS have successfully discriminated genetic variants for distinct CAD phenotypes—suggestive of distinct underlying pathological mechanisms (Reilly et al. 2011). This opens up avenues for patient stratification and personalized treatments.

Conversely, for other complex diseases that have previously been regarded as distinct clinical entities, GWAS findings may point to common underlying disease processes and a shared pathogenesis. To illustrate, a GWAS of Crohn's disease and ulcerative colitis found a sizeable overlap in susceptibility loci for the two diseases. With such a degree of sharing of genetic risk, it seems likely that many of the biological mechanisms involved in one disease have some other role in the other (Jostins et al. 2012). These studies provide a rich source of clues to the biological mechanism underlying complex disease. However, the translation of such GWAS discoveries into prognostic and therapeutic benefit will need greater insights into the relation between each loci and the different phenotypes.

Risk communication

The extent to which genetic information will play a role in 'personalized medicine' will depend not only on whether predictive accuracy beyond conventional measure can be attained, but also on whether there are interventions whose effectiveness is improved by knowledge of a genetic result. A key challenge will therefore be to find effective ways to communicate information on small incremental risk increases conferred by genetic markers for disease risk, in order to prompt changes in lifestyle that will ultimately confer health gain to individuals. At present, it is not clear if personal genomic information is likely to beneficially impact public health through motivation of lifestyle behavioural change. Risk perception and risk communication are complex processes, and are shaped by social and cultural values, assumptions, and beliefs held at the individual and societal level.

An individual's ability to make sense of genetic information will involve an understanding of the concepts of genes, genetics, ancestry, and heritability, in a manner that is culturally relevant and acceptable. Most of the information we have on genetic susceptibility to disease is, at best, incomplete as discussed earlier in this chapter; health professionals' advice to people will have to change as more robust evidence on risk factors comes

to light, sometimes seeming to conflict with, or even to contradict, previous knowledge. Conveying this uncertainty is a difficult and challenging task.

The creation of personal genomic risk profiles, that predict an individual's lifetime risk of various common diseases, raises the possibility that public health promotion could potentially be directed at susceptible individuals/groups of people in the population based on their genomic risk profile. This marks a departure from conventional public health practice which has tended to focus on social and environmental determinants of disease and health, and allows a focus on preventing disease, disability, and mortality among people with specific genotypes that increase their disease risk. It is tempting to think that knowledge of individual risk of a given disease, as predicted by genome-wide profiling, will motivate individuals to change their behaviour and promote greater adherence to a healthy lifestyle or treatment regimen in order to reduce their risk; however, human behaviour is complex and risk estimates are difficult to communicate meaningfully to patients (Loos 2011). Recent studies, including a Cochrane review, seem to indicate that genetic testing may not result in individuals changing their behaviour in any significant or sustained way (Marteau et al. 2010a; Bloss et al. 2011).

Conclusions

It is clear that public health practice cannot ignore the impact of genetic epidemiology and the knowledge derived from genomics research. The results from genetic epidemiological research have two broad translational applications which have the potential to contribute to population health and well-being. First, as our understanding of disease aetiology and its underlying biological mechanisms increases, there is the potential for new drug development and improved therapeutic strategies to manage disease. Second, in parallel, there is growing interest in the use of these genetic variants to predict individual disease risk over and above classical risk factors, and to develop stratified and personalized approaches to diagnosis and disease management. The public health community has a central leadership role to play in critically and systematically evaluating these discoveries for their potential use in disease prediction, prevention, and treatment response. Furthermore, public health practice must ensure that the benefits of genomics are applied with equity across the population.

As our understanding of the molecular basis of disease increases, there is a need for public health practitioners to have a degree of knowledge about genetic susceptibility and to understand the scientific basis of genetic epidemiology and the benefits it can afford, as well as to be aware of the limitations of this knowledge. The ultimate goal of public health is to promote and improve the health and well-being of the population. A public health approach to genetic epidemiology should assess the contribution of genomics to health in the context of wider determinants of disease and health, and provide a robust evidence base to inform patients, healthcare policymakers, health service commissioners, and providers.

REFERENCES

Altshuler, D., Daly, M.J., and Lander, E.S. (2008). Genetic mapping in human disease. *Science*, **322**, 881–8.

Amaral, P.P. and Mattick, J.S. (2008). Noncoding RNA in development. *Mammalian Genome*, **19**, 7–8.

Balding, D.J. (2006). A tutorial on statistical methods for population association studies. *Nature Reviews Genetics*, **7**, 781–91.

Bamshad, M.J., et al. (2011). Exome sequencing as a tool for Mendelian disease gene discovery. *Nature Reviews Genetics*, **12**, 745–55.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, **57**, 289–300.

Bloss, C.S., Schork, N.J., and Topol, E.J. (2011). Effect of direct-to-consumer genomewide profiling to assess disease risk. *The New England Journal of Medicine*, **364**, 524–34.

Bochud, M. (2012). Genetics for clinicians: from candidate genes to whole genome scans (technological advances). *Best Practice & Research Clinical Endocrinology & Metabolism*, **26**, 119–32.

Bomba L., Walter K., and Soranzo N. (2017). The impact of rare and low-frequency genetic variants in common disease. *Genome Biology*, **18**, 77.

Boomsma, D., Busjahn, A., and Peltonen, L. (2002). Classical twin studies and beyond. *Nature Reviews Genetics*, **3**, 872–82.

British Medical Association Board of Science (2005). *Population Screening and Genetic Testing. A Briefing on Current Programmes and Technologies*. London: British Medical Association.

Burke, W., et al. (2010). Extending the reach of public health genomics: what should be the agenda for public health in an era of genome-based and 'personalized' medicine? *Genetics in Medicine*, **12**, 785–91.

Campbell, M.C. and Tishkoff, S.A. (2008). African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping. *Annual Review of Genomics and Human Genetics*, **9**, 403–33.

Cheeseman, I.H., et al. (2012). A major genome region underlying artemisinin resistance in malaria. *Science*, **336**, 79–82.

Choi, M., et al. (2009). Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 19096–101.

Cirulli, E.T. and Goldstein, D.B. (2010). Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nature Reviews Genetics*, **11**, 415–25.
Clayton, D.G. (2009). Prediction and interaction in complex disease genetics: experience in type 1 diabetes. *PLoS Genetics*, 5, e1000540.
Collins, F.S., Morgan, M., and Patrinos, A. (2003). The Human Genome Project: lessons from large-scale biology. *Science*, **300**, 286–90.
Cooper, G.M., et al. (2008). A genome-wide scan for common genetic variants with a large influence on warfarin maintenance dose. *Blood*, **112**, 1022–7.
Cordell, H.J. and Clayton, D.G. (2005). Genetic association studies. *The Lancet*, **366**, 1121–31.
Daly, A.K. (2010). Genome-wide association studies in pharmacogenomics. *Nature Reviews Genetics*, **11**, 241–6.
Danesh, J. and Erqou, S. (2009). Risk factors: lipoprotein(a) and coronary disease—moving closer to causality. *Nature Reviews Cardiology*, **6**, 565–7.
Davey Smith, G. and Ebrahim, S. (2003). 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *International Journal of Epidemiology*, **32**, 1–22.
Donnelly, P. (2008). Progress and challenges in genome-wide association studies in humans. *Nature*, **456**, 728–31.
Ebrahim, S. and Davey Smith, G. (2008). Mendelian randomization: can genetic epidemiology help redress the failures of observational epidemiology? *Human Genetics*, **123**, 15–33.
Emerging Risk Factors Collaboration, et al. (2009). Lipoprotein(a) concentration and the risk of coronary heart disease, stroke, and nonvascular mortality. *Journal of the American Medical Association*, **302**, 412–23.
Freedman, M.L., et al. (2004). Assessing the impact of population stratification on genetic association studies. *Nature Genetics*, **36**, 388–93.
Ge, D., Fellay, J., et al. (2009). Genetic variation in IL28B predicts hepatitis C treatment-induced viral clearance. *Nature*, **461**, 399–401.
Geneletti, S., Gallo, V., Porta, M., Khoury, M.J., and Vineis, P. (2011). Assessing causal relationships in genomics: from Bradford-Hill criteria to complex gene-environment interactions and directed acyclic graphs. *Emerging Themes in Epidemiology*, **8**, 5.
Grosse, S.D. and Khoury, M.J. (2006). What is the clinical utility of genetic testing? *Genetics in Medicine*, **8**, 448–50.
Gurdansani, D., et al (2015). The African Genome Variation Project shapes medical genetics in Africa. *Nature* 2015,517, 327–32.
Harrow, J., et al. (2012). GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Research*, **22**, 1760–74.
Hashem, I. A. T., Yaqoob, I., Anuar, N.B., Mokhtar, S., Gani, A., and Ullah Khan, S. (2015). The rise of 'big data' on cloud computing: review and open research issues. *Information Systems*, **47**, 98–115.
Hatch, E., Wynne, H., Avery, P., Wadelius, M., and Kamali, F. (2008). Application of a pharmacogenetic-based warfarin dosing algorithm derived from British patients to predict dose in Swedish patients. *Journal of Thrombosis and Haemostasis*, **6**, 1038–40.
Hattersley, A.T. and McCarthy, M.I. (2005). What makes a good genetic association study? *The Lancet*, **366**, 1315–23.
Hill, A.V. (2006). Aspects of genetic susceptibility to human infectious diseases. *Annual Review of Genetics*, **40**, 469–86.
Hirschhorn, J.N. and Daly, M.J. (2005). Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics*, **6**, 95–105.
Hirschhorn, J.N. and Gajdos, Z.K.Z. (2011). Genome-wide association studies: results from the first few years and potential implications for clinical medicine. *Annual Review of Medicine*, **62**, 11–24.

Hoggart, C.J., et al. (2003). Control of confounding of genetic associations in stratified populations. *American Journal of Human Genetics*, **72**, 1492–504.
Horne, B.D., et al. (2012). Pharmacogenetic warfarin dose refinements remain significantly influenced by genetic factors after one week of therapy. *Thrombosis and Haemostasis*, **107**, 232–40.
Hunter, D.J. (2005). Gene–environment interactions in human diseases. *Nature Reviews Genetics*, **6**, 287–98.
Hunter, D.J. and Kraft, P. (2007). Drinking from the fire hose—statistical issues in genomewide association studies. *The New England Journal of Medicine*, **357**, 436–9.
International HapMap Consortium (2007). A second generation human haplotype map of over 3.1 *million SNPs. Nature*, **449**, 851–62.
International HapMap 3 Consortium (2010). Integrating common and rare genetic variation in diverse human populations. *Nature*, **467**, 52–8.
International Human Genome Mapping Consortium (2001). A physical map of the human genome. *Nature*, **409**, 934–41.
International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
International SNP Map Working Group (2001). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, **409**, 928–33.
International Warfarin Pharmacogenetics Consortium, et al. (2009). Estimation of the warfarin dose with clinical and pharmacogenetic data. *New England Journal of Medicine*, **360**, 753–64.
Ioannidis, J.P.A., Thomas, G., and Daly, M.J. (2009). Validating, augmenting and refining genome-wide association signals. *Nature Reviews Genetics*, **10**, 318–29.
Ioannidis, J.P.A., Trikalinos, T.A., and Khoury, M.J. (2006). Implications of small effect sizes of individual genetic variants on the design and interpretation of genetic association studies of complex diseases. *American Journal of Epidemiology*, **164**, 609–14.
Ioannidis, J.P.A., Trikalinos, T.A., Ntzani, E.E., and Contopoulos-Ioannidis, D.G. (2003). Genetic associations in large versus small studies: an empirical assessment. *The Lancet* 361, 567–71.
Jorgensen, A.L., Fitzgerald, R.J., Oyee, J., Pirmohamed, M., and Williamson, P.R. (2012). Influence of CYP2C9 and VKORC1 on patient response to warfarin: a systematic review and meta-analysis. *PLoS One*, 7, e44064.
Jostins, L., et al. (2012). Host–microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature*, **491**, 119–24.
Kaikkonen, M.U., Lam, M.T.Y., and Glass, C.K. (2011). Non-coding RNAs as regulators of gene expression and epigenetics. *Cardiovascular Research*, **90**, 430–40.
Kamstrup, P.R., Tybjaerg-Hansen, A., Steffensen, R., and Nordestgaard, B.G. (2009). Genetically elevated lipoprotein(a) and increased risk of myocardial infarction. *Journal of the American Medical Association*, **301**, 2331–9.
Khoury, M.J., et al. (2011a). Current priorities for public health practice in addressing the role of human genomics in improving population health. *American Journal of Preventive Medicine*, **40**, 486–93.
Khoury, M.J., Davis, R., Gwinn, M., Lindegren, M.L., and Yoon, P. (2005). Do we need genomic research for the prevention of common diseases with environmental causes? *American Journal of Epidemiology*, **161**, 799–805.
Khoury, M.J., Gwinn, M., Clyne, M., and Yu, W. (2011b). Genetic epidemiology with a Capital E, ten years after. *Genetic Epidemiology*, **35**, 845–52.
Kingsmore, S.F., Lindquist, I.E., Mudge, J., Gessler, D.D., and Beavis, W.D. (2008). Genome-wide association studies: progress and potential for drug discovery and development. *Nature Reviews Drug Discovery*, **7**, 221–30.
Klann, T.S., Black, J.B., and Gersbach, C.A. (2018). CRISPR-based methods for high-throughput annotation of regulatory DNA. *Current Opinion in Biotechnology*, **52**, 32–41.

Kraft, P., et al. (2009). Beyond odds ratios—communicating disease risk based on genetic profiles. *Nature Reviews Genetics*, **10**, 264–9.
Kruglyak, L. (2008). The road to genome-wide association studies. *Nature Reviews Genetics*, **9**, 314–18.
Langmead, B. and Nellore, A. (2018). Cloud computing for genomic data analysis and collaboration. *Nature Reviews Genetics*, **19**, 208–19.
Lawlor, D.A., Harbord, R.M., Sterne, J.A.C., Timpson, N., and Davey Smith, G. (2008). Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Statistics in Medicine*, **27**, 1133–63.
Li, Y., Willer, C., Sanna, S., and Abecasis, G. (2009). Genotype imputation. *Annual Review of Genomics and Human Genetics*, **10**, 387–406.
Lohmueller, K.E., Pearce, C.L., Pike, M., Lander, E.S., and Hirschhorn, J.N. (2003). Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nature Genetics*, **33**, 177–82.
Loos, R.J.F. (2011). Genome-wide risk profiles—will they change your life(style)? *Nature Reviews Endocrinology*, **7**, 252–4.
Manolio, T.A., et al. (2009). Finding the missing heritability of complex diseases. *Nature*, **461**, 747–53.
Marchini, J., Cardon, L.R., Phillips, M.S., and Donnelly, P. (2004). The effects of human population structure on large genetic association studies. *Nature Genetics*, **36**, 512–17.
Marteau, T.M., et al. (2010a). Effects of communicating DNA-based disease risk estimates on risk-reducing behaviours. *Cochrane Database of Systematic Reviews*, 6, CD007275.
Marteau, T.M., et al. (2010b). Trial protocol: using genotype to tailor prescribing of nicotine replacement therapy: a randomised controlled trial assessing impact of communication upon adherence. *BMC Public Health*, **10**, 680.
McCarthy, M.I., et al. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics*, **9**, 356–69.
Metzker, M.L. (2010). Sequencing technologies—the next generation. *Nature Reviews Genetics*, **11**, 31–46.
National Center for Biotechnology Information (NCBI dbSNP Build 151). *dbSNP Database*. [Online] Available at: https://www.ncbi.nlm.nih.gov/SNP/snp_summary.cgi
Ng, S.B., et al. (2010). Exome sequencing identifies the cause of a mendelian disorder. *Nature Genetics*, **42**, 30–5.
Ng, S.B., al. (2009). Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*, **461**, 272–6.
Novembre, J. and Ramachandran, S. (2011). Perspectives on human population structure at the cusp of the sequencing era. *Annual Review of Genomics and Human Genetics*, **12**, 245–74.
Palmer, L.J. and Cardon, L.R. (2005). Shaking the tree: mapping complex disease genes with linkage disequilibrium. *The Lancet*, **366**, 1223–34.
Park, J.H., et al. (2010). Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nature Genetics*, **42**, 570–5.
Pharoah, P.D.P., Antoniou, A.C., Easton, D.F., and Ponder, B.A.J. (2008). Polygenes, risk prediction, and targeted prevention of breast cancer. *The New England Journal of Medicine*, **358**, 2796–803.
Phyo, A.P., et al. (2012). Emergence of artemisinin-resistant malaria on the western border of Thailand: a longitudinal study. *The Lancet*, **379**, 1960–6.
Price, A.L., Zaitlen, N.A., Reich, D., and Patterson, N. (2010). New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics*, **11**, 459–63.
Pussegoda, K. (2010). Exome sequencing: locating causative genes in rare disorders. *Clinical Genetics*, **78**, 32–3.

Reilly, M.P., et al. (2011). Identification of ADAMTS7 as a novel locus for coronary atherosclerosis and association of ABO with myocardial infarction in the presence of coronary atherosclerosis: two genome-wide association studies. *The Lancet*, **377**, 383–92.

Roberts, L., Davenport, R.J., Pennisi, E., and Marshall, E. (2001). A history of the Human Genome Project. *Science*, **291**, 1195.

Roman, T. S., and Mohlke, T.S. (2018). Functional genomics and assays of regulatory activity detect mechanisms at loci for lipid traits and coronary artery disease. *Current Opinion in Genetics & Development*, **50**, 52–9.

Romanens, M., et al. (2010). Improvement of cardiovascular risk prediction: time to review current knowledge, debates, and fundamentals on how to assess test characteristics. *European Journal of Cardiovascular Prevention and Rehabilitation*, **17**, 18–23.

Sandhu, M., Wood, A., and Young, E. (2010). Genomic risk prediction. *The Lancet*, **376**, 1366–7.

Sandhu, M.S., Debenham, S.L., Barroso, I., and Loos, R.J. (2008). Mendelian randomisation studies of type 2 diabetes: future prospects. *Diabetologia*, **51**, 211–13.

Sanseau, P., et al. (2012). Use of genome-wide association studies for drug repositioning. *Nature Biotechnology*, **30**, 317–20.

Tabor, H.K., Risch, N.J., and Myers, R.M. (2002). Candidate-gene approaches for studying genetic traits: practical consideration. *Nature Reviews Genetics*, **3**, 1–7.

Takeuchi, F., et al. (2009). A genome-wide association study confirms VKORC1, CYP2C9, and CYP4F2 as principal genetic determinants of warfarin dose. *PLoS Genetics*, 5, e1000433.

Talmud, P.J., et al. (2010). Utility of genetic and non-genetic risk factors in prediction of type 2 diabetes: Whitehall II prospective cohort study. *British Medical Journal*, 340, b4838.

Tatarunas, V., Lesauskaite, V., Veikutiene, A., Grybauskas, P., Jakuska, P., and Benetis, R. (2012). The combined effects of clinical factors and CYP2C9 and VKORC1 gene polymorphisms on initiating warfarin treatment in patients after cardiac valve surgery. *Journal of Heart Valve Disease*, **21**, 628–35.

Teare, M.D. and Barrett, J.H. (2005). Genetic linkage studies. *The Lancet*, **366**, 1036–44.

Teichert, M., et al. (2009). A genome-wide association study of acenocoumarol maintenance dosage. *Human Molecular Genetics*, **18**, 3758–68.

The 1000 Genomes Project Consortium (2010). A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–73.

The 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 56–65.

The ENCODE Project Consortium (2011). A user's guide to the Encylopedia of DNA Elements (ENCODE). *PLos Biology*, 9, e1001046.

Thomas, D.C. and Clayton, D.G. (2004). Betting odds and genetic associations. *Journal of the National Cancer Institute*, **96**, 421–3.

Thomas, D.L., et al. (2009). Genetic variation in IL28B and spontaneous clearance of hepatitis C virus. *Nature*, **461**, 798–801.

Tishkoff, S.A., et al. (2009). The genetic structure and history of Africans and African Americans. *Science*, **324**, 1035–44.

Triglyceride Coronary Disease Genetics Consortium and Emerging Risk Factors Collaboration, et al. (2010). Triglyceride-mediated pathways and coronary disease: collaborative analysis of 101 studies. *The Lancet*, **375**, 1634–9.

Trusheim, M.R., Berndt, E.R., and Douglas, F.L. (2007). Stratified medicine: strategic and economic implications of combining drugs and clinical biomarkers. *Nature Reviews Drug Discovery*, **6**, 287–93.

Uhlemann, A.C. and Fidock, D.A. (2012). Loss of malarial susceptibility to artemisinin in Thailand. *The Lancet*, **379**, 1928–30.

Visscher, P.M., Brown, M.A., Mccarthy, M.I., and Yang, J. (2012). Five years of GWAS discovery. *American Journal of Human Genetics*, **90**, 7–24.

Voight, B.F., et al. (2012). Plasma HDL cholesterol and risk of myocardial infarction: a mendelian randomisation study. *The Lancet*, **380**, 572–80.

Wacholder, S., Chanock, S., Garcia-Closas, M., El Ghormli, L., and Rothman, N. (2004). Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. *Journal of the National Cancer Institute*, **96**, 434–42.

Wacholder, S., Rothman, N., and Caporaso, N. (2000). Population stratification in epidemiologic studies of common genetic variants and cancer: quantification of bias. *Journal of the National Cancer Institute*, **92**, 1151–8.

Wald, N.J., Hackshaw, A.K., and Frost, C.D. (1999). When can a risk factor be used as a worthwhile screening test? *British Medical Journal*, **319**, 1562–5.

Ware, J.H. (2006). The limitations of risk factors as prognostic tools. *The New England Journal of Medicine*, **355**, 2615–17.

Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**, 661–78.

Wood, A.M. and Greenland, P. (2009). Evaluating the prognostic value of new cardiovascular biomarkers. *Disease Markers*, **26**, 199–207.

Wray, N.R., Yang, J., Goddard, M.E., and Visscher, P.M. (2010). The genetic interpretation of area under the ROC curve in genomic profiling. *PLoS Genetics*, 6, e1000864.

Xu, J., et al. (2009). Estimation of absolute risk for prostate cancer using genetic markers and family history. *Prostate*, **69**, 1565–72.

Yang, Q. and Khoury, M.J. (1997). Evolving methods in genetic epidemiology. III. Gene–environment interaction in epidemiologic research. *Epidemiologic Reviews*, **19**, 33–43.

Zhang, J., Chiodini, R., Badr, A., and Zhang, G. (2011). The impact of next-generation sequencing on genomics. *Journal of Genetics and Genomics*, **38**, 95–109.

Ziv, E. and Burchard, E.G. (2003). Human population structure and genetic association studies. *Pharmacogenomics*, **4**, 431–41.

SECTION 9

Prevention and control of public health hazards

9.1 Tobacco
Tai Hing Lam and Sai Yin Ho

9.2 Substance use and misuse: considerations on global public health
9.3 Giang Le Minh and Steve Shoptaw Alcohol
Robin Room

9.4 Injury prevention and control: the public health approach
9.5 Corinne Peek-Asa and Adnan A. Hyder Interpersonal violence
Rachel Jewkes

9.6 Collective violence: war
Barry S. Levy

9.1

Tobacco

Tai Hing Lam and Sai Yin Ho

Introduction to tobacco use

Please respond to the following statements before reading further:

About 1 out of 20 smokers (i.e. 5% or less), if they continue to smoke, will eventually be killed by smoking.
To reduce deaths caused by smoking, preventing young people from smoking is the most important action.

Response: strongly agree/agree/disagree/strongly disagree.

Most people will choose agree or strongly agree for these two statements. However, in fact, smoking kills one in two smokers and quitting smoking prevents more deaths more quickly.

Tobacco is the world's most important and most avoidable public health problem. Tobacco killed 100 million people in the last century, mostly in developed countries, and could kill 1 billion this century, mostly in low- and middle-income countries (LMICs) especially China. Currently, tobacco kills 5.4 million people every year. Unless urgent action is taken, tobacco will kill 8 million per year by 2030, 80% of whom will be in developing countries. 'Of the more than 1 billion smokers alive today, around 500 million will be killed by tobacco' (WHO 2008). But millions of lives have been saved since the 1960s, mostly in developed countries (notably the United States and United Kingdom) which have progressively and successfully implemented effective tobacco control policies.

Tobacco control is the most remarkable example of global public health actions against non-communicable diseases (NCDs). It is unique in having an international public health treaty, the World Health Organization (WHO) Framework Convention on Tobacco Control (FCTC), which has been

ratified by 176 parties covering 88% of the world's population (as of 30 November 2012). It is one of the most rapidly ratified United Nations (UN) treaties of all time but two of the world's largest countries have not ratified the FCTC, namely the United States and Indonesia (WHO 2012a).

In the September 2011 Political Declaration of the UN on NCDs, tobacco use tops the list of four major risk factors that also include the harmful use of alcohol, unhealthy diet, and physical inactivity (UN 2011).

Globally, the epidemic of tobacco-induced diseases, mortality, and related social and economic burden is still expanding. Using the host–agent–vector model of infectious diseases, the tobacco epidemic is an epidemic of the agent, that is, tobacco. It is sometimes also described as an epidemic of smoking and its harms to others. These two descriptions tend to blame the agent (tobacco) and the host (smokers) or focus on the victims (passive smokers).

However, some might also consider that the tobacco epidemic is really a tobacco industry epidemic. The agent is produced massively, its use by the host is heavily promoted, and control measures are aggressively obstructed by the vector, the tobacco industry. There will be no tobacco epidemic without the tobacco industry. Tobacco control must therefore target the vector. Helping the hosts, that is, smokers, to quit smoking and preventing non-smokers, especially young people and women, from starting smoking are also important, but the impact will be limited to those who can be and are being helped effectively. Smoking cessation campaigns and services cannot produce substantial public health benefits when the tobacco industry is not under strict control. The success of tobacco industry control and other control measures will lead to more quitting, less relapse, and less uptake of smoking.

Effective control measures are stated in the FCTC and translated into six policies of WHO's MPOWER package as follows (WHO 2008):

Monitor tobacco use and prevention policies.
Protect people from tobacco smoke.
Offer help to quit tobacco use.
Warn about the dangers of tobacco use.
Enforce comprehensive restrictions on tobacco advertising, promotion, and sponsorship.
Raise taxes on tobacco.

Note that these six policies are not ranked, and all six are needed as a comprehensive approach against the tobacco industry and tobacco use. However, their implementation varies greatly even among countries that have ratified and are parties of the FCTC. WHO evaluates the progress and publishes global status reports regularly, which is part of the M (‘monitor’) under MPOWER.

Tobacco kills up to two in three smokers

At least one in two smokers killed

It is important to understand and warn about the dangers of tobacco use: first, the absolute risk of death from all causes combined, and second, the large number of resultant deaths globally, regionally, and locally. This absolute risk, that a smoker will have a 50% chance of being killed by tobacco, is easy to understand and more relevant to individuals than relative risks (RRs) or number of deaths. RRs only tell the exposed people that they have higher risks than those who are unexposed, but not the exact probability of harm, the absolute risk. Large RRs do not mean large absolute risks if the incidence rate in the unexposed is low. Telling smokers that millions have died from tobacco elsewhere might be alarming to some, but difficult to comprehend to others.

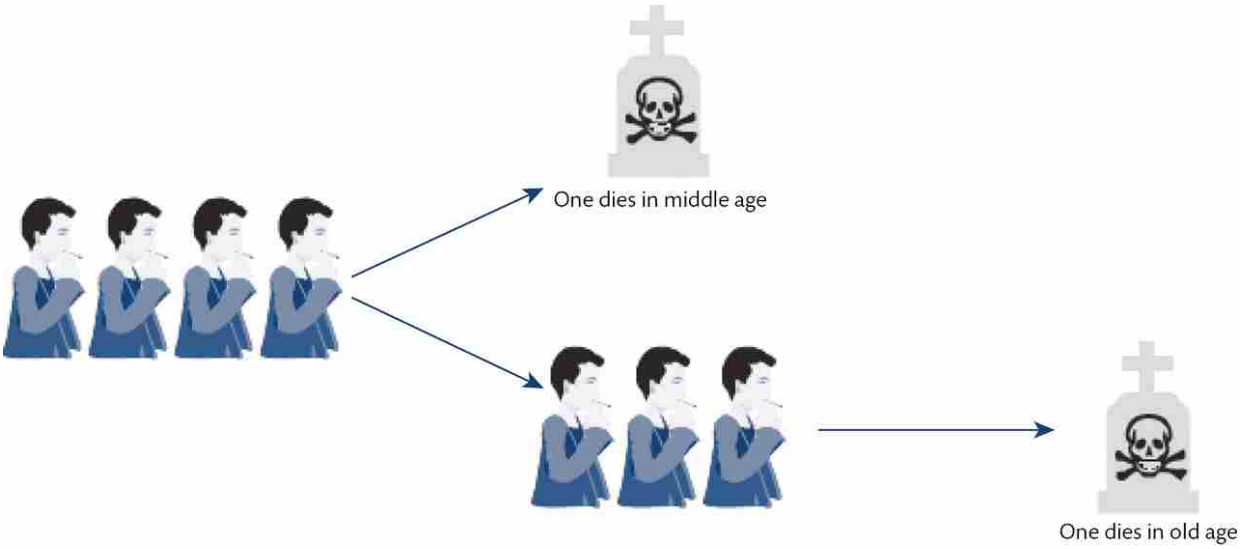


Fig. 9.1.1 One in two smokers will be killed by smoking.

The WHO statement ‘Tobacco kills up to one in every two users’ (WHO 2008) is derived from Peto, who stated ‘about half of teenagers who keep smoking steadily will eventually be killed by tobacco (about a quarter in old age plus a quarter in middle age)’ (Peto 1994) (see Fig. 9.1.1). This was based on the translation of the RR of two in the prospective British Doctors Study by Doll and Peto with 40 years of follow-up (Doll et al. 1994). This is a unique example of how an RR can be translated into absolute risk for public health actions and communication of risks.

From relative risk to absolute risk

The RR is the absolute risk of disease or mortality in the exposed group (smokers) divided by that in the unexposed (non-smokers). Translating the RR into an absolute risk in the exposed without knowing the absolute risk in the unexposed is only feasible for total or all-cause mortality. This is based on the attributable fraction (AF) in the exposed (AF = (RR − 1)/RR). With an RR of two for total mortality in smokers compared with non-smokers, the AF in smokers is (2 − 1)/2 = 0.5. That means of all smokers who died, half were due to smoking. Because all smokers will eventually die, the denominator of all smoker deaths can be turned into all current smokers who are alive, and hence of all current smokers, 50% or one in two will eventually be killed by smoking (Lam 2012).

Up to two in three smokers killed

Previously reported RRs could have been underestimated. The RR is based on a comparison of mortality rates of current smokers and non-smokers after some years of follow-up. Current smokers recruited at baseline are survivors as their weaker counterparts might have died. Non-smokers can be either never smokers or former smokers. The latter could have quit smoking while still being healthy (healthy quitters), or they had to quit because of ill health (ill quitters). Hence, RR can be underestimated as current smokers tend to be healthier (lower mortality rate in the numerator), and non-smokers can include some ill quitters (greater mortality rate in the denominator). This bias is particularly strong when a cohort includes older people.

Another bias could be due to insufficient duration of follow-up, especially for diseases that take many years to develop. A longer follow-up would yield more deaths and allow subgroup analysis. The results from Doll and Peto's study, after 50 years of follow-up, showed that if smoking started at a young age, two in three smokers (RR of 3) will be killed by smoking (Doll et al. 2004). That two in three very young smokers will be killed by smoking is particularly relevant for controlling smoking in children and youth, but this statement has not yet been publicized widely. Latest findings based on large contemporary cohorts in the United States and United Kingdom also revealed all-cause mortality RRs approaching 3 in male and female current smokers compared with never smokers (Jha et al. 2013; Pirie et al. 2013; Thun et al. 2013).

Not many cohort studies, especially those from LMICs, have been followed up for as long as the British doctor cohort studies and hence, their RRs of mortality due to smoking tend to be underestimated. In a small cohort of 1,696 people followed up for 20 years, Lam et al. (1997) found an RR of about 2, which is similar to that from Doll and Peto but greater than those of earlier (Yuan et al. 1996) and more recent cohort studies from China (Elliott et al. 2009) with shorter duration of follow-up.

The total exposure to tobacco in smokers varies by countries and cohort studies. Another explanation for the smaller RRs of cohort studies in China and other LMICs at an earlier stage of the tobacco epidemic is that their exposure to tobacco could be lower. Smokers in these cohort studies tended to have started smoking at an older age and could not afford to smoke as much as those in high-income countries.

A recent meta-analysis by Gellert et al. (2012) on smoking and all-cause mortality focused on elderly people (aged 60 or above). Based on 17 cohort studies from seven countries, the RR was 1.85 in current smokers. This means that the one in two deaths is still applicable to older smokers.

Public health applications

Such epidemiological findings can be used for more effective public health advocacy, education, and publicity campaigns. The AF of 50% can be translated into the number of people in a specific population who will be killed by smoking. If there are 4 million adults, and the prevalence of smoking is 50%, the number of people to be killed is 4 million × 0.5 × 0.5 = 1 million. The same can be applied to children and young people. A statement that among today's 4 million children in our country/city/district, about 1 million will be killed by tobacco, if the current prevalence of smoking is not reduced, is more striking. On the other hand, such figures can be used positively, saying that prevention of smoking and promotion of quitting can save millions of lives. As there is a trend of increased smoking in children and youth in many LMICs, if the RR of 3 is used, the estimates will be even more alarming.

For unknown reasons, the 'one in two deaths' message is not used in many websites that help smokers quit, including the US Centers for Disease Control and Prevention (CDC) website. The US Surgeon General's (USSG's) report (US Department of Health and Human Services (2004) does not include total mortality. Existing pictorial warnings in cigarette packs of more than 50 jurisdictions also do not show or highlight this risk. Even recent warnings in Canada and Australia in plain packaging do not have it (Physicians for a Smoke-Free Canada n.d.). The 'one in two risk' has seldom been used as a key campaign message until recently. In 2011, Ireland was the first country to use it for a major social marketing campaign resulting in wide media coverage and great increases in calls to the quitline and demand for smoking cessation services (Howell and HSE Social Marketing Working Group on Tobacco 2012). The Hong Kong Council on Smoking and Health launched a campaign on 'one in two' with a video for TV in October 2012. For public health information and education, the mortality risk of 'one in two' to 'two in three' should be highlighted consistently and widely (Lam 2012). For randomized controlled trials on smoking cessation advice or

counselling, there is also no report about using such absolute risk warning, and new trials are warranted.

Clinical applications

The AF of 50% can also be used to motivate clinicians or other healthcare professionals (HCPs) to be more active in helping smokers to quit. If they can help two smokers to quit, they would have saved one life. However, the AF does not tell when the smokers will be killed by smoking. The message to the smokers can be simple: 'If you and your good friend continue to smoke, one of you will be killed by smoking prematurely. We cannot predict who and when, but the premature death can come quickly and suddenly, and when it comes, it will be too late. Quit smoking now.'

Note that AF in the exposed can be used in clinical epidemiology for individuals with a disease and is exposed to the risk (causal) factor of the disease. If the RR of lung cancer due to smoking is 20 (Aubin et al. 2011), AF in the exposed is (20 − 1)/20 = 95%. Clinically, one can tell the smoker who has lung cancer that the probability that his lung cancer is due to his smoking is about 95%. This probability will be greater if the smoker has smoked heavily for a long duration, as the RR refers to an 'average' smoker, and there is a dose–response relation between RR and the amount of exposure (i.e. pack-years).

Specific harms of tobacco: active smoking

Causation versus association

Of the four major risk factors, tobacco use is the one confirmed to be a causal factor of the largest number of diseases. Hence, the term 'cause' or 'causal factor' should be used for smoking instead of risk factor, which implies association rather than causation. Similarly, the terms 'tobacco-induced diseases' or 'tobacco-caused diseases' are clearer than terms such as 'tobacco-related' or 'smoking-related' diseases. The USSG reports are the longest and most comprehensive series of reviews, with the latest on active smoking published in 2004 (US Department of Health and Human Services 2004). Updated reports are published regularly often with new harms added to the previous lists. Globally, different review panels use quite similar

criteria for causal inference. Despite the remarkable consensus on many diseases, disagreements remain as to whether the evidence is sufficient (or only suggestive) to reach a conclusion on causation for some diseases. More evidence to prove or confirm causation of additional diseases is of scientific interest and can add to the total disease burden estimates, but is not essential for tobacco control.

International Agency for Research on Cancer

The International Agency for Research on Cancer (IARC), a WHO agency, is the world authority in the evaluation of carcinogenic risks to humans. The 1986 and 2002 IARC monographs classify tobacco smoking and tobacco smoke as a Group 1 carcinogen, meaning that there is sufficient evidence that it can cause cancer in human. The most updated and comprehensive reviews released by IARC in 2009 conclude that tobacco smoking is the single largest cause of cancer worldwide, and 'Tobacco smoking causes cancers of the lung, oral cavity, naso-, oro- and hypopharynx, nasal cavity and accessory sinuses, larynx, oesophagus, stomach, pancreas, colorectum, liver, kidney (body and pelvis), ureter, urinary bladder, uterine cervix and ovary (mucinous), and myeloid leukaemia. Also, a positive association has been observed between tobacco smoking and cancer of the female breast. For cancers of the endometrium (post-menopausal) and of the thyroid, there is evidence suggesting lack of carcinogenicity' (IARC 2012). Note that two important cancers are added, that is, cancer of the colon and ovary, for which the evidence is now considered sufficient, but that for breast cancer is still limited (Secretan et al. 2009; IARC 2012). The report also concludes that bidi (hand-rolled cigarette commonly used in South Asia) smoking increases the risk for cancers of the oral cavity, oropharynx, hypopharynx, larynx, lung, oesophagus, and stomach, and smokeless tobacco causes cancer of the oral cavity, oesophagus, and pancreas.

The 2010 USSG report

The 2010 USSG report focused on mechanisms rather than causal inference. The major conclusions are: (1) no risk-free exposure level; (2) tobacco smoke causes adverse health outcomes through DNA damage, inflammation, and oxidative stress; (3) risk is related to the duration and level of exposure;

(4) exposures are due to powerful nicotine addiction in the brain; (5) low levels of exposure, including second-hand smoking, lead to rapid and sharp increases in endothelial dysfunction and inflammation, which are implicated in acute cardiovascular events and thrombosis; and (6) evidence is insufficient that modifications in tobacco production can reduce risks. Moreover, it has important new conclusions that cigarette smoking produces insulin resistance and chronic inflammation, which can accelerate macro- and microvascular complications. Hence, for diabetes: (1) smoking increases the risk of type 2 diabetes; (2) people with diabetes need more insulin; and (3) smokers with diabetes have more complications including heart and kidney disease, amputation, retinopathy, and peripheral neuropathy. The report has a very readable booklet for the public that is highly recommended (Centers for Disease Control and Prevention 2010).

Discrepancies and potential confusions

The conclusions on causal relations for cancers from the IARC are mostly similar to the 2004 USSG report. The most notable exceptions are that the 2002 IARC publication has included liver, nasal cavities and nasal sinuses, and nasopharyngeal cancer as causally related to smoking and cancers of the *tongue and lip* are specifically stated under oral cavity. The 2004 USSG report, which appeared later, does not. The 2012 IARC report has now added that smoking causes colon and ovarian cancer. Another notable discrepancy is that although the statement that smoking can cause erectile dysfunction (ED) is often shown in pictorial or other health warnings on cigarette packs, the 2004 USSG report considers that the evidence for causation is only suggestive but not sufficient.

Important new results

Since the 2004 USSG report, there has been no new meta-analysis or systematic review on ED. New results based on cross-sectional surveys have shown a consistent association between smoking and ED in Australia (Millett et al. 2006), Boston (Kupelian et al. 2007), Hong Kong (Lam et al. 2006), and China (He et al. 2007), and a Swiss survey showed that smoking was associated directly with premature ejaculation and indirectly with ED in young men (Mialon et al. 2012). Results from a randomized controlled trial

showed that quitting smoking improved ED (Chan et al. 2010). A smoking cessation clinic study showed that quitters had better erectile function than those who relapsed (Harte and Meston 2012). All three meta-analyses in 2007 concluded that smoking is associated with an increased risk of active tuberculosis (TB) disease and mortality (mostly due to reactivation of previous TB infection) and the evidence is strong and consistent (Bates et al. 2007; Lin et al. 2007; Slama et al. 2007). A mathematical model predicted that 'smoking would produce an excess of 18 million tuberculosis cases and 40 million deaths from tuberculosis between 2010 and 2050, if smoking trends continued along current trajectories. The effect of smoking was anticipated to increase the number of tuberculosis cases by 7% (274 million v 256 million) and deaths by 66% (101 million v 61 million), compared with model predictions that did not account for smoking' (Basu et al. 2011). The causal relation also suggests that tobacco control must be integrated with TB control in TB endemic countries. Public health advocacy has to be based on precise and sometimes conservative estimates of disease burden. With new results, systematic reviews by different authorities will come up with revised or updated conclusions. Public health professionals need to keep updated and critically appraise the original sources of information.

Specific harms of tobacco: passive smoking

Numerous reviews

No other substance like second-hand smoke (SHS) has been subject to so many systematic reviews within such a short period since the first two reports of an association with lung cancer were published in 1981 (Hirayama 1981; Trichopoulos et al. 1981). The massive and aggressive criticisms from the tobacco industry and its scientists were also unprecedented.

The 2006 USSG report

The 2006 USSG report is the most comprehensive review focusing on SHS (US Department of Health and Human Services 2006). It concludes that the evidence is sufficient that SHS causes: (1) premature death and disease in children and adults who do not smoke; (2) sudden infant death syndrome, acute respiratory infections, ear problems and more severe asthma in

children, and smoking by parents causes respiratory symptoms and slows lung growth in their children; and (3) immediate adverse effects on the cardiovascular system, coronary heart disease (CHD) (25–30% increased risk) and lung cancer (20–30% increased risk) in adults. The conclusions that have great public health implications are: (1) there is no risk-free level of exposure to SHS; (2) many millions of children and adults are still exposed; and (3) eliminating smoking in indoor spaces fully protects non-smokers from SHS, but separating smokers from non-smokers and cleaning the air and ventilating buildings cannot eliminate exposure. The list of diseases with suggestive but insufficient evidence is much longer, and most notably includes childhood cancer (leukaemia, lymphoma, and brain tumour), onset of childhood asthma, and adult-onset asthma, breast cancer, nasal sinus cancer, stroke, atherosclerosis, and chronic obstructive pulmonary disease (COPD).

The 2002 IARC monograph

The 2002 IARC monograph concludes that involuntary smoking is a Group 1 carcinogen based on more than 50 studies in many countries, causing lung cancer with excess risk from spousal exposure of 20% for women and 30% for men (Aubin et al. 2011). Excess risk from workplace exposure is 16–19%. The US Cal/EPA 2006 report concludes that SHS can induce asthma and can cause premenopausal breast cancer with an excess risk of 80–120% and these are different from the 2006 USSG report. The 2009 IARC report has added that parental smoking can cause hepatoblastoma in offspring, that there is limited evidence for cancer of larynx and pharynx, and the evidence for female breast cancer remains inconclusive.

The estimation of deaths from SHS is less straightforward. It depends on what diseases are to be included, and what criteria (sufficient or suggestive evidence for causation) or reviews or meta-analyses are being used. Reliable SHS exposure data are not available in many places, and definitions of exposure vary in different surveys, which may not be based on a sample representative of a whole country. The WHO Global Burden of Disease Study estimates that SHS has caused 603,000 premature deaths in never smokers in 2004, including 166,000 from lower respiratory infections, 1,100 from asthma in children, 35,800 from asthma, 21,000 from lung cancer, and 379,000 from ischaemic heart disease in adults. Of all deaths from SHS,

28% occur in children and 47% in women. The report notes that current smokers may be especially susceptible because of high and close exposure or existing smoking-caused diseases (Oberg et al. 2011).

National or local estimates of mortality in non-smokers due to SHS are critically important for advocacy for legislation to ban smoking completely. The 2006 USSG report estimates that in 2005, SHS killed over 3,000 Americans from lung cancer, 46,000 from CHD, and 430 newborns from sudden infant death syndrome. The 2007 China Tobacco Control report estimates that SHS killed more than 100,000 Chinese non-smokers. In Hong Kong, the results that SHS killed 150 catering workers from lung cancer and CHD (Hedley et al. 2006), and 1,324 people (from stroke, heart disease, chronic lung disease, and lung cancer) each year (McGhee et al. 2005), had been used successfully in campaigns leading to the 2006 law amendments extending the total ban of smoking to all indoor workplaces, restaurants and bars, and many other public places. New results on SHS will contribute important new evidence for causal inference for asthma in children (Burke et al. 2012), stroke (Oono et al. 2011), TB disease (Lin et al. 2007; Slama et al. 2007; Leung et al. 2010), COPD (Yin et al. 2007; He et al. 2012), and acute respiratory symptoms (Monto and Ross 1978). The uncertainty about SHS and breast cancer remains (Pirie et al. 2008).

Adverse effects of SHS in current smokers

There is a paucity of data on the adverse effects from SHS in current smokers, mainly because of the exclusion of current smokers in most, if not all, studies on SHS. In Hong Kong, SHS exposure was reported to be strongly associated with increased acute respiratory symptoms and recent outpatient service use in adult current smokers (Lam et al. 2005) and to be associated with increased risks for persistent respiratory symptoms among adolescent current smokers (Lai et al. 2009). The main purpose of banning smoking completely in indoor workplaces is to protect workers, which is an occupational health issue. Opponents and the tobacco industry have advocated for smoking areas or smoking rooms, but workers who need to serve or clean inside smoking rooms are heavily exposed to SHS. A further argument is that smoker-employees can be deployed to work in smoking areas. Results on additional harms of SHS to active smokers can be used as a counterargument. Also, both adult and youth smokers should be warned to

avoid SHS, which may motivate some to support a totally smoke-free legislation and to quit smoking.

Chemicals and carcinogens

Tobacco smoke

Tobacco smoke is a very complex mixture containing more than 7,000 chemical compounds (Rodgman and Perfetti 2009). Hundreds of these compounds are toxic and at least 69 are carcinogens. The list is expected to further increase. Also new additives have been and will continue to be added to cigarettes by the tobacco industry. These include various flavours (including menthol and fruits), and herbs or herbal medicine to attract special groups of smokers, such as children and youth, and people who believe in herbal medicine (Eriksen et al. 2012). The effectiveness of using such information to warn about the harm of smoking needs to be evaluated. The main problem is that smokers do not feel any acute toxic effects from such chemicals because the level of exposure to each compound is low.

Second-hand smoke

SHS is a complex mixture of mainstream smoke (smoke inhaled and exhaled by the smoker) and side-stream smoke (from the burning end of the cigarette). Many of the toxic substances have a higher concentration in side-stream smoke than mainstream smoke, and it has often been mistakenly publicized that SHS is more harmful than active smoking. In terms of toxicity, this is not true because the toxic substances in SHS are greatly diluted by the ambient air before inhalation by the non-smokers. Hence, the exposure level of various compounds from SHS in a non-smoker is much lower than that in a smoker. In terms of the number of people affected to varying extents from mild irritation to deaths due to lung cancer or CHD, this statement can be true because of the widespread SHS exposure in non-smokers, and the absolute number of non-smokers, including children, is always greater than the number of smokers. Public health professionals should be mindful that messages are based on strong evidence and are effective.

Third-hand smoke

Third-hand smoke (THS) refers to the residual tobacco smoke pollutants on dust, surfaces, clothes, and human body after SHS has cleared. It can be re-emitted as gases as well as react with other pollutants such as nitrous acid to form tobacco-specific nitrosamines, some of which are carcinogenic (Burton 2011). Although the phenomenon of smelling residual smoke is not new, the coining of THS (Winickoff et al. 2009) has triggered much research and public interest. THS is accumulated with each cigarette smoked and remains for weeks. Young children are particularly at risk of exposure as they explore their home environment. However, no studies have linked THS exposure to physical harm and more studies on the chemistry and health risk of THS are warranted (Matt et al. 2011).

Risk perception of the public

The risk perception and tolerance of the public is often irrational and their resulting actions are not proportional to the magnitude of risks. Most people have zero tolerance to fewer toxic contaminants in foods, such as malachite green in fish, but relatively few are aware of the chemicals and carcinogens in tobacco smoke. More awareness of the harms of SHS has led to the rapid decline in tolerance to SHS. Non-smokers take various actions to avoid exposure and to complain about smokers who smoke near them. The past few decades have witnessed a change from total acceptance of smoking to zero tolerance among many non-smokers, especially in countries with comprehensive smoking bans.

The protection of innocent non-smokers, including children and women and employees, has been the major driver for smoke-free legislation. In China and many LMICs, awareness is low and even non-smoking physicians are not concerned about their own exposure to SHS (Lam et al. 2011). Tobacco control advocates need to learn from successful experiences in using SHS and in comparing the risk of tobacco smoke with other toxic compounds to change risk perception and motivate actions from non-smokers.

Although the harm of THS has yet to be shown, parents who are aware of that potential risk are more likely to adopt home smoking bans (Winickoff et al. 2009). This represents an opportunity to promote a totally smoke-free

environment in the private home, the last frontier in the battle of tobacco control.

Warn about the dangers

Numerous surveys reveal that the knowledge about tobacco harms is often superficial. The earliest health warning required by governments or law was mostly a small and inconspicuous statement that smoking is harmful to health on tobacco advertisements (started in 1971 in the United States) or cigarette packs. Unfortunately, this is still the situation in many countries such as China.

Canada was the first country to require pictorial health warnings by law on cigarette packs in 2001. Its success has spurred many countries to follow suit. The Canadian Cancer Society 2012 report showed that 63 countries or jurisdictions and territories, covering 40% of the world's population, require pictorial warnings (Canadian Cancer Society 2012).

The FCTC Article 11 requires that health warnings 'should be 50% or more of the principal display areas but shall be no less than 30%' (WHO 2003) and may be in the form of, or include, picture warnings. Many countries have done more than this minimum, with Australia having some of the largest warnings in the world at 75% of the package front and 90% of the back. From December 2012, Australia has also implemented plain packaging (Fig. 9.1.2) to prohibit tobacco company colours, logos, and design elements on the brand part of the package, and to standardize the shape and format of the package. The Constitutional challenges launched by the tobacco companies against the Australian Government have been rejected by the High Court of Australia. The United Kingdom was the second country to implement these measures in 2016. Not all governments have the political will or the legal and other resources to defend against these. Nevertheless, the Australian success has served as a role model and the Australian court ruling will help governments, at least in common law countries, to defend and fight against new legal challenges.

Strong and increasing evidence shows that pictorial warnings are highly cost-effective in raising awareness and reducing tobacco use. The bigger the size of the warnings, the more effective they are. Some countries have also switched to the use of shocking images of diseased or damaged organs or dead bodies, including dead foetuses or babies. It is also important to start

using the most striking pictures, have a set of more warnings (Australia has now a set of 14 for plain packaging), and to require frequent rotations, with the printing costs totally borne by the tobacco companies and with minimal delay (as tobacco companies will produce many more packs before the law is passed and argue that they need many months or years to print new packs). The law should provide the health ministry the authority to change or update the pictures and warnings without having to amend the law. The warnings must be based on evidence that is considered sufficient for causation by most authorities. Information on pictorial warnings from all countries can be obtained from the Tobacco Labelling Resource Centre (n.d.) and the WHO warnings database (n.d.).

The tobacco industry has been fighting strongly in every country against such warnings, citing comments from smokers that they will take no notice of such warnings, and argues that printing such warnings increases production costs and affects the tobacco companies' rights to use their intellectual property to show their brand names and logos, jeopardizing freedom of trade and expression. In the United States, despite the Food and Drug Administration Rule in June 2011 requiring pictorial warnings to cover 50% of the package front and back as from September 2012, the implementation is on hold due to the successful legal challenge of the tobacco industry. The outcome of future legal proceedings is uncertain.

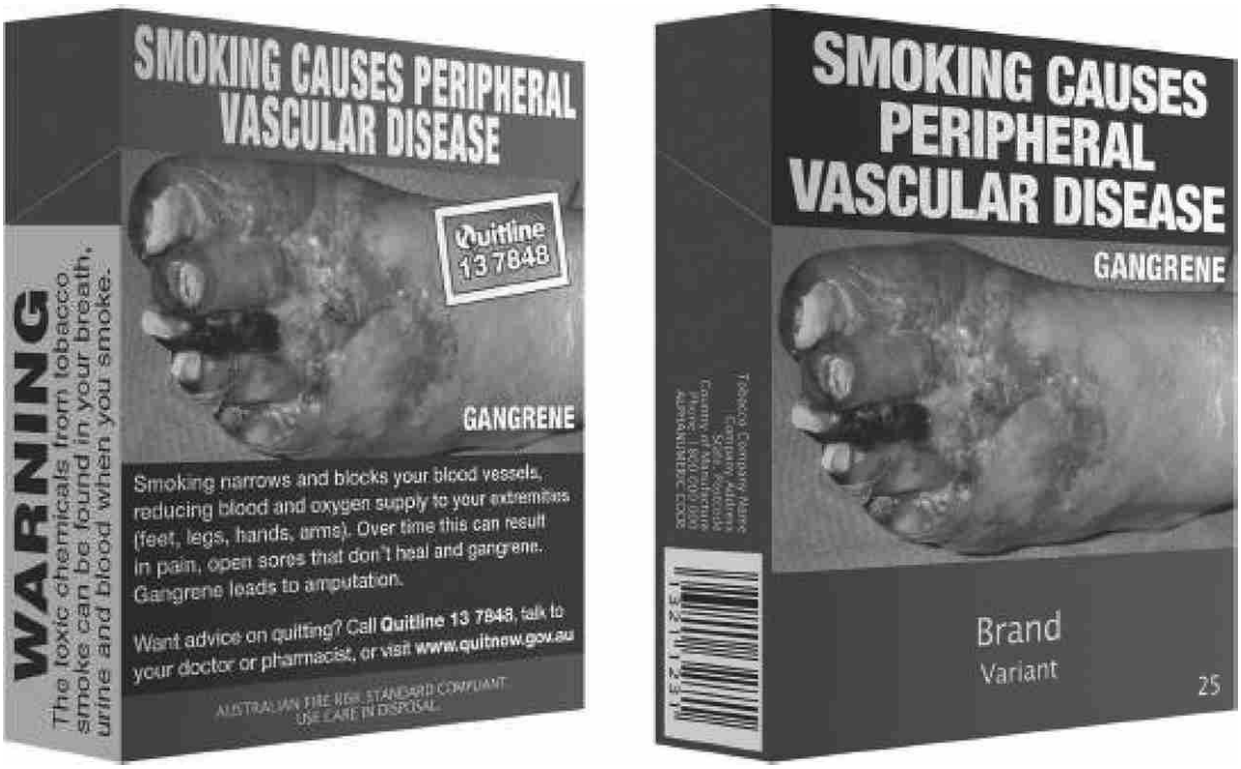


Fig. 9.1.2 Plain packaging and pictorial health warnings on Australian cigarette packs.
Copyright © Commonwealth of Australia. Reproduced with permission.

FCTC Article 11 (a) requires parties to ensure that 'tobacco product packaging and labelling do not promote a tobacco product by any means that are false, misleading, deceptive or likely to create an erroneous impression', and that 'These may include terms such as "low tar," "light," "ultra-light," or "mild" '(WHO 2003). Many countries still have tobacco products not complying with this article. The development and promotion of Chinese-style cigarettes as safer and healthier by the Chinese tobacco industry is particularly objectionable.

Although research has shown clearly that there is no completely safe form of tobacco, the tobacco industry continues to promote 'safer' cigarettes. There is evidence to show that such efforts have gained traction. The Global Adult Tobacco Survey in China (Chinese Center for Diseases Control and Prevention 2012) shows that more professionals and those with higher education believe that 'light' or 'low tar' cigarettes are safer. Because of the traditional belief in herbal medicine, adding such substances into cigarettes has been a major new effort by the Chinese tobacco industry to promote safer Chinese-style cigarettes.

Media campaigns

Mass media campaigns can be highly effective to raise awareness, motivate smokers to quit, and prevent young people from starting smoking. The effectiveness increases with the intensity of the campaigns, as well as the content and messages. Well-planned and sustained TV campaigns can reach most people, resulting in behavioural changes. Many antismoking videos have been systematically evaluated and confirmed to be effective in developed countries. However, the *WHO Report on the Global Tobacco Epidemic, 2011* showed a paucity of mass media campaigns against tobacco in many countries, especially in LMICs (Eriksen et al. 2012). The high costs of producing videos and broadcasting them on TV may be a factor, but could be overcome through global sharing of videos that have been shown to be effective, and by using alternative media such as YouTube and mobile phones. Further research, development, and evaluation are warranted.

Media campaigns are also needed to solicit public support for legislation and other government actions. Even when there is majority support for tax increases and smoke-free areas, objections from a small percentage of the population, sometimes leveraged on by the tobacco industry, can put great pressures on governments and legislators. Media campaigns to solicit public support need to be justified strongly by warnings that emphasize how many people, especially young people, will be killed by smoking, and such justification can further raise awareness.

Alternative ways to collaborate with the media should also be explored. Press conferences on local study results usually attract much media coverage, free of charge. The WHO's World No Tobacco Day is an annual event that can be linked to other broad-based activities that can garner substantial media reporting and commentaries. International news, for example, on important findings from medical journals, or on resolutions from the WHO or the FCTC can trigger the interest of local mass media. Public health advocates need to learn what the media want and how to work with them. For sustained effects, media campaigns need to be well-planned and repeated, with updated results and new activities.

Monitoring tobacco use and prevention

Prevalence of smoking

Globally, about 20% of all adults smoke cigarettes, consuming 5.9 trillion cigarettes in 2009. Even if smoking prevalence remains the same, tobacco consumption as a whole will increase due to the rising world population, especially in LMICs. China is the world's largest producer and consumer of cigarettes, accounting for 38% of global consumption. Because of its huge population and high smoking prevalence, China has the largest number of smokers (adults and youth) and smoking-induced deaths. The second to fifth highest cigarette-consuming countries are the Russian Federation, United States, Indonesia, and Japan.

Monitoring tobacco use is fundamental to our understanding of the extent of the epidemic, assessing the current and projecting the future burden of disease, and evaluating the effectiveness of prevention polices. The *Tobacco Atlas*, 4th edition (Eriksen et al. 2012), shows comprehensive world tables including prevalence in men and women, and in boys and girls. Sex-specific data are important since males and females have great differences in tobacco use. Time-trend data will become available for more countries. Interpretation of trend data where available should take into account variations in survey methods over time, and sampling errors.

In 1998, the WHO, US CDC, and Canadian Public Health Association initiated the Global Tobacco Surveillance System (GTSS) (Warren et al. 2009a) to help countries monitor tobacco use, using standardized methods in school-based and household surveys. The Global Youth Tobacco Survey (GYTS), Global School Personnel Survey (GSPS), Global Health Professions Student Survey (GHPSS), and Global Adult Tobacco Survey (GATS) started in 1999, 2000, 2007, and 2007 respectively (Warren et al. 2009b). The surveys are in progress, with greatest coverage for the GYTS as it started early. The GATS is a nationally representative household survey of men and women aged 15 or older. It includes: (1) tobacco use; (2) knowledge attitudes and perceptions; (3) cessation; (4) SHS exposure; (5) economics; (6) media; and (7) background characteristics. These surveys also provide essential data for monitoring prevention policies.

Other tobacco use

The tobacco epidemic started in the United Kingdom and United States with the invention of machines (in 1881 in the United States) for mass production of cigarettes, and the aggressive advertising and promotion of cigarettes by

the tobacco companies, before the serious health hazards were widely known. Tobacco use, however, also includes kreteks (Indonesia), roll-your-own cigarettes (Europe and New Zealand), bidis (South Asia, mainly India), pipes (worldwide), sticks (Papua New Guinea), water pipes (North Africa, the Mediterranean region, and parts of Asia), cigars (worldwide), and various forms of smokeless tobacco, including chewing tobacco (worldwide), moist snuff (Scandinavia and United States), dissolvable smokeless tobacco products (high-income countries), and dry snuff (Europe) (Eriksen et al. 2012). Since the mid-2010s, e-cigarettes or vaping has become increasingly prevalent, with 41 million users worldwide in 2018, up from 7 million in 2011. The long-term effects on health are currently uncertain.

The International Tobacco Control Policy Evaluation Project

The International Tobacco Control (ITC) Project is the first international cohort study on the psychosocial and behavioural impact of key national level policies of the FCTC (International Tobacco Control Policy Evaluation Project n.d.). ITC includes 20 countries with 50% of the world's population and 70% of the world's tobacco users and covers:

Health warning labels and package descriptors
Smoke-free legislation
Pricing and taxation of tobacco products
Communication and education
Cessation
Tobacco advertising and promotion.

Stages of tobacco epidemic

The patterns of smoking and tobacco-induced diseases, particularly lung cancer, vary greatly in different countries. Lopez et al. (1994) defined four stages of the tobacco epidemic, based on developed countries like the United Kingdom, United States, and Australia. Smoking prevalence rose rapidly since the 1900s, first in men peaking in the 1950s, followed by women about two decades later reaching a lower peak, and then declined since the 1960–1970s, after large numbers of smokers stopped. Tobacco deaths in men lagged this, peaking about three decades later in the 1980s and then declined

as a result of increased cessation, whereas in women, tobacco deaths continued to increase or plateau into the present century. This four-stage model highlights the long delay (about three decades) between widespread smoking and its full effects on mortality. It has been applied to other countries or regions, with sub-Saharan Africa in Stage 1, Latin America in Stage 2, Eastern Europe in Stage 3, and the United Kingdom in Stage 4.

A recent review of the model after two decades showed that in developed countries, the prevalence of smoking has continued to decline in both sexes and the proportion of deaths due to smoking has declined in men and continues to increase or plateau in women (Thun et al. 2012). Latest findings also support that tobacco deaths in the United States have plateaued in men but continue to rise in women (Thun et al. 2013). The model generally applies to men in developing countries but cannot predict when (or if) women will begin smoking in large numbers. Thus, modified criteria and separate models are needed to describe the epidemic entirely for men and women in developing countries. The long lag between peak tobacco deaths and peak tobacco consumption is a key reason for the lack of urgent action among governments and politicians.

Smoking in women

A limitation of this model is that it does not take into account the history and timing of tobacco control measures. Data from Hong Kong (Lai 2012) show that its downward male smoking trend (from 39.7% in 1982 to 19.9% in 2010) appears to be similar to Australia and the United States, but its female trend from 5.6% to 3.0% does not follow the model. About three decades ago, tobacco control advocates in Hong Kong predicted a great rise in female smoking prevalence, which did not happen. This could be due, at least partly, to the early introduction of tobacco control measures including a 300% increase in tobacco tax in 1983 (and 100% in 1991), and legislation restricting (1982, 1990, 1992, 1998, 1999, and 2007) and eventually banning tobacco advertising (2008).

Enforce bans on tobacco advertising

It is recognized that a comprehensive ban on advertising, promotion, and sponsorship would reduce the consumption of tobacco products. Before

restrictions on tobacco advertising in the mid-1960s, many high-income countries, such as the United Kingdom and United States, witnessed aggressive advertising especially targeting women. Now, 103 countries have some restrictions on tobacco marketing, but only 24 countries (10% of the world population) have banned all forms of direct and indirect advertising (WHO 2013a).

Partial restrictions are ineffective as the tobacco industry has many ways to bypass them. In Hong Kong, where all forms of tobacco advertising are banned but the display of cigarettes at retailing shops and newspaper stands is allowed, the tobacco companies pay large sums of money to expand the display area with spotlights on individual cigarette packs. The government has difficulty prosecuting this as newspaper stand and shop owners will protest such actions. LMICs should aim to move directly to a total ban. Some African countries have done so before the multinational tobacco companies have gained a foothold. Other countries like Thailand have banned the display of cigarettes.

Banning of indirect advertising is very difficult as some cigarettes share the same brand as non-tobacco products (brand stretching). In China, a famous and very expensive cigarette brand from Shanghai was named, many decades ago, ‘Zhong Hua’ meaning China. The tobacco company launched a campaign of ‘I love Zhong Hua’, featuring two Chinese characters in the same format as the cigarette logo. Some Chinese brands are named after a famous place or an animal (such as Panda). Bypassing advertising bans in this manner is increasing, but many governments have no experience or resources to prosecute them or the confidence that they will win in court.

Banning of sponsorship is the most difficult of all. The FCTC defines this as ‘any form of contribution to any event, activity, or individual with the aim, effect or likely effect of promoting a tobacco product or tobacco use either directly or indirectly’ (WHO 2003). Victoria in Australia was the first to use tobacco tax proceeds to establish a health foundation in 1987 to replace tobacco sponsorship. This successful example has been cited by tobacco control advocates in many countries, but few have succeeded (such as Thailand). Many organizations’ events and activities, politicians, legislators, researchers, and other individuals are receiving tobacco sponsorships/money. They are unlikely to support stringent tobacco control measures, and may join the tobacco alliance directly or indirectly. Governments need to ban offering and receipt of all forms of tobacco

sponsorships. Interim measures could include legislation requiring disclosure and policies forbidding such sponsorships. To help implement and enforce comprehensive bans on tobacco advertising, promotion, and sponsorship, the WHO has developed a 3-day training workshop for LMICs (World Health Organization and International Union Against Tuberculosis and Lung Disease 2011).

Protect people from tobacco smoke

A major breakthrough in tobacco control was the designation of smoke-free areas to protect people from SHS, and the related non-smokers' rights movement starting in the 1980s in the United States. Banning of smoking in public places by law has been introduced, often progressively from a partial to total ban, and is expanding to include all public places (from indoor to outdoor) and all indoor workplaces (public and private). Further bans were introduced in outdoor areas in some countries, such as busy streets in Japan. With overwhelming public support, Singapore has recently extended the smoking bans to the common areas of residential buildings, covered walkways, pedestrian bridges, a 5-metre radius from the edge of bus stops, and outdoor compounds of hospitals. The number of people protected by comprehensive smoke-free laws has doubled from 2008 to 2010, but 89% of the world's population remain unprotected (Eriksen et al. 2012).

Public support is strong only in countries where awareness of the harms of SHS is high together with strong political will of governments. The tobacco industry has worked to oppose this, building alliances with smokers' rights groups and others who may be adversely affected, such as the hotel and catering industry. Some restaurant workers or their unions are mobilized by their employers, with support from tobacco companies, to protest. The tobacco industry will lobby governments and legislators and repeatedly argue that banning smoking in restaurants and bars will result in loss of customers, revenue, and jobs. However, in all jurisdictions, implementation of smoke-free laws has mostly resulted in no changes or increased restaurant sales, and savings in cleaning costs and reduction of fire risks.

A total ban is needed to eliminate SHS exposure. However, the tobacco industry is often successful in exploiting loopholes in such laws (such as no legal liability or penalty for venue owners or managers if they allow smoking in their statutory smoke-free areas in Hong Kong). Many restaurant

owners still believe that their business will suffer from total smoking bans, particularly in LMICs. Public health advocates must distinguish between the tobacco companies and venue owners even when they form an alliance. Mass media and other campaigns, such as motivating students and other non-smokers to demand smoke-free dining areas, are effective ways to change the minds of some restaurant owners and workers. Some will never be convinced but all will want a level playing field such that any new laws should equally apply to all venues; allowing some restaurants to have smoking areas but not others will not work.

A national smoke-free law is needed but in China, only some major cities have started to introduce local smoke-free legislation. However, enforcement is weak, penalties for violation low, and compliance is expectedly unsatisfactory. Mass media campaigns are needed before the implementation of the new law. Initial law enforcement with substantial penalties is essential to convince the public and venue owners about the government's seriousness. Effective enforcement of comprehensive smoke-free legislation saves lives by reducing heart attacks, increasing quitting in the population, and reducing smoking among employees (Glasgow et al. 1997).

Raise taxes on tobacco

Raising taxes is the most effective way to reduce tobacco consumption quickly. Repeated and sufficiently high increases in tax drove the decline in consumption especially in high-income countries a few decades ago. On the other hand, with economic development and increasing incomes in LMICs such as China, the reduction in real tobacco prices has substantially contributed to increasing consumption. High prices are particularly effective in preventing the uptake of smoking in the young and reducing consumption among low-income people, with more quitting in both groups. Strong evidence has come from both high- and LMICs, such as the United Kingdom, South Africa, Morocco, and Israel (Eriksen et al. 2012). A more recent example is from France, where tripling the price of cigarettes has halved consumption and doubled government revenue in 2003–2004 (Gallus et al. 2006).

The seminal World Bank report in 1999 states that all evidence from high- and LMICs shows that price increases are highly effective in reducing demand. The statement that 'On average, a price rise of 10% on a pack of

cigarettes would be expected to reduce demand for cigarettes by about 4% in high-income countries and by about 8% in low- and middle-income countries' (World Bank 1999) has been widely cited. The *Tobacco Atlas*, 4th edition (Eriksen et al. 2012) shows different degrees of reduction, 2–6% in high- and 2–8% in LMICs. The model in the World Bank report shows that raising prices by 10% worldwide would result in 40 million smokers quitting in 1995 and preventing at least 10 million deaths. This estimate is conservative because if one in two smokers will be killed by tobacco, 40 million current smokers quitting would prevent about 20 million deaths. As price increases will also reduce uptake of smoking in young people, there will be further reduction of deaths in the more distant future.

Big variations in the excise tax as a percentage of cigarette prices are shown in the *Tobacco Atlas*. The WHO recommends that excise tax should reach at least 70% of the retail price but only five nations have achieved this standard (Eriksen et al. 2012). The World Bank proposes that taxes should account for two-thirds to four-fifths of the retail price (Mackay et al. 2006). Tobacco tax increase can also generate additional income for governments. As of 2010, at least 166 WHO Member States have imposed a tobacco excise tax, and at least 27 of them used a portion of the tobacco tax revenue for health purposes. Numerous economic analyses have shown that the economic costs of tobacco, including lost productivity and health and other expenditures from tobacco-induced disease burden, are greater than tobacco tax income. The 2012 Asian Development Bank report estimated that in China, India, Philippines, Thailand, and Vietnam, a 50% price increase would reduce tobacco deaths by over 27 million, and generate over $27 billion additional revenue annually (Asian Development Bank 2012).

Given these data, it is not clear why most governments have not implemented the WHO recommendation of at least 70%, even after they have ratified the FCTC. Collusion between the tobacco industry and government officials, the loss of tobacco tax revenue if consumption falls sharply, or inadequate funding allocation for tobacco control in LMICs are possible reasons. More research is needed to find the root causes for this, and effective ways to convince governments to do so.

Any tax increase is not welcomed in general by the public and may lead to dissatisfaction towards governments, which could result in protests and unrest. The economic argument that reduction of tobacco use can save healthcare expenditures has little direct immediate impact because such

savings are only realized in the future. A very strong argument from opponents is that high taxes will result in increased smuggling, making cheaper illicit cigarettes more available to many low-income smokers, especially the young. However, a recent report showed that tax increases in the United Kingdom have not led to increased smuggling.

Tobacco industry

In celebrating the fifth anniversary of the FCTC, WHO Director General Dr Margaret Chan said, 'the tobacco industry is ruthless, devious, rich and powerful. As we all know, neither WHO nor public health is rich, but with the Framework Convention now in place, we are indeed powerful' (WHO 2010). It is the only industry that has been openly and repeatedly discredited, condemned, and shamed by WHO.

The tobacco industry is dominated by several multinational tobacco companies. The China National Tobacco Corporation is the largest state-owned enterprise producing the greatest number of cigarettes in the world. Despite the progress in tobacco control in some countries, and the FCTC, the tobacco companies remain some of the most profitable in the world (Eriksen et al. 2012).

Tobacco company documents

Research and numerous reports have exposed the tobacco industry's various antipublic health measures. Most of the findings are based on the analysis of millions of pages of internal tobacco company documents exposed since 1994 and during litigation proceedings against the tobacco companies in the 1994 seminal judgement in the United States, now deposited in Minnesota, United States and Guildford, United Kingdom (Mackay et al. 2006; Shafey et al. 2009). These documents are still under-researched but are now easily accessible for public health professionals and researchers, particularly those interested in what the tobacco industry had said about their countries or regions and strategies including the targeting of children or women. They provide insights into the varied strategies employed by the tobacco industry to promote the continued use of tobacco among different communities and population segments (Mackay et al. 2006; Shafey et al. 2009; Eriksen et al. 2012; WHO 2013b).

Tobacco corporate social responsibility

The tobacco companies spend huge sums of money in funding smokers' groups, tobacco associations and institutes, and other alliances to oppose tobacco control measures; funding researchers to challenge the evidence of tobacco harms particularly from SHS; funding research but suppressing or distorting unfavourable results; and funding government and non-government organizations and various events or programmes on education, arts, culture and sports, environmental protection, and medical, health, or rehabilitation programmes to promote cigarette brands and corporate image and to buy in supporters or silence potential opponents (WHO 2013b).

Youth Smoking Prevention (YSP) programmes, funded by the tobacco industry, are particularly deceptive and troubling. These are ineffective as they focus on young people only and portray smoking as an adult habit thus making it more appealing for adolescents. YSP programmes have attracted NGOs interested in adolescent health, community leaders, and schools who do not understand their motivation. This has created much confusion and frustration for tobacco control advocates, who need to be extremely cautious about how to expose the tobacco connections and ineffectiveness of YSP programmes, yet not to alienate those who have been deceived. The WHO booklet from Western Pacific Regional Office on YSP is an invaluable resource (WHO 2003). How should youth programmes be organized that can be distinguished from YSP? In Hong Kong, the strategy is to promote the WHO booklet (Ho et al. 2010) and to launch youth quitting programmes instead of smoking prevention programmes. It is useful to bear in mind and emphasize that the great majority of young people do not smoke. Overemphasis on youth smoking may result in overestimation of peer smoking in adolescents, which can increase adolescent smoking (Ho et al. 2010).

The tobacco industry has not changed, as evidenced by their current efforts to use litigation and legal threats against governments that are planning or have introduced new control measures. The *Tobacco Atlas*, 4th edition, lists legal challenges against tobacco control measures by the tobacco industry in 18 countries (Eriksen et al. 2012). The tobacco companies will launch more of these in more countries. Some tobacco control advocates in the West have called for government to buy the tobacco companies to control them. However, the example of China's national tobacco monopoly indicates that the result could be even worse. Some

tobacco control advocates in China have called for the separation of the administration arm, which can then monitor and control the enterprise arm more independently.

Should governments sue the tobacco industry? A major landmark was the success of the US government's 1999 lawsuit which resulted in the major tobacco companies paying settlements of US$206 billion over 25 years to 46 states to compensate for government spending on smoking-related healthcare (Gruber 2001). The number of antitobacco legal cases has been increasing but only slowly. The ultimate legal defeat of the tobacco industry would be to make tobacco illegal.

WHO FCTC

The FCTC is an evidence-based treaty and a paradigm shift in developing a regulatory strategy on demand reduction and supply issues (WHO 2003). The success of the FCTC has prompted and inspired public health advocates to call for similar treaties, such as a framework convention on alcohol control (FCAC) (Casswell and Thamarangsi 2009; Lam and Chim 2010). The FCTC is a readable, action-oriented document that can be used by governments, public health professionals, and the public. It has led to a global movement and is a strong lobbying document to pressurize governments to comply with the international treaty that they have ratified. The FCTC only stipulates the minimum that governments must do, and Australia's plain packaging in 2012 clearly shows that more can be done.

Countries that have not ratified the FCTC are under great pressure to ratify. Many governments have not yet fulfilled their obligations to comply fully with the FCTC, and these are reported regularly in WHO MPOWER reports.

The FCTC is governed by the Conference of the Parties (COP), which comprises all Parties (countries or jurisdictions) to the Convention. It reviews and makes decisions to promote its effective implementation, and adopts protocols, annexes, and amendments to the Convention. Starting from COP3, the regular sessions of COP are held biennially.

The Protocol to Eliminate Illicit Trade in Tobacco Products, in pursuance of FCTC Article 15, is the first protocol adopted by the COP as effective, appropriate, and comprehensive domestic and international responses to

eliminate all forms of illicit trade (WHO Framework Convention on Tobacco Research 2012).

Framework Convention Alliance

The FCTC is not just for health ministries or governments. NGOs have played key roles from the beginning of its negotiation, and their influence is expanding. The Framework Convention Alliance (FCA) created in 1999 has over 350 organizations from more than 100 countries. It is a civic society alliance whose mission is to help develop and implement the FCTC as the basis for effective global tobacco control (FCA n.d.). The FCTC has become the benchmark for the evaluation of the performance and effectiveness of tobacco control measures of parties, countries, or regions. For example, those who have been performing badly are being shamed by The Dirty Ashtray Award by civil society; China received one such award for making excuses not to print warning pictures on cigarette packing at COP3 in 2008.

FCTC and UN Political Declaration on NCDs

The FCTC has laid the foundation for the 2011 UN Political Declaration for the control of NCDs, and now the reverse is happening. As a follow-up to the UN Political Declaration, WHO is consulting on the development of the NCD Global Monitoring Framework (WHO 2012b) and the Global Action Plan on NCDs (2013–2020) (WHO 2012c). After lobbying by the FCA, the draft Global Action Plan on NCDs (10 October 2012) now explicitly promotes the FCTC and urges all WHO Member States to accelerate its implementation. While the NCD Global Monitoring Framework includes a tobacco smoking target (voluntary, to be determined), FCA further advocates to monitor specific process and impact indicators on the implementation of FCTC policy measures. There will be rapid and major developments on all these fronts.

Research and evaluation

The FCTC is based on solid research conducted mostly in high-income countries. Evidence from LMICs is inadequate. Progress in policies and implementation in some countries has been rapid and great, and previous

experiences and evidence can become outdated while the impact of new measures needs to be rapidly evaluated to guide further action. In countries with slow progress, the lack of funding and research is very serious. Their needs for local data and evidence are more urgent and should be most relevant for local policies and implementation. Further and relevant research is needed for all countries to support and evaluate the implementation of the FCTC. The Global Network of the Society for Research on Nicotine and Tobacco (SRNT) has published a series of papers in a themed issue of *Nicotine & Tobacco Research* on the research requirements and needs for all the FCTC articles to support specific policies and practices (Leischow et al. 2012). This is a most thorough analysis of the state of science to provide new directions and priorities for research. A commentary by Mackay emphasizes that 'advocates and activities MUST keep their feet firmly based in science, and never stray from that science' but 'The need to bridge the existing gaps between research, policy, and practice appears to be a global phenomenon' (Mackay 2013).

Offer help to quit tobacco use

Quitting smoking saves lives and the earlier the better. In cancer patients, increasing evidence shows that quitting smoking can improve the effectiveness of treatment and survival and more successful trials are needed to motivate HCPs to help such patients to quit (McBride and Ostroff 2003; Parsons et al. 2010).

Most successful former smokers quit without pharmaceutical therapy or professional services ('cold turkey'). They do have different experiences including past failures that deserve further study. On the other hand, effective treatments are available which have helped many to quit. The pursuit for more effective pharmaceutical treatments, including nicotine vaccination will go on, as the existing treatments are expensive and failures common.

The 1990 USSG report on smoking cessation was the first comprehensive review about the benefits of quitting (US Department of Health and Human Services 1990). The excess risk of ischaemic heart disease is reduced by half after stopping for 1 year, but it takes 10 years of quitting for lung cancer risk to reduce to 30–50% of continuing smokers. The 2007 IARC report on risk reversal, the most recent comprehensive review, concludes that the evidence

is adequate for: (1) lower risk in former smokers; and (2) lower risk with prolonged abstinence for lung, laryngeal, oral, oesophageal (squamous cell), pancreatic, bladder, and cervical cancer (seven items), and CHD incidence and death in subjects with or without the disease, cerebrovascular incidence, and death in those without established disease (three items), and cough and phlegm, decline in forced expiratory volume in 1 second in healthy subjects, and those with mild to severe disease and COPD mortality (five items). For overall mortality, those 'stopping at 60, 50, 40 or 30 years of age gain, respectively, about 3, 6, 9 or almost the full 10 years of life expectancy, in comparison with those who continue to smoke' (IARC 2007). Those who stopped smoking before middle age, at about 20–30, avoid nearly all of the future mortality risks from tobacco (Doll et al. 2004; Pirie et al. 2013).

Quitting or preventing uptake?

Uptake prevention and quitting have different target subjects and need different strategies. Most people intuitively support very strongly the prevention of smoking in children and young people and believe that this is the *most* important measure. They want more health education against youth smoking and consider health education in schools the most important action. Such a belief is exploited by the tobacco industry's YSP programme.

The 1999 World Bank report states that if quitting results in reducing adult cigarette consumption by half by the 2020s, about one-third of tobacco-related deaths in the 2020s would be prevented and almost half would be prevented thereafter, preventing 20 or 30 million tobacco deaths in the first quarter and 100 million in the second quarter of this century. If young smokers are reduced by half by the 2020s, hundreds of millions of tobacco deaths would be prevented after 2050. This would prevent none of the 150 million tobacco deaths in the 2020s, about 10–20 million out of 300 million in the second quarter of the century. Hence, quitting will prevent more deaths more quickly in the first half of this century than that from preventing young people from starting, although the latter is also important for preventing deaths in the second half of the century.

Nicotine dependence or habit

Smoking is often wrongly described as a habit or a personal choice. Nicotine is strongly addictive and most smokers are both nicotine and psychosocially dependent. The Fagerstrom test is the most commonly used test to measure nicotine dependence (Heatherton et al. 1991). But the term 'nicotine addict' is seldom used to describe heavily nicotine-dependent smokers and the term 'treatment of nicotine dependence' is much less frequently used than 'smoking cessation therapy'. HCPs need to understand why smokers frequently fail in their attempts to quit, or relapse after a short time.

Smoking cessation interventions include smoking cessation clinics (which vary from simple counselling to specialized in-hospital treatment) or telephone quitlines, smoking cessation advice, and counselling during routine consultation, and pharmacological aids and therapies. Most of these have been confirmed by numerous trials to be effective and more cost-effective than other medical treatments. The Cochrane Library should be consulted for updated reviews and conclusions (Cochrane Review Groups n.d.). Effectiveness means greater quit rates after interventions than no interventions, but failures occur more often than successes.

The seminal randomized controlled trial by Russell (1979) that brief advice by general practitioners in the United Kingdom increased the quit rate from 0.3% to 3.3% has led to a strong argument that if all physicians can spare a little time advising patients to quit, even though the absolute quit rate is low, many smokers will benefit, resulting in major community impact. Thirty years later, the review by Aveyard and Foulds (2009) showed that a brief intervention of at least 1 minute is effective (versus no intervention). However, most clinicians are too busy to spare even 1 minute and many feel incompetent. They also do not have any incentive, support, or pressure to do so. Such barriers are still common in both high- and LMICs. To address these issues, a simple 30-second pilot intervention emphasizing a 'one in two risk' was recently conducted in China with promising results (Lin et al. 2013).

Do most smokers want to quit?

The 2008 MPOWER report states 'Among smokers who are aware of the dangers of tobacco, three out of four want to quit' (WHO 2008) without specifying the country. But the reference (Jones 2006) indicates that this statement is based on a survey of US adults in 2006 and it should be noted

that the denominator is *not* all smokers. While this may be true in countries at the late stage of the tobacco epidemic, the Global Adult Tobacco Use Surveys in China and other LMICs show that most smokers do not intend to quit (Cooper et al. 2010; Chinese Center for Diseases Control and Prevention 2012). A deeper understanding of the smoker's intention and the strength of their motivation or determination to quit or reduce smoking can guide public health policies and planning on cessation services. Social support from family, friends, co-workers, and employers is needed.

Clinical or public health approaches?

During the early years of tobacco control, the emphasis was mostly on public health measures such as legislation and taxation. These measures were deemed more effective than clinical services that target individual smokers, especially since few physicians or HCPs were prepared to act. The high costs of clinical services also rendered them less cost-effective. The controversy over these two approaches has become more complex as effective pharmaceutical treatments have become available (Chapman and MacKenzie, 2010; West et al. 2010). While the Bloomberg Initiative to Reduce Tobacco Use does not support smoking cessation services for individuals, these services are rapidly developing elsewhere, such as in the United Kingdom. More quitlines and international quitting campaigns such as Quit and Win have also been developed. The WHO has provided detailed guidelines for setting up and running national quitlines, especially in LMICs (WHO 2011).

Quit rates for all the existing nicotine replacement therapies and drugs are not high as most users fail to quit after receiving treatment. The costs are high and adherence is low. The high failure rate is demoralizing. The 5As in the United States (ask, advise, assess, assist, arrange) has been one of the most referenced guidelines (Cornuz and Willi 2008) but these are too demanding for busy HCPs. Simpler guidelines or methods are recommended such as ABC in New Zealand (McRobbie et al. 2008). Development and randomized controlled trials of shorter advice of less than 1 minute are needed (Lam 2012; Lam et al. 2012).

Public health professionals need to understand that the clinician's job is primarily treatment. If clinicians are frustrated by failures of smokers to quit despite their efforts, they should be made more aware and become more

motivated to support or join in public health advocacy campaigns. Clinicians and their professional organizations are strong allies for public health in general and tobacco control in particular. Public health professionals should collaborate with clinicians to develop more effective measures to increase cessation in individual smokers and in the population.

Pilot cessation clinics

In LMICs, many physicians are smokers, and they seldom advise their patients to quit smoking. Some smoke quite openly, if smoking is allowed in the healthcare settings, such as hospitals. Even non-smoking HCPs are mostly inactive (Lam et al. 2011). In recent years in China, smoking cessation clinics have been established in many hospitals as a requirement for designation as smoke-free hospitals. This was also done to comply with the FCTC. But most clinics have attracted few smokers, from the hospitals and the community. Opening a clinic without strong support from top management, including adequate funding, capacity building, and incentives for HCPs, will not work. Lacking strong tobacco control measures in the greater environment, and funding to promote the cessation clinics, the wait for smokers to come will be in vain.

As a first step, introducing a pilot smoking cessation clinic with provision of training and support can motivate a hospital and its management and some HCPs to become interested, firstly in treatment, and then in tobacco control in general. The pilot can act as a springboard, bringing tobacco control advocates into a hospital to train interested HCPs in both smoking cessation intervention and tobacco control advocacy. With a core of trained HCPs, the clinic can become a focal point for further actions, such as implementation of smoke-free hospital policy, and extending beyond the clinic and the hospital into the community, such as supporting smoke-free workplaces and tobacco tax increases or public education campaigns in the neighbourhood. The clinic can also motivate the smoking HCPs to quit smoking, or to discourage smoking within the hospital, and become a training site for other staff and medical and other students.

On the other hand, public health professionals should work closely with clinicians to set up strategies for proactive approaches to help smokers outside the clinic, such as in the wards and outpatient departments and to promote quitting and tobacco control within and beyond the hospital to

develop the cessation clinics into tobacco control advocacy centres. Systematic evaluation of the impacts and sharing of experiences are essential for further improvements and development.

Smoking cessation can also act as the ‘carrot’ for the ‘stick’ of government proposed policies to expand smoke-free places and increase tobacco tax. Tobacco control opponents often argue that the policies are unfair to smokers and unkind to the poor, leaving the smokers nowhere to smoke and depriving the poor of their only affordable pleasure. Many government officers have difficulty responding to such allegations. If government-subsidized free smoking cessation services are offered, the simple answer is that smokers are encouraged to quit using the pilot clinics. If they quit, they will not need to find permissible places to smoke and will save a lot of money for more healthy and pleasurable commodities for the whole family.

Some government officers also feel that because there are so many smokers, they will not have the resources or capacity to open an adequate number of cessation clinics to cope with the demand. Some even feel that smokers should pay for the services, using the money saved from buying cigarettes. In reality, only a small percentage of smokers will visit cessation clinics, and many more will quit by themselves, often with social support from the family and others. A few cessation clinics or a quitline should therefore be sufficient and the sudden increase in visits and calls serves as an indicator of the effectiveness of the new control measures. Furthermore, if all or even a percentage of the excess revenue arising from the increase in tobacco tax is used to support smokers to quit and to promote quitting, governments will have a strong justification that part of the money obtained from smokers will be used to benefit the smokers.

Quitting or smoking reduction?

Many smokers who claim that they have reduced smoking may not be reporting the truth, and total abstinence is the only proven way for the benefits to be realized. Smoking reduction may be a distraction from total abstinence. Evidence for the benefits of smoking reduction and organized or regular services or campaigns focusing only on smoking reduction are scarce. But there are many smokers who do not want to quit completely but are willing to reduce smoking. New evidence has shown that helping such

smokers can increase quit rate and reduction (Chan et al. 2011). Smoking cessation campaigns and services will not attract such smokers. New approaches are needed to assess the needs of smokers who are not willing to quit and to recruit and help them to reduce, and to follow-up and evaluate further whether such approaches are effective in eventually helping the smokers to quit.

Controversies of harm reduction

Evidence, mainly from Sweden, that smokeless tobacco users have a lower risk of lung cancer and other diseases than cigarette smokers, has triggered debates about switching from cigarettes to smokeless tobacco as a harm reduction measure (WHO 2003). Opponents to such switching argue that it does not reduce the harm substantially, and promotion of switching can be seen as promotion of smokeless tobacco use resulting in reduced motivation to quit. It may also increase uptake of smokeless tobacco use in young people. More broadly, such strategies effectively entail the replacement of nicotine from tobacco use with nicotine-containing products (Foulds and Branstetter 2012). The debate is continuing, but the tobacco companies are moving fast to promote smokeless tobacco and develop other nicotine products. The UK National Institute for Health and Care Excellence (NICE) has released draft guidance for consultation, which includes nicotine replacement therapy (NRT) and e-cigarettes but excludes smokeless tobacco (NICE 2013).

Endgame or endless war?

The rapid reduction of smoking prevalence in some countries, the increasing impact of the FCTC, and strong political will of some governments have led to optimism and aspirations that a smoke-free country has become reachable within 20 years. The Bhutanese 2004 law aims to end tobacco sale, the 2010 Finland Tobacco Act aims to end tobacco use, and the US government has the ‘vision of a society free from tobacco-related death and disease’ (US Department of Health and Human Services 2012). In March 2011, the New Zealand government committed to become an essentially smoke-free nation by 2025 and the website has much information on policies and actions (Smokefree New Zealand 2025 n.d.).

On the other hand, the war against tobacco and the tobacco industry has just begun in many LMICs, where political will is lacking to fight wholeheartedly. For the pessimists, new tobacco control advocates, and younger public health professionals, this appears to be an endless war, sometimes with more battles lost than won. However, a careful review of past battles, globally and in many countries with more advanced tobacco control measures, indicates many more victories, every one of which had been fought initially by only a few dedicated and courageous people (Lam 2013). Fighting a battle today anywhere is much easier than it was a few decades ago. We now have the FCTC and another new international treaty under the FCTC (COP5) and more will be forthcoming. We also have the UN Declaration against NCDs (and tobacco). When we are facing the aggressive challenges from the tobacco industry including suing governments, we know Big Tobacco is desperate and that eventually, the public health goals will prevail, despite occasional loss, compromises, and setbacks. For public health professionals, our experiences in tobacco control will guide us to tackle the other three risk factors. If tobacco cannot be controlled successfully, the chances for success in controlling the other risk factors will be slim. Fighting against the risk factors and NCDs will be an endless war, as promoting health and well-being of all people will be our never-ending mission.

ACKNOWLEDGEMENTS

We thank Dr L. Xu for her help in literature search, referencing, and preparing the figures. Our sincere gratitude goes to Professors J. Mackay and R. Peto, who have been inspirational and supportive to our research and public health advocacy on tobacco control.

REFERENCES

Asian Development Bank (2012). *Tobacco Taxes: A Win–Win Measure for Fiscal Space and Health*. Mandaluyong City: Asian Development Bank.

Aubin, H.J., Benyamina, A., Karila, L., Luquiens, A., and Reynaud, M. (2011). [Current strategies for treatment of alcohol problems]. *La Revue du Praticien*, **61**, 1373–7.

Aveyard, P. and Foulds, J. (2009). *The 2008 US Clinical Practice Guideline: The Key Recommendations and a Commentary.* Ask the Experts. [Online] Available at: http://www.treatobacco.net

Basu, S., Stuckler, D., Bitton, A., and Glantz, S.A. (2011). Projected effects of tobacco smoking on worldwide tuberculosis control: mathematical modelling analysis. *British Medical Journal*, 343, d5506.

Bates, M.N., Khalakdina, A., Pai, M., Chang, L., Lessa, F., and Smith, K.R. (2007). Risk of tuberculosis from exposure to tobacco smoke: a systematic review and meta-analysis. *Archives of Internal Medicine*, **167**, 335–42.

Burke, H., et al. (2012). Prenatal and passive smoke exposure and incidence of asthma and wheeze: systematic review and meta-analysis. *Pediatrics*, **129**, 735–44.

Burton, A. (2011). Does the smoke ever really clear? Thirdhand smoke exposure raises new concerns. *Environmental Health Perspectives*, 119, A70–4.

Canadian Cancer Society (2012). *Cigarette Package Health Warnings: International Status Report* (3rd ed.). Canadian Cancer Society. [Online] Available at: http://www.tobaccolabels.ca/wp/wp-content/uploads/2014/04/Cigarette-Package-Health-Warnings-InternationalStatus-Report-English-CCS-Oct-2012.pdf

Casswell., S. and Thamarangsi., T. (2009). Reducing harm from alcohol: call to action. *The Lancet*, **373**, 2247–57.

Centers for Disease Control and Prevention (2010). *A Report of the Surgeon General: How Tobacco Smoke Causes Disease … And What it Means to You*. [Online] Available at: http://www.cdc.gov/tobacco/data_statistics/sgr/2010/consumer_booklet/pdfs/consumer.pdf

Chan, S.S., et al. (2010). Smoking-cessation and adherence intervention among Chinese patients with erectile dysfunction. *American Journal of Preventive Medicine*, **39**, 251–8.

Chan, S.S., Leung, D.Y., Abdullah, A.S., Wong, V.T., Hedley, A.J., and Lam, T.H. (2011). A randomized controlled trial of a smoking reduction plus nicotine replacement therapy intervention for smokers not willing to quit smoking. *Addiction*, **106**, 1155–63.

Chapman, S. and MacKenzie, R. (2010). The global research neglect of unassisted smoking cessation: causes and consequences. *PLoS Medicine*, 7, e1000216.

Chinese Center for Diseases Control and Prevention (2012). *Global Adult Tobacco Survey (GATS): China 2010 Country Report*. Beijing: Chinese Center for Diseases Control and Prevention.

Cochrane Review Groups (n.d.). *Cochrane Tobacco Addiction Group*. [Online] Available at: http://onlinelibrary.wiley.com/o/cochrane/clabout/articles/TOBACCO/frame.html

Cooper, J., et al. (2010). To what extent do smokers make spontaneous quit attempts and what are the implications for smoking cessation maintenance? Findings from the International Tobacco Control Four Country Survey. *Nicotine & Tobacco Research*, 12, S51–7.

Cornuz, J. and Willi, C. (2008). Nonpharmacological smoking cessation interventions in clinical practice. *European Respiratory Review*, **17**, 187–91.

Doll, R., Peto, R., Boreham, J., and Sutherland, I. (2004). Mortality in relation to smoking: 50 years' observations on male British doctors. *British Medical Journal*, **328**, 1519.

Doll, R., Peto, R., Wheatley, K., Gray, R., and Sutherland, I. (1994). Mortality in relation to smoking: 40 years' observations on male British doctors. *British Medical Journal*, **309**, 901–11.

Elliott, P., et al. (2009). Genetic loci associated with C-reactive protein levels and risk of coronary heart disease. *Journal of the American Medical Association*, **302**, 37–48.

Eriksen, M., Mackay, J., and Ross, H. (2012). *Tobacco Atlas* (4th ed.). Atlanta, GA: American Cancer Society; New York: World Lung Foundation. [Online] Available at: http://www.TobaccoAtlas.org

Foulds, J. and Branstetter, S. (2012). Tobacco harm reduction. In: Riley, D. and Pates, R. (eds.) *Drugs and Harm Reduction*. West Sussex: Wiley-Blackwell Publisher, pp. 213–28.

Framework Convention Alliance (FCA) (n.d.). Website. [Online] Available at: http://www.fctc.org

Gallus, S., Schiaffino, A., La Vecchia, C., Townsend, J., and Fernandez, E. (2006). Price and cigarette consumption in Europe. *Tobacco Control*, **15**, 114–19.

Gellert, C., Schottker, B., and Brenner, H. (2012). Smoking and all-cause mortality in older people: systematic review and meta-analysis. *Archives of Internal Medicine*, **172**, 837–44.

Glasgow, R.E., Cummings, K.M., and Hyland, A. (1997). Relationship of worksite smoking policy to changes in employee tobacco use: findings from COMMIT. Community Intervention Trial for Smoking Cessation. *Tobacco Control*, 6(Suppl. 2), S44–8.

Gruber, J. (2001). Tobacco at the crossroads: the past and future of smoking regulation in the United States. *The Journal of Economic Perspectives*, **15**, 193–212.

Harte, C.B. and Meston, C.M. (2012). Recreational use of erectile dysfunction medications and its adverse effects on erectile function in young healthy men: the mediating role of confidence in erectile ability. *Journal of Sexual Medicine*, **9**, 1852–9.

He, J., et al. (2007). Cigarette smoking and erectile dysfunction among Chinese men without clinical vascular disease. *American Journal of Epidemiology*, **166**, 803–9.

He, Y., et al. (2012). Secondhand smoke exposure predicted chronic obstructive pulmonary disease and other tobacco related mortality in a 17-years cohort study in China. *Chest*, 142(4), 909–18.

Heatherton, T.F., Kozlowski, L.T., Frecker, R.C., and Fagerstrom, K.O. (1991). The Fagerstrom Test for Nicotine Dependence: a revision of the Fagerstrom Tolerance Questionnaire. *British Journal of Addiction*, **86**, 1119–27.

Hedley, A.J., et al. (2006). Risks for heart disease and lung cancer from passive smoking by workers in the catering industry. *Toxicological Sciences*, **90**, 539–48.

Hirayama, T. (1981). Non-smoking wives of heavy smokers have a higher risk of lung cancer: a study from Japan. *British Medical Journal (Clinical Research Ed.)*, **282**, 183–5.

Ho, S.Y., et al. (2010). Comprehensive smoke-free legislation and displacement of smoking into the homes of young children in Hong Kong. *Tobacco Control*, **19**, 129–33.

Howell, F. and HSE Social Marketing Working Group on Tobacco (2012). 1 in every 2 smokers will die of a tobacco related disease: can you live with that? In: *15th World Conference on Tobacco or Health*, Singapore.

International Agency for Research on Cancer (IARC) (2007). *Tobacco Control: Reversal of Risk after Quitting Smoking.* IARC Handbooks of Cancer Prevention, Volume 11. Lyon: IARC.

International Agency for Research on Cancer (IARC) (2009). *Evaluating the Effectiveness of Smoke-free Policies*. Lyon: IARC. [Online] Available at: http://www.iarc.fr/en/publications/pdfs-online/prev/handbook13/handbook13.pdf

International Agency for Research on Cancer (IARC) (2012). *A Review of Human Carcinogens. Part E: Personal Habits and Indoor Combustions.* IARC Monographs on the Evaluation of Carcinogenic Risks to Humans. Volume 100E. A Review of Human Carcinogens. Lyon: IARC. [Online] Available at: http://monographs.iarc.fr/ENG/Monographs/vol100E/mono100E.pdf

International Tobacco Control Policy Evaluation Project (n.d.). *About ITC*. [Online] Available at: http://www.itcproject.org/about

Jha, P., et al. (2013). 21st-century hazards of smoking and benefits of cessation in the United States. *The New England Journal of Medicine*, **368**, 341–50.

Jones, J.M. (2006). *Smoking Habits Stable; Most Would Like to Quit*. [Online] Available at: http://www.gallup.com/poll/23791/Smoking-Habits-Stable-Most-Would-Like-Quit.aspx

Kupelian, V., Link, C.L., and McKinlay, J.B. (2007). Association between smoking, passive smoking, and erectile dysfunction: results from the Boston Area Community Health (BACH) Survey. *European Urology*, **52**, 416–22.

Lai, H.K., Ho, S.Y., Wang, M.P., and Lam, T.H. (2009). Secondhand smoke and respiratory symptoms among adolescent current smokers. *Pediatrics*, **124**, 1306–10.

Lai, V.W. (2012). Stages of the cigarette epidemic on entering its second century. Invited commentary. *Tobacco Control*, **21**, 101–2.

Lam, T.H. (2012). Absolute risk of tobacco deaths: one in two smokers will be killed by smoking: comment on ‘Smoking and all-cause mortality in older people’. *Archives of Internal Medicine*, **172**, 845–6.

Lam, T.H. (2013). Conversation with Judith Mackay. *Addiction*, **108**, 1897–904.

Lam, T.H., Abdullah, A.S., Ho, L.M., Yip, A.W., and Fan, S. (2006). Smoking and sexual dysfunction in Chinese males: findings from men's health survey. *International Journal of Impotence Research*, **18**, 364–9.
Lam, T.H., Chan, S.S., Abdullah, A.S., Wong, V.T., Chan, A.Y., and Hedley, A.J. (2012). Smoking reduction intervention for smokers not willing to quit smoking: a randomised controlled trial. *Hong Kong Medical Journal*, 18(Suppl. 3), 4–8.
Lam, T.H. and Chim, D. (2010). Controlling alcohol-related global health problems. *Asia-Pacific Journal of Public Health*, **22**, 203–8S.
Lam, T.H., He, Y., Li, L.S., He, S.F., and Liang, B.Q. (1997). Mortality attributable to cigarette smoking in China. *Journal of the American Medical Association*, **278**, 1505–8.
Lam, T.H., et al. (2005). Secondhand smoke and respiratory ill health in current smokers. *Tobacco Control*, **14**, 307–14.
Lam, T.H., Jiang, C., Chan, Y.F., and Chan, S.S. (2011). Smoking cessation intervention practices in Chinese physicians: do gender and smoking status matter? *Health and Social Care in the Community*, **19**, 126–37.
Leischow, S.J., Ayo-Yusuf, O., and Backinger, C.L. (2012). Converging research needs across framework convention on tobacco control articles: making research relevant to global tobacco control practice and policy. *Nicotine & Tobacco Research*, 15(4), 761–6.
Leung, C.C., et al. (2010). Passive smoking and tuberculosis. *Archives of Internal Medicine*, **170**, 287–92.
Lin, H.H., Ezzati, M., and Murray, M. (2007). Tobacco smoke, indoor air pollution and tuberculosis: a systematic review and meta-analysis. *PLoS Medicine*, 4, e20.
Lin, P.R., Zhao, Z.W., Cheng, K.K., and Lam, T.H. (2013). The effect of physician's 30 s smoking cessation intervention for male medical outpatients: a pilot randomized controlled trial. *Journal of Public Health*, **35**, 375–83.
Lopez, A.D., Collishaw, N.E., and Piha, T. (1994). A descriptive model of the cigarette epidemic in developed countries. *Tobacco Control*, **3**, 242–7.
Mackay, J. (2013). The role of research on the development and implementation of policy. *Nicotine & Tobacco Research*, 15(4), 757–60.
Mackay, J., Eriksen, M., and Shafey, O. (2006). *Tobacco Atlas* (2nd ed.). New York: American Cancer Society.
Matt, G.E., et al. (2011). Thirdhand tobacco smoke: emerging evidence and arguments for a multidisciplinary research agenda. *Environmental Health Perspectives*, **119**, 1218–26.
McBride, C.M. and Ostroff, J.S. (2003). Teachable moments for promoting smoking cessation: the context of cancer care and survivorship. *Cancer Control*, **10**, 325–33.
McGhee, S.M., et al. (2005). Mortality associated with passive smoking in Hong Kong. *British Medical Journal*, **330**, 287–8.
McRobbie, H., et al. (2008). New Zealand smoking cessation guidelines. *New Zealand Medical Journal*, **121**, 57–70.
Mialon, A., Berchtold, A., Michaud, P.A., Gmel, G., and Suris, J.C. (2012). Sexual dysfunctions among young men: prevalence and associated factors. *Journal of Adolescent Health*, **51**, 25–31.
Millett, C., et al. (2006). Smoking and erectile dysfunction: findings from a representative sample of Australian men. *Tobacco Control*, **15**, 136–9.
Monto, A.S. and Ross, H.W. (1978). The Tecumseh study of respiratory illness. X. Relation of acute infections to smoking, lung function and chronic symptoms. *American Journal of Epidemiology*, **107**, 57–64.
National Institute for Health and Care Excellence (NICE) (2013). *Tobacco: Harm-Reduction Approaches to Smoking*. NICE. [Online] Available at: https://pathways.nice.org.uk/pathways/smoking-tobacco-harm-reduction-approaches
Oberg, M., Jaakkola, M.S., Woodward, A., Peruga, A., and Pruss-Ustun, A. (2011). Worldwide burden of disease from exposure to second-hand smoke: a retrospective analysis of data from

192 countries. *The Lancet*, **377**, 139–46.
Oono, I.P., MacKay, D.F., and Pell, J.P. (2011). Meta-analysis of the association between secondhand smoke exposure and stroke. *Journal of Public Health*, **33**, 496–502.
Parsons, A., Daley, A., Begh, R., and Aveyard, P. (2010). Influence of smoking cessation after diagnosis of early stage lung cancer on prognosis: systematic review of observational studies with meta-analysis. *British Medical Journal*, 340, b5569.
Peto, R. (1994). Smoking and death: the past 40 years and the next 40. *British Medical Journal*, **309**, 937–9.
Physicians for a Smoke-Free Canada (n.d.). *Picture Based Cigarette Warnings*. [Online] Available at: http://www.smoke-free.ca/warnings
Pirie, K., Beral, V., Peto, R., Roddam, A., Reeves, G., and Green, J. (2008). Passive smoking and breast cancer in never smokers: prospective study and meta-analysis. *International Journal of Epidemiology*, **37**, 1069–79.
Pirie, K., Peto, R., Reeves, G.K., Green, J., Beral, V., and Million Women Study Collaborators (2013). The 21st century hazards of smoking and benefits of stopping: a prospective study of one million women in the UK. *The Lancet*, **381**, 133–41.
Rodgman, A. and Perfetti, T.A. (2009). *The Chemical Components of Tobacco and Tobacco Smoke*. Boca Raton, FL: CRC Press.
Russell, M.A. (1979). Tobacco dependence: is nicotine rewarding or aversive? *NIDA Research Monograph*, 100–22.
Secretan, B., et al. (2009). A review of human carcinogens—Part E: tobacco, areca nut, alcohol, coal smoke, and salted fish. *Lancet Oncology*, **10**, 1033–4.
Shafey, O., Eriksen, M., Ross, H., and MacKay, J. (2009). *Tobacco Atlas* (3rd ed.). Atlanta, GA: American Cancer Society.
Slama, K., et al. (2007). Tobacco and tuberculosis: a qualitative systematic review and meta-analysis. *International Journal of Tuberculosis and Lung Disease*, **11**, 1049–61.
Smokefree New Zealand 2025 (n.d.) *SMOKEFREE 2025*. [Online] Available at: http://smokefree.org.nz/smokefree-2025
Thun, M., Peto, R., Boreham, J., and Lopez, A.D. (2012). Stages of the cigarette epidemic on entering its second century. *Tobacco Control*, **21**, 96–101.
Thun, M.J., et al. (2013). 50-year trends in smoking-related mortality in the United States. *The New England Journal of Medicine*, **368**, 351–64.
Tobacco Labelling Resource Centre (n.d.) *Health Warnings*. [Online] Available at: http://www.tobaccolabels.ca/healthwarningsinfo
Trichopoulos, D., Kalandidi, A., Sparros, L., and MacMahon, B. (1981). Lung cancer and passive smoking. *International Journal of Cancer*, **27**, 1–4.
United Nations (UN) (2011). *Resolution 66/2. Political Declaration of the High-level Meeting of the General Assembly on the Prevention and Control of Non-communicable Diseases*. New York: United Nations. [Online] Available at: http://www.who.int/nmh/events/un_ncd_summit2011/political_declaration_en.pdf
US Department of Health and Human Services (1990). *The Health Benefits of Smoking Cessation: A Report of the Surgeon General*. [Online] Available at: http://www.surgeongeneral.gov/library/reports/
US Department of Health and Human Services (2004). *The Health Consequences of Smoking: A Report of the Surgeon General*. [Online] Available at: http://www.cdc.gov/tobacco/data_statistics/sgr/2004/complete_report/index.htm
US Department of Health and Human Services (2006). *The Health Consequences of Involuntary Exposure to Tobacco Smoke: A Report of the Surgeon General*. [Online] Available at: https://www.cdc.gov/tobacco/data_statistics/sgr/2006/index.htm
US Department of Health and Human Services (2012). *Ending the Tobacco Epidemic: Progress Toward a Healthier Nation*. Washington, DC: US Department of Health and Human Services.

Warren, C.W., Asma, S., Lee, J., Lea, V., and Mackay, J. (2009a). *Global Tobacco Surveillance System: The GTSS Atlas*. Atlanta, GA: CDC Foundation. [Online] Available at: http://www.cdc.gov/tobacco/global/gtss/tobacco_atlas/

Warren, C.W., et al. (2009b). Evolution of the Global Tobacco Surveillance System (GTSS) 1998–2008. *Global Health Promotion*, **16**, 4–37.

West, R., et al. (2010). Should smokers be offered assistance with stopping? *Addiction*, **105**, 1867–9.

Winickoff, J.P., et al. (2009). Beliefs about the health effects of 'thirdhand' smoke and home smoking bans. *Pediatrics*, 123, e74–9.

World Bank (1999). *Curbing the Epidemic—Governments and the Economics of Tobacco Control*. Washington, DC: World Bank. [Online] Available at: http://documents.worldbank.org/curated/en/1999/05/437174/curbing-epidemic-governments-economics-tobacco-control

World Health Organization (WHO) (2003). *WHO Framework Convention on Tobacco Control*. Geneva: WHO. [Online] Available at: http://www.who.int/tobacco/framework/WHO_FCTC_english.pdf

World Health Organization (WHO) (2008). *WHO Report on the Global Tobacco Epidemic, 2008: The MPOWER Package*. Geneva: WHO. [Online] Available at: http://www.who.int/tobacco/mpower/mpower_report_full_2008.pdf

World Health Organization (WHO) (2010). Opening remarks on the fifth anniversary of the WHO Framework Convention on Tobacco Control. In: *Convention on Tobacco Control: A Triumph for Public Health.* Geneva, Switzerland, 26 February 2010.

World Health Organization (WHO) (2011). *Developing and Improving National Toll-Free Tobacco Quit Line Services. A World Health Organization Manual*. Geneva: WHO. [Online] Available at: http://www.who.int/tobacco/publications/smoking_cessation/quit_lines_services/en/index.html

World Health Organization (WHO) (2012a). *Status of the WHO Framework Convention on Tobacco Control (FCTC)*. Geneva: WHO.

World Health Organization (WHO) (2012b). *A Comprehensive Global Monitoring Framework, Including Indicators, and a Set of Voluntary Global Targets for the Prevention and Control of Noncommunicable Diseases*. [Online] Available at: http://www.who.int/nmh/events/2012/discussion_paper3.pdf

World Health Organization (WHO) (2012c). *Development of an Updated Action Plan for the Global Strategy for the Prevention and Control of Noncommunicable Diseases*. [Online] Available at: http://www.who.int/nmh/events/2012/action_plan_20120726.pdf

World Health Organization (WHO) (2013a). *WHO Report on the Global Tobacco Epidemic. Enforcing Bans on Tobacco Advertising, Promotion and Sponsorship*. [Online] Available at: http://www.who.int/tobacco/global_report/2013/en/

World Health Organization (WHO) (2013b). *Tobacco Industry and Corporate Responsibility*. Geneva: WHO. [Online] Available at: http://www.who.int/tobacco/publications/industry/CSR_report.pdf

World Health Organization (WHO) (n.d.) *Tobacco Free Initiative (TFI): WHO FCTC Health Warnings Database*. [Online] Available at: http://www.who.int/tobacco/healthwarningsdatabase/en/index.html

World Health Organization and International Union Against Tuberculosis and Lung Disease (2011). *Tobacco Advertising, Promotion and Sponsorship: Enforcing Comprehensive Bans*. Geneva: WHO. Available at: http://www.who.int/tobacco/publications/building_capacity/training_package/adv_promotion_sponsorship

World Health Organization Framework Convention on Tobacco Research (2012). *Conference of the Parties to the WHO Framework Convention on Tobacco Control*. Fifth session. Seoul,

Republic of Korea, 12–17 November 2012. [Online] Available at: http://apps.who.int/gb/fctc/PDF/cop5/FCTC_COP5(1)-en.pdf

Yin, P., et al. (2007). Passive smoking exposure and risk of COPD among adults in China: the Guangzhou Biobank Cohort Study. *The Lancet*, **370**, 751–7.

Yuan, J.M., Ross, R.K., Wang, X.L., Gao, Y.T., Henderson, B.E., and Yu, M.C. (1996). Morbidity and mortality in relation to cigarette smoking in Shanghai, China. A prospective male cohort study. *Journal of the American Medical Association*, **275**, 1646–50.

9.2

Substance use and misuse

Considerations on global public health

Giang Le Minh and Steve Shoptaw

Psychoactive drug use and public health: introduction

Use and misuse of psychoactive drugs present major public health problems in most countries. Nicotine (in the form of tobacco products) and alcohol are the most commonly used and misused psychoactive drugs, and account for the greatest mortality and morbidity related to psychoactive drug use. Since nicotine and alcohol are discussed in Chapters 9.1 and 9.3, this chapter provides an overview of the epidemiology of psychoactive drugs, in particular cannabis, opiates, amphetamine-type stimulants (ATS), and cocaine.

The United Nations Office on Drugs and Crime (UNODC) compiles annual prevalence estimates of the use of various illicit psychoactive drugs and assesses trends over time. In the most recent *World Drug Report* published in 2019 the UNODC estimated that 271 million people, or 5.5% of the adult population (persons aged 15–64), used an illicit drug in 2017 (UNODC 2019). Of these, about 13% of persons using drugs were considered to use at levels that cause problems or drug use disorders (UNODC 2019). Prevalence of drug use disorders globally was relatively stable from 2009 to 2016; however, a small increase was observed in 2017 in large part due to findings of increased drug use in Nigeria and India (UNODC 2019). With respect to the use of specific illicit drugs, the UNODC estimated that cannabis was by far the most commonly used drug in 2017 with 3.8% of adults having used it in the previous year, followed by opioids (1.1%), ATS (0.6%), and cocaine (0.4%). Yet production continues to increase worldwide. In 2017, opium production reached 10,500 tonnes,

breaking the record since the beginning of twenty-first century (according to UNODC estimates). Cocaine manufacture also reached its highest level ever—1,410 tonnes in 2016. The quantities of other drugs such as amphetamines or new psychoactive drugs intercepted worldwide in the same year increased several times as well.

Involvement in substance use introduces serious consequences to health of individuals and societies among all nations. The Global Burden of Disease (GBD) study (https://www.theLANCET.com/gbd) estimates the numbers of deaths and disability-adjusted life years (DALYs), which is the sum of years of life lost due to premature death and years lived with disability by specific countries and by specific diseases. In 2017, there were approximately 585,000 deaths and 42 million years of 'healthy' life lost as a result of the use of drugs worldwide. Years of life lost due to premature death were attributed to approximately half of DALYs. Globally, the number of DALYs and deaths attributed to drug use has increased by 134 and 81%, respectively, over the past three decades (UNODC 2019). Despite the modest numbers of adults who use illicit drugs compared to alcohol and tobacco, they generate many costs to society, including lost productivity, increased disability and deaths, disruption of family and other social relationships, and costs of increased law enforcement and incarceration.

Research over the last two decades has increased understanding of the mechanisms of development and sustained expression of substance use disorder over the lifespan. This includes: (1) there is great individual variation in reactions to drugs; (2) repeated drug use leads to changes in brain cells (neurons) and the ways they communicate, particularly in systems related to reward, reinforcement and decision-making; (3) outcomes to treatment of substance use disorders (SUDs) involve ceasing (or cutting down) substance use *and* returning to abstinence following lapse or relapse; (4) SUDs are chronic and relapsing—there is no 'cure'—though effective treatments exist; (5) globally and across cultures there are strong similarities in how SUDs present; and (6) drug use is contemporary, with always changing complexities of the specific drugs used, their combinations, and the interventions that ameliorate harmful effects of drugs on the brain and behaviour. While knowledge of mechanisms underlying SUDs is likely to continue to increase rapidly, it is unrealistic to expect that there will be any 'magic bullet' breakthroughs for either the treatment or prevention of SUDs in the near future. As such, the public health challenge

is to support policymakers in understanding and responding to SUDs using models for managing other chronic diseases that in parallel fashion require ongoing treatment over the lifespan.

Global epidemiology and social determinants of substance use and substance use disorders

Geography is not necessarily destiny with respect to what drugs a person might use, but there are strong geographical patterns in the use of different drugs. The information on geographic differences in use of specific drugs that follows is summarized from the *World Drug Report* (UNODC 2019). The *World Drug Report* is based on reports from various participating countries producing variation in data quality; consequently, these data should be interpreted as estimates. Law enforcement information, particularly on seizures of drugs, is also used in assessing trends in international drug use. Despite some limitations, the *World Drug Report* is the best available single source on global use of illicit drugs.

Cannabis

Cannabis is the most prevalent of illicit substances used globally with 188 million individuals estimated to use cannabis worldwide in 2017. Cannabis products are most often smoked, but can be taken orally. In addition to the use of cannabis for its psychoactive properties, many around the world use cannabis and related products (e.g. cannabidiol) for relief of a range of medical complaints and conditions. Cannabis cultivation has increased in recent years, although many of the increases are small-scale growing operations. While the cannabis herb is prevalent in most countries, cannabis resin, or ‘hashish’, is more common in locations including Afghanistan, the Middle East, and small pockets of Western Europe and northern Africa. Among adults, the highest level of cannabis use is reported in North America (13.8%), followed by Oceania (10.9%), and West and Central Africa (10.0%). The prevalence of cannabis use among adults in Asia is low (1.8%); however, given the very large population of the region, it remains the largest population of individuals using cannabis worldwide. The relationship between cannabis exposure and health and psychosocial

deficits is documented, though a causal relationship is uncertain (Hasin 2018). For these reasons, and in the face of the failure of prohibition, decriminalization and legalization of cannabis use are increasingly advocated for in Europe and North America in the face of the failure of prohibition (van Ours 2012).

A novel health threat to cannabis and tobacco use involves lung health problems linked to using ‘vapes’, equipment that vaporizes oils and waxes for inhalation instead of burning combustible materials to produce smoke. Marketers of these devices promote fewer health risks to vaping cannabis and tobacco as smoke is not inhaled into the lungs. Even when considering the negative consequences to smoking combustible cannabis and tobacco, observations of severe lung disease, lung infections, and deaths that correspond with use of vaping devices elevate health concerns. These vaping instruments and cannabis and tobacco products are all unregulated complicating attribution of cause for severe illnesses and deaths linked with their use. In light of these factors, current medical opinion is to avoid vaping to administer cannabis and tobacco products (Christiani 2020).

Opioids

Among the 53 million people who used opioids worldwide in 2017, risk for overdose and death is high with two-thirds of deaths from drugs occurring in people who use opioids. Opioids are prescribed to relieve pain (predominantly oral medications), but opioids also can be injected or smoked for their euphoric properties (predominantly heroin). The highest levels of opioid use among adults were in North America (4.0%), Oceania (2.5%), South-West Asia and the Near and Middle East (2.3%), and South Asia (1.8%), all of which were higher than average global prevalence. It is important to note that those using prescription opioids outnumber those using heroin in North America and Australia, while in South-West Asia and the Near and Middle East and South Asia heroin is the main opioid of use. Deaths from the opioid epidemic in the US are alarmingly high and are now accounted mostly by increased use of synthetic opioids (such as fentanyl), often in combination with cocaine and methamphetamine (Hedegaard et al. 2018). In Europe, heroin use was the opioid used for most of the individuals who sought drug treatment.

Amphetamine-type stimulants (ATS)

Illicit use of ATS, which includes amphetamine, methamphetamine, and methylenedioxymethamphetamine (MDMA, ecstasy) is difficult to measure globally, complicated by small-scale local producers of these drugs, but was estimated to involve almost 50 million adults globally in 2017. The quantity of ATS seized from 1998 to 2017 increased more than tenfold, substantially more than cannabis, opioids, and cocaine, though, opioid seizures dominated drugs seizures from 2013 to 2017 (UNODC 2019).

ATS (excluding ecstasy) is most prevalent in Oceania—especially Australia (Farrell et al. 2019)—and North America, while there is evidence of increasing use in Asia, especially given increased seizure of methamphetamine in the region (UNODC 2019). ATS can be taken orally, smoked (e.g. 'ice'/crystal methamphetamine) or injected (excepting ecstasy, which is taken orally), and with strong positive attributes (especially for cocaine and methamphetamine) that include inducing euphoria, brightening mood, dampening appetite, increasing energy, sharpening body experiences (especially during sex), and focusing attention. Methamphetamine is considered to be the most worrisome ATS due to its neurotoxicity, addictiveness and the high burden of care associated with methamphetamine use disorder (Courtney and Ray 2014; Farrell et al. 2019). Despite lack of data in East and South-East Asia, there is a growing concern about methamphetamine use and methamphetamine-related issues in this region. East and Southeast Asia are the largest manufacture regions of methamphetamine worldwide. Trade routes often used for cocaine trafficking are increasingly also being used for ATS trafficking, with evidence of spillover into West Africa. At a prevalence of 0.4%, ecstasy use globally is comparable to cocaine use in the adult population, but higher rates are observed in specific regions including Oceania (1.7%), North America (0.9%), and Western and Central Europe (0.9%). Ecstasy, in particular, tends to be prevalent in young people and has recently seen a resurgence in use among the younger demographic after a period of stability or even decline from 2007 to 2012.

Cocaine

Cocaine consumption among adults has remained relatively stable with approximately 18 million users worldwide in 2017 (UNODC 2019). Cocaine is taken intranasally (powder) or smoked (crack), but also can be injected. Cocaine use is most prevalent among adult populations in North America (2.1%), Western and Central Europe (1.3%), and Oceania (1.7%). While cocaine use decreased or stabilized in the Americas and Europe by 2010, there have been increases in these regions as of 2017. Cocaine use increases correspond with increases in cocaine bush cultivation in South America a region where cocaine use accounts for a third of patients in treatment for SUDs. While limited data exist in Africa and parts of Asia, there are indications that there is increasing and emerging use of cocaine in those regions; this may reflect the increased trafficking of cocaine through these areas. Although increased cocaine use globally may be occurring, treatment for cocaine use makes up a small proportion of all drug admissions in all locales except Central and South America.

New psychoactive substances

New psychoactive substances (NPS), relative potency plant-based, or synthetic chemical compounds that mimic the effects of most types of drugs such as cannabis, opioids, hallucinogens, or stimulants, first emerged since the early 2000s in Western countries and late 2000s in Asia (Zawilska and Andrzejczak 2015; UNODC 2019). Examples of NPS include cathinones, phenethylamines, aminoindanes, and bath salts. NPS are worrisome for their neurotoxicity and lethality (Hohmann et al. 2014). Currently, East and Southeast Asia are perceived as the production factory of global NPS (UNODC 2019). Although the share of NPS in the drug market might be still limited (they account for only 3% of drug seizure cases), global NPS have been increasing in number, type and availability (Zawilska and Andrzejczak 2015). Although most countries have tried to regulate NPS, a significant number of these drugs remain uncontrolled (Lee et al. 2018). NPS are developed and appear on the market frequently with changes to molecular structure that present challenges to legal actions due to the sheer number and types of new compounds.

Poly drug use

Poly drug use is a frequent condition worldwide, often with mixtures of drugs used specifically for the subjective effects of the drugs together. Speedballs or goofballs are combination opioids (usually heroin) taken with cocaine, amphetamine, or methamphetamine, which brings energy to the mellow opioid 'high' experience. Marijuana often is used with stimulants to dampen the harshness of coming down from these drugs. Concurrent use of alcohol and cocaine metabolizes in the body to produce, cocaethylene, a compound that produces a high that is subjectively distinct from either cocaine or alcohol alone. As permutations vary by drug availability and preference, prevalence of poly drug use is difficult to measure in distinct combinations. Instead, estimates count those who report poly drug use and tally the constituents of the reported poly drug mixtures, which may bias prevalence estimates for each of the constituent drugs in the mixture.

Historical interpretation and implications for public health

There are identifiable patterns in drug use in different regions of the world. Some patterns reflect cultural differences, when drug is used in cultures and national societies for specific purposes, with acceptance and approval of those drugs (Jay 2010). Production, distribution networks and markets also affect ways drugs function culturally and reflect the way in which the reinforcing nature of the drug is woven into spiritual, religious, or hedonistic purposes. As drug use is integrated into culture, production and distribution of the drug can also be protected by local culture, including corruption of local authorities.

Supply-side responses at the level of the society to interrupt and prosecute global drug distribution networks have little impact on overall use (Csete et al. 2016). Where there are societal, cultural, and/or individual factors that produce a demand for drug, a distribution system and market will arise to meet the demand. On the other hand, demand-side efforts at prevention, harm reduction, and treatment are sorely needed to reduce the demand for drugs drives production and distribution.

Social determinants of substance use disorders

SUDs are complex with multiple etiological and risk factors that can be classified into three levels: individual, interpersonal, and contextual.

Individual risk factors include biological predispositions, mental health disorders (major depression, anxiety), family characteristics, and socioeconomic status that correspond with SUDs (Stone et al. 2012). People with drug use disorders are almost four times more likely to have major depressive disorder (OR 3.80, 95% CI 3.02–4.78) and three times more likely to have an anxiety disorder (OR 2.91, 95% CI 2.58–3.28) (Lai et al. 2015). Interpersonal risk factors for SUD include family dysfunction ranging from parenting styles to childhood trauma including child physical and sexual abuse. Important contextual factors that facilitate the development of SUDs include level of disorganization for the neighbourhood and social environment (Stone et al. 2012; Ford et al. 2017).

These factors heavily shape social norms for young adults and adolescents during development. Social networks that develop around youth who grow up in proximity to peer substance users can normalize perceptions of substance use, increasing likelihood youths will use substances (Arbour-Nicitopoulos et al. 2010).

Policies in law and taxation that shape availability of specific drugs strongly affect the prevalence of drug use and drug use disorders, especially in youth and others susceptible to socioeconomic forces. Laws that restrict legal drinking age and when alcohol can be sold can reduce alcohol consumption (Stone et al. 2012). Increasing prices of drugs leads to long-term reductions in SUDs, even among abusers (Pacula and Chaloupka 2001).

Public health consequences of substance use and misuse

Differentiation between substance use and substance use disorders

There are different adverse consequences of substance use and misuse, for individuals, their families, and for society as a whole. Though most people do not use substances or use at levels that do not cause problems, for those who occasionally use substance heavily or in hazardous situations can experience negative consequences that are aversive enough to modify behaviour—most cut down or stop using substance. The range of negative consequences with heavy use of substances include individual (e.g. physical

and psychological discomfort following binge use—a 'hangover'), interpersonal (e.g. a family member complains substance use is causing problems), employment (e.g. a supervisor fires the employee who repeatedly misses Mondays) and legal consequences (e.g. citation for driving under the influence). For the minority of people who continue to engage in chronic, heavy levels of substance use, adverse consequences usually increase in frequency and severity. Still more people will cut down or stop using the substance in response. Only a minority continues to use substance at levels that define addiction, especially in light of negative consequences.

Defining the substance use disorder diagnosis

There are no biological markers that confirm substance use disorder diagnoses. Tests that screen for substances or their metabolites in urine, saliva, breath, blood, or nails reliably document recent substance use, which confirms the key symptom underlying substance use disorder: active substance use. Instead, SUDs are defined by behavioural criteria. Major diagnostic manuals are the *Diagnostic and Statistical Manual of Mental Disorders* (DSM) of the American Psychiatric Association and the *International Classification of Diseases* (ICD) system of the World Health Organization (WHO). The criteria for defining SUDs have undergone considerable evolution in both of these classification systems since their first presentation. The current DSM-5 (American Psychiatric Association 2013) and the ICD-11 (WHO 2018) are the latest versions that detail the diagnostic criteria of SUDs and substance dependence, respectively.

A major diagnostic change occurred in 2013 when the DSM-5 integrated two diagnostic categories, abuse and dependence, into one diagnosis that represents a spectrum disorder that ranges in severity determined by number of diagnostic criteria present. The DSM-5 now defines substance use disorder as having at least two of the following characteristics over a 12-month period that cause *clinically significant impairment or distress*: (1) hazardous use, (2) continued use despite social or interpersonal problems, (3) neglect of major roles in order to use, (4) withdrawal, (5) tolerance, (6) using in larger amounts or for longer periods than intended, (7) repeated efforts to stop or control use, (8) time spent getting, using and recovering

from drug, (9) continued use despite knowledge of negative consequences, (10) important activities given up in order to use, and (11) craving. Those with 2–3 criteria are classified as mild SUD, 4–5 as moderate SUD, and 6 or more as severe SUD.

The ICD-11 was released in 2018 and will take effect in 2021. The ICD-11 substance dependence diagnosis is based on three central features that are shared with DSM-5 and is used when there is evidence of at least two of these central features over a 12-month period—or continuous use over a one-month period: (1) impaired control over substance use; (2) increasing central role of substance use over other activities and responsibilities; and (3) physiological features as manifested by (i) tolerance, (ii) withdrawal symptoms following cessation or reduction in substance use, or (iii) repeated substance use (or pharmacologically similar substance) to prevent or alleviate withdrawal symptoms.

Public health and clinical impact of using these diagnostic systems have strengths in informing research and clinical efforts that contribute to scientific knowledge and treatment delivery. Globally, severity of diagnoses for SUD corresponds with likelihood to succeed with treatments for SUD, with those with mild disorder being more likely to meet substance goals in treatment than those with severe SUD. Use of diagnoses confirms a medical understanding of addiction, which confers far less stigma than non-medical model concepts. By doing so, rigour and precision in measuring fidelity of treatment delivery and rigour in measuring outcomes from treatment increase.

Infectious diseases

Substance-related infectious disease transmission occurs when persons who use substances engage in behaviours to get (e.g. exchange sex) or use those drugs (e.g. share injection drug equipment). Injection drug behaviours, including sharing of injection equipment and paraphernalia used to prepare drug for injection have high efficiency in transmitting infections during drug administration, particularly for Hepatitis C virus. Substance use can facilitate infectious disease transmission when psychoactive properties inhibit inhibitions or otherwise enhance condomless sexual experiences such as ‘chemsex’ (sex under influence of substance), extreme sexual

behaviours (multiple partner), or exchange sex). While lower transmission efficiency compared to injection drug use, far more people globally use their drugs but do not inject, which increases risks for sexually transmitted infections (Shoptaw et al. 2013).

HIV/AIDS

Given the efficiency of infectious disease transmission when sharing injection drug use equipment and paraphernalia, it is not surprising that in the beginning of the US HIV epidemic, people who inject drugs (PWID) represented a group with high incidence and prevalence. Early studies in New York City showed approximately half of PWID in the city had HIV infection (Des Jarlais et al. 1996). Sharp reductions in HIV incidence and prevalence began in the 1990s among PWID that corresponded with several developments. Public health officials recognized effective medications for opioid use disorder (MOUD), particularly agonist (methadone) and partial agonist (buprenorphine, buprenorphine/naloxone) medications reduced heroin injection and thereby reduced transmission events for this group. There are solid reasons for this action: methadone promotes ART adherence among PWID (Uhlmann et al. 2010), even among PWID living with HIV/AIDS who are homeless (Palepu et al. 2011). Since 2005, methadone and buprenorphine are included on the WHO essential medications list for their effectiveness as treatments for opioid use disorder and their ability to prevent infectious disease transmission (WHO 2019).

Also occurring during the 1990s, there were thrilling developments in treating HIV/AIDS, with development of multiple classes of medications that, when used in combination, dramatically slowed mortality, dampened transmission, and improved health for those living with HIV/AIDS. Policymakers recognizing this ensured access to HIV medications to all, including PWID living with HIV/AIDS. Concurrently, civil societies (grassroots groups of PWID and their allies) mounted effective campaigns at local, national, and global levels to ensure access to harm reduction (NSP—needle and syringe programmes) for individuals who use injection drugs. Modelling studies show that simultaneous access to these three elements (opioid agonist treatment, ART, and NSPs) will slow or stop new HIV transmissions among PWID (Degenhardt et al. 2010). Recognition of these elements and six other elements were compiled into a global consensus

document that promotes a framework for countries to use to prevent HIV transmission by ensuring access and by measuring outcomes of HIV prevention among PWID (World Health Organization, UNODC, United Nations AIDS 2012). Interventions also included HIV testing and counselling, prevention, and treatment of sexually transmitted infections, condom distribution programmes for PWID and their sexual partners, targeted information, education and communication, prevention, vaccination, diagnosis, and treatment of viral hepatitis infection and prevention diagnosis and treatment of tuberculosis. With these effective prevention interventions, there is no good reason for an additional incident case of HIV among PWID. Indeed, additional new cases identify a lack of political will to ensure PWID have access to the combination prevention services framework.

Sexual transmission of HIV among persons who use drugs

Around the world, far more people use non-injection methods to use drugs. Risks for HIV transmission linked with non-injection use of substances are indirect in that their use facilitates condomless sex with HIV-positive or status unknown sex partners. Non-injection substances can increase HIV transmission along four domains. Drugs (methamphetamine, cocaine) can directly interact with characteristics of cells to alter immune function and enhance HIV transmission dynamics. Substances can alter HIV transmission behaviours as well. Binge alcohol drinking (beer goggles) diminishes inhibitions; stimulants (cocaine, methamphetamine, ecstasy) enables sex at the extremes (longer, rougher sex sessions) and enable ‘chemsex’ or ‘high fun’, a unique sexual experience that occurs only when under the influence of these drugs among men who have sex with men (MSM). Substances can interact with individuals to support behaviours that enhance risks for HIV transmission (exchange sex) and that promote sex in high-risk environments (venues where sexually transmitted infections are common).

Groups of users around the world have preferences that can enhance risks for HIV transmission. Binge alcohol drinking is the predominant substance linked with HIV transmission behaviours among MSM in South America among others. Use of stimulants by MSM (especially methamphetamine) significantly corresponds with HIV incidence (Plankey et al. 2007). Non-

injection substance use interferes with consistent ART adherence, which can interfere with sustained viral suppression among those who are living with HIV/AIDS.

Over the past decade, HIV transmission was slowed or halted using pre-exposure prophylaxis (PrEP) medications (emtricitabine/tenofovir) for HIV-negative persons among high-risk groups (serodiscordant couples, MSM). Landmark research showed that people living with HIV with sustained undetectable viral load levels were unable to transmit HIV to their partners (Cohen et al. 2011). Undetectable equals untransmittable. Efforts to test biomedical HIV prevention among PWID have shown that providing opioid substitution therapy for PWID living with HIV not only reduces mortality and morbidity for the individual, but prevents HIV transmission to sexual partners (Miller et al. 2018). Research findings also show that providing PrEP to PWID who are HIV-negative prevents HIV infection in parallel fashion to groups who do not use drugs (Choopanya et al. 2013). Research has yet to show findings in parallel fashion to PWID for persons who use non-injection substances. Conceivably, non-injection use of these substances interferes with consistent adherence to ARVs among those who are living with HIV (treatment) or who are high-risk HIV-negative individuals who use substances (prevention).

Hepatitis C virus (HCV) among persons who inject drugs

While HIV and HCV are transmitted through sharing of drug injection equipment, HCV is dramatically more infectious than HIV. Less efficient transmission of HCV can occur when coming into direct contact with blood (sharing a toothbrush used by someone with bleeding gums) and when having condomless sex with multiple sexual partners. A minority of individuals exposed to HCV spontaneously clear the infection. About 75 to 85% of persons with HCV progress to chronic phase with about 20% developing end-stage liver disease. Given this, it is not surprising that more PWID are living with HCV infection than are living with HIV—and co-infection with HCV and HIV can accelerate HIV disease progression.

Untreated HCV infection represents the leading risk for morbidity and mortality for PWID. In meta-analysis, the time to significant liver damage is 26–38 years after time of infection, with cirrhosis 34–46 years afterwards (Smith et al. 2015). In the decade prior, treatment for HCV involved use of

ribavirin and pegylated interferon. This combination produced sustained viral suppression (cure) for some individuals, with better outcomes for persons with select genotypes of HCV and less or no efficacy for others. Another drawback to the ribavirin/interferon combination involved significant side effects including moderate to severe depression and exhaustion that led many to abandon the treatment. In 2013 direct acting antiretrovirals (DAAs) were developed. These medications act by disrupting viral replication and infection processes. The result is these medications shortened treatment periods, reduced side effects of the treatment and produced cure rates for most persons, independent of HCV genotype. Concerns over reinfection among PWID following successful treatment for HCV are not borne out by data (Martinello et al. 2017). The key drawback to using DAAs is their high cost—so high as to likely interfere with their use by PWID. Use of DAAs early in HCV infection for PWID, while expensive, prevents liver damage and corresponding higher-cost treatments later in life. As well, HCV cure for PWID earlier in life prevents potential for HCV transmission via drug or sexual transmission behaviours, which further reduces costs for healthcare. Despite the high efficacy for DAAs as treatment for HCV, the high costs will likely continue to dampen policymakers' willingness to scale them up even in the US and across the world for use among PWID until costs come down. One hopeful sign is that for most states in the US, DAAs are covered by Medicaid formularies, which portends access to these drugs for PWID.

Drug-induced (overdose) deaths

In 2015, 450,000 persons were determined to have died from drug use. Deaths from overdoses are a unique and extreme major adverse consequence to drug use. In North America, especially the US, drug overdose is the leading cause of accidental death. Over 72,000 fatalities were tallied in 2017 in the US (Ahmad et al. 2018) of which 47,600 were attributed to opioid overdose (Statistics 2018). This opioid epidemic has become North America's most widespread behavioural public health problem, with a higher number of deaths due to drug overdose in 2016 compared to deaths due to HIV at the peak of the AIDS epidemic in the US (CDC 2017). Cause of death due to overdose is attributed by medical

examiners based on results from drug screening tests and recorded on certificates of death. Reports on overdose deaths that rely on death records have some inherent unreliability as autopsies are not conducted on all deaths. When poly substance use is present at death, it may be difficult to determine which substance is the cause of death. Stimulant drugs such as cocaine and methamphetamine can also induce deaths, especially when these drugs are combined with opioids, benzodiazepines, alcohol, and other drugs.

Methodological toolbox for studying illicit substance use

As illicit drug use is, by definition, illegal there are many difficulties in collecting rigorous data on the topic. The difficulties include: (1) social desirability responding (Pauls and Stemmler 2003) in which persons who use drugs may deny (or minimize) their drug use in order to present a desirable image to the researchers and/or to themselves; (2) locating survey participants, as many persons who use drugs at high levels may be homeless or unstably housed; and (3) the effects of drugs on memory (though greatest memory problems may be associated with alcohol use). Various techniques have been developed to reduce these problems including: (1) audio-computer assisted self-interviewing (ACASI) (Des Jarlais et al. 1996), in which questionnaires are programmed into computers so that the research participant listens to the questions over headphones and responds to the questions using either the keyboard or a touch screen. ACASI does not require participants to be literate and the questions can be presented identically to all individuals; (2) respondent-driven sampling (Heckathorn 1997), in which participants are reimbursed for recruiting others. Restrictions are placed on the numbers of participants that any individual can recruit, and additional data is collected in order to increase generalizability to the underlying population; and (3) timeline follow back (Sobell and Sobell 1992), a method in which major events in an individual's life are used to anchor to the recall of events over long time periods.

Surveys and sampling strategies

There are several major methods for studying the epidemiology of illicit drug use. These include household surveys of the general population,

school surveys of drug abuse among young adults, and special studies aimed at examining drug use among particular groups, such as those who use drugs at high frequencies.

Household surveys of the general population

Surveys of random samples of household members provide what is usually considered to be the best estimates of the use of different drugs in a country. Such surveys are quite expensive but provide data that cannot be obtained through any other method. The US National Survey of Drug Use and Health is conducted annually with a sample of over 67,500 respondents age 12 years or older. In that year, 10.6% of adults reported illicit drug use in the previous month (the measure of current drug use). The most commonly used illicit drug reported was marijuana followed by opioids. Specifically, there were an estimated 11.5 million people who reported non-medical use of pain relievers in 2016; a fivefold increase from 2012.

School surveys

Surveys can also be conducted in schools. School-age youth are of particular interest because illicit drug use typically begins during adolescence, and trends in the use of different drugs are often much more pronounced in this age group. Thus, school surveys may be particularly useful for generating predictions about future drug use. ‘Monitoring the Future’ is the largest and longest running school survey in the US (Johnston et al. 2006). In the 2018 survey of the 44,400 students who were interviewed, 38.8% reported using illicit drugs in the previous year, with marijuana being the most commonly used illicit drug, (35.9%), and narcotics (other than heroin) being the second most commonly used illicit drug (6.0%).

Administrative data

Persons who use illicit drugs frequently come into contact with both law enforcement and health agencies. They may be arrested for possession of drugs and/or they may seek treatment for both their substance abuse disorder and for related health problems, such as overdose and infections. Administrative data can be useful to detect temporal trends in criminal

justice and treatment services. For example, According to the Treatment Episode Data Set (TEDS) (TEDS 2017), there were slightly more than 2 million admissions to publicly funded drug and alcohol treatment programmes in the US in 2017. Among those aged 12 and older who sought treatment, five substances accounted for 93% of admissions: opiates (34%, primarily heroin), alcohol (29%), marijuana/hashish (13%), stimulants (12%, primarily methamphetamine), and cocaine (5%). Over the past decade, administrative data using TEDS highlighted the way opiates exceeded alcohol as the primary driver of treatment admissions in the US with stimulants/methamphetamine taking the spot from cocain06 in the rankings of publicly funded treatment admissions for these five illicit substances.

Special studies

While carefully framed surveys can provide data about the breadth of drug use in the general and school populations, and administrative data can provide useful information about problems associated with illicit drug use, these types of studies do not provide the depth of information needed to understand the problems associated with specific forms of illicit drug use. Special studies usually recruit persons who use drugs at high frequencies. PWID with or at risk for HIV infection are an example. Special studies may also recruit subjects from clinical (drug treatment programme) settings from the community. They may be cross-sectional (one-time data collection) or cohort studies (with the same subjects followed over time).

This review of survey and method tools provides guidance and initial recommendations on how to use methods to collect data at the population level and within special populations that to advise prevalence, incidence, and potential interventions with important groups of persons who use substances.

Interventions: approaches to treatment and prevention of substance use disorders

While there are effective methods for reducing transmission of infectious pathogens and for reducing overdoses among persons who use drugs, the

ideal method for addressing these problems would be to eliminate drug use among persons with SUDs. Current treatments can be used to produce periods of abstinence from weeks to months, but there are currently no treatments that can produce permanent abstinence. This reality best frames SUDs not only as spectrum disorders, but as disorders that are chronic and relapsing over the lifespan. Multiple treatments often are required before abstinence is achieved.

For prevention, there exists an accepted framework for PWID drafted by the WHO, UNODC, and the Joint United Nations Program on HIV/AIDS (UNAIDS). This technical guide outlines a comprehensive package for prevention and treatment for PWID (World Health Organization, UNODC, United Nations AIDS 2012) that includes methods for setting country targets and for measuring progress in providing these services. Services include needle and syringe programmes, opiate substitution treatment, HIV testing and counselling, antiretroviral treatment, prevention and treatment of sexually transmitted infections, condom programmes for PWID and their sexual partners, targeted information, education, and communication for PWID and their sexual partners, prevention, vaccination and treatment of viral hepatitis, and prevention, diagnosis, and treatment of tuberculosis.

Opiate substitution treatment/medication for opioid use disorder

Effective pharmacological treatment is available only for opioid dependence. Solid evidence supports agonist maintenance treatments including methadone and buprenorphine (mono formula or in combination with naloxone) (Connery 2015). Still, as SUDs interrelated to various psychological and social factors, evidence suggests medication should be combined with other treatment components including cognitive behavioural therapies, motivational interviewing, contingency management, harm reduction, social support services, and integrated care to achieve optimal outcomes.

Methadone. Scale-up of methadone maintenance treatment (MMT) has been a major turning point in global public health for PWID. Methadone is a full opioid agonist (as is heroin) that activates opioid mu receptors on nerve cells located in areas of the brain that are the basis for analgesic and pleasure effects from opioids. This binding of methadone at mu receptors

prevents opioid withdrawal symptoms, largely without producing euphoria. Methadone stops (or reduces) use of supplemental opioids, allowing the treated patient to avoid behaviours of drug use and to engage prosocial behaviours. Notable country outcomes following rapid scale-up of MMT in China and Vietnam demonstrate dramatic improvements in drug-related public health and social issues, especially reductions in HIV transmission. For most countries a drawback to providing MMT include requirements to deliver the care in highly regulated, specialty care settings to prevent diversion of the drug. There are successful models where methadone is prescribed at outpatient pharmacies with few medical or social harms, though policymakers are slow to engage this option. MMT is a primary HIV prevention strategy for PWID: 40% of reduction in the risk of incident HIV from sharing of injection drug and equipment (MacArthur et al. 2012; Karki et al. 2016). Given the strong advantages to MMT, retention in care is the primary treatment goal (Karki et al. 2016)

Buprenorphine. Buprenorphine is an optional treatment to methadone approved in the US since 2002. Buprenorphine is an opioid partial mu agonist that is manufactured as a mono-product and a co-formulated product containing a naloxone tag that reduces concerns over injection use. The mono and combo products are available in oral (daily) form. Injection (weekly, monthly) and implant (6 months) formulations of buprenorphine are available in high-income countries. Buprenorphine tightly binds at the mu opioid receptor (preventing withdrawal symptoms), but binds so tightly that buprenorphine effectively blocks effects of supplemental opioid use. A key advantage to buprenorphine is the 'partial' agonist blockade function that is so effective that it protects against most risks for overdose (especially respiratory distress) when treated patients use supplemental opioids. This safety profile allows the medication to be prescribed in outpatient clinic settings. Systematic reviews that compare MMT with buprenorphine and placebo therapy show that both MMT and buprenorphine maintenance therapy (BMT) are more effective and cost-effective than no treatment and MMT is slightly more effective in retaining patients and suppressing illicit opioid use than buprenorphine (Mattick et al. 2014).

Naltrexone. Naltrexone is a long acting opioid antagonist that tightly binds at the mu receptor sites. It is available in oral (daily), injection (monthly) and implant (6 months) formulations. The antagonist action of naloxone is complete and blocks any opioids from producing euphoria,

which stops opioid use completely. The binding of naltrexone at the mu receptor is so strong that it will displace any opioid from the mu receptor site that may be remaining in the body (precipitating opioid withdrawal). This effect also is so strong that a liquid formulation (naloxone—injection, nasal spray) is used widely to reverse opioid overdoses in emergency settings. Naloxone also is distributed widely (often by syringe and needle programmes) among friends/family members of opioid users to use in cases of overdose. While naltrexone is on board, the medication blocks any opioid use, preventing relapse to opioids. When the medication has worn off, however, risks for overdose exist as individuals become overwhelmed by prior doses of opioids as tolerance is lost with sustained treatment by antagonists (Miotto et al. 1997). Outcomes show use of oral naltrexone yields no difference from treatment with placebo (Minozzi et al. 2011). The main failure of oral naltrexone is a high dropout rate (Minozzi et al. 2011). In 2010, the extended released injectable naltrexone was approved in the US, attempting to solve the issue of the oral formula. This monthly injection provides patients with opioid dependence with more freedom in managing their condition. It has been proved to be equally effective as buprenorphine/naloxone treatment for patients who have been detoxified (Jarvis et al. 2018; Lee et al. 2018).

There are no medications approved for use to treat stimulants. Recent developments in medication development have yielded promising results for use of mirtazapine for methamphetamine use disorders (Colfax et al. 2011; Coffin et al. 2019). These trials replicated a statistically significant and meaningful reduction of methamphetamine use by 15–18% compared to placebo over a period of 12 weeks. These effects were observed in the context of regular counselling sessions, underscoring the likelihood that any medication for stimulant use disorder will require both medication and psychosocial interventions. Improvements in trial design and conduct in evaluating putative medication candidates for stimulant use disorder signal a strong possibility of development of one or more medications for cocaine and methamphetamine use disorders in the next decade.

Cognitive behavioural therapy

Cognitive behavioural therapy (CBT), a multisession psychoeducation therapy, teaches persons with substance use disorder steps to initiate

substance abstinence (dispose of drug and paraphernalia, set a quit date), to respond to cravings and urges using psychological and imagery skills and to return to abstinence in the setting of lapse or relapse. Most models of CBT stem from the original work of Marlatt and Gordon's model of relapse prevention (Marlatt and Donovan 2005).

The signal size of CBT for drug use disorder is in the moderate range (d=0.45). CBT has larger effect sizes when evaluated for treatment of cannabis use disorder with small effect size for poly substance use disorder (McHugh et al. 2010), findings verified in a second meta-analysis (Magill and Ray 2009). Moreover, CBT effects are measurable in trials at up to one-year follow-up evaluations (Rawson et al. 2004). CBT is a mature behavioural treatment for substance use disorder—so much so that elements of CBT are widely integrated into many ecumenical and 12-step approaches to counselling for SUDs.

Contingency management

Contingency management (CM) is a treatment for SUDs that involves providing voucher incentives in increasing value for successive biological samples documenting drug abstinence. As such, the treatment is based in operant conditioning (Skinner 1938). Its application began in delivery of methadone maintenance therapy by providing take-home medications to individuals who provided evidence of no illicit opioids while in treatment (Stitzer et al. 1977). The treatment was first used to treat cocaine use disorder in the early 1990s (Higgins et al. 1993). The treatment is used with a range of SUDs successfully. Several meta-analyses reliably show effect sizes for CM range between d = 0.4 to d = 0.6 (Prendergast et al. 2006; Dutra et al. 2008; Benishek et al. 2014; Davis et al. 2016).

CM can be delivered using two distinct methods. One provides vouchers of increasing value for successive samples documenting drug abstinence while the other provides opportunities to draw from a fishbowl for chances to win a small, medium, or large size prize. With successive samples marking drug abstinence, additional draws from the fishbowl are earned. The differing procedures have roughly equal efficacy.

There are no treatments for stimulant use disorder with greater effect sizes than CM, yet adoption of the treatment remains lacklustre. In the US, the Veterans Administration Healthcare System has recently adopted

providing the treatment for veterans diagnosed with cocaine use disorder. Early outcomes from this work indicate that the treatment is feasible and shows similar effects as in efficacy trials (DePhilippis et al. 2018). Evidence for increasing scale-up of CM more broadly is provided in a recent meta-analysis showing significantly greater effects in producing outcomes in treatment for stimulant use disorder than CBT or other modalities (De Crescenzo et al. 2018).

Motivational interviewing

Motivational interviewing (MI) is an approach based in client-centred therapy that has the basic assumption that people change thinking and behaviour about their substance use along a series of stages (Miller and Rollnick 2013). A hallmark of this assumption is that individuals can enter into treatment at any of these stages. The therapy style for MI emphasizes the psychological basis of addiction being ambivalence—the simultaneous thoughts that substance use both is and is not a problem. Individuals frequently shift between beliefs that substance use is not a problem (followed by substance use) and beliefs that substance use is a problem (followed by an attempt to quit or cut down). Therapists trained in MI techniques elicit a process of 'natural' change (i.e. change that is identified, initiated, and maintained by the individual) using non-directive interventions. Outcomes show that MI engages individuals in longer term treatment and promotes specific behaviour changes. From this approach, 'denial' is not confronted by therapists. Instead, MI techniques focus on articulating ambivalent beliefs in individuals about their substance use and work to develop discrepancies in their beliefs with the goal of developing 'change talk'. Statements from individuals that emphasize disadvantages of the status quo, that emphasize need to change, that express optimism for change and that disclose intention to change are defined as change talk and correspond with positive outcomes in treatment (Magill et al. 2019).

Meta-analysis of MI outcomes for substance use disorder treatment show weak to moderate effect sizes (Sayegh et al. 2017). As with CBT, stronger effect sizes for MI are observed for cannabis use disorder and weaker effect sizes for alcohol, tobacco, and poly substance use.

Detoxification

For most who become addicted to opioids, detoxification becomes a goal at some point. Many fail during multiple attempts at detoxification before a minority of individuals successfully maintain opioid abstinence. As noted, this raises risks for overdose following return to opioid use following the detoxification attempt. Some seeking detoxification services may not wish to become abstinent, but instead reduce their tolerance level to a more affordable level. Treatment for detoxification for opioid use disorder involves starting individuals on methadone then tapering methadone dose over a period of 2–3 weeks. During some point of the taper, withdrawal symptoms will cause many to discontinue early and return to opioid use, though often at a level somewhat lower than at the start of the taper). Despite the many individuals worldwide who start detoxification, the literature describing outcomes using this approach is uneven and poorly articulated, presumably because the high early discontinuation interferes with drawing conclusions in randomized controlled trials. One meta-analysis on opioid detoxification concludes that variable and poor quality studies prevent comparisons and instead recommends patients be counselled toward opioid maintenance treatment (Amato et al. 2004). Detoxification outcomes can be improved some, however, by integrating psychosocial counselling approaches while conducting the taper during detoxification programmes (Amato et al. 2008).

Residential treatment

Residential treatment for SUDs involves providing direct service with multiple components within a licensed facility that can evaluate, diagnose, and treat the symptoms and disabilities linked with substance use disorder. Individuals in residential treatment receive services 24 hours as needed. Services range in intensity, from mild through intensive levels, with focus on instilling abstinence, preventing relapse, coping with triggers, mood management, 12-step involvement, physical exercise, congregant living skills, and activities of daily life (Reif et al. 2014b). Individuals who begin residential treatment generally receive more intensive services and as abstinence is established, mood improves and impulsive behaviour inhibited, intensity of services decrease allowing the individual to engage

independent behaviours. Lengths of stay at residential treatment facilities range from 28 days through several months. Toward discharge from residential treatments, individuals have opportunities to establish drug-free routines outside the facility. Many engage work, family, and other social activities, then return to the residential treatment facility to end the day.

Efficacy of the residential treatment approach shows that the approach has value, though the quality of the evidence prevents arriving at statistics that describe the extent of that value. Still this essential treatment approach for those with severe SUDs merits consideration as policymakers determine the mix of service intensities and modalities that are available to persons seeking treatment. Recovery housing provides a stepped down level of care that still provides direct service with multiple components that includes supervised, short-term housing to individuals with SUDs or co-occurring mental and SUDs after residential treatment. Recovery housing for persons with SUDs have positive effects on long-term substance use outcomes and improvements in functioning, including employment and criminal activity (Reif et al. 2014a).

12-step and related social programmes

The 12 Steps of Alcoholics Anonymous is a fellowship of people who share the problems that have become unmanageable due inability to control use of alcohol (AA), narcotics (Narcotics Anonymous—NA), cocaine (Cocaine Anonymous—CA), Crystal Meth Anonymous—(CMA) and other substances. Twelve Step programmes rely on the original or derivative programme of recovery set forth by the *Alcoholics Anonymous: The Big Book* (Anonymous 2002a) and the *Twelve Steps and Twelve Traditions* (Anonymous 2002b). A key distinction to these 12-step programmes is that they are not considered treatment for substance use disorder formally. Instead 12-step groups are a social fellowship in which persons with addiction meet, work the programme set forth in the Big Book, and provide and receive social support. As these programmes fiercely protect their anonymous nature, there are no randomized controlled trials that would describe their efficacy in producing or sustaining abstinence from substance. Despite efficacy indicators, the ubiquity of these groups around the world undoubtedly have the greatest effectiveness of any therapy or programme designed to instil and maintain substance abstinence worldwide.

The programme of the 12-Steps of Alcoholics Anonymous is built around a concept of a ‘God’ that can be defined in many ways that is central to admitting powerlessness over control of substance use. Discomfort with the God-concept among many has contributed to development of secular versions of the 12-Steps. These include: Rational Recovery, SMART Recovery, Women for Sobriety, Secular Organizations for Sobriety, and Moderation Management. These groups use differing conceptualizations about addiction and its treatment from the 12-Steps and from each other. Many of these emphasize CBT and related psychological constructs to support recovery. Some work toward a secular adaptation of 12-Step groups.

Harm reduction

Harm reduction among people who use drugs grew out of shared recognition among these individuals and their allies that due to stigma and lack of political support, health promoting and harm reducing strategies and interventions would not happen unless they banded together to demand them. The initial work of this grassroots movement was helping to institute and scale-up syringe and needle programmes worldwide. With the advent of the HIV/AIDS epidemic and the efficiency of transmission of virus via injection methods, efforts to broaden access to syringe and needle programmes gained acceptance in communities (Des Jarlais et al. 1996). This grassroots movement professionalized and developed organizations, leaders, and impact, especially on the global public health stage. Today, this involves a well-supported and effective ‘civil society’ movement. This movement has been able to organize civil disobedience and to build alliances with academic and community partners to achieve major accomplishments in ensuring harm reduction and health promoting interventions are available, worldwide, for persons who use drugs. In addition to syringe and needle programmes, harm reduction approaches have developed to include supervised injection facilities, primarily in response to the growing worldwide spread of fentanyl among PWID. Fentanyl is 50–100 times more potent than heroin, so the risks for overdose are now higher than perhaps any time in history. The supervised injection facility ensures that rescue medications and medical personnel are onsite for reversing unintended overdose. These harm reduction spaces also are

designed to be spaces that provide sterile injection supplies, that provide access and/or linkages to medical care (especially wound/abscess care, treatment for sexually transmitted infections) and linkages to formal treatment for SUDs.

Systematic review of trials on efficacy of syringe and needle programmes show these programmes provide social and physical environments that mitigate drug-related harms and increased access to social and material resources (McNeil and Small 2014). The specific elements of this support include refuge from street-based drug use settings, promotion of safer injecting in providing equipment and guidance for this objective, provide linkages to healthcare services and resources and protect from law enforcement. Given the stigmatized and victimized experiences common to PWID, provision of these safe spaces continue to be at the same level of priority as providing MOUD and access to naloxone reversal kits for persons living with opioid use disorder.

REFERENCES

Ahmad, F.B., Rossen, L.M., Spencer, M.R., Warner, M., and Sutton, P. (2018). Provisional drug overdose death counts. National Center for Health Statistics. [Online] Available at: https://www.cdc.gov/nchs/nvss/vsrr/drug-overdose-data.htm

Amato, L., Davoli, M., Ferri, M., and Ali, R. (2004). Methadone at tapered doses for the management of opioid withdrawal. *Cochrane Database of Systematic Reviews*, (4), CD003409.

Amato, L., Minozzi, S., Davoli, M., Vecchi, S., Ferri, M.M., and Mayet, S. (2008). Psychosocial and pharmacological treatments versus pharmacological treatments for opioid detoxification. *The Cochrane Database of Systematic Reviews*, (3), CD005031.

American Psychiatric Association (2013). DSM-5. *Diagnostic and Statistical Manual of Mental Disorders*. Washington, DC: American Psychiatric Publishing.

Anonymous (2002a). *Alcoholics Anonymous: The Big Book:* A.A.W.S. Center City, MN: Hazelden Distributed Titles.

Anonymous (2002b). *Twelve Steps and Twelve Traditions:* A.A.W.S. Center City, MN: Hazelden Distributed Titles.

Arbour-Nicitopoulos, K.P., Kwan, M.Y.W., Lowe, D., Taman, S., and Faulkner, G.E.J. (2010). Social norms of alcohol, smoking, and marijuana use within a Canadian university setting. *Journal of American College Health*, 59(3), 191–6.

Benishek, L.A., et al. (2014). Prize-based contingency management for the treatment of substance abusers: a meta-analysis. *Addiction (Abingdon, England)*, 109(9), 1426–1436.

Centers for Disease Control and Prevention (CDC) (2017). *Drug Overdose Death Data*. [Online] Available at: https://www.cdc.gov/drugoverdose/data/statedeaths.html

Choopanya, K., Martin, M., Suntharasamai, P., Sangkum, U., Mock, P.A., Leethochawalit, M., and the Bangkok Tenofovir Study Group (2013). Antiretroviral prophylaxis for HIV infection in injecting drug users in Bangkok, Thailand (the Bangkok Tenofovir Study): a randomised, double-blind, placebo-controlled phase 3 trial. *The Lancet*, 381(9883), 2083–90.

Christiani, D.C. (2020). Vaping-induced lung injury. *The New England Journal of Medicine*, **382**, 960–2.
Coffin, P., et al. (2019, 16 June). *Mirtazapine for Methamphetamine Use Disorder Among Men and Transwomen Who Have Sex with Men: A Randomized Placebo-Controlled Trial*. Paper presented at the 81st Annual Scientific Meeting for the College on Problems of Drug Dependence, San Antonio, TX.
Cohen, M.S., et al.; HPTN 052 Study Team. (2011). Prevention of HIV-1 infection with early antiretroviral therapy. *New England Journal of Medicine*, 365(6), 493–505.
Colfax, G.N., et al. (2011). Mirtazapine to reduce methamphetamine use: a randomized controlled trial. *Archives of General Psychiatry*, 68(11), 1168–75.
Connery, H.S. (2015). Medication-assisted treatment of opioid use disorder. *Harvard Review of Psychiatry*, 23(2), 63–75.
Courtney, K.E. and Ray, L.A. (2014). Methamphetamine: an update on epidemiology, pharmacology, clinical phenomenology, and treatment literature. *Drug and Alcohol Dependence*, **143**, 11–21.
Csete, J., et al. (2016). Public health and international drug policy. *The Lancet (London, England)*, 387(10026), 1427–80.
Davis, D.R., Kurti, A.N., Skelly, J.M., Redner, R., White, T.J., and Higgins, S.T. (2016). A review of the literature on contingency management in the treatment of substance use disorders, 2009–2014. *Preventive Medicine*, **92**, 36–46.
De Crescenzo, F., et al. (2018). Comparative efficacy and acceptability of psychosocial interventions for individuals with cocaine and amphetamine addiction: a systematic review and network meta-analysis. *PLoS Medicine*, 15(12), e1002715.
Degenhardt, L., Mathers, B., Vickerman, P., Rhodes, T., Latkin, C., and Hickman, M. (2010). Prevention of HIV infection for people who inject drugs: why individual, structural, and combination approaches are needed. *The Lancet*, **376**, 285–301.
DePhilippis, D., Petry, N.M., Bonn-Miller, M.O., Rosenbach, S.B., and McKay, J.R. (2018). The national implementation of Contingency Management (CM) in the Department of Veterans Affairs: attendance at CM sessions and substance use outcomes. *Drug and Alcohol Dependence*, **185**, 367–73.
Des Jarlais, D.C., et al. (1996). HIV incidence among injecting drug users in New York City syringe-exchange programmes. *Lancet (London, England)*, 348(9033), 987–91.
Dutra, L., Stathopoulou, G., Basden, S.L., Leyro, T.M., Powers, M.B., and Otto, M.W. (2008). A meta-analytic review of psychosocial interventions for substance use disorders. *American Journal of Psychiatry*, 165(2), 179–87.
Farrell, M., et al. (2019). Responding to global stimulant use: challenges and opportunities. *Lancet (London, England)*, 394(10209), 1652–67.
Ford, J.A., Sacra, S.A., and Yohros, A. (2017). Neighborhood characteristics and prescription drug misuse among adolescents: the importance of social disorganization and social capital. *The International Journal on Drug Policy*, **46**, 47–53.
Hasin, D.S. (2018). US epidemiology of cannabis use and associated problems. *Neuropsychopharmacology*, 43(1), 195–212.
Heckathorn, D.D. (1997). Respondent-driven sampling: a new approach to the study of hidden populations. *Social Problems*, 44(2), 174–99.
Hedegaard, H., Miniño, A.M., and Warner, M. (2018). *Drug Overdose Deaths in the United States, 1999–2017*. Hyattsville, MD: CDC. [Online] Available at: https://www.cdc.gov/nchs/data/databriefs/db329-h.pdf
Higgins, S.T., Budney, A.J., Bickel, W.K., Hughes, J.R., Foerg, F., and Badger, G. (1993). Achieving cocaine abstinence with a behavioral approach. *American Journal of Psychiatry*, 150(5), 763–9.

Hohmann, N., Mikus, G., and Czock, D. (2014). Effects and risks associated with novel psychoactive substances: mislabeling and sale as bath salts, spice, and research chemicals. *Deutsches Arztebllatt International*, 111(9), 139–47.

Jarvis, B.P., et al. (2018). Extended-release injectable naltrexone for opioid use disorder: a systematic review. *Addiction*, 113(7), 1188–209.

Jay, M. (2010). *High Society: Mind-Altering Drugs in History and Culture*. London: Thames and Hudson.

Johnston, L.D., O'Malley, P.M., Bachman, J.G., and Schulenberg, J.E. (2006). Demographic subgroup trends for various licit and illicit drugs, 1975–2005. (Monitoring the Future Occasional Paper No. 63). Ann Arbor, MI: Institute for Social Research. [Online] Available at: http://monitoringthefuture.org/

Karki, P., Shrestha, R., Huedo-Medina, T.B., and Copenhaver, M. (2016). The impact of methadone maintenance treatment on HIV risk behaviors among high-risk injection drug users: a systematic review. *Evidence-based Medicine & Public Health*, 2, e1229.

Lai, H.M.X., Cleary, M., Sitharthan, T., and Hunt, G.E. (2015). Prevalence of comorbid substance use, anxiety and mood disorders in epidemiological surveys, 1990–2014: a systematic review and meta-analysis. *Drug and Alcohol Dependence*, **154**, 1–13.

Lee, J.D., et al. (2018). Comparative effectiveness of extended-release naltrexone versus buprenorphine-naloxone for opioid relapse prevention (X:BOT): a multicentre, open-label, randomised controlled trial. *The Lancet*, 391(10118), 309–18.

MacArthur, G.J., et al. (2012). Opiate substitution treatment and HIV transmission in people who inject drugs: systematic review and meta-analysis. *BMJ (Clinical Research Ed.)*, 345, e5945–5.

Magill, M., et al. (2019). Do what you say and say what you are going to do: a preliminary meta-analysis of client change and sustain talk subtypes in motivational interviewing. *Psychotherapy Research*, 29(7), 860–9.

Magill, M. and Ray, L.A. (2009). Cognitive-behavioral treatment with adult alcohol and illicit drug users: a meta-analysis of randomized controlled trials. *Journal of Studies on Alcohol and Drugs*, 70(4), 516–27.

Marlatt, G.A. and Donovan, D.M. (2005). *Relapse Prevention: Maintenance Strategies in the Treatment of Addictive Behaviors*. New York: Guilford Press.

Martinello, M., Hajarizadeh, B., Grebely, J., Dore, G.J., and Matthews, G.V. (2017). HCV cure and reinfection among people with HIV/HCV coinfection and people who inject drugs. *Current HIV/AIDS Reports*, 14(3), 110–21.

Mattick, R.P., Breen, C., Kimber, J., and Davoli, M. (2014). Buprenorphine maintenance versus placebo or methadone maintenance for opioid dependence. *The Cochrane Database of Systematic Reviews*, (2), CD002207.

McHugh, R.K., Hearon, B.A., and Otto, M.W. (2010). Cognitive behavioral therapy for substance use disorders. *The Psychiatric Clinics of North America*, 33(3), 511–25.

McNeil, R. and Small, W. (2014). 'Safer environment interventions': a qualitative synthesis of the experiences and perceptions of people who inject drugs. *Social Science & Medicine*, **106**, 151–8.

Miller, W.C., et al. (2018). A scalable, integrated intervention to engage people who inject drugs in HIV care and medication-assisted treatment (HPTN 074): a randomised, controlled phase 3 feasibility and efficacy study. *The Lancet*, 392(10149), 747–59.

Miller, W.R. and Rollnick, S. (2013). *Motivational Interviewing: Helping People Change* (3rd ed.). New York: Guilford Press.

Minozzi, S., Amato, L., Vecchi, S., Davoli, M., Kirchmayer, U., and Verster, A. (2011). Oral naltrexone maintenance treatment for opioid dependence. *Cochrane Database of Systematic Reviews*, (4), CD001333.

Miotto, K., McCann, M.J., Rawson, R.A., Frosch, D., and Ling, W. (1997). Overdose, suicide attempts and death among a cohort of naltrexone-treated opioid addicts. *Drug and Alcohol Dependence*, 45(1–2), 131–4.

Pacula, R.L. and Chaloupka, F.J. (2001). The effects of macro-level interventions on addictive behavior. *Substance Use & Misuse*, 36(13), 1901–22.

Palepu, A., Milloy, M.J., Kerr, T., Zhang, R., and Wood, E. (2011). Homelessness and adherence to antiretroviral therapy among a cohort of HIV-infected injection drug users. *Journal of Urban Health*, 88(3), 545–55.

Pauls, C.A. and Stemmler, G. (2003). Repressive and defensive coping during fear and anger. *Emotion*, 3(3), 284–302.

Plankey, M.W., et al. (2007). The relationship between methamphetamine and popper use and risk of HIV seroconversion in the multicenter AIDS cohort study. *Journal of Acquired Immune Deficiency Syndrome*, 45(1), 85–92.

Prendergast, M., Podus, D., Finney, J., Greenwell, L., and Roll, J. (2006). Contingency management for treatment of substance use disorders: a meta-analysis. *Addiction*, 101(11), 1546–60.

Rawson, R.A., et al. (2004). A multi-site comparison of psychosocial approaches for the treatment of methamphetamine dependence. *Addiction (Abingdon, England)*, 99(6), 708–17.

Reif, S., et al. (2014a). Recovery housing: assessing the evidence. *Psychiatric Services (Washington, D.C.)*, 65(3), 295–300.

Reif, S., et al. (2014b). Residential treatment for individuals with substance use disorders: assessing the evidence. *Psychiatric Services (Washington, D.C.)*, 65(3), 301–12.

Sayegh, C.S., Huey, S.J., Zara, E.J., and Jhaveri, K. (2017). Follow-up treatment effects of contingency management and motivational interviewing on substance use: a meta-analysis. *Psychology of Addictive Behaviors*, 31(4), 403–14.

Shoptaw, S., et al. (2013). Not just the needle: the state of HIV-prevention science among substance users and future directions. *Journal of Acquired Immune Deficiency Syndrome*, 63(Suppl 2), S174–8.

Skinner, B.F. (1938). *The Behaviour of Organisms: An Experimental Analysis*. New York: Appleton-Century.

Smith, D.J., Combellick, J., Jordan, A.E., and Hagan, H. (2015). Hepatitis C virus (HCV) disease progression in people who inject drugs (PWID): a systematic review and meta-analysis. *The International Journal on Drug Policy*, 26(10), 911–21.

Sobell, L.C. and Sobell, M.B. (1992). Timeline follow-back: a technique for assessing self-reported alcohol consumption. In: Litten, R.Z. and Allen, J.P. (eds.) *Measuring Alcohol Consumption: Psychosocial and Biochemical Methods*. Totowa, NJ: Humana Press, pp. 41–72.

Statistics, N.C.f.H. (2018). *National Drug Overdose Deaths Involving Prescription Opioids, Number Among All Ages, 1999–2017*. [Online] Available at: https://www.cdc.gov/nchs/

Stitzer, M., Bigelow, G., Lawrence, C., Cohen, J., D'Lugoff, B., and Hawthorne, J. (1977). Medication take-home as a reinforcer in a methadone maintenance program. *Addictive Behaviors*, 2(1), 9–14.

Stone, A.L., Becker, L.G., Huber, A.M., and Catalano, R.F. (2012). Review of risk and protective factors of substance use and problem use in emerging adulthood. *Addictive Behaviors*, 37(7), 747–75.

Substance Abuse and Mental Health Services Administration, Center for Behavioral Health Statistics and Quality. Treatment Episode Data Set (TEDS) (2017) (2019). *Admissions to and Discharges from Publicly-Funded Substance Use Treatment*. Rockville, MD: Substance Abuse and Mental Health Services Administration.

Uhlmann, S., et al. (2010). Methadone maintenance therapy promotes initiation of antiretroviral therapy among injection drug users. *Addiction*, 105(5), 907–13.

United Nations Office on Drugs and Crime (UNODC) (2019). *World Drug Report* [Online] Available at: https://wdr.unodc.org/wdr2019/

Van Ours, J.C. (2012). The long and winding road to cannabis legalization. *Addiction*, 107(5), 872–3.

World Health Organization (WHO) (2018). *International Classification of Diseases for Mortality and Morbidity Statistics* (11th ed.). Geneva: WHO.

World Health Organization (WHO) (2019). *World Health Organization Model List of Essential Medicines: 21st List 2019*. Geneva: WHO. [Online] Available at: https://www.who.int/medicines/publications/essentialmedicines/en/

World Health Organization, UNODC, United Nations AIDS. (2012). *WHO, UNODC, UNAIDS Technical Guide for Countries to Set Targets for Universal Access to HIV Prevention, Treatment and Care for Injecting Drug Users*. Geneva: WHO.

Zawilska, J.B. and Andrzejczak, D. (2015). Next generation of novel psychoactive substances on the horizon—a complex problem to face. *Drug & Alcohol Dependency*, **157**, 1–17.

9.3

Alcohol

Robin Room

Alcohol and its effects

Alcoholic beverages have been consumed in most, if not all, human societies since the beginning of recorded history. Beverages containing ethanol (C_2H_5OH) can be fermented from a large number of organic materials that comprise carbohydrates, and in one part of the world or another, these products are prepared from fruits, berries, various grains, plants, honey, or milk. Under most circumstances, such fermented beverages can contain up to 14% ethanol. The most widely commercialized fermented beverages are beer prepared from barley or other grains (usually 3–7% ethanol), apple and other fruit ciders (usually 3–7% ethanol), and grape wine (usually 8–14% ethanol). Other fermented beverages are also common in particular cultures, often from home production or in commercial form: for example, sorghum or millet beers in Eastern and Southern Africa, palm wine toddy in West Africa and the Indian subcontinent, pulque (prepared from the maguey cactus) in Mexico, and rice wine (sake) in Eastern Asia.

Distilled beverages, in which ethanol is concentrated by evaporation and condensation from a fermented liquid, came to Europe from Arabia in the Middle Ages. In Europe, at first, their use was primarily medicinal, but by the 1600s, popular use as a social beverage spread rapidly. Distilled beverages can be made up of almost-pure ethanol, but those sold for drinking usually contain between 25% and 50% ethanol. Distilled alcohol is also added to wine, producing ‘fortified wines’ with about 20% ethanol. Because distilled beverages and fortified wines do not readily spoil, they could be shipped over long distances, even before refrigeration and airtight packaging became available, and played a particularly important part in

commerce and exploitation in the age of the European empires. Different cultures consume varying strengths of alcoholic beverages, often with water or a ‘mixer’ being added to distilled beverages and, in some cultures, also to wine and other fermented beverages.

Use-values of alcohol

Ethanol has many uses in human life. These include non-beverage uses as a fuel and as a solvent. Important beverage-related use-values include use as a medicine, as a religious sacrament, as a foodstuff, and as a thirst-quencher (Mäkelä 1983). But alcoholic beverages have received special attention as a public health hazard not only as a risk factor for chronic disease but also because of their psychoactive properties. These properties carry with them another set of use-values: in terms of psychopharmacology, ethanol is classified as a depressant, although its effect can also be stimulating (Hendler et al. 2011), and alcoholic beverages have long been used to affect mood and feelings, whether in a calming or in a stimulating direction. With enough consumption, alcohol becomes an anodyne and, indeed, an anaesthetic; distilled spirits were used as an anaesthetic in surgical practice before the mid-nineteenth century. Many drinkers seek and appreciate the levels of intoxication which lie between mild mood alteration at one end of the spectrum and being comatose at the other.

The decisions to drink and how much to drink are, however, often not made by the individual in isolation. Drinking is usually a social act, and the pace and level of drinking are frequently subject to collective influence, with drinking together being seen as an expression of solidarity and community. Although drunkenness may be sought to relieve misery or loneliness, it is more commonly associated with sociable celebration.

Adverse effects

Alcohol consumption can have a variety of adverse effects, some of which are acute effects associated with the particular drinking event. Drinking progressively impairs physical coordination, cognition, and attention, resulting in an increased risk of accidents and injury. Above a threshold level, drinking can also affect intention and judgement, so intoxication potentially plays a causal role in violent behaviour and crime (Graham et al.

1998). This relation appears to be culturally mediated, because there is a substantial variation between cultures in the association of intoxication with violence and crime (Room 2001). A sufficient amount of alcohol may result in a potentially fatal overdose, by interrupting various autonomic bodily functions.

Other adverse effects of alcohol consumption are chronic effects related to a repeated pattern of drinking. Alcohol consumption can adversely affect nearly every organ of the body, although some effects are not common. Chronic conditions in which alcohol is implicated as an important cause include liver cirrhosis; cancers of the upper digestive tract, liver, bowel, and breast; cardiomyopathy; and gastritis and pancreatitis (Rehm et al. 2017). Through a variety of mechanisms, alcohol is also implicated in the incidence and course of infectious diseases (Parry et al. 2009).

Repeated heavy drinking can also adversely affect mental health; specific neurological disorders are associated with sustained heavy drinking. More common concomitants include depression and affective disorders. Alcoholism—the experience of loss of control over drinking, along with other psychological and physical sequelae—has also been considered a mental disorder in modern times. In current nosologies, alcoholism has been replaced by the terms *alcohol use disorder* (in *Diagnostic and Statistical Manual of Mental Disorders* [DSM-5] terminology) and *alcohol dependence* (in the *International Classification of Diseases* [ICD-11]) (Saunders 2017).

The impairment of coordination and judgement produced by drinking potentially impacts on bystanders and the drinkers' acquaintances, friends, and family, as well as on the drinker. The effects can be through impairment of coordination or judgement during the drinking event, resulting in injury or distress, or through impairment of performance in family, friendship, work, and other social roles as a result of recurring drinking episodes (Laslett et al. 2011). The actual and potential adverse effects on others have historically been the primary justification for alcohol controls and other societal responses to problematic drinking (Room 1996); the effects on the adult drinker's own health have been of much less importance in determining public policy on alcohol.

Positive effects

For the drinker, and sometimes for those around, alcohol consumption can have positive effects. We have already mentioned the different use-values of alcohol—which mean that drinkers are usually willing to pay more than just the cost of production and distribution of the beverage. Apart from its valued effects on mental state, alcohol use has some positive outcomes on health. By far the most important of these, in terms of public health, is its potential for preventing particular forms of cardiovascular disease (CVD). Drinking is also often bad for the heart (Chadwick and Goode 1998; Poikolainen 1999). Data from countries in the former Soviet Union, where a pattern of intermittent intoxication is common, support the strong adverse effect of binge drinking on heart disease mortality (Zaridze et al. 2009; Rehm and Room 2009). A fairly consistent finding is that, in terms of net effect on CVD, drinking small amounts regularly protect against CVD—although relatively little is known about how any protective effect interacts with or overlaps other protective factors, such as regular exercise, diet, or taking aspirin (acetyl salicylic acid [ASA]) or other pharmaceuticals (Iestra et al. 2005). The net protective effect of light drinking on CVD is indeed still in question (Zhao et al. 2017), including from a Mendelian randomization meta-analysis of data from 56 epidemiological studies which found no net beneficial cardiovascular effects even for light drinkers (Holmes et al. 2014). Some protective effect of light drinking has also been found for Type-2 diabetes (Rehm et al. 2017).

Research on cumulative effects of alcohol

Effects on the drinker's health

The relationship between amount of drinking and overall mortality was long thought of as a J-shaped curve, with abstainers, and often, very light drinkers showing a higher mortality than those drinking a little more. It was clear that any net protective effect was at older ages: studies limited to younger cohorts typically found a monotonic relationship between amount of drinking and mortality (Andréasson et al. 1991; Rehm and Sempos 1995). In the 2010s, however, there has been increasing argument against a net protective effect at any level of drinking: that apparent J-curve results reflected a number of confounding factors, including the selective nature and often limited follow-up period of many cohort studies. After finding

repeated attenuation of apparent protective effects as other factors were controlled, Knott et al. (2015) concluded that 'beneficial associations between low intensity alcohol consumption and all-cause mortality may in part be attributable to inappropriate selection of a referent group and weak adjustment for confounders'.

There is no disagreement that beyond relatively low cumulative volumes of drinking there are net adverse effects of drinking on physical health. A study in 7 European countries with diverse drinking patterns found that the risk of dying with drinking as a cause rose above 1 in 1,000 at levels of drinking ranging between 7 gm and 20 gm of ethanol per day—that is, at or below two drinks per day (Shield et al. 2017). The pattern of drinking is also a potentially important factor in mortality due to alcohol. Although this has long been obvious in injury deaths, there is growing recognition of its significance in other causes of death. Variations among cultures in drinking habits may partly explain the variations in the relation of volume of drinking to mortality in different countries (Shield et al. 2017).

In recent years, advice on 'low-risk drinking' or 'safe drinking' guidelines has been formulated and publicised in a number of countries (e.g. Chief Medical Officers 2016). Whereas earlier guidelines were inclined to be stated only in terms of the volume of drinking, in line with the measurement methods of medical epidemiology literature, more recent guidelines have also emphasized limits on the amount consumed on a given occasion or day (Stockwell and Room 2012).

Current literature on the cumulative effects of drinking on health relies substantially on summations of prospective epidemiological literature, following the tradition set up by English et al. (1995). Using meta-analysis on studies of the relationship between volume of drinking and specific causes of death in which alcohol was either a risk or protective factor, studies following this tradition derive attributable fractions for different levels of the volume of drinking and apply these fractions to segments of the population at each level in order to arrive at estimates on total lives and life-years lost and gained, subtracting the projected protective effects of alcohol from the negative burden. In addition to life-years lost, the study's most comprehensive indicator, disability-adjusted life years (DALYs), includes a projection of the burden of disability attributable to alcohol.

According to estimates for 2010, 5.0% of the total burden of disease globally (as measured in DALYs) is attributable to alcohol (Lim et al.

2012). This compares with 6.3% for tobacco and 1.0% for illegal drugs. The burden is considerably higher for males than for females; in the 2016 estimates, alcohol ranked fourth for men as a risk factor for overall health, but nineteenth for women (Risk Factors Collaborators 2017). As earlier estimates showed (Ezzati et al. 2002), the alcohol share of the burden is highest in high-income societies, including Eastern Europe and Northern Asia, as well as in Latin America. The relative position of alcohol among risk factors is actually highest in middle-income countries, where it ranks first; although the alcohol share of all DALYs is lower in other developing regions, this fraction is calculated on the basis of a higher total burden of disease and disability in these countries.

Alcohol as an issue in public health

Shifting societal responses to problematic drinking

Efforts to control problematic drinking date back to the beginning of recorded history. These efforts have been many-sided, including informal responses in the family and community, as well as governmental controls. Religious teachings and movements have often been directed against drinking or intoxication: Muslims are forbidden by their faith to drink at all, and drinking is also discouraged or forbidden in at least some branches of all the major world religions.

In the last few centuries, European and Europe-derived societies have been hosts to conflicting trends in terms of alcohol issues. The production of alcoholic beverages became an important part of European economies and of imperial domination and trade in the age of European colonization. Alcohol production and exports took on political importance not only in the wine cultures of Southern Europe, but also in such countries as the Netherlands and Britain. In the British colonies as well as in America, in the late-eighteenth century, distilled spirits were the only profitable way to get grain to market (Rorabaugh 1979). In recent decades, alcohol beverage industries have become increasingly internationalized and concentrated (Jernigan 2010), and multinational companies, mostly based in Europe or North America, have pressed with considerable success to open up global markets for alcohol.

Starting in the early 1800s, there were substantial waves of popular, and eventually, governmental response to the problems that were resulting from the very heavy consumption of alcoholic beverages in English-speaking and Northern European societies (Blocker 1989; Levine 1991). As a culmination of decades of popular temperance movements, in the early twentieth century, alcohol prohibition was adopted in 13 countries (Schrad 2010) and imposed on indigenous people in colonial territories, and stringent controls were placed on the availability of alcohol in some other countries. Although alcohol's impact on public order and morals and on family life were more central to the temperance movement thinking than the public health issues, mainstream thought in medicine and public health acknowledged the substantial adverse impacts of alcohol on health (Emerson 1932), and prohibition or an alternative, stringent controls on the availability of alcohol (Catlin 1931), were often identified with the public health interest.

In the United States, and some other societies which had prohibited or greatly restricted alcohol availability, there was a strong reaction against it by the early 1930s, with middle-class youth in the lead (Room 1984a, 1984b). In this cultural–political context, as the new generation moved into professional and research positions, adverse effects of alcohol were downplayed or denied (Herd 1992; Katcher 1993), and alcohol issues almost disappeared from public health textbooks and discourse. Any problems with drinking were seen as attributable to a relatively small cadre of alcoholics, unable to control their drinking because of a mysterious predisposing factor. As late as 1968, the main emphasis of the American Public Health Association was on building treatment capacity for alcoholism (Cross 1968).

The 'new public health' approach

The last three decades of the twentieth century saw the rise of what has been termed in the alcohol literature the 'new public health' approach (Beauchamp 1976; Tigerstedt 1999) to alcohol issues (although its approach is much less individualistic than what has been called the 'new public health approach' in other fields). This approach brought together several strands of research and philosophy. In contrast to a concept in terms of a unitary disease, 'alcoholism', the approach was premised on a

disaggregated approach: there was a diversity of alcohol-related problems, fairly widely distributed among the drinking population (Knupfer 1967; World Health Organization, Expert Committee on Problems Related to Alcohol Consumption 1980). It was noted that for many problems, the heaviest drinkers accounted for only a minority, because there were so many more drinking at somewhat lower levels (Moore and Gerstein 1981); picking up Rose's (Rose 1981) phrase, Kreitman (Kreitman 1986) termed this the 'preventive paradox'. Attention was, therefore, paid not only to the heaviest drinkers, but also to the whole range of drinking levels, and indeed, to the distribution of consumption in the population (Ledermann 1956; De Lint and Schmidt 1968). What happens with moderate drinkers, it was argued, influences the social climate for heavy drinking, because drinking is largely a social activity, marked by mutual influences and norms of reciprocity (Bruun et al. 1975a; Skog 1985). In a given population, it was found that rates of alcohol-related problems tend to rise and fall with changes in the level of consumption (Seeley 1960). Controls on the availability of alcohol, including taxes, affect the level of consumption, and thus, also the rates of alcohol-related problems (Seeley 1960; Terris 1967; Popham et al. 1976). The level of alcohol consumption in a population, and controls on alcohol availability, is thus seen as a public health concern, and part of a society's overall 'alcohol policy' (Bruun et al. 1975a).

In enumerating the elements of the approach, we have given references for early statements of each element. It will be seen that the strands of this approach were woven together gradually over a period of some years. A 1975 report by an international group of researchers (Bruun et al. 1975a) became a pivotal document for the approach. A few years later, the approach was given an authoritative endorsement in the United States by a committee of the National Academy of Sciences (Moore and Gerstein 1981). The most recent restatement of the approach by an international group of scholars appeared in 2010 (Babor et al. 2010), with a somewhat parallel analysis oriented to the developing world (Room et al. 2002).

The approach has had considerable influence on World Health Organization (WHO) programmes in the field of alcohol (Room 2005; WHO 2007). For instance, action on the availability and marketing of alcohol and on pricing policies are key elements of WHO's Global Strategy to Reduce the Harmful Effects of Alcohol, adopted in 2010 (WHO 2010). At national levels, there has been a considerable variation in its influence on

policy. In Sweden, where it is known as the Total Consumption Model, it attained dominance as the basis for official policy (Sutton 1997). However, policies based on this approach have been eroded as a consequence of Sweden's accession to the European Union (Holder et al. 1998). The approach also has had considerable backing in other Nordic countries.

In English-speaking countries, it has encountered substantial resistance in the cultural–political realm. Those allied with the alcoholic beverage industry have strongly attacked the approach, both in analyses and polemics (e.g. Mott 1991; Grant and Litvak 1998) and through direct political action to remove official proponents (Room 1984c). An approach that contemplates government regulation and influencing of private consumer choices is also unwelcome to those committed to consumer sovereignty and the primacy of individual choice (e.g. Peele 1987). Often, proponents of approaches seeking to 'domesticate' drinking—to reduce problems from drinking by integrating it into everyday life—have portrayed the new public health approach as antithetical to this (Olsson 1990), although some researchers have noted that there is no necessary antithesis (Whitehead 1979).

In terms of its influence on policy, the approach has undoubtedly had some effect in strengthening the defence of existing control structures and regulations. But efforts to get the approach adopted as the practical base for policy have met with resistance and failure in a number of countries (Hawks 1993; Room 2004). One response to this resistance has been some calls for an alternative approach (Stockwell et al. 1997), arguing that policy measures directed at heavy and problematic drinkers are more politically acceptable than measures directed at all drinkers. But the main public health response, as in the WHO's Global Strategy on alcohol (WHO 2010) and its 'best buys' on alcohol policy for tackling noncommunicable diseases (WHO 2017), has been to emphasize whole-of-population policies such as on availability and taxes.

The policy approach offered as an alternative is a focus on harm reduction, primarily by reducing instances of intoxication or insulating the drinkers from harm (Plant et al. 1997; Stimson et al. 2007). However, there is in fact usually no conflict between approaches aimed at total consumption and approaches aiming to reduce harm from heavy drinking. As Stockwell et al. (1997) noted, 'aggregate consumption levels are in fact likely to fall if effective (harm reduction) strategies are introduced'.

Conversely, many measures that affect the whole drinking population—taxation being a good example—are especially hard on heavier drinkers. Nor are targeted harm-reduction measures necessarily more politically acceptable than measures that affect all drinkers. Old systems of rationing and individual buyer surveillance (Järvinen 1991), which were directed specifically at restraining heavy drinking, are now politically unacceptable in any high-income society, although rationing, at least, was highly effective as a targeted prevention measure (Norström 1987).

Beyond its specific features, the controversy over the new public health approach replicates familiar patterns of controversy over public health approaches in general, particularly when those approaches impinge on familiar and valued patterns of behaviour, with substantial economic interests at stake. At the level of the knowledge base, the approach has had considerable success: the empirical evidence underlying the approach has considerably strengthened since the approach was first put forward. Thus the approach has remained dominant in public health thinking and advice, as in the WHO's Global Strategy on alcohol (WHO 2010) and its 'best buys' on alcohol policy for tackling noncommunicable diseases (WHO 2017). At a political level, however, the approach has had only limited success, and primarily in areas peripheral to its main focus—that is, in drinking–driving and minimum-age limits for drinkers.

Strategies of prevention and control and their effectiveness

Simplifying them, there are seven main strategies to minimize alcohol problems. One strategy is to educate or persuade people not to use alcohol or about ways to use it so as to limit harm. A second strategy, a kind of negative persuasion, is to deter drinking-related behaviour with the threat of penalties. A third strategy, in the positive direction, is to provide alternatives to drinking or to drink-related activities. A fourth strategy is in one way or another to insulate the user from harm. A fifth strategy is to regulate availability of alcohol or the conditions of its use; prohibition of supply may be regarded as a special case of such regulation. A sixth strategy is to work with social or religious movements oriented to reducing alcohol problems. And, a seventh strategy is to treat or otherwise help

people who have in trouble with their drinking habits. We will consider, in turn, these strategies and the evidence on their effectiveness.

Education and persuasion

In principle, education can be offered to any segment of the population in a variety of venues, but it is usually education of youth in schools that first comes to mind in the prevention of alcohol problems. Community-based prevention programmes, which are often also directed at adults, also may include an educational component.

Education offers new information or ways of thinking and leaves it to the listener to draw conclusions concerning beliefs and behaviour. However, most alcohol education programmes go beyond this. A common theme in North American evaluative literature on alcohol education is that 'knowledge-only' approaches do not result in changes in behaviour (Botvin 1995). School-based alcohol education has, thus, usually had a persuasional element, aiming to influence students in a particular direction.

Persuasion is directly concerned with changing beliefs or behaviours, and may or may not also offer information. Mass-media campaigns aimed at persuasion have been a very common component of prevention programmes for alcohol-related problems, but this can also be pursued through other media and modalities.

In most societies, public-health-oriented persuasion about alcohol must compete with a variety of other persuasive messages, including those intended to sell alcoholic beverages. The evidence that alcohol advertising influences teenagers and young adults towards increased drinking and problematic drinking is becoming stronger (Casswell et al. 2002). Even where alcohol advertising is not allowed in the mass media, these messages are often conveyed to consumers and potential consumers in a variety of other ways.

Evidence on effectiveness

The literature on effectiveness of educational approaches is dominated by studies on school-based education from the United States. This means that alcohol education has usually been in the context of drug and tobacco education and that the emphasis has been on abstention (Beck 1998) or at

least on delaying the start of drinking, in cultural circumstances where the median age of actually starting to drink is about 13 years although the minimum legal drinking age is 21 years. In general, despite the best efforts of a generation of researchers, this literature has had difficulty showing substantial and lasting effects (Foxcroft et al. 2003; Gorman et al. 2007). There is a good argument from general principles for alcohol education in the context of the consumer's health, but there is little evidence from the formal evaluation literature at this point of its effectiveness beyond the short term (Babor et al. 2010).

Persuasive media campaigns have also been a favourite modality in many places in recent decades. In general, evaluations of such campaigns have been able to demonstrate impacts on knowledge and awareness about substance use problems, but show little success in affecting attitudes and behaviours (Babor et al. 2010). As with school education approaches, there are hints in the literature that more success may come from influencing the community around the drinker—in terms of attitudes of significant others or popular support for alcohol policy measures—than from directly persuading the drinker himself or herself. Thus, media messages can be effective as agenda-setting mechanisms in the community, increasing or sustaining public support for other preventive strategies (Casswell et al. 1989).

Deterrence

In its broadest sense, deterrence means simply the threat of negative sanctions or disincentives for behaviour—a form of negative persuasion. Criminal laws deter in two ways: by general deterrence, which is the effect of the law in preventing a prohibited behaviour in the population as a whole; and specific deterrence, which is the effect of the law in discouraging those who have been caught from doing it again (Ross 1982). A law will have a greater preventive effect and be cheaper to administer to the extent it has a strong general deterrence effect.

Prohibitions on driving after drinking more than a specified amount are now in effect in most nations (WHO 2014, pp. 67–70). In many societies, there have also been laws against drinking in public places, against public drunkenness (being in a public place while intoxicated) and against obnoxious behaviour while intoxicated. Other common prohibitions are concerned with producing or selling alcoholic beverages outside state-

regulated channels and with aspects of drinking under a specified minimum age.

Evidence on effectiveness

Drinking–driving legislation, such as ‘per-se’ laws outlawing driving while at or above a defined blood-alcohol level, has been shown to be effective in changing behaviour and reducing rates of alcohol-related problems (Ross 1982; Hingson 1996; Babor et al. 2010). The effect is through both general and specific deterrence. The quickness and certainty of punishment as well as its severity are important in the deterrent value (too much severity tends to undercut the quickness and certainty). Drinking–driving is an ideal area for applying general deterrence, as the gains from breaking the law are limited and automobile drivers typically have something to lose by being caught.

Many English-speaking and Scandinavian countries have had a tradition of criminalizing drinking in public places or public drunkenness as such. The trend in the 1970s and following was to decriminalize public drunkenness, although in many places it remains illegal. In the 1990s and following, there has been some trend to criminalize drinking in specific public places (Pennay and Room 2012). Criminalization of such behaviours has some effect in moving behaviour around, but is not very effective in changing the behaviour of marginalized heavy drinkers who have little to lose.

Providing and encouraging alternative activities

Another strategy, in principle involving positive incentives, is to provide and seek to encourage activities that are an alternative to drinking or to activities closely associated with drinking. This includes initiatives such as making soft drinks available as an alternative to alcoholic beverages, providing locations for sociability as an alternative to taverns, and providing and encouraging recreational activities as an alternative to leisure activities involving drinking. Job creation and skill development programmes are other examples.

Evidence on effectiveness

‘Boredom’ and ‘because there’s nothing else to do’ are certainly among the reasons given by some for drinking. And, there are often good reasons of general social policy for providing and encouraging alternative activities. But as has been noted, the problem with alternatives to drinking is that drinking combines so well with many of them. Soft drinks are indeed an alternative to alcoholic beverages for quenching thirst, but they may also serve as a mixer in an alcoholic drink. Involvement in sports may go along with drinking as well as replace it. The few evaluation studies of providing alternative activities, again from a restricted number of societies, have generally not shown lasting effects on drinking behaviour (Moskowitz et al. 1983; Norman et al. 1997), although they undoubtedly often serve a general social purpose in broadening opportunities for the disadvantaged (Carmona and Stewart 1996).

Insulating use from harm

A major social strategy for reducing alcohol-related problems in many societies has been measures to separate the drinker, and particularly heavy drinkers, from potential harm. This separation can be physical (in terms of distance or walls), it can be temporal, or it can be cultural (e.g. defining the drinking occasion as ‘time out’ from normal responsibilities). These ‘harm-reduction’ strategies, as they are called in the context of illicit drugs, are often built into cultural arrangements around drinking, but can also be the object of purposive programmes and policies (Moore and Gerstein 1981), such as promotion of ‘designated drivers’, where one person in a social group is chosen to abstain and drive in the particular social situation (DeJong and Hingson 1998).

A variety of modifications to the driving environment positively affect casualties associated with drinking and driving, along with other casualties. These include mandatory use of seat belts, airbags, and improvements in the safety of vehicles and roads. Many other practical measures to separate intoxication episodes from casualties and other adverse consequences have been put into practice, although usually without formal evaluation.

Evidence on effectiveness

Drinking–driving countermeasures are a prime example of an approach in terms of insulating drinking behaviour from harm, as they seek to reduce alcohol-related traffic casualties without necessarily stopping or reducing alcohol use (Evans 1991). There is substantial evidence on the success of a range of such countermeasures, including environmental change approaches as well as deterrence (DeJong and Hingson 1998; Babor et al. 2010). Some environmental measures that reduce traffic casualties in general—such as requiring the wearing of seat belts in cars or providing sidewalks separated from the road—may prevent casualties associated with intoxication even more than other casualties.

Regulating the availability and conditions of use

In terms of the substantial harm to health and public order they can cause, alcoholic beverages are not ordinary commodities. Governments have, thus, often actively intervened in the markets for such beverages, far beyond usual levels of state intervention in markets for commodities.

Total prohibition can be viewed as an extreme form of regulation of the market. In this circumstance, where no one is licensed to sell alcohol, the state has no formal control over the conditions of the sales that occur nevertheless, and there are no legal sales interests, controlled through licensing, to cooperate with the state in market regulation.

With a general prohibition, typically, the consumption of alcohol does fall in the population, and there are also declines in the rates of the direct consequences of drinking such as cirrhosis or alcohol-related mental disorders (Moore and Gerstein 1981; Teasley 1992). But prohibition also brings with it characteristic negative consequences, including the emergence and growth of an illicit market and the crime associated with this. Partly for this reason, alcohol prohibition at a national level is now a live option only in Islamic societies, where the law reflects religious belief, which helps to sustain compliance. Local-area alcohol prohibitions are more widespread.

The features of alcohol control regimes that regulate the legal market in alcohol vary greatly. Special taxes on alcohol are very common, imposed often as much for revenue as for public health considerations. Most societies have minimum-age limits forbidding sales to underage customers and many have regulations forbidding sales to the already intoxicated.

Often, the regulations include limiting the number of sales outlets, restricting hours and days of sale, and limiting sales to special stores or drinking places. Rationing of alcohol purchases—limiting the amount individuals can buy in a given time period—has also been used as a means of regulating availability. Regulations restricting or forbidding advertising of alcoholic beverages attempt to limit or channel efforts by private interests to increase demand for particular alcoholic beverages. Such regulations potentially complement education and persuasion efforts. State monopolization on sales of some or all alcoholic beverages at the retail and/or wholesale level has also been commonly been used as a mechanism to minimize alcohol-related harm (Room 1993).

Effectiveness of specific types of regulation on availability

The decades since the 1970s have seen the development of a burgeoning literature on the effects of alcohol control measures. Reference guides for communities, summarizing the research evidence and attuned to particular national or regional conditions, are becoming available (e.g. Neves et al. 1998; Grover 1999). Specific types of regulation of the alcohol market, and the evidence on their effectiveness, are discussed as follows:

Minimum-age limits: A minimum-age limit is a partial prohibition, applied to one segment of the population. There is a strong evaluation literature showing the effectiveness of establishing and enforcing minimum-age limits in reducing alcohol-related problems (Babor et al. 2010). However, this literature is mostly based on North America, focusing mostly on youthful driving casualties and evaluating reduction from and increases to age 21 years as the limit, a minimum-age limit higher than in most societies. There is limited evidence on the applicability of the literature's findings in other societies and where youth cultures may be less automobile-focused (but see Møller 2002; Kypri et al. 2006).

Taxes and other price increases: Generally, consumers show some response on the price of alcoholic beverages, as on all other commodities. If the price goes up, the drinker will drink less; data from high-income societies suggests this is at least as true of the heavy drinker as of the occasional drinker (Babor et al. 2010). Studies have found that alcohol tax increases reduce the rates of traffic casualties, of cirrhosis mortality, and of incidents of violence (Cook 2007).

An alternative policy for price increases is the imposition of a minimum unit price or floor price—a requirement that alcoholic beverages not be sold at less than a specified price per unit of ethanol. Such a policy tends to differentially affect off-premise drinking—in homes or public places—since on-premise drinking includes servers' wages, etc. Heavier drinkers are more likely than moderate drinkers to be buying the cheapest alcohol (Callinan et al. 2015), so instituting a floor price is more targeted at problematic drinking than a tax increase. There is good evidence that setting a floor price reduces levels of alcohol-related hospitalizations and other problems (Stockwell et al. 2013).

Limiting sales outlets, and hours and conditions of sale: A substantial body of literature shows that levels and patterns of alcohol consumption, and rates of alcohol-related casualties and other problems, are influenced by such sales restrictions (Livingston 2013; Wilkinson et al. 2016; Sanchez-Ramirez and Voaklander 2018), which typically make the purchase of alcoholic beverages slightly inconvenient, influence the setting of and after drinking, or shorten late-night drinking sessions. If they are substantially enforced, rules influencing 'house policies' in drinking places on not serving intoxicated customers, for example, have also been shown to have some effect (Saltz 1997).

Monopolizing production or sale: Studies of the effects of privatizing retail alcohol monopolies have often shown some increase in levels of alcohol consumption and problems (Hahn et al. 2012), in part because the number of outlets and hours of sale typically increase with privatization (Her et al. 1999) and partly also because the new private interests typically exert political influence for further increases in availability. From a public health perspective, it is the retail level that is important, although monopolization of the production or wholesale level may facilitate revenue collection and effective control of the market.

Rationing sales: Rationing the amount of alcohol sold to an individual potentially directly impacts on heavy drinkers and has been shown to reduce levels both of intoxication-related problems such as violence and of drinking-history-related problems such as cirrhosis mortality (Schechter 1986; Norström 1987). But, although a form of rationing—the medical prescription system—is well accepted in most societies for psychoactive medications, general rationing of alcohol has been politically unacceptable nowadays for alcoholic beverages in high-income societies. Individual

control measures are coming back in play, but mostly in the form of individual bans on purchasing or drinking imposed by the justice system (Room 2012). Such bans, enforced by ankle bracelets measuring any alcohol consumption, have been shown to be effective in preventing drink-driving and alcohol-related intimate violence (Kilmer et al. 2013).

Advertising and promotion restrictions: Many societies have regulations on advertising and other promotion of sales of alcoholic beverages (WHO 2014, pp. 75–80). Although it is well accepted that advertising can strongly affect consumer choices on products in the market, it has proved difficult to measure the effects of advertising on demand for alcoholic beverages as a whole, partly because the effects are likely to be cumulative and long-term, making them difficult to measure. Bans of advertising in specific media do not seem to have much effect on overall consumption levels (Nelson 2010), perhaps because the resources which were spent on marketing through that channel will just be diverted to promotion on other media. Even the relatively restrictive French Évin Law on alcohol advertising does not appear to have shielded teenagers from the industry's marketing efforts (Gallopel-Morvan et al. 2017). However, there is some evidence of a relation between trends in teenage drinking and changes in the exposure of teenagers to alcohol advertising (White et al. 2017), and a systematic review of 20 studies found a 'relatively consistent association between alcohol promotion (predominantly advertising) and drinking outcomes in adolescence' (Scott et al. 2016).

Social and religious movements and community action

Substantial reductions in alcohol-related problems have often been the result of spontaneous social and religious movements, which put a major emphasis on quitting intoxication or drinking. Less common have been reductions in drinking without large-scale social movements, although such a reduction has happened among teenagers in many high-income countries in the early 2000s, with some variation in timing (de Looze et al. 2015: Kraus et al. 2018; Pennay et al. 2018). In recent decades, efforts have also been made to form partnerships between state organizations and non-governmental groups to work on alcohol problems, often at the level of the local community. There has been an active tradition of evaluated community action projects on alcohol problems, often using a range of

prevention strategies (Giesbrecht et al. 1990; Greenfield and Zimmerman 1993; Holmila 1997; Holder 1998; Room 2017). School-based prevention efforts have also moved increasingly to try to involve the community, in line with general perceptions that such multifaceted strategies will be more effective (Paglia and Room 1999).

Although some of the biggest historical reductions in alcohol problem rates have resulted from spontaneous and autonomous social or religious movements, support or collaboration from a government can easily be perceived as official co-optation or manipulation (Room 1997). Thus, there is considerable question about the extent to which such movements can or should become an instrument of government prevention policies. But adopting public health-oriented alcohol policies may become more feasible, and such policies can build on and support such changes.

Evidence on effectiveness

In the short term, movements of religious or cultural revival can be highly effective in reducing levels of drinking and of alcohol-related problems. Alcohol consumption in the United States fell by about half in the first flush of temperance enthusiasm in 1830–45 (Moore and Gerstein 1981). Rates of serious crime are reported to have fallen for a while to a fraction of their previous level in Ireland in the wake of Father Mathew's temperance crusade (Room 1983). The enthusiasm that sustains such movements tends to decay over time, although they often leave behind new customs and institutions of much longer duration. For instance, although the days when the historic temperance movement in English-speaking societies was strong are long gone, the movement had the long-lasting effect of largely removing drinking from the workplace in these societies. The community action projects in recent decades can show some success (Midford and Shakeshaft 2017), but their approaches have often been limited by a lack of power at the community level to implement policy change (Room 2017).

Particularly in the developing world, religious or cultural renewal movements oriented to reducing or prohibiting drinking are often a strong avenue of preventive action (Room et al. 2002). For instance, a reform movement among poor indigenous people in a region of Ecuador touched off in 1987 by religious renewal movements and an earthquake appears to have had lasting effects on popular sobriety (Butler 2006).

Treatment and other help

Providing effective treatment or other help for drinkers who find they cannot control their drinking can be regarded as an obligation of a just and humane society. The help can take several forms: a specific treatment system for alcohol problems, professional help in general health or welfare systems, or non-professional assistance in mutual-help movements such as Alcoholics Anonymous. To the extent such help is effective, it is also a means of preventing or reducing future alcohol-related problems in the person helped, although its effectiveness at a population level is less clear.

Treatments for alcohol problems need not be complex or expensive. The evaluation literature suggests that brief outpatient interventions aimed at changing cognitions and behaviour around drinking are as effective in most circumstances as longer and more intensive treatment (Long et al. 1998). Positive results from such interventions in primary healthcare settings have been clearly established (Kaner et al. 2018), including in a diverse range of countries in a WHO study (Babor et al. 1994).

Evidence on effectiveness

In terms of the effects of treatment on those who come for it, there is good evidence on the effectiveness of treatment for alcohol problems. Typically, the improvement rate from a single episode of treatment is about 20% higher than the no-treatment condition: further treatment episodes are often needed. Brief treatment interventions or mutual-help approaches usually result in net savings in social and health costs associated with the heavy drinker (at least where healthcare is not self-paid), as well as improving the quality of life (Holder et al. 1992; Holder and Cunningham 1992).

The effectiveness of providing treatment as a strategy for reducing rates of alcohol problems in a society is more equivocal. In a North American context, it has been argued that the steep increase in alcohol problems treatment provision and mutual-help group membership in recent decades contributed to reducing alcohol problems rates (Smart and Mann 1990). But the strength of the evidence for this contention is disputed (Holder 1997; Smart and Mann 1997). A treatment system for alcohol problems is an important part of an integrated national alcohol policy, but as an instrument of prevention—of reducing societal rates of alcohol problems—it is unlikely to be very cost-effective (Anderson et al. 2009).

Building integrated alcohol policies

Alcohol policy at a community or societal level

Often, the different strategies for preventing alcohol problems appear to be synergic in their effects (DeJong and Hingson 1998). For instance, controls on availability are more likely to be adopted, continued, and respected when the public has been successfully persuaded of their effects and effectiveness. But strategies can also work at cross-purposes: a prohibition policy, for instance, makes it difficult to pursue measures that insulate drinking from harm.

In a society where alcohol is a regular item of consumption, in view of the resulting rates of alcohol-related social and health problems, there is a strong justification for adopting a comprehensive policy concerning alcohol, taking into account production, marketing, and consumption, and the prevention and treatment of alcohol-related problems. In recent years, the idea that there should be an integrated alcohol policy at community or national levels, reaching across the many sectors of government and civil society that deal with alcohol issues, has become a common public health aim, although accomplishing this has often proved difficult (Smart and Mann 1997; Crombie et al. 2007; WHO 2010).

In terms of strategies we have reviewed for managing and reducing the rates of alcohol problems in the society, there is a clear evidence for effectiveness and cost-effectiveness of measures regulating the availability and conditions of use, and measures which insulate use from harm. With respect to some aspects of alcohol problems, notably drinking–driving, deterrence measures also fall in the same category. Along with availability (taxes and hours and days of sale), the WHO also identifies control of advertising and other marketing as a ‘best buy’ for public health policy on alcohol (WHO 2017). Despite their perennial popularity, evidence of the cost-effectiveness of education or persuasion and treatment strategies in reducing societal rates of problems is limited at best. Education and treatment are good things for a society and a government to be doing about alcohol problems, but they do not constitute in themselves a public health policy on alcohol (Babor et al. 2010). These strategies will nevertheless be pursued in most societies, and they can best be pursued with attention to

using cost-effective methods, and to integrating targets and messages with other aspects of alcohol policy.

Alcohol policy in a global perspective

Apart from agreements made a century ago among the European colonial powers about control of the spirits trade in Africa (Bruun et al. 1975b), there is little tradition of collaboration on alcohol policy at the international level. It has been largely up to each nation to cope on its own with the serious social and health problems associated with drinking. Although alcohol smuggling has a long history, the nation-state could usually rely on distances and traditional trade barriers to keep alcohol issues largely a matter within its borders, in terms of the supply as well as of the problems.

Since the 1980s, an accelerated rate of economic globalization has increasingly rendered obsolete the assumption that alcohol issues are local issues. This globalization affects alcohol issues in three main ways. The first of these is the influence of a global ideology of free markets. In its sweep, this ideology has caught up and dismantled a variety of market arrangements that served to hold down and to structure alcohol consumption. State and provincial alcohol monopolies in North America were weakened or dismantled (Her et al. 1999). In Eastern Europe and the countries in transition, alcohol monopolies were swept away along with most other governmental intrusions in the market (Moskalewicz 1993). Many of the municipally run beer halls in Southern African countries were privatized (Jernigan 1997). In line with the general ideology, privatization of alcohol production and distribution has been often suggested, abetted, or imposed on developing countries by international development agencies (White and Batia 1998). Where the market is organized in a state licensing system for private interests, free-market ideology has often overriden any concerns about public health risks or community need (Room 2004).

Second, trade agreements, trade dispute mechanisms, and the growth of new sales media have effectively reduced the ability of national and subnational governments to control their local alcohol markets. The influence of trade agreements and trade dispute decisions in breaking down alcohol controls, including control of price through taxation, has been most fully documented for North America (Room et al. 2006) and Europe (Tigerstedt 1990; Holder et al. 1998), but these mechanisms also operate in

the developing world. For instance, average taxes on alcoholic beverages in South Korea were lowered in 2000 as a result of complaints to the World Trade Organization by the European Union and the United States (Kim 2000). In 2010–2012, the EU and US joined by Australia, Canada, and five other countries, argued that Thailand's proposed warning labels on alcohol bottles would violate trade agreements (O'Brien 2013), and Thailand has been dissuaded from taking this step by the implied threat of a trade dispute. Global alcohol producers, acting through compliant countries, are successfully pushing for limits in trade agreements on the placement of warnings on the main label of bottles or cans (O'Brien et al. 2017).

Third, alcohol production, distribution, and marketing became increasingly globalized (Room et al. 2002; Jernigan 2010). Transnational alcohol companies expanded rapidly into low- and middle-income countries in search of new markets, benefiting from weak policy environments and the sweeping tide of market liberalization. Although most alcoholic beverages are still produced in the country in which they are sold, industrially produced beverages were increasingly produced in plants owned, co-owned, or licensed by multinational firms. To promote increased sales, these firms have been able to transform and step up the marketing techniques used in the national market, bringing to bear all the marketing resources and expertise they have developed in other markets.

In light of these converging trends, there is a growing need for mechanisms to express public health interests in alcohol issues at the international level, both in trade agreements and settlements of trade disputes, and in creating mutual obligations for one nation to back up rather than subvert the alcohol regulations and policies of another. There are growing calls for a framework convention on alcohol control to be negotiated under the WHO auspices or otherwise, on the model of the tobacco convention (A framework convention on alcohol control 2007; Room et al. 2008), although there has been some debate in public health about the advisability of this path (Taylor and Dhillon 2013; Room 2013).

Conclusion

The comparative risk analysis in the estimations of the global burden of disease has underlined the substantial role of alcohol in death and disability,

particularly in middle- and high-income countries. Over the last 30 years, a substantial literature has emerged which allows a differentiation of prevention strategies and policies in terms of their effectiveness. Some strategies—for example, school education, public information campaigns, and provision of alternatives—are often politically popular but have limited or no effect. A few strategies that have proved effective—notably, drinking–driving countermeasures—have been applied in a number of countries. Other effective strategies—especially controls on price and availability—have been widely resisted in the political process, and undercut by trade agreements. In a globalized world, control of the alcohol market in the interest of public health is needed not only at the local and national level, but also internationally. This need has led to calls for a framework convention on alcohol control, which would also provide an impetus and template for actions at national and subnational levels.

REFERENCES

A framework convention on alcohol control [editorial] (2007). *The Lancet*, 370(9593), 1102.

Anderson, P., Chisholm, D., and Fuhr, D.C. (2009). Effectiveness and cost-effectiveness of policies and programmes to reduce the harm caused by alcohol. *The Lancet*, 373(9682), 2234–46.

Andréasson, S., Romelsjö, A., and Allebeck, P. (1991). Alcohol, social factors, and mortality among young men. *British Journal of Addiction*, **86**, 877–87.

Babor, T., et al. (2010). *Alcohol: No Ordinary Commodity: Research and Public Policy* (2nd ed.). Oxford: Oxford University Press.

Babor, T.F., et al. (1994). Randomized clinical trial of brief interventions in primary health care: summary of a WHO project (with commentaries and a response). *Addiction*, **89**, 657–78.

Beauchamp, D. (1976). Exploring new ethics for public health: developing a fair alcohol policy. *Journal of Health Politics, Policy, and Law*, **1**, 338–54.

Beck, J. (1998). 100 years of 'just say no' versus 'just say know'. *Evaluation Review*, **22**, 15–45.

Blocker, J. (1989). *American Temperance Movements: Cycles of Reform*. Boston, MA: Twayne Publishers.

Botvin, G.J. (1995). Principles of prevention. In: Coombs, R.H. and Ziedonis, D. (eds.) *Handbook on Drug Abuse Prevention: A Comprehensive Strategy to Prevent the Abuse of Alcohol and Other Drugs*. Boston, MA: Allyn and Bacon, pp. 19–44.

Bruun, K., et al. (1975a). *Alcohol Control Policies in Public Health Perspective*. Helsinki: Finnish Foundation for Alcohol Studies, vol. 25.

Bruun, K., Rexed, I., and Pan, L. (1975b). *The Gentlemen's Club*. Chicago, IL: University of Chicago Press.

Butler, B.Y. (2006). *Holy Intoxication to Drunken Dissipation: Alcohol Among Quichua Speakers in Otavalo, Ecuador*. Albuquerque, NM: University of New Mexico Press.

Callinan, S., Room, R., Livingston, M., and Jiang, H. (2015). Who purchases low-cost alcohol in Australia? *Alcohol and Alcoholism*, 50(6), 647–653.

Carmona, M. and Stewart, K. (1996). *Review of Alternative Activities and Alternatives Programs in Youth-oriented Prevention*. Rockville (MD): Center for Substance Abuse Prevention. CSAP Technical Report 13.
Casswell, S., et al. (1989). Changes in public support for alcohol policies following a community-based campaign. *British Journal of Addiction*, **84**, 515–22.
Casswell, S., Pledger, M., and Pratap, S. (2002). Trajectories of drinking from 18 to 26 years: identification and prediction. *Addiction*, 97(11), 1427–37.
Catlin, G.E.G. (1931). *Liquor Control*. London: Thornton Butterworth.
Chadwick, D.J. and Goode, J.A. (eds.) (1998). *Alcohol and Cardiovascular Diseases*. Chichester: John Wiley & Sons.
Chief Medical Officers (2016). *UK Chief Medical Officers' Low Risk Drinking Guidelines 2016*. London: UK Department of Health. [Online] Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/545937/UK_CMOs__report.pdf
Cook, P.J. (2007). *Paying the Tab: The Costs and Benefits of Alcohol Control*. Princeton NJ: Princeton University Press.
Crombie, I.K., et al. (2007). How do public health policies tackle alcohol-related harm? A review of 12 developed countries. *Alcohol and Alcoholism*, 42(5), 492–9.
Cross, J.N. (1968). *Guide to the Community Control of Alcoholism*. New York: American Public Health Association.
De Lint, J. and Schmidt, W. (1968). The distribution of alcohol consumption in Ontario. *Quarterly Journal of Studies on Alcohol*, **29**, 968–73.
De Looze, M., et al. (2015). Decreases in adolescent weekly alcohol use in Europe and North America: evidence from 28 countries from 2002 to 2010. *European Journal of Public Health*, 25(suppl_2), 69–72.
DeJong, W. and Hingson, R. (1998). Strategies to reduce driving under the influence of alcohol. *Annual Review of Public Health*, **19**, 359–78.
Emerson, H. (ed.) (1932). *Alcohol and Man: The Effects of Alcohol on Man in Health and Disease*. New York: Macmillan.
English D.R., et al. (1995). *The Quantification of Drug Caused Morbidity and Mortality in Australia (2 vol.)*. Canberra: Australian Government Publishing Service.
Evans, L. (1991). *Traffic Safety and the Driver*. New York: Van Nostrand Reinhold.
Ezzati M., et al. (2002). The Comparative Risk Assessment Collaborating Group. Selected major risk factors and global and regional burden of disease. *The Lancet*, **360**, 1347–60.
Foxcroft D.R., et al. (2003). Longer-term primary prevention for alcohol misuse in young people: a systematic review. *Addiction*, 98(Suppl 4), 397–411.
Gallopel-Morvan, K., Spilka, S., Mutatayi, C., Rigaud, A., Lecas, F., and Beck, F. (2017). France's Évin Law on the control of alcohol advertising: content, effectiveness and limitations. *Addiction*, **112**, 86–93.
Giesbrecht, N., et al. (eds.) (1990). *Research, Action and the Community: Experiences in the Prevention of Alcohol and Other Drug Problems*. Rockville, MD: Office of Substance Abuse Prevention. DHHS Publication No. (ADM) 89–1651.
Gorman, D.M., Conde, E., and Huber, Jr. J.C. (2007). The creation of evidence in 'evidence-based' drug prevention: a critique of the Strengthening Families Program Plus Life Skills Training evaluation. *Drug and Alcohol Review*, 26(6), 585–93.
Graham K., et al. (1998). Current directions in research on understanding and preventing intoxicated aggression. *Addiction*, **93**, 659–76.
Grant, M. and Litvak, J. (1998). Introduction: beyond per capita consumption. In: Grant, M., and Litvak, J. (eds.) *Drinking Patterns and Their Consequences*. Washington, DC: Taylor and Francis, pp. 1–4.

Greenfield, T. and Zimmerman, R. (eds.) (1993). *Experiences with Community Action Projects: New Research in the Prevention of Alcohol and Other Drug Problems*. Rockville, MD: Center for Substance Abuse Prevention. DHHS Publication No. (ADM) 93–1976.

Grover, P.T. (ed.) (1999). *Preventing Problems Related to Alcohol Availability—Environmental Approaches: Reference Guide*. Washington (DC): Center for Substance Abuse Prevention. DHHS Publication No. SMA 99–3298. [Online] Available at: http://text.nlm.nih.gov/ftrs/dbaccess/csap

Hahn, R.A., et al. (2012) Effects of alcohol retail privatization on excessive alcohol consumption and related harms: a community guide systematic review. *American Journal of Preventive Medicine*, 42(4), 418–27.

Hawks, D. (1993). The formulation of Australia's national health policy on alcohol. *Addiction*, 88(Suppl), 19S–26S.

Hendler, R.A., Ramchandani, V.A., Gilman, J., and Hommer, D.W. (2011). Stimulant and sedative effects of alcohol. In: Sommer, W. and Spanagel, R. (eds.) *Behavioral Neurobiology of Alcohol Addiction*. Berlin & Heidelberg: Springer, pp. 489–509.

Her, M., et al. (1999). Privatizing alcohol sales and alcohol consumption: evidence and implications. *Addiction*, **94**, 1125–39.

Herd, D. (1992). Ideology, history, and changing models of liver cirrhosis epidemiology. *British Journal of Addiction*, **87**, 179–92.

Hingson, R. (1996). Prevention of drinking and driving. *Alcohol Health and Research World*, **20**, 219–26.

Holder, H.D. (1998). *Alcohol and the Community: A Systems Approach to Prevention*. Cambridge: Cambridge University Press.

Holder, H. (1997). Can individually directed interventions reduce population-level alcohol-involved problems? *Addiction*, **92**, 5–7.

Holder, H.D. and Cunningham, D.W. (1992). Alcoholism treatment for employees and family members: its effect on health care costs. *Alcohol Health and Research World*, **16**, 149–53.

Holder, H.D., Lennox, R.D.L., and Blose, J.O. (1992). Economic benefits of alcoholism treatment: a summary of twenty years of research. *Journal of Employee Assistance Research*, **1**, 63–82.

Holder, H.D., Kühlhorn E., Nordlund S. et al. (1998). *European integration and Nordic alcohol policies*. Aldershot: Ashgate.

Holmes, M.V., et al., on behalf of The InterAct Consortium (2014). Association between alcohol and cardiovascular disease: Mendelian randomisation analysis based on individual participant data. *British Medical Journal*, 349, g4164.

Holmila, M., (ed.) (1997). *Community Prevention of Alcohol Problems*. Basingstoke: Macmillan.

Iestra, J.A., Kromhout, D., Van der Schouw, Y.T., Grobbee, D.E., Boshuizen, H.C., and Van Staveren, W.A. (2005). Effect size estimates of lifestyle and dietary changes on all-cause mortality in coronary artery disease patients: a systematic review. *Circulation*, 112(6), 924–34.

Järvinen, M. (1991). The controlled controllers: women, men, and alcohol. *Contemporary Drug Problems*, **18**, 389–406.

Jernigan, D.H. (1997). *Thirsting for Markets: The Global Impact of Corporate Alcohol*. San Rafael, CA: Marin Institute for the Prevention of Alcohol and Other Drug Problems.

Jernigan, D.H. (2010). The extent of global alcohol marketing and its impact on youth. *Contemporary Drug Problems*, **37**, 57–89.

Kaner, E.F.S., et al. (2018) Effectiveness of brief alcohol interventions in primary care populations. *Cochrane Database of Systemic Reviews*, (2), CD004148.

Katcher, B.S. (1993). The post-repeal eclipse in knowledge about the harmful effects of alcohol. *Addiction*, **88**, 729–44.

Kilmer, B., Nicosia, N., Heaton, P., and Midgette, G. (2013). Efficacy of frequent monitoring with swift, certain, and modest sanctions for violations: insights from South Dakota's 24/7 Sobriety Project. *American Journal of Public Health*, 103(1), e37–e43.

Kim, H-R. (2000). Revised liquor taxes leave soju makers in lurch. *Korea Herald*, 13 March.

Knott, C.S., Coombs, N., Stamatakis, E., and Biddulph, J.P. (2015). All cause mortality and the case for age specific alcohol consumption guidelines: pooled analyses of up to 10 population based cohorts. *British Medical Journal*, 350, h384.

Knupfer, G. (1967). The epidemiology of problem drinking. *American Journal of Public Health*, **57**, 973–86.

Kraus, L., et al. (2018) 'Are the times a-changin'? Trends in adolescent substance use in Europe. *Addiction*, **113**, 1317–32.

Kreitman, N. (1986). Alcohol consumption and the preventive paradox. *British Journal of Addiction*, **81**, 353–63.

Kypri, K., et al. (2006). Minimum purchasing age for alcohol and traffic crash injuries among 15- to 19-year-olds in New Zealand. *American Journal of Public Health*, 96(1), 126–31.

Laslett A.-M., Room R., Ferris J., Wilkinson C., Livingston M., and Mugavin J. (2011). Surveying the range and magnitude of alcohol's harm to others in Australia. *Addiction*, 106(9), 1603–11.

Ledermann, S. (1956). *Alcool, Alcoolisme, Alcoolisation*. Paris: Presses Universitaires de France. INED Cahier No. 29.

Levine, H.G. (1991). Temperance cultures: concern about alcohol problems in Nordic and English-speaking cultures. In: Lader, M., Edwards, G., and Drummond, D.C. (eds.) *The Nature of Alcohol and Drug-related Problems*. Oxford: Oxford University Press, pp. 15–36.

Lim S.S., et al. (2012) A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010. *The Lancet*, **380**, 2224–60.

Livingston, M. (2013) To reduce alcohol-related harm we need to look beyond pubs and nightclubs. *Drug & Alcohol Review*, 32(2), 113–14.

Long C.G., Williams M., and Hollin C.R. (1998) Treating alcohol problems: a study of program effectiveness and cost-effectiveness according to length and delivery of treatment. *Addiction*, **93**, 561–71.

Mäkelä, K. (1983). The uses of alcohol and their cultural regulation. *Acta Sociologica*, **26**, 21–31.

Midford, R. and Shakeshaft, A. (2017). Community-focused approaches to the prevention of alcohol-related harms: from past experience to future possibilities. In: Kolind, T., Thom, B., and Hunt, G. (eds.) *The SAGE Handbook of Drug & Alcohol Studies: Social Science Approaches*. Los Angeles, CA: SAGE, pp. 213–37.

Møller, L. (2002). Legal restrictions resulted in a reduction of alcohol consumption among young people in Denmark. In: Room R. (ed.) *The Effects of Nordic Alcohol Policies: What Happens to Drinking When Alcohol Controls Change?* Helsinki: Nordic Council for Alcohol and Drug Research. NAD Publication No. 42, pp. 93–100. Available at: http://www.drugslibrary.stir.ac.uk/documents/nad42.pdf

Moore, M.H. and Gerstein, D.R. (eds.) (1981). *Alcohol and Public Policy: Beyond the Shadow of Prohibition*. Washington, DC: National Academy Press.

Moskalewicz, J. (1993). Privatization of the alcohol arena in Poland. *Contemporary Drug Problems*, **20**, 63–275.

Moskowitz J.M., et al. (1983). Evaluation of a junior high school primary prevention program. *Addictive Behaviors*, **8**, 393–401.

Mott, G. (1991). The anti-alcohol network. *Moderation Reader*, 5(5), 6–20.

Nelson, J.P. (2010). Alcohol advertising bans, consumption and control policies in seventeen OECD countries, 1975–2000. *Applied Economics*, 42(7), 803–23.

Neves P., et al. (1998). *Communities Take Action! A Practical Guide for Municipalities, Enforcement Agencies, Community Groups, and Others Concerned About the Impact of Alcohol on Public Health and Safety*. Toronto: Addiction Research Foundation.

Norman E., et al. (1997). Prevention programs reviewed: what works? In: Norman, E. (ed.) *Drug-free Youth: A Compendium for Prevention Specialists*. New York: Garland Publishing, pp. 22–45.

Norström, T. (1987). Abolition of the Swedish rationing system: effects on consumption distribution and cirrhosis mortality. *British Journal of Addiction*, **82**, 633–41

O'Brien, P. (2013). Australia's double standard on Thailand's alcohol warning labels. *Drug & Alcohol Review*, 32(1), 5–10.

O'Brien, P., Gleeson, D., Room, R., and Wilkinson, C. (2017) New directions in trade law for alcohol labelling: Implications of the Trans-Pacific Partnership Agreement for using public health information to reduce alcohol-related harm. *Melbourne University Law Review*, **41**, 341–91.

Olsson, B. (1990). Alkoholpolitik och alkoholens fenomenologi: uppfattningar som artikulerats i pressen [Alcohol policy and the phenomenology of alcohol: concepts articulated in the press]. *Alkoholpolitik*, **7**, 184–95.

Paglia, A. and Room, R. (1999). Preventing substance use problems among youth: a literature review and recommendations. *Journal of Primary Prevention*, **20**, 3–50.

Parry C., Rehm J., Poznyak V., and Room, R. (2009). Alcohol and infectious disease: an overlooked causal linkage? *Addiction*, 104(3), 331–2.

Peele, S. (1987). The limitations of control-of-supply models for explaining and preventing alcoholism and drug addiction. *Journal of Studies on Alcohol*, **48**, 61–77.

Pennay A. and Room R. (2012). Prohibiting public drinking in urban public spaces: a review of the evidence. *Drugs: Education, Prevention and Policy*, 19(2), 91–101.

Pennay, A., Holmes, J., Törrönen, J., Kraus, L., and Room, R. (2018). Researching the decline in adolescent drinking: the need for a global and generational approach. *Drug and Alcohol Review*, 37(Suppl. 1), S115–19.

Plant M., Single E., and Stockwell T. (eds.) (1997). *Alcohol: Minimizing the Harm—What Works?* London and New York: Free Association Books.

Poikolainen, K. (1999). It can be bad for the heart, too — drinking patterns and coronary heart disease. *Addiction*, **93**, 1757–9.

Popham R., Schmidt W., and de Lint J. (1976). The effects of legal restraint on drinking. In: Kissin B. and Begleiter H. (eds.) *The Biology of Alcoholism: vol. 4. Social Aspects of Alcoholism*. New York and London: Plenum, pp. 579–625.

Rehm, J., et al. (2017). The relationship between different dimensions of alcohol use and the burden of disease—an update, *Addiction*, 112(6), 968–1001.

Rehm J. and Sempos C.T. (1995). Alcohol consumption and all-cause mortality. *Addiction*, **90**, 471–80.

Rehm J. and Room R. (2009). A case study in how harmful alcohol consumption can be. *The Lancet*, **373**, 2176–77.

Risk Factors Collaborators (2017). Global, regional, and national comparative risk assessment of 84 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *The Lancet*, **390**, 1345–22.

Room, R. (1983). Alcohol and crime: behavioral aspects. In: Kadish, S. (ed.) *Encyclopedia of Crime and Justice*. Vol. 1. New York: Free Press, pp. 35–44.

Room, R. (1984a). Alcohol control and public health. *Annual Review of Public Health*, **5**, 293–317.

Room, R. (1984b). A ‘reverence for strong drink’: the lost generation and the elevation of alcohol in American culture. *Journal of Studies on Alcohol*, **45**, 540–6.
Room, R. (1984c). Former NIAAA directors look back: policy makers on the role of research. *Drinking and Drug Practices Surveyor*, **19**, 38–42.
Room, R. (1993). The evolution of alcohol monopolies and their relevance for public health. *Contemporary Drug Problems*, **20**, 169–87.
Room, R. (1996). Alcohol consumption and social harm—conceptual issues and historical perspectives. *Contemporary Drug Problems*, **23**, 373–88.
Room, R. (1997). Voluntary organizations and the state in the prevention of alcohol problems. *Drugs and Society*, **11**, 11–23.
Room, R. (2001). Intoxication and bad behaviour: understanding cultural differences in the link. *Social Science and Medicine*, **53**, 189–98.
Room, R. (2004). Disabling the public interest: alcohol policies and strategies for England. *Addiction*, **99**, 1083–9.
Room, R. (2005). Alcohol and the World Health Organization: the ups and downs of two decades. *Nordisk Alkohol- & Narkotikatidskrift*, **22** (English suppl), 146–62.
Room, R. (2012). Individualised control of drinkers: back to the future? *Contemporary Drug Problems*, 39(2), 311–43.
Room, R. (2013). Healthy is as healthy does: where will a voluntary code get us on international alcohol control? *Addiction* 108(3), 456–457.
Room, R. (2017). Prevention of alcohol-related problems in the community context. In: Giesbrecht, N. and Bosma, L. (eds.) *Preventing Alcohol-Related Problems: Evidence and Community-based Initiatives*, pp. 19–32. Washington, DC: American Public Health Association.
Room, R., et al. (2002). *Alcohol and Developing Societies: A Public Health Approach*. Helsinki: Finnish Foundation for Alcohol Studies.
Room, R., Giesbrecht, N., and Stoduto, G. (2006). Trade agreements and disputes. In: Giesbrecht, N., et al. (eds.) *Sober Reflections: Commerce, Public Health, and the Evolution of Alcohol Policy in Canada, 1980–2000*. Montreal and Kingston: McGill-Queen’s University Press, pp. 74–96.
Room R., Schmidt L., Rehm J., and Mäkelä P. (2008). International regulation of alcohol. *British Medical Journal*, 337(6), a2364.
Rorabaugh, W.J. (1979). *The Alcoholic Republic*. New York: Oxford University Press.
Rose, G. (1981). Strategy of prevention: lessons from cardiovascular disease. *British Medical Journal*, **282**, 1847–51.
Ross, H.L. (1982). *Deterring the Drinking Driver: Legal Policy and Social Control*. Lexington, MA: Lexington Books.
Saltz, R.F. (1997). Prevention where alcohol is sold and consumed: server intervention and responsible beverage service. In: Plant, M., Single, E., and Stockwell, T. (eds.) *Alcohol: Minimizing the Harm—What Works*? New York: Free Association Books, pp. 72–84.
Sanchez-Ramirez, D.C., and Voaklander, D. (2018). The impact of policies regulating alcohol trading hours and days on specific alcohol-related harms: a systematic review. *Injury Prevention*, 24(1), 94–100.
Saunders, J.B. (2017). Substance use and addictive disorders in DSM-5 and ICD 10 and the draft ICD 11. *Current Opinion in Psychiatry*, 30(4), 227–37.
Schechter, E.J. (1986). Alcohol rationing and control systems in Greenland. *Contemporary Drug Problems*, **13**, 587–620.
Schrad, M. (2010). *The Political Power of Bad Ideas: Networks, Institutions, and the Global Prohibition Wave*. Oxford: Oxford University Press.

Scott, S., Muirhead, C., Shucksmith, J., Tyrrell, R., and Kaner, E. (2016). Does industry-driven alcohol marketing influence adolescent drinking behaviour? A systematic review. *Alcohol and Alcoholism*, 52(1), 84–94.

Seeley, J.R. (1960). Death by liver cirrhosis and the price of beverage alcohol. *Canadian Medical Association Journal*, **83**, 1361–6.

Shield, K.D., et al. (2017). Life-time risk of mortality due to different levels of alcohol consumption in seven European countries: implications for low-risk drinking guidelines. *Addiction*, 112(9), 1535–44.

Skog, O.-J. (1985). The collectivity of drinking cultures: a theory of the distribution of alcohol consumption. *British Journal of Addiction*, **80**, 83–99.

Smart, R.G. and Mann, R.E. (1990). Are increased levels of treatment and Alcoholics Anonymous large enough to create the recent reduction in liver cirrhosis? *British Journal of Addiction*, **85**, 1385–7.

Smart R.G. and Mann R.E. (1997). Interventions into alcohol problems: what works? *Addiction*, **92**, 9–13.

Stimson G., et al. (eds.) (2007). *Drinking in Context: Patterns, Interventions, and Partnerships*. New York and Abingdon: Routledge.

Stockwell, T. and Room, R. (2012). Constructing and responding to low-risk drinking guidelines: conceptualisation, evidence and reception. *Drug and Alcohol Review*, 31(2), 121–5.

Stockwell T., et al. (1997). Sharpening the focus of alcohol policy from aggregate consumption to harm and risk reduction. *Addiction Research*, **5**, 1–9.

Stockwell, T., et al. (2013). Minimum alcohol prices and outlet densities in British Columbia, Canada: estimated impacts on alcohol-attributable hospital admissions. *American Journal of Public Health*, 103(11), 2014–20.

Sutton, C. (1997). *Swedish Alcohol Discourse: Constructions of a Social Problem*. Uppsala: Acta Universitatis Upsaliensis, Studia Sociologica Upsaliensia, p. 45.

Taylor A.L., and Dhillon I.S. (2013). Alternative legal strategies for alcohol control: not a framework convention—at least not right now. *Addiction*, 108(3), 450–5.

Teasley, D.L. (1992). Drug legalization and the 'lessons' of prohibition. *Contemporary Drug Problems*, **19**, 27–52.

Terris, M. (1967). Epidemiology of cirrhosis of the liver: national mortality data. *American Journal of Public Health*, **57**, 2076–88.

Tigerstedt, C. (1990). The European Community and the alcohol policy dimension. *Contemporary Drug Problems*, **17**, 461–79.

Tigerstedt, C. (1999). Alcohol policy, public health, and Kettil Bruun. *Contemporary Drug Problems*, **26**, 209–35.

White O.C. and Batia A. (1998). *Privatization in Africa*. Washington, DC: World Bank.

White, V., et al. (2017). Adolescents' exposure to paid alcohol advertising on television and their alcohol use: exploring associations over a 13-year period. *Addiction*, **112**, 1742–51.

Whitehead, P.C. (1979). Prevention of alcoholism. In: Robinson, D. (ed.) *Alcohol Problems*. New York: Holmes and Meier.

Wilkinson, C., Livingston, M., and Room, R. (2016). Impacts of changes to trading hours of liquor licences on alcohol-related harm: a systematic review 2005–2015. *Public Health Research and Practice*, 26(4), e2641644.

World Health Organization, Expert Committee on Problems Related to Alcohol Consumption (1980). *Problems Related to Alcohol Consumption*. Geneva: WHO. Technical Report Series 650.

World Health Organization, Expert Committee on Problems Related to Alcohol Consumption (2007). *Second Report*. Geneva: WHO. Technical Report Series 944.

World Health Organization (WHO) (2010). *Global Strategy to Reduce the Harmful Effects of Alcohol*. Geneva: WHO.

World Health Organization (WHO) (2014). *Global Status Report on Alcohol 2014*. Geneva: WHO. [Online] Available at: http://www.who.int/substance_abuse/publications/global_alcohol_report/en/

World Health Organization (WHO) (2017) . *Tackling NCDs: 'Best Buys' and Other Recommended Interventions for the Prevention and Control of Noncommunicable Diseases*. Geneva: WHO. [Online] Available at: http://www.who.int/iris/handle/10665/259232

Zaridze D., et al. (2009). Alcohol and cause-specific mortality in Russia: a retrospective case-control study of 48,557 adult deaths. *The Lancet*, **373**, 2201–14.

Zhao, J., Stockwell, T., Roemer, A., Naimi, T., and Chikritzhs, T. (2017). Alcohol consumption and mortality from coronary heart disease: an updated meta-analysis of cohort studies. *Journal of Studies on Alcohol and Drugs*, 78(3), 375–86.

9.4

Injury prevention and control

The public health approach

Corinne Peek-Asa and Adnan A. Hyder

Introduction

Injuries contribute significantly to premature life lost and years lived with disability for all countries, all regions of the world, and all age groups, and disproportionately impact the young. According to analyses of the 2016 Global Burden of Disease data, injuries cause over 4.6 million deaths per year, accounting for nearly 8.4% of all deaths and 10.7% of disability-adjusted life years (Global Burden of Disease Study 2016). For children between 5 and 14 years of age, road traffic crashes, drowning, conflict, and terrorism, falls, and interpersonal violence are all in the leading 15 causes of death. Deaths represent just a small proportion of the many injuries that cause potentially lifelong disability, and injuries also cause significant psychological trauma and financial loss. The burden of traumatic injuries necessitates that injury prevention be considered an international public health priority.

Because the majority of people will suffer multiple minor injuries throughout their lives, most of which cause only minor discomfort or inconvenience, populations and decision-makers may be lulled into believing that injuries are just part of life. However, a serious injury or the traumatic death of a loved one can completely change the course of the lives of those affected. The toll of these injuries is also extremely deleterious to families, communities, and societies.

Injuries can be prevented, and global investment in trauma and injury prevention will have significant long-term health and financial benefits. For high-income countries, progress has been made in reducing the toll of

injuries. In the United States, for example, road traffic crashes per million vehicle miles travelled decreased by 37% between 1994 and 2014 (National Highway Traffic Safety Administration 2017). Mortality rates have decreased for deaths from drowning, residential fires, homicide, poisoning, among others. However, these rates remain unacceptably high knowing that many effective prevention strategies have not yet been widely implemented. The World Health Organization (WHO) anticipates that, if current trends continue, road traffic injuries, self-harm and falls will all be among the leading 20 causes of death by the year 2030 (WHO 2018).

The burden of injury worldwide is disproportionately concentrated in lower-income countries. The WHO anticipates that, if current trends continue, road traffic injuries, violence and self-inflicted injuries will all be among the leading 20 causes of disability-adjusted life years lost by the year 2030 (WHO 2002, 2018). Road traffic crashes, which in 1990 ranked as the tenth leading cause of disability-adjusted life years lost, is ranked sixth in 2016 (GBD 2016). Operations of war will rise from the rank of 16 in 1990 to 8 by 2020, and interpersonal violence will rise from rank 19 to 12. Despite some successes in many areas of injury prevention, new risks, changing environments, and increasing population size constantly challenge injury control efforts.

Effective injury prevention strategies will require organization of the public health response and increased integration of professionals from many backgrounds. Modern injury control research combines ideas and skills from public health, biomechanics, engineering, behavioural sciences, law, law enforcement, medicine, and urban planning, among others. Research that identifies how effective interventions in high-income countries can be translated in low- and middle-income settings is a priority, but injury prevention strategies will need to be appropriate for local environments. Injury prevention activities that integrate multiple approaches within an organized public health response have a stronger chance of success.

This chapter will present the current state of knowledge regarding the burden of injuries and will introduce the basic concepts of building injury prevention infrastructure and capacity.

Causal model of injuries

Injuries are generally divided into the two broad categories of intentional and unintentional. Intentional injuries are those in which there was an intent to commit harm, either to oneself or someone else. Unintentional injuries occur without a direct intent to commit harm, even if gross negligence was involved. For example, a motor vehicle occupant death caused by a drunk driver would be considered an unintentional injury even though in many countries the driver could be prosecuted for a crime.

Injuries are further classified by their cause, such as motor vehicle occupant injuries, drowning, suicide/attempted suicide, homicide/assault, or residential fire injuries. Until the late 1990s, cause and intent were coded together, so that, for example, a poisoning death would not be distinguished as intentional or unintentional. In the 1990s, collaboration between the US Centers for Disease Control and Prevention and the American Public Health Association recommended that cause and intent be considered as separate components of describing an injury, and most data are now coded accordingly (Centers for Disease Control and Prevention 1997). Through this collaboration, a matrix to code injuries by intent and cause using the International Classification of Diseases, 9th revision, was developed. This matrix served as a template to create codes for injury cause and intent that are now included in the 10th revision of this coding system.

Injury causes are very broad and represent a diverse range of physical harm, and thus can be difficult to tie together conceptually. The traditional epidemiologic model for infectious diseases provides a framework for the epidemiologic study of traumatic injury (Fig. 9.4.1). At the centre of the causal pathway for injuries is the agent-host interaction. The agent, which in the case of injuries is energy, is absorbed by the host to cause injury. Energy can take many forms, such as mechanical, electrical, chemical, radiation, and thermal. An example of an agent-host relationship is a motor vehicle crash, in which the energy exerted on the individual is mechanical. The reservoir is the place in the environment where the agent is found. The potential for energy transfer exists everywhere, but its potential to cause injury is limited to specific conditions. For instance, the potential energy in a motor vehicle exists only when the car is being driven, and causes injury only when the vehicle crashes.

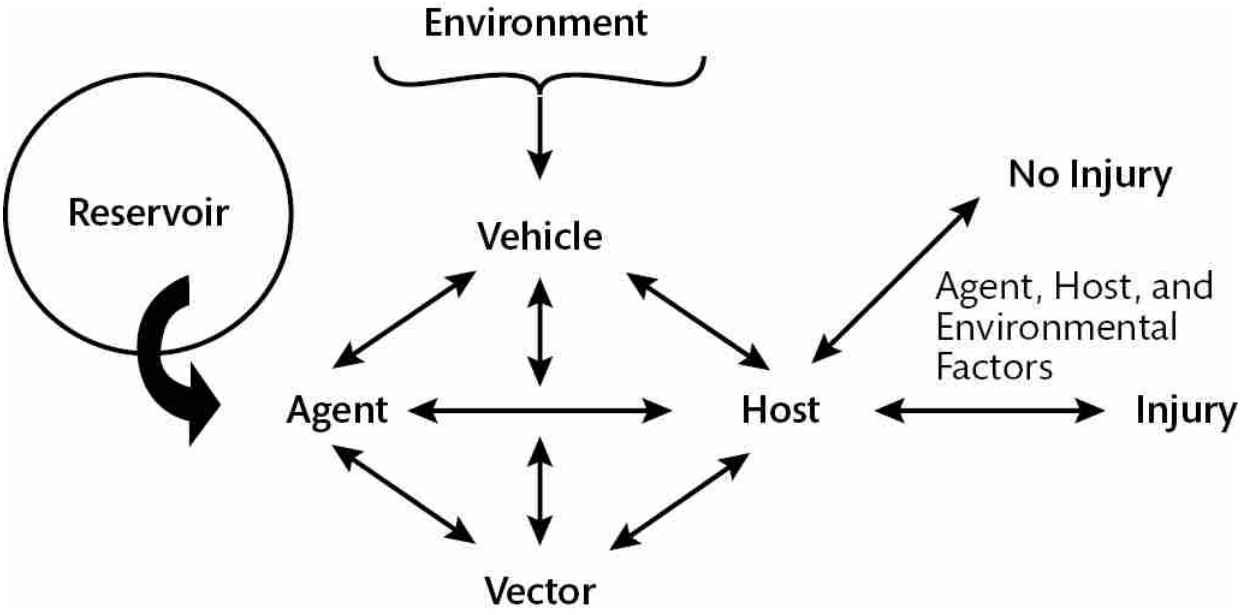


Fig. 9.4.1 Framework for the epidemiologic study of traumatic injury.

Vehicles and vectors are mechanisms which transport energy from the reservoir to the host. A vehicle can be an inanimate object, such as a motor vehicle; or animate, such as a dog biting a child. For many injury causes, vehicles and vectors are both involved in energy transfer, such as when one individual (vector) stabs another with a knife (vehicle). The injury outcome is the trauma or injury sustained by the individual, and is influenced by host responses to the energy. Only energy transmitted beyond a host's tolerance causes an injury, and therefore not all exposures to energy result in noticeable injury. A human has some resistance to energy which can be increased through exercise or protective devices, or reduced through changes in intrinsic factors such as existing medical conditions or age and through extrinsic factors such as fatigue and alcohol.

This causal model is important when considering injury prevention strategies. Injury prevention aims to prevent energy transfer or to reduce the amount of energy that is transferred. Prevention activities can focus on the host, on the environment, on vehicles or vectors, or combine all of these components. A theoretical approach to injury prevention based on this model is presented later in this chapter.

Injury Data Sources

Globally, improved data infrastructures and increased access to data has aided in the rising awareness of the burden of injury and violence as a health problem. Data that provide information related to injuries can be found from a wide variety of sources, both within and outside of the health sector (Institute of Medicine 1999; Hyder and Morrow 2006). Mortality data, most frequently collected through vital statistics reporting, is the most commonly used source for measuring the burden of injury. Most countries have some structure for death reporting that includes the cause of death, although the quality of these data varies markedly around the world.

Efforts to combine global mortality data to compare incidence, years of potential life lost, and disability-adjusted life years have improved our knowledge of the burden of injury. The Global Burden of Disease Study attempted to collect consistent and internationally comparable health data (GBD 2016) and is one of the most comprehensive sources of injury mortality data. The study is going to be updated each year and so estimates will change annually—including prior estimates which will be updated at the same time. The data, however, have limitations in terms of available data at country level, quality of reporting, and coding injury types, and thus are may be an underestimate of overall injuries. The recent 2016 Global Burden of Diseases, Injuries, and Risk Factors study is the most recent effort to compare incidence and risk factors for injuries and diseases (GBD 2016, 2017. This effort made several key advances in global data comparisons which included expanded and clarified definitions of causes of death; the integration of risk factors and risk behaviours; and, improved estimates to reduce uncertainty.

The WHO also reports injury data in both global and disease specific outlets. For example, WHO's Global Health Estimates are the main death and disability data generated by WHO and updated periodically (see: http://www.who.int/healthinfo/global_burden_disease/en/). In addition, road safety data is reported periodically through their Global Status Reports on Road Safety including data provided by countries and those modelled by WHO (see: http://www.who.int/violence_injury_prevention/road_safety_status/2015/en/).

Although these global estimates are extremely valuable for comparisons, they often provide little detail about contextual risk factors and circumstances for injuries. Detailed surveillance systems that focus on

specific types of injuries have been very helpful to identify and prioritize prevention approaches. For example, the Fatal Accident Reporting System of the United States Department of Transportation provides details of road traffic deaths and has helped improve car and road design and has assisted in development and implementation of safety policies. The Census of Fatal Occupational Injuries maintained by the United States Bureau of Labor Statistics has also helped compare workplace fatalities by occupation and industry and prioritized workplace hazards.

In comparison with mortality data, population-based data of non-fatal injuries and their outcomes are more challenging to collect and far less available, especially in the developing world. In high-income countries, population-based data describing non-fatal injuries can be found within hospital reporting data, often used to monitor national health indicators. Sources that include injuries as one of many health outcomes often lack detail to inform injury prevention approaches. Other sources are specific to certain types of injuries, such as trauma registries. Efforts to pool trauma data, such as the United States' National Trauma Data Bank or Canada's National Trauma Registry, are helpful in describing larger patient samples, but trauma registries can rarely be tied to a population base. Many data sources outside the health sector also cover a wide spectrum of injury-related events, such as data from law enforcement, the transportation sector, criminal and justice sectors, and insurance companies. This diversity of data sources makes the field of injuries and violence unique in terms of the intersectoral nature of the information (Norton et al. 2006).

In many instances, data sources can be linked together to provide a more complete description of the injury. Motor vehicle crash injuries, for example, often require data from traffic enforcement to describe the cause and nature of the crash, and data from the health sector to describe the injuries and their severity. Linking these data sources can be impeded by issues of privacy and access to identifying information for linkage, as well as data quality issues. In low- and middle-income countries these data systems, when they exist are often not computerized and have seldom been evaluated for quality, making such linkages even more challenging. However, information from multiple sources is often necessary to examine causal hypotheses about injuries.

A growing emphasis on improving health behaviour has led to global efforts to collect data on risk factors. Many countries conduct surveys of health behaviours, such as the United States' Youth Risk Behavior Survey

(Eaton et al. 2012). More recently, multicountry surveys have been launched to compare risk behaviours across countries. One example is the Health Behaviour in School-aged Children study, which surveys adolescents in more than 40 countries (Roberts et al. 2009). Evaluations of large scale prevention efforts can also provide important risk factor data, such as a collection of city-wide observational road safety risk factor studies conducted undertaken as part of the Bloomberg Initiative for Global Road Safety (Gupta et al. 2017). These surveys help establish ties between the prevalence of risk behaviours and experiences with injury and violence.

Despite improvements in data infrastructures that have been possible with technological advances, access to high-quality and detailed data on injury and violence remains an important limitation in most countries. Data limitations are particularly problematic in low-income countries. In many low-income countries, death reporting is the only major source of information about the burden of injuries, and these data can have gaps in reporting and coding. The public health infrastructure and routine collection of health information in the developing world has been fragile, especially in regions such as sub-Saharan Africa and South Asia. It is thus not surprising that there has been little tradition of developing specific information sources for injuries. Population-based studies from low- and middle-income countries, though, consistently conclude that the injury burden is higher than reported in national official statistics and that injuries are significantly underreported in these regions.

Global burden of injury

According to the 2016 Global Burden of Disease Study, over 4.6 million deaths occur as a result of injuries around the world (Table 9.4.1) (GBD 2016). More than 25% of these deaths are caused by road traffic injuries (RTIs), with self-inflicted injuries (self-harm) and falls comprising a further 17.7 and 14.7% respectively. Likewise, injury death rates are highest for RTIs, followed by self-inflicted injuries and falls; similar patterns are observed for non-fatal health outcomes using disability-adjusted life year (DALY) rates per 100,000 population (Fig. 9.4.2). Unintentional injuries, including RTIs, falls, drowning, fire burns, and poisoning, together result in more than 2.5 million deaths in 2016.

Table 9.4.1 Deaths from injuries worldwide, 2016

	Total	Females	Males
All injuries	**4,610,991**	**1,395,618**	**3,215,373**
Road injuries	1,342,284 (29.1)	318,141 (22.8)	1,024,143 (31.9)
Self-harm	817,148 (17.7)	257,624 (18.5)	559,523 (17.4)
Falls	678,458 (14.7)	294,501 (21.1)	383,957 (11.9)
Interpersonal violence	390,794 (8.5)	73,112 (5.2)	317,682 (9.9)
Drowning	302,932 (6.6)	87,836 (6.3)	215,096 (6.7)
Forces of nature, conflict, and terrorism	161,895 (3.5)	48,019 (3.4)	113,876 (3.5)
Exposure to mechanical forces	154,844 (3.4)	38,431 (2.8)	116,412 (3.6)
Fire, heat, and hot substances	132,084 (2.9)	64,708 (4.6)	67,376 (2.1)
Adverse effects of medical treatment	126,733 (2.7)	60,822 (4.4)	65,911 (2.0)
Foreign body	106,266 (2.3)	41,457 (3.0)	64,809 (2.0)
Other unintentional injuries	98,288 (2.1)	17,157 (1.2)	81,131 (2.5)
Other transport injuries	95,006 (2.1)	18,585 (1.3)	76,421 (2.4)
Animal contact	91,588 (2.0)	38,824 (2.8)	52,764 (1.6)
Poisonings	57,077 (1.2)	21,023 (1.5)	36,053 (1.1)
Environmental heat and cold exposure	55,596 (1.2)	15,375 (1.1)	40,221 (1.3)

Source: Global Burden of Disease Study 2016.

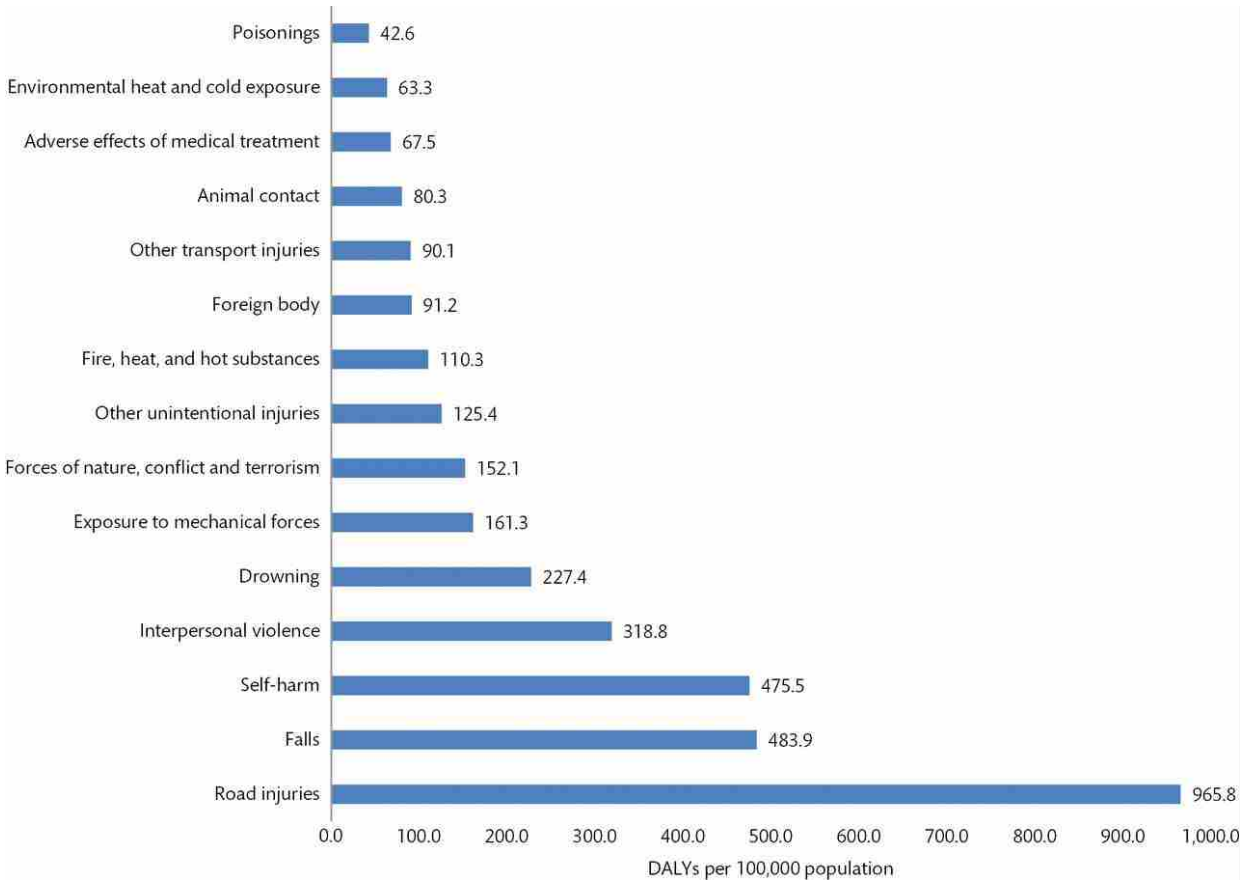


Fig. 9.4.2 DALY rates per 100,000 population.

RTIs alone kill over 1 million people every year. In 2016, RTIs became the ninth leading cause of death worldwide, down from eighth in 1990—a per cent change of 20.2%. Males account for more than three-fourth of these deaths. According to the WHO's *Global status report on road safety*, the majority of those killed are from low- or middle-income countries (LMICs) (WHO 2015). The absolute number of fatalities and the mortality rate resulting from RTIs vary considerably across countries, and although all age groups are affected young adults, particularly males, are most at risk of loss of life. Since this age group corresponds to the most economically productive segment of the population, RTIs have serious implications for national economies and also account for more than 71 million DALYs—some 965.8 DALYs per 100,000 population.

Falls are the second leading cause of unintentional injury deaths and there were over 678,000 fall-related deaths in 2016. This constitutes a mortality rate of 9.2 per 100,000 population. More males die from falls than females

(approximately 56.6% and 43.3% respectively). In addition, falls account for over 35 million DALYs or 483.9 DALYs per 100,000 population.

Drowning and fire burns are also significant causes of fatal injury—contributing 6.6% and 2.9% to the global burden of unintentional injury deaths respectively. Drowning amounts for over 300,000 deaths in 2016—more than 50,000 of these drowning deaths were among children under 5 years of age. Males comprise the majority (60%) of drowning deaths across all age groups till 70 years of age. The mortality rate from drowning is approximately 4.1 per 100,000. Non-fatal drowning events result in more than 16 million DALYs lost per year.

Well over 132,000 deaths in 2016 were caused by fire burns, which result in more than 8 million DALYs—110.3 DALYs per 100,000. Slightly more males than females die from fire burns (compare 51% and 49% respectively). In 2016, there were more than 57,000 fatal poisonings were estimated, which correlates to a mortality rate of nearly 0.77 per 100,000 population. Other unintentional injuries including unintentional exposure to firearms, and non-venomous animal contact, i.e. bites or attack, among others contributed 13.7% of the global burden of injuries and result in nearly 630,000 deaths worldwide.

In 1990 self-harm ranked fourteenth in the top 25 causes of worldwide deaths; and in 2016, they ranked seventeenth and are the second leading cause of injury death after RTIs, and the leading cause of intentional injury deaths. Self-inflicted injuries, including suicides, attempted suicide, self-destructive behaviours, and self-mutilation, cause the deaths of over 800,000 people globally, resulting in more than 35 million DALYs lost. Self-harm or self-inflicted injury deaths peak in the 20–24 years age group and more than 65% of self-inflicted injury deaths in 2016 were among males (68.4%). Although self-harm ranks twenty-first on the top 25 causes of death among females, overall, it ranks sixth in the Eastern Europe, tenth in Central Asia and South Asia, twelfth in Oceania, thirteenth in Southern Sub-Saharan Africa, and fifteenth in East Asian regions. On the other hand, self-harm mortality rates among females are highest in South Asia and Eastern Europe (12.7 and 1.4 per 100,000 population respectively).

Interpersonal violence, including assault by firearm, sharp objects or other means cause more than 390,000 deaths worldwide; and females account for less than 20% of these deaths. The mortality rate due to interpersonal violence is approximately 8.5 and 2.0 per 100,000 population for males and

females, respectively. However, the extent to which intimate partner violence (IPV) is accurately reported as interpersonal violence is unknown given varying cultural norms and reporting tendencies around the world. For example, a case of apparent self-immolation may in fact be domestic violence involving fire. Women in countries who have conducted programmes to screen and respond to IPV, including victim assistance programmes and social marketing campaigns against IPV, may be less inhibited to divulge their experiences with violence. Thus, the use of country or regionally specific demographic surveys as a supplement to global aggregate data is essential to determine the actual burden.

Conflict and terrorism were estimated to cause more than 150,000 deaths worldwide in 2016, more than 70% are males. Conflict and terrorism results in some 10,000,000 DALYs lost.

Economic and societal burden

The average annual cost of road traffic crashes was historically estimated at about 1% of the gross national product in developing countries, 1.5% in countries in economic transition, and 2% in highly motorized countries. The annual economic cost of RTIs globally was estimated at about US$518 billion (Jacobs et al. 2000). At that time, the annual cost of road traffic crashes in low- and middle-income countries was about US$65 billion, exceeding the total annual amount received then by these countries in development assistance. More recent estimates have been regional or national—and in general indicate that the cost of RTIs ranges from 1% to 3% of annual gross domestic product (or gross national product). Policies that reduce deaths and injuries from RTIs can save millions of dollars regardless of the income or development level of the country.

Studies exploring the economic and social costs of RTIs in low-income countries show that males who provided the majority of the household income in India and Bangladesh were the most common victims of road traffic fatalities and the consequences included reduced household income and reduced food consumption for the victim’s family (Aeron-Thomas et al. 2004). In addition, the poor were found to spend a much greater proportion of their income on funeral and/or medical costs than the non-poor. A recent qualitative study of the intangible costs of RTI in Mexico acknowledges similar findings but also emphasizes the loss of a family member to RTI can

impact not only a family's household economy, but the very stability of the familial unit—in whichever form—as well as the individual family member's sense of her or his life's purpose (Pérez-Núñez et al. 2012). A multicountry study from South Asia found that household with a road traffic injury patient tend to incur significantly high out-of-pocket costs on health-related expenditures, hospital care, and medicines, thus high lighting the need for mechanisms to provide financial risk protection for such households (Alam and Mahal 2016).

Estimates of the cost of violence in the United States reach over 3% of the gross domestic product. In England and Wales, the total costs from violence amount to an estimated US$40 billion annually. The economic effects of interpersonal violence are expected to be more severe in poorer countries, and yet, there is a scarcity of studies on the costs of violence in low- and middle-income countries (WHO 2004). However, estimates from low- and middle-income countries indicate that the overall costs of violence are substantial, ranging up to 25% of annual gross domestic product. Comparisons with high-income countries are complicated by the fact that economic losses related to productivity tend to be undervalued in lower-income countries. For example, child abuse results in US$94 billion in annual costs to the US economy (1% of GDP) and IPV costs the US economy nearly US$13 billion (0.1% of GDP); this can be compared to 1.6% of GDP in Nicaragua and 2% of GDP in Chile. The cost of gun violence also has been calculated at US$155 billion annually in the United States.

Theoretical basis for injury prevention

The causal model for injuries allows the injury process to be categorized into distinct phases which are important for prevention (Haddon 1970, 1972). The *pre-injury* phase is the period prior to the energy transfer, the *injury* phase is the often millisecond period in which the energy is transferred to the host, and the *post-injury* phase is the period of recovery and rehabilitation. Prevention approaches affect the injury process in one or more of these three injury phases.

Primary injury prevention includes strategies that aim to prevent the transfer of injury to the host, and thus act in the pre-injury phase. Examples of successful primary prevention strategies include roadway designs that

reduce motor vehicle collisions, child-resistant caps on medication bottles to prevent ingestion and poisoning, and pool fences to prevent submersion and drowning (Graham 1993; Thompson and Rivara 2000).

Secondary injury prevention includes strategies that reduce the amount of energy that is transferred to the host. While these strategies are often put in place long before the injury event, their function acts to reduce energy transferred during the injury phase. Examples of secondary prevention strategies include seat belts and air bags (Evans 1995), motorcycle and bicycle helmets (US General Accounting Office 1991; Thomas et al. 1994). Seat belts and helmets do not prevent the crash itself, but they reduce the amount of energy transmitted to the host during the crash. Secondary prevention strategies might not prevent injury completely. For example, lower extremity injuries are not prevented by either seat belts or helmets. However, studies have consistently shown that when these devices are present, the risk for mortal injury is much less and injury severity is reduced.

Tertiary prevention strategies act in the post-injury phase to help with recovery and rehabilitation once an injury has occurred. One example is the development of trauma systems, which help with triage and transport of an injured individual to reduce the time between injury and definitive medical care (Mann et al. 1999; Nathens et al. 2000). Optimizing resources to enable recovery and rehabilitation does not prevent the injury or reduce the amount of energy transfer, but tertiary prevention strategies can have enormous impact on improving survival, function, and quality of life following an injury.

The Haddon Matrix is a framework that combines these injury phases with the major components of the injury causal model to help identify prevention approaches (Table 9.4.2). This matrix, developed by Dr William Haddon in the United States, was the foundation for the study of motor vehicle crashes and countermeasures for highway safety and continues to be an applicable theoretical framework for injury prevention (Haddon 1972). Using the three injury phases, Haddon categorized prevention approaches into those that affect the host, vehicles, and vectors, the physical environment, and the sociocultural environment. An example of the Haddon Matrix with examples from motor vehicle occupant protection is included in Table 9.4.2.

Table 9.4.2 The Haddon Matrix for motor vehicle safety (illustrative)

Phases	Factors			
	Human	**Vehicle**	**Physical environment**	**Sociocultural environment**
Pre-injury	Reduce alcohol intoxication, programmes to increase defensive driving and decrease road rage	Increase vehicle stability, increased visibility	Improvements in road structure and traffic controls, traffic calming measures	Support safety programmes and increase consumer awareness of safety issues
Injury	Use of seat belts and car seats, proper placement of car seats, booster seats	Increase energy absorbed by the vehicle frame, safety features such as air bags, head rests, shatterproof windshields, and collapsible steering columns	Install energy-absorbing guard rails	Programmes to provide car seats, educational programmes about car seat installation and placement
Post-injury	Stabilize serious injuries, reduce bleeding and other complications	Design for easier extrication	Enhanced emergency medical systems and field care	Support infrastructure of trauma care, including 911 system, emergency and trauma care, and rehabilitation services

The Haddon Matrix provides a framework for identifying individual approaches to injury prevention. However, one single intervention strategy is unlikely to be highly effective. Multiple collaborative approaches need to be combined to maximize success. For example, the success of seat belt use in reducing motor vehicle occupant fatalities in the United States required a combination of engineering, education, policy, and enforcement that led to the current use rate of 80% (National Highway Traffic Safety Administration 2007). Although seat belts were developed and available from the early 1900s, they were not required to be a standard feature on passenger cars in the United States until 1968. The steps to getting seat belts installed as a regular feature in passenger cars required considerable advocacy and policy initiatives. However, once the seat belts were installed, use rates without any occupant incentive or education were about 11%. Efforts in the areas of

legislation, enforcement, and public education were necessary to achieve the high use rates currently observed. Stricter legislation, for example, is associated with higher use, as evidenced by the higher use rates consistently observed among states with primary (not wearing a seat belt is by itself a citable offence) compared with secondary (not wearing a seat belt can be cited only along with another citation) laws in the United States (National Highway Traffic Safety Administration 2007).

Risk factors for injuries

The Social-Ecological Model is helpful in conceptualizing the relationship of injury risk factors and also the level at which different types of prevention strategies work. The most common Social-Ecological Model used in injury prevention includes four levels: individual, relationship, community, and societal (Fig. 9.4.3) (Centers for Disease Control and Prevention 2009; Dahlberg and Krug 2002). This model posits that individual behaviour is largely determined by the influence of interpersonal relationships, community norms, and societal factors such as culture and political influence.

Fig. 9.4.3 Social-ecological model.

Individual-level variables that influence injury risk include age, gender, socioeconomic status, health status, and personality characteristics such as risk tolerance and the practice of safety behaviours. Many examples of the influence of individual-level characteristics on injury risk have been documented. For example, men have higher rates of injury than women for almost all causes of injury, and this difference is found consistently across the world. However, women appear to be at greater risk of fire-related burn injuries compared with men, and burn-related injuries account for a much higher proportion of injuries in young children when compared with other age groups (Jie and Ren 1992; Liu et al. 1998). Individual characteristics have been widely documented to influence the risk for motor vehicle-related deaths. Driver speeding, drinking, distraction, or drowsiness have been identified almost universally around the world as risk factors (Liu et al. 2003; WHO 2004). Declining resiliency and poor health status is an important individual risk factor for injuries among the elderly, particularly falls (Carpenter et al. 2009). Low bone density; poor nutritional status; low body mass index; low calcium intake; comorbid conditions (hypertension,

diabetes); poor performance in activities of daily living; poor cognitive function; poor vision; environmental factors affecting balance; family history of hip fracture; and alcohol consumption have all been identified as individual characteristics associated with increased fall risk (Cummings et al. 1995; Dargent-Molina et al. 1996; Clark et al. 1998; Boonyaratavej et al. 2001). Individual-level interventions often focus on increasing knowledge and encouraging behaviour change.

Relationship-level factors that predict injury include family and interpersonal relationship influences, including the environment in which these relationships occur. Interpersonal relationship factors can increase or decrease risk for injury. For children, parents are a key factors in safety; and limited adult supervision has been associated with pedestrian, drowning and poisoning injury risk in young children (Azizi et al. 1993; Soori 2001; Brenner 2003; Chadbunchachai et al. 2012). Parent communication style has been tied to risky driving attitudes of teens, and parent driving may influence the driving behaviour of their children (Mirman and Kay 2012; Yang et al. 2013). As we mature, peers and long-term partners become strong influences. The home environment can also influence injury risk. Low socioeconomic status, poor home conditions, and failure to remove home hazards are all factors associated with high and differential rates of home injury (Edelman 2007; Nyberg et al. 2012). Relationship-level interventions often focus on influential communication or changes in the home environment.

Community factors include schools, workplaces, neighbourhoods, all of which are influenced by community social norms and environments. Globally, community factors and characteristics vary more than individual characteristics. For example, the majority of drowning incidents in high-income countries are associated with recreational or occupational activities, while in most low- and middle-income countries they are associated with everyday activities near natural bodies of water (Kobusingye et al. 2001; Brenner 2003; Hyder et al. 2008). Rural and urban environments also have distinct and different patterns of injury, with injury rates from such causes as road traffic, drowning, poisoning, and occupational-related causes higher in rural communities (Peek-Asa et al. 2004). Prevention strategies usually are designed to impact the climate, processes, and policies in a given local community.

Societal factors encompass the cultural norms and socio-political characteristics that influence safety specifically, such as the safety culture. Broad social factors such as population health, socioeconomic status, and political stability also influence injury risk (CDC 2009). For example, injury death rates are far higher in low-income countries when compared to high-income countries. For road traffic deaths, motorization rates rise with income to a certain level (Kopits and Cropper 2003). In high-income countries, high motorization rates are not strongly correlated with road traffic deaths (Graham 1993). However, a growing number of LMIC that are experiencing very rapid increases in motorization have had a disproportionate increase in road traffic fatality rates (Ghaffar et al. 1999). Much of this increase is due to the lack of an organized safety culture, lack of road traffic safety laws and their enforcement, and governing agencies not prioritizing road safety. Prevention strategies at the societal level often focus on policy implementation and enforcement as well as efforts to change social norms.

Successful injury prevention strategies

Reducing the burden of injury requires multifaceted and comprehensive efforts. As demonstrated through the Haddon Matrix and the Socio-ecological Model, injuries occur as a complex interplay between individuals and their physical and sociocultural environments. Injury rates have very clear patterns and determinants that are affected by individual, family, community, geographic, and sociocultural factors. These factors and their interplay must be understood for effective intervention and prevention. Although individual, one-time programmes can have impact by increasing knowledge, encouraging behaviour change, or modifying an environment, long-term and sustainable reductions in injury usually require changes at the sociocultural level.

As the evidence base of successful injury prevention and intervention strategies increases, so do opportunities to translate these programmes to new populations. Translational research, both at the clinic and community level, is a promising priority for injury prevention. Considerations for translation of programmes to new populations include the priority of the injury problem in the new population; human and economic resources to implement and sustain the programme; and cultural relevance of the

programme (Woolf 2008). Injury prevention research, which tends to have a relatively short period between action and measurable impact, will likely gain in priority as nations realign research investments to focus on high-impact projects (Zerhouni 2005).

Resources to identify successful injury prevention programmes

A growing evidence base helps guide decisions on the most effective injury prevention strategies. In addition to information to measure the scope, burden, and cost of injuries, several good resources are available to identify evidence-based injury prevention approaches.

The Cochrane Collaboration provides scientific evidence-based reviews of healthcare interventions through the Cochrane Library. The library has more than 100 reviews on the prevention, treatment, and rehabilitation of traumatic injuries, including topics such as fall-related injuries to older persons, pool fencing to prevent drowning in children, and interventions for promoting smoke alarm ownership and function. Abstracts are available free of charge (http://www.cochrane.org/reviews/index.htm) and full reviews are available by subscription, with most major libraries holding subscriptions. Cochrane reviews are highly weighted towards evidence from randomized controlled trials, which excludes many of the observational evaluations conducted for injury interventions. The search strategy protocol requires inclusion of international findings as well as unpublished data, and when sufficient data are present meta-analyses are conducted. Thus, these reviews are meant to be applicable to an international audience. Recent additions on injury prevention include home safety education, alcohol interventions, and a synthesis of evaluations of the WHO Safe Communities programme (Spinks et al. 2009; Kendrick et al. 2012; Muckle et al. 2012).

The Campbell Collaboration, begun in 1999, includes systematic reviews of social intervention, divided into the categories of education, crime and justice, international development, and social welfare. The goal of the collaboration is to provide evidence for policy decisions regarding social issues. While the majority of injury prevention-related reviews are found within the crime and justice reviews, many reviews in the other categories are strongly related to injury and violence prevention. The Campbell Collaboration includes evidence from all types of study designs, and has methods groups to ensure systematic interpretation of findings. Recent

injury-related reviews address safety interventions to reduce workplace injury, community policing, and school-based bullying prevention programmes (Mazerolle et al. 2012; Farrington et al. 2008). Reviews are available at: http://www.campbellcollaboration.org/.

The Guide to Community Preventive Services: Systematic Reviews and Evidence-Based Recommendations (the *Guide*) was developed by the Task Force on Community Preventive Services, established by the United States Centers for Disease Control and Prevention. The *Guide's* recommendations are primarily based on evidence of effectiveness, including the suitability of the study design, but they also assess the applicability of the intervention to other populations or settings, the economic impact, barriers observed in implementing the interventions, and if the intervention had other beneficial or harmful effects. The *Guide* then provides a recommendation as to whether the approach is 'strongly recommended', 'recommended', has 'insufficient evidence' or is 'discouraged'. The Guide has injury-related reviews in the categories of alcohol, motor vehicle, physical activity, substance abuse, worksite, mental health, social environment, and violence. Reviews are available at http://www.thecommunityguide.org/default.htm.

The *Disease Control Priorities Network*, funded by the Bill and Melinda Gates Foundation, integrates economic evaluation to identify health priorities globally. The *Network* provides evidence on intervention impact, including cost-effectiveness, for the most prevalent global health conditions. This effort expands prior strategies by integrating economic elements and by placing a higher emphasis on evidence-based policies. The *Network* publishes the *Disease Priorities Handbook*, now in its third edition, that provides a synthesis of programme effectiveness and priorities by world region (Disease Control Priorities Network 2018).

Many global sources of information provide guidance on a variety of injury issues. The National Center for Injury Prevention and Control in the United States (http://www.cdc.gov/injury/) provides information about many injury prevention programmes and includes WISQARS, an injury and violence data query tool (http://www.cdc.gov/injury/wisqars/index.html). The Department of Violence and Injuries Prevention of the WHO (https://www.who.int) offers a diversity of guidelines and manuals that provide assistance in implementing programmes or conducting research. Examples include guidelines for establishing injury surveillance systems, conducting community-based surveys and evaluating prehospital care.

International non-governmental organizations also offer a variety of assistance and expert members on specific issues. Examples include the Society for the Advancement of Violence and Injury Research (http://www.savirweb.org/), the Safe States Alliance (http://safestates.org/), and the United Nations Road Safety Collaboration (http://www.who.int/roadsafety/en/).

Successful injury prevention strategies: road traffic injuries

With increased knowledge about the burden of RTIs globally there is growing attention to making roadways safer. Over the last several decades, many high-income countries have experienced decreases in motor vehicle occupant death rates, even while the total miles driven, and thus exposure to the roadway, have increased. In contrast, middle- and low-income countries have experienced large increases in road traffic fatality rates. Nearly 90% of road traffic deaths now occur in LMICs, although these countries have only 54% of the world's registered vehicles (WHO 2015). Worldwide, almost half of road traffic deaths are among pedestrians, cyclists, and riders of 2-wheeled vehicles. As the fleet of vehicles on these roadways gets more complicated, especially with a growing number of passenger and industrial vehicles, there is potential for continued increases in fatality rates. Efforts to reduce RTIs throughout the world are a good case study of how an integrated approach can lead to successful prevention.

The main reason for success in high-income countries has been a comprehensive approach to intervention that has included integration of environmental, engineering, legislative, and educational approaches. These changes have evolved over decades, leading to the development of a safety culture around road traffic. A "culture of safety" exists when safety is embraced as a fundamental goal at multiple levels: government, community, and individual. The paradigm shift to a traffic safety culture was first formalized in 1997 by the Swedish government with their Vision Zero initiative (http://www.visionzeroinitiative.com/). Since then, many other countries have adopted similar initiatives which include a traffic safety culture as the foundation.

Historically, as the number of passenger vehicles increased with industrialization, many high-income countries focused efforts on protecting vehicle occupants, and these efforts have shown many successes. In the

United States, The National Highway Traffic Safety Administration and the Federal Highway Safety Administration estimate that 243 000 lives were saved between 1966 and 1990 as a result of highway, traffic, and motor vehicle safety programmes (Institute of Medicine 1999). Road modifications mandated by the Federal Highway Traffic Safety Act of 1966 led to safer roadway infrastructures which allowed more cars to travel at higher speeds while reducing crash risk (Graham 1993). At the same time, changes to the motor vehicle itself (technological advancements to crashworthiness) increased the likelihood of survival in a crash. In addition to the implementation of seat belts and then air bags, legislation required modifications that improved vehicle crash worthiness. Some examples of these modifications include shatter-resistant windshields, collapsible steering columns, crash-friendly dashboard surfaces, frames that are resistant to passenger space intrusion, increased strength of the vehicle frame, seat, and doors, lap and shoulder belts, and airbags (Haddon and Baker 1981). Studies have found that these safety improvements to the motor vehicle led to a 24–40% decrease in fatal crashes and saved over 9,000 lives each year (Robertson 1981; Institute of Medicine 1999).

Most of these changes were due in part to legislative efforts, and these efforts were themselves multifaceted in nature. For example, the implementation of many of these laws required a successful engineering development as a foundation—improved crash worthiness of cars was based on concentrated research to identify the most successful design features, which were then required legislatively. Other legislative approaches are more behavioural in nature, such as laws that require seat belt use, driver training, speed limits, and laws that aim to reduce drunk driving. Implementation of these laws requires successful advocacy and education of legislators. However, their effectiveness varies dramatically. For example, most countries have laws against drinking and driving, yet countries have wide variation in the rate of alcohol-involved crashes as well as the number of drinking drivers, drivers cited for drinking and driving, and punished for drinking and driving. While this variation has much to do with factors associated with patterns of alcohol consumption, access to transportation, and the driving culture, this variation is also due to factors associated with the laws themselves. The effectiveness of legislative approaches depends on the level of enforcement that accompanies them and on the perception that the legislated consequences will actually occur (Mann et al. 2003). For

example, the effectiveness of drinking and driving laws is influenced by the level of enforcement, the ability of courts to impose penalties, and that the nature of the penalties serves as a deterrent.

The focus on motor vehicle occupants without concomitant integration of urban planning, sustainable transportation, and pedestrian and bicycle safety may be problematic in the long run. More recently, countries throughout Europe as well as Australia and the United States have introduced an approach called ‘traffic calming’ that aims to control motor vehicle traffic and speed in neighbourhoods, and thus increase safety and walkability in communities (Ewing 1999). The concept of traffic calming dates back to the Dutch city of Delft in the 1960s, when residents angry with fast-moving cars cutting through their local streets blocked them with gardens, social areas, and playgrounds (Ewing 1999). Modern traffic calming involves a variety of approaches to reduce traffic volume and speed in residential areas using visually appealing strategies.

Success in reducing road traffic deaths has depended on the development of an infrastructure that can identify and respond to road hazards, and a wide variety of approaches to address these hazards. In some developed and many developing countries, the car culture is growing without concomitant safety cultures (Bishai et al. 2003; WHO 2015). Highway design standards are often outdated or translated too directly from an industrialized country, and insufficient resources for roadway maintenance complicate the issue (Transport and Road Research Laboratory 1991). Poorly designed and maintained roadways are even further stressed by an ever-increasing number and variety of vehicles. In addition to a growing number of passenger cars and a high proportion of commercial vehicles and buses, developing countries’ roadways include animal-powered vehicles, small engine vehicles (such as scooters and motorcycles) that carry multiple people, bicycles, human-powered transport vehicles (such as a rickshaw), and pedestrians. This complex mix of vehicles contributes to increased deaths on the road.

In order to avert the growing toll of road traffic deaths, the WHO has called for a ‘systems approach’ in which the interaction between vehicles, roads, road users, and their physical, social, and economic environments form the basis for a multisectoral response (WHO 2004). The implementation of comprehensive traffic safety laws has been identified as one of the most important actions to reduce global RTIs, and these laws must include enforcement with appropriate penalties and concomitant public

awareness campaigns. The WHO has also called on governments to identify lead agencies in traffic safety, and to give these agencies a clear mission and authority for creating safer roads. Governments must also encourage multisectoral collaboration and support ongoing collection of data. These steps all fit within defined models of creating a global traffic safety culture.

Successful injury prevention strategies: drowning

Drowning prevention efforts often focus on three types of strategies: (1) remove, reduce, or change the water hazard; (2) change behaviour in risk taking, supervision or skills around water; and (3) prevent contact between people and the water environment (Wilson and Rogmans 2006). While these strategies remain relevant in all settings, drowning is experienced differently in HIC and LMIC settings. In HICs drowning events are often associated with recreational swimming, in swimming pools, or along the coast; while drowning in LMICs often occurs as a result of everyday activities near open bodies of water, ditches, or uncovered wells (Hyder et al. 2008). Therefore, successful drowning prevention strategies are often not transferable from HIC to LMICs.

Since the 1970s, HICs have implemented well-established preventive measures to prevent drowning events including: mandating fencing and self-closing or self-latching gates around domestic and public swimming pools; requiring supervision by certified lifeguards at public pools and swimming beaches; and recommending the use of personal flotation devices for boaters and novice swimmers (Thompson and Rivara 2000; Pearn et al. 2008; Centers for Disease Control and Prevention 2012). For adults, basic swimming instruction is an important skill, and deters the likelihood of drowning particularly in domestic or public swimming pools. Educational campaigns, such as one conducted in Australia to improve beachgoer recognition of rip currents to prevent surfing related drowning, are also effective preventive measures (Hatfield et al. 2011). The knowledge of basic life support or resuscitative measures, i.e. cardiopulmonary resuscitation, among non-medical personnel is also crucial to treat drowning victims. Alcohol is an emerging and important risk factor for drowning that ought to be banned whether at swimming pools, at swimming beaches, or on watercrafts including recreational boats through appropriate laws.

Many of these drowning prevention strategies are applicable to children though swim instruction thus far has been understudied and has yielded limited effectiveness (Asher et al. 1995). Attentive supervision of children in and around domestic swimming pools, swimming beaches, and around the home is crucial. And yet, a recent Cochrane review of home interventions found there was a lack of evidence that home safety interventions were effective in preventing children being left alone in the bath, which is a common cause of drowning among children (Kendrick et al. 2012).

In LMICs, the epidemiology of drowning is quite different as children comprise the largest burden. In Bangladesh, for example, drowning is one of the most common causes of death in children aged 1–4 years (Borse et al. 2011). More importantly, the critical risk factor–water–abounds; a single district alone might have 50,000 or more ditches and ponds[2]. Obviously, fencing off these bodies of water is not feasible in this context; however, other barrier-based interventions for drowning prevention in children less than 5 years have been piloted in rural Bangladesh (Callaghan et al. 2010). Door barriers (which prevent young children from leaving the home) and playpens (which limit movement of young children) appear to work in pilot settings, especially for children aged 1–3 years at high risk of drowning when their mothers are busy performing their daily household tasks. In addition, *crèches* (day care centres) have also shown promise by removing children from high risk contexts during the highest risk time of day in similar settings (Rahman et al. 2010). These interventions might be effective ways to reduce the number of drowning events and to ease the burden of supervision on mothers. Producing such barriers in the community or making them available at little or no cost to low-income families in LMICs, particularly in regions where drowning is most endemic, is an important agenda for global public health. However, large scale studies are needed to determine the true effectiveness of these interventions before they can be widely recommended.

Successful injury prevention strategies: Burns

There are sufficient data available from HICs to support the claim that burns can be successfully prevented using education, engineering changes, enforcement of legislative protection and environmental modifications

(Gielen et al. 2018). A range of successful burn prevention strategies such as smoke detectors, fire sprinklers, regulating hot water heater temperatures, mandating flame-resistant children's sleepwear, and enforcing safe and proper electrical wiring of residences and businesses have been implemented across HICs (Gielen et al. 2018). Installing smoke detectors in the United States, for example, has been associated with a 61% reduction in the risk of death from residential house fires (Marshall et al. 1998). This reduction has been made possible by programmes such as the Smoke Alarm Installation and Fire Safety Education of the National Center for Injury Prevention and Control (Centers for Disease Control and Prevention) which targets low-income households since 1998 and has installed approximately 553,000 smoke alarms with 3,755 lives potentially saved (Kress et al. 2012). Further, should someone in a HIC suffer a burn, she or he is likely to be transported to a dedicated burn centre in a timely fashion to receive specialized care necessary to reduce damage and restore functioning.

Many of these strategies are impractical for LMICs given certain institutional and environmental challenges and a different epidemiology of burns. Common risk factors are poor living conditions, overcrowding, lack of education, lack of infrastructure to minimize burn risk and damage and limited access to burn care and rehabilitation (Clark et al. 2018). For example, childhood burns risk factors include: use of kerosene, cooking in open spaces, overcrowding, and lack of parental supervision (Mashreky et al. 2010; Van Niekerk et al. 2010). Interventions that have been proposed include separating the cooking area from the living area, increasing height of cooking surfaces, making separate cooking stations outside the house, reducing the amount of flammable substances available in the household, use of safer oil lamps, and enhancing child supervision (WHO 2008).

Two vulnerable groups can be identified in LMICs—young infants and children and adolescent/adult females. For younger children, a partitioned area imposed during cooking hours, shelves at appropriate heights, supervision to avoid young children carrying hot objects are being suggested. In addition, timely outpatient assessment and treatment is vital to mitigating the impact of such burns, and creation of social awareness has been used for mothers and caretakers (Mukerji et al. 2001). Interventions for the second risk group (young women) include safe, inexpensive, and efficient kerosene lamps and stoves (see next) placed above the floor level (McLoughlin 1995). Other efforts at encouraging women (through

grassroots education and multimedia) to wear appropriate clothing to reduce clothes catching fire and using flame-resistant materials in the manufacturing of local dresses are being explored.

Some of these interventions have been reiterated by the WHO's world report on child injury prevention which also promotes the use of safe kerosene lamps, separating cooking areas from living areas, raising cooking facilities off the ground, and banning of fireworks and flame-retardant clothing materials (WHO 2008). In LMIC households, cookstoves are commonly used for cooking and heating and unfortunately, poor manufacturing standards, low quality control, and a lack of safety features have made this a key risk factor for burns (Peck et al. 2008). In response, the Global Alliance for Clean Cookstoves, led by the UN Foundation, seeks to encourage 100 million homes in LMICs to adopt clean and safe cook stoves by 2020 as well as support a clean and innovative cook stove industry. Such cook stoves will dramatically reduce fuel consumption, exposure to harmful emissions and smoke, and the risk of fire burns (http://www.cleancookstoves.org).

Successful injury prevention strategies: acute care

Many injuries lead to death because of inadequate emergency and trauma care, and prehospital deaths are disproportionately large in low- and middle-income countries. In low-income countries, prehospital mortality is extremely high. For example, one study found that 81% of trauma deaths occurred in the prehospital setting in Ghana, a low-income country, compared with 72% for Mexico, a middle-income country and 59% in the United States (Mock et al. 1998).

Improvements in emergency medical response and trauma care have had a major impact in preventing deaths from injuries. Reducing the time between injury and definitive treatment leads to improved outcomes (IOM 1999). The most effective way to reduce time to treatment is through the implementation of a trauma system, which changes in the social environment to include better coordination between agencies, triage, and transport protocols, assignment of treatment levels to trauma centres, and improved training and certification at all levels of emergency care delivery.

Early evaluations of trauma systems conducted in high-income countries focused on the number of deaths that could be prevented, and estimates were

in the range of 20% to 40% (Mackenzie et al. 1992; Mullins et al. 1994; Shackford et al. 1986). One meta-analysis of the effectiveness of prehospital trauma systems in low- and middle-income countries concluded that injury mortality could be reduced by approximately a quarter (Henry and Reingold 2012). Effects were strongest for middle-income countries, largely because they had stronger health infrastructures on which to build trauma systems. One study estimated that if injury case fatality rates in low- and middle-income countries could be reduced to the same level as those in high-income countries, more than 1.7 million deaths, over a third of all injury deaths, could be reduced globally (Mock et al. 2012).

Developing systems for coordinated care, especially when out-of-hospital coordination is involved, is a challenge. In many countries, no organized system to identify and transport patients to the hospital exist (Nielsen et al. 2012). In low-income countries and in rural areas, patients are more likely to arrive at the hospital in commercial (e.g. minibus, taxi) and private vehicles other than in an ambulance (Young et al. 2003; Nielsen et al. 2012). The private citizens who provide emergency transport are also the most likely to provide first aid in the absence of trained emergency response personnel. Because building emergency transportation infrastructures would be expensive, experts have proposed formalizing relationships with commercial vehicles and local citizens to create an organized approach (Mock et al. 2012).

Lack of trained specialists in emergency care, both in the field and in hospitals, is another barrier to reducing trauma death. Many countries lack trained medical personnel, equipment, and infrastructure (Kobusingye et al. 2005). An evaluation of trauma care capabilities in Mexico, Vietnam, India, and Ghana indicated that even when a sufficient number of healthcare professionals were trained, ‘brain drain’ from rural to urban and from low- to higher-income communities led to widespread provider shortages (Mock et al. 2006). In order to develop a stronger trauma response system, improvements must be made in the areas of increased human resources, physical resources, security for healthcare personnel, transportation systems to get patients to definitive treatment, and administrative infrastructure with effective leadership and governance (Kobusingye et al. 2005; Mock et al. 2006).

Although many of these improvements require new resources, system-wide changes are possible through effective planning even in the absence of

significantly increased resources. For example, several low-cost trauma training programmes have been developed and used effectively in low-income countries, and these have contributed to decreased injury mortality (Kobusingye et al. 2005). Administrative changes, such as making necessary diagnostic equipment available for critical trauma cases, are also low-cost. However, identifying and implementing low-cost strategies cycles back to the need for data. Finding the low-cost strategies that will work in a healthcare facility will be enhanced by data systems that can track trends in patient treatment and be used for quality assurance. These systems are lacking in many healthcare organizations throughout the world.

Improving emergency medical systems and trauma treatment capacity could greatly reduce injury mortality worldwide, and also reduce the physical, emotional, and financial consequences of severe injuries. These efforts need to be in collaboration with general worldwide efforts to improve access to public health and medical services. Over the past few years, WHO has released guidelines for prehospital and trauma care with a special focus on low- and middle-income countries (https://www.who.int), and approaches to improved acute care are increasingly evaluated in different countries.

Building capacity for injury prevention

In order to identify and implement successful injury prevention strategies, countries, states, and local communities need to develop necessary infrastructure and capacity to conduct the essential activities of an injury prevention programme. Infrastructure includes the designation or creation of agencies and leadership to oversee and integrate essential injury control activities within an intersectoral strategic framework. Capacity includes the availability of human, financial, relational, and structural resources. Both components are vital to the sustainability of any injury prevention and control programme. Since there is a lack of trained injury prevention professionals in many LMICs, capacity development efforts are critical in the developing world to address their high burden of injuries and unique challenges.

The WHO Department of Violence and Injury Prevention and Disability responded to this need by outlining three key domains for such efforts in *Capacity Building for Preventing Injuries and Violence: Strategic Plan 2009–2013*. These domains include:

Human resources: people and the knowledge and skills they require;
Institutional and infrastructural capacity: the systems and structures necessary to allow for effective human resources; and
Networks and partnerships: a means by which capacities can be strengthened within and across settings (http://www.who.int/violence_injury_prevention/capacitybuilding/strategic_plan/Strategic_plan.pdf).

In order to facilitate such efforts, the WHO has developed two training programmes, Teach Violence and Injury Prevention (TEACH-VIP) and Mentor Violence and Injury Prevention (MENTOR-VIP). TEACH-VIP, is a comprehensive injury prevention and control curriculum which is delivered via PowerPoint presentations and supplementary information structured around a set of lessons—some 'core' and others 'advanced' (http://www.who.int/violence_injury_prevention/capacitybuilding/teachvip/en/print.html). Whereas the TEACH-VIP training curriculum focuses on the transfer of knowledge, MENTOR-VIP offers an opportunity for junior investigators and injury prevention professionals to further develop key skills. MENTOR-VIP is a 12-month mentorship programme which matches mentees with leading injury prevention experts to accomplish specific goals (http://www.who.int/violence_injury_prevention/capacitybuilding/mentor_vip/en/).

The development of a critical mass of trained injury professionals within a country or region to understand the local environment, develop, implement, and evaluate appropriate prevention programmes is a crucial step. The Fogarty International Center of the National Institutes of Health in the United States has provided opportunities for twinning between US-based and developing country institutions for capacity development which targets all three WHO domains (http://www.fic.nih.gov/Programs/Pages/trauma-injury.aspx). These programmes cover collaborations in countries like China, Colombia, and Pakistan; topics from traumatic brain injuries, violence, to emergency care; and promote sustainable research and dissemination of results to influence policy and investments in each country.

Individual institutions and projects have also been successful in promoting capacity development initiatives including: online training programmes from universities (e.g. https://publichealth.gwu.edu/programs/graduate-summer-institute); funded foundation programs (e.g. Bloomberg Philanthropies);

advocacy training by the Global Alliance of NGOs for Road Safety (http://roadsafetyngos.org/); and others.

Conclusion

Despite the high burden of injuries throughout the world, investment in safety and injury prevention infrastructure has been low. In many countries, the amount invested in prevention is far surpassed by the amount spent on the treatment of traumatic injuries, and thus studies that focus on cost-benefits are badly needed. International assistance has also not prioritized safety or injury prevention. External assistance to the health sectors of low-income countries was US$2–3 per DALY lost to infectious diseases, but only US$0.06 per DALY lost to injuries (Mock et al. 2004).

Since 90% of the world's population lives in low- and middle-income countries, and the burden of injuries is predicted to increase in these countries, it is imperative that future research and development activities focus especially on their needs. Basic research to describe the existing burden, causes, and distribution of injuries is still needed in LMIC. Trials of injury interventions have largely not been conducted in LMIC, and there is a great need to modify, adapt, and test existing, as well as new interventions in these specific settings. Work is also required to assess barriers to implementation of such interventions globally.

The role of international organizations (such as WHO and the World Bank) and national agencies (such as ministries of health or medical research councils) needs to be emphasized in moving ahead the injury prevention agenda. The international movements currently underway for violence prevention and RTIs prevention as promulgated through the two World Reports are examples of how joint global-national partnerships are needed for making change.

Progress will also be supported if a growing number of professionals become advocates for safety. Professionals from public health, medicine, engineering, social services, urban planning, law, and law enforcement, among others, can all have a strong voice in raising awareness for safety programmes and to encourage individuals to make safe choices.

Key points

Injuries and violence are a leading cause of mortality and morbidity worldwide.

Injuries disproportionately affect young and vulnerable populations, and injury rates are disproportionately high among low- and middle-income countries.

Prevention of injuries is very feasible and works best with a multidisciplinary and integrated approach.

A growing body of research and resources can help identify potentially effective approaches to prevent and reduce the consequences of injury.

Building capacity for injury prevention should be a priority for all countries, but is an urgent need for low- and middle-income countries.

REFERENCES

Aeron-Thomas, A., Jacobs, G.D., Stexton, B., Gururaj, G., and Rahmann, F. (2004). *The Involvement and Impact of Road Crashes on the Poor: Bangladesh and India Case Studies*. Published Project Report PRP 010. Crowthorne: Transport Research Laboratory LTD.

Alam K and Mahal A. (2016). The economic burden of road traffic injuries on households in South Asia. *PLoS One*, 11(10), e0164362.

Asher, K.N., Rivara, F.P., Feliz, D., Vance, L., and Dunne, R. (1995). Water safety training as a potential means of reducing risk of young children's drowning. *Injury Prevention*, **1**, 228–33.

Azizi, B.H., Zulkifli, H.I., and Kasim, M.S. (1993). Risk factors for accidental poisoning in urban Malaysian children. *Annals of Tropical Paediatrics*, 13(2), 183–8.

Bishai, D., Hyder, A.A., Ghaffar, A., Morrow, R.H., and Kobusingye, O. (2003). Rates of public investment for road safety in developing countries: case studies of Uganda and Pakistan. *Health Policy and Planning*, 18(2), 232–5.

Boonyaratavej, N., Suriyawongpaisal, P., Takkinsatien, A., Wanvarie, S., Rajatanavin, R., and Apiyasawat, P. (2001). Physical activity and risk factors for hip fractures in Thai women. *Osteoporosis International*, 12(3), 244–8.

Borse, N., Hyder, A.A., Streatfield, P.K., Arifeen, S.E., and Bishai, D. (2011). Childhood drowning and traditional rescue measures: case study from Matlab, Bangladesh. *Archives of Disease in Childhood*, 96(7), 675–80.

Brenner, R.A. (2003). Prevention of drowning in infants, children and adolescents. *Pediatrics*, 112(2), 440–5.

Callaghan, J.A., Hyder, A.A., Khan, R., Blum, L.S., Arifeen, S., and Bacqui, A.H. (2010). Child supervision practices for drowning prevention in rural Bangladesh: a pilot study of supervision tools. *Journal of Epidemiology & Community Health*, 64(7), 645–7.

Carpenter, C.R., Scheatzle, M.D., D'Antonio, J.A., Ricci, P.T., and Coben, J.H. (2009). Identification of fall risk factors in older adult emergency department patients. *Academic Emergency Medicine*, 16(3), 211–19.

Centers for Disease Control and Prevention (CDC) and the Injury Control and Emergency Health Services Section of the American Public Health Association (1997). Recommended framework for presenting injury mortality data. *MMWR Recommendations & Reports*, 46(RR-14), 1–30. [Online] Available at: http://www.cdc.gov/mmwr/preview/mmwrhtml/00049162.htm

Centers for Disease Control and Prevention (CDC) (2009). Violence prevention: the social-ecological model: a framework for violence prevention. [Online] Available at: http://www.cdc.gov/ViolencePrevention/overview/social-ecologicalmodel.html

Centers for Disease Control and Prevention (2012). Drowning—United States, 2005–2009. *MMWR Morbity & Mortality Weekly Report*, 61(19), 344–7.

Chadbunchachai, W., Suphanchaimaj, W., Settasatien, A., and Jinwong, T. (2012). Road traffic injuries in Thailand: current situation. *Journal of the Medical Association of Thailand*, **95** (Suppl 7), S274–81.

Clark, A.T., Campbell, S., and Arnoldo, B.D. (2018). Prevention of burn injuries. Chapter 4. In: Herndon, D.N. (ed.) *Total Burn Care* (5th ed.). Edinburgh: Elsevier Inc.

Clark, P., de la Pena, F., Gomez Garcia, F., Orozco, J.A., and Tugwell, P. (1998). Risk factors for osteoporotic hip fractures in Mexicans. *Archives of Medical Research*, 29(3), 253–7.

Cummings, S.R., et al. (1995). Risk factors for hip fracture in white women. *New England Journal of Medicine*, 332(12), 767–73.

Dahlberg, L.L., and Krug, E.G. (2002). Violence—a global public health problem. In: Krug, E., Dahlberg, L.L., Mercy, J.A., Zwi, A.B., and Lozano, R. (eds.) *World Report on Violence and Health*. Geneva: WHO, pp. 1–56.

Dargent-Molina, P., et al. (1996). Fall-related factors and risk of hip fracture: The EPIDOS Prospective Study. *The Lancet*, 348(9021), 145–9.

Disease Control Priorities Network (2018). *Disease Priorities Handbook*, 3rd edition. Seattle, WA: University of Washington.

Eaton, D.K., et al. (2012). Youth risk behavior surveillance—United States. Centers for Disease Control and Prevention (CDC). *MMWR Surveillance Summaries*, 61(4), 1–162.

Edelman, L. S. (2007). Social and economic factors associated with the risk of burn injury. *Burns*, 33(8), 958–65.

Evans, L. (1995). Restraint effectiveness, occupant ejection from cars, and fatality reductions. *Accident Analysis & Prevention*, **22**, 167–75.

Ewing, R.H. (1999). *Traffic Calming: State of the Practice. Federal Highway Administration FHWA-RD-99-135*. Washington, DC. [Online] Available at: http://www.ite.org/traffic/tcstate.htm#tcsop

Farrington, D., Baldrey, A.C., and Kyvsgaard. (2008). Effectiveness of programs to prevent school bullying. *Victims & Offenders*, 2(2), 183–204.

Ghaffar, A., Hyder, A.A., Mastoor, M.I., and Shaikh, I. (1999). Injuries in Pakistan: directions for future health policy. *Health Policy Planning*, 14(1), 11–17.

Gielen, A.C., Frattaroli, S., Pollack, K.M., Peek-Asa, C., and Young, J.G. (2018). How the science of injury prevention contributes to advancing home fire safety: successes and opportunities. *Injury Prevention*, *24*(Suppl 1), i7–i13.

Global Burden of Disease Study (GBD) (2016). *Global Burden of Disease Study 2016 (GBD 2016) Results by Cause 1990–2016*. Seattle, WA: Institute for Health Metrics and Evaluation (IHME). [Online] Available at: http://www.healthdata.org/gbd

Global Burden of Disease (GBD) 2016 DALYs and HALE Collaborators (2017). Global, regional, and national disability-adjusted life-years (DALYs) for 333 diseases and injuries and healthy life expectancy (HALE) for 195 countries and territories, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *The Lancet*, 390(10100), 1260–344.

Graham, J.D. (1993). Injuries from traffic crashes: meeting the challenge. *Annual Review of Public Health*, **14**, 515–43.

Gupta, S., Paichadze, N., Gritsenko, E., Klyavin, V., Yurasova, E., and Hyder, A.A. (2017). Evaluation of the five-year Bloomberg Philanthropies Global Road Safety Program in the Russian Federation. *Public Health*, 144S, S5–S14.

Haddon, W., Jr. (1970). On the escape of tigers: an ecologic note. *American Journal of Public Health*, **60**, 2229–34.

Haddon, W., Jr. (1972). A logical framework for categorizing highway safety phenomena and activity. *Journal of Trauma*, **12**, 193–207.

Haddon, W., and Baker, S.P. (1981). Injury control. In: Clark, D.W. and McMahaon, B. (eds.) *Preventive and Community Medicine* (2nd ed.). Boston, MA: Little, Brown, and Company, pp. 109–40.

Hatfield, J., Williamson, A., Sherker, S., Brander, R., and Hayen, A. (2011). Development and evaluation of an intervention to reduce rip current related beach drowning. *Accident Analysis & Prevention*, **46**, 45–51.

Henry, J.A. and Reingold, A.L. (2012). Prehospital trauma systems reduce mortality in developing countries: a systematic review and meta-analysis. *Journal of Trauma & Acute Care Surgery*, 73(1), 261–8.

Hyder, A.A., Borse, N.N., Blum, L., Khan, R., El Arifeen, S., and Baqui, A.H. (2008). Childhood drowning in low- and middle-income countries: urgent need for intervention trials. *Journal of Paediatric Child Health*, 44(4), 221–7.

Hyder, A.A., and Morrow, R.H. (2006). Measures of health and disease in populations. In: Black, R.B., Merson, M., and Mills, A. (eds.) *International Public Health: Diseases, Programs, Systems and Policies* (2nd ed.). Boston, MA: Jones & Bartlett Publishers,.

Institute of Medicine (IOM) (1999). *Reducing the Burden of Injury: Advancing Prevention and Treatment*. Bonney, R.J., Fulco, C.E., and Liverman, C.T. (eds.). Washington, DC: National Academy Press.

Jacobs, G., et al. (2000). *Economic Cost of Road Traffic Injuries*. London: Transport Research Laboratories.

Jie, X., and Ren, C.B. (1992). Burn injuries in the Dong Bei area of China: a study of 12,606 cases. *Burns*, 18(3), 228–32.

Kendrick, D., et al. (2012). Home safety education and provision of safety equipment for injury prevention. *Cochrane Database of Systematic Reviews*, 12(9), CD005014.

Kobusingye, O., Guwatudde, D., and Lett, R. (2001). Injury patterns in rural and urban Uganda. *Injury Prevention*, 7(1), 46–50.

Kobusingye, O.C., Hyder, A.A., Bishai, D., Hicks, E.R., Mock, C., and Joshipura, M. (2005). Emergency medical systems in low- and middle-income countries: recommendations for action. *Bulletin of the World Health Organization*, 83s(8), 626–31.

Kopits, E., and Cropper, M. (2003). *Traffic Fatalities and Economic Growth*, Policy Research Working Paper 3035. Washington, DC: World Bank.

Kress, H.C., Noonan, R., Freire, K., Marr, A., and Olson, A. (2012). Top 20 violence and injury practice innovations since 1992. *Journal of Safety Research*, 43(4), 257–63.

Liu, E.H., Khatri, B., Shakya, Y.M., and Richard, B.M. (1998). A 3 year prospective audit of burn patients treated at the Western Regional Hospital of Nepal. *Burns*, 24(2), 129–33.

Liu, G.F., et al. (2003). Driver sleepiness and risk of car crashes in Shenyang, a Chinese northeastern city: population-based case-control study. *Biomedical & Environmental Sciences*, 16(3), 219–26.

MacKenzie, E., Steinwachs, D., and Bone, L. (1992). Inter-rater reliability of preventable death judgments. The Preventable Death Study Group. *Journal of Trauma*, 33(2), 292–302.

Mann, N.C., Mullins, R.J., MacKenzie, E.J, Jurkovich, G.J., and Mock, C.N. (1999). A systematic review of published evidence regarding trauma system effectiveness. *Journal of Trauma*, 47, S25–33.

Mann, R.E., et al. (2003). The effects of drinking-driving laws: a test of the differential deterrence hypothesis. *Addiction*, 98(11), 1531–6.

Marshall, S.W., Runyan, C.W., Bangdiwala, S.I., Linzer, M.A., Sacks, J.J., and Butts, J.D. (1998). Fatal residential fires: who dies and who survives?. *Journal of the American Medical Association*, 279(20), 1633–7.

Mashreky, S.R., Rahman, A., Khan, T.F., Svanström, L., and Rahman, F. (2010). Determinants of childhood burns in rural Bangladesh: a nested case-control study. *Health Policy*, 96(3), 226–30.

Mazerolle, L., Davis, J., Higginson, A., Mengersen, K., Bedford, L., and Somerville, A. (2012). Community-oriented policing's impact on interpersonal violent crime in developing countries. *The Campbell Collaboration*. [Online] Available at: https://core.ac.uk/download/pdf/18442374.pdf

McLoughlin, E. (1995). A simple guide to burn prevention. *Burns*, **21**, 226–9.

Mirman, J.H., and Kay, J. (2012). From passengers to drivers: parent perceptions about how adolescents learn to drive. *Journal of Adolescent Research*, 27(3), 401–24.

Mock, C., Joshipura, M., Arreola-Risa, C., and Quansah, R. (2012). An estimate of the number of lives that could be saved through improvements in trauma care globally. *World Journal of Surgery*, **36**, 959–63.

Mock, C.N., Jurkovich, G.J., nii-Amon-Kotei, D., Arreola-Risa, C., and Maier, R.V. (1998). Trauma mortality patterns in three nations at different economic levels: implications for global trauma system development. *Journal of Trauma*, 44(5), 804–12.

Mock, C., Nguyen, S., Quansah, R., Arreola-Risa, C., Viradia, R., and Joshipura, M. (2006). Evaluation of trauma care capabilities in four countries using the WHO-IATSIC Guidelines for Essential Trauma Care. *World Journal of Surgery*, 30(6), 946–56.

Mock, C., Quansah, R., Krishnan, R., Arreola-Risa C., and Rivara, F. (2004). Strengthening the prevention and care of injuries worldwide. *The Lancet*, 363(9427), 2172–9.

Muckle, W., Muckle, J., Welch, V., and Tugwell, P. (2012). Managed alcohol as a harm reduction intervention for alcohol addiction in populations at high risk for substance abuse. *Cochrane Database of Systematic Reviews*, 12(12), CD006747.

Mukerji, G., Chamania, S., Patidar, G.P., and Gupta, S. (2001). Epidemiology of paediatric burns in Indore, India. *Burns*, 27(1), 33–8.

Mullins, R.J., et al. (1994). Outcome of hospitalized injured patients after institution of a trauma system in an urban area. *Journal of the American Medical Association*, 271(24), 1919–24.

Nathens, A., Jurkovich, G., Cummings, P., Rivara, F., and Maier, R. (2000). The effect of organized systems of trauma care on motor vehicle crash mortality. *Journal of the American Medical Association*, 283(15), 1990–4.

National Highway Traffic Safety Administration (2017). *Fatal Analysis Reporting System Encyclopedia, 1994–2016*. [Online] Available at: https://www-fars.nhtsa.dot.gov/Main/index.aspx

National Highway Traffic Safety Administration (2007). Seat belt use in 2006—use rates in states and territories. Traffic Safety Facts, DOT-HS-810-960. [Online] Available at: http://www-nrd.nhtsa.dot.gov/pdf/nrd-30/NCSA/RNotes/2007/810690.pdf

Nielsen, K., Mock, C., Joshipura, M., Rubiano, A.M., Zakariah, A., and Rivara, F. (2012). Assessment of the status of prehospital care in 13 low- and middle-income countries. *Prehospital Emergency Care*, 16(3), 381–9.

Norton, R., Hyder, A.A., and Gururaj, G. (2006). Unintentional injuries and violence. In: Black, R.B., Merson, M., and Mills, A. (eds.) *International Public Health: Diseases, Programs, Systems and Policies* (2nd ed.). Silver Boston, MA: Jones & Bartlett Publishers.

Nyberg, C., Schyllander, J., Stark Ekman, D., and Janson, S. (2012). Socio-economic risk factors for injuries in Swedish children and adolescents: a national study over 15 years. *Global Public Health*, 7(10), 1170–84.

Pearn, J.H., Nixon, J.W., Franklin, R.C., and Wallis, B. (2008). Safety legislation, public health policy and drowning prevention. *International Journal of Injury Control & Safety Promotion*, 15(2), 122–3.

Peck, M.D., Kruger, G.E., van der Merwe, A.E., Godkaumbura, W., and Ahuja, R.B. (2008). Burns and fires from non-electric domestic appliances in low and middle income countries Part I. The scope of the problem. Burns, 34(3), 303–11.

Peek-Asa, C., Zwerling, C., and Stallones, L. (2004). Acute traumatic injuries in rural populations. *American Journal of Public Health*, 94(10), 1689–93.

Pérez-Núñez, R., Pelcastre-Villafuerte, B., Híjar, M., Avila-Burgos, L., and Celis, A. (2012). A qualitative approach to the intangible cost of road traffic injuries. *International Journal of Injury Control & Safety Promotion*, 19(1), 69–79.

Rahman, A., Miah, A.H., Mashreky, S.R., Shafinaz, S., Linnan, M., and Rahman, F. (2010). Initial community response to a childhood drowning prevention programme in a rural setting in Bangladesh. *Injury Prevention*, 16(1), 21–5.

Roberts, C., et al. (2009). The Health Behaviour in School-aged Children (HBSC) study: methodological developments and current tensions. *International Journal of Public Health*, 54, Suppl 2, 140–50.

Robertson, L.S. (1981). Automobile safety regulations and death reductions in the United States. *American Journal of Public Health*, 71(8), 818–22.

Shackford, S.R., Hollingworth-Fridlund, P., Cooper, G.F., and Eastman, A.B. (1986). The effect of regionalization upon quality of trauma care as assessed by concurrent audit before and after institution of a trauma system: a preliminary report. *Journal of Trauma*, 26(9), 812–20.

Soori, H. (2001). Developmental risk factors for unintentional childhood poisoning. *Saudi Medical Journal*, 22(3), 227–30.

Spinks, A., Turner, C., Nixon, J., and McClure, R.J. (2009). The 'WHO Safe Communities' model for the prevention of injury in whole populations. *Cochrane Database of Systematic Reviews*, 8(3), CD004445.

Thomas, S., Acton, C., Nixon, J., Battistutta, D., Pitt, W.R., and Clark, R. (1994). Effectiveness of bicycle helmets in preventing head injury in children: case-control study. *British Medical Journal*, **308**, 173–7.

Thompson, D.C. and Rivara, F.P. (2000). Pool fencing for preventing drowning in children. *Cochrane Database of Systematic Review*, (2), CD001047.

Transport and Road Research Laboratory (1991). *Towards Safer Roads in Developing Countries: A Guide for Engineers and Planners*. Berkshire: Overseas Development Administration.

U.S. General Accounting Office (1991). *Motorcycle Helmet Laws Save Lives and Reduce Costs to Society*. Washington, DC: U.S. General Accounting Office, GAO/RCED-91-170.

Van Niekerk, A., Menckel, E., and Laflamme, L. (2010). Barriers and enablers to the use of measures to prevent pediatric scalding in Cape Town, South Africa. *Public Health Nursing*, 27(3), 203–20.

Wilson, J. and Rogmans, M. (2006). Purposes and scope of prevention of drowning. In: Bierens, J.J.L.M. (ed.) *Handbook on Drowning: Prevention, Rescue, Treatment*. Berlin: Springer, pp. 87–92.

Woolf, S.H. (2008). The meaning of translational research and why it matters. *Journal of the American Medical Association*, 299(2), 211–13.

World Health Organization (WHO) (2002). *The World Health Report 2002: Reducing Risks, Promoting Health Life*. Geneva: WHO.

World Health Organization (WHO) (2004). *The Economic Dimensions of Interpersonal Violence*. Geneva: WHO.

World Health Organization (WHO) (2008). *World Report on Child Injury Prevention*. Geneva: WHO.

World Health Organization (WHO) (2015). *Global Status Report on Road Safety 2015*. Geneva: WHO. [Online] Available at: http://www.who.int/violence_injury_prevention/road_safety_status/2015/en/

World Health Organization (WHO) (2018). *Projections of Mortality and Causes of Death, 2015 and 2030*. Geneva: WHO. [Online] Available at: http://www.who.int/healthinfo/global_burden_disease/projections/en/

Yang, J., Campo, S., Ramirez, M., Krapfl, J.R., Cheng, G., and Peek-Asa, C. (2013). Family communication patterns and teen drivers' attitudes toward driving safety. *Journal of Pediatric Health Care*, 27(5), 334–41.

Young, T., Torner, J., Sihler, K., Hansen, F., Peek-Asa, C., and Zwerling, C. (2003). Factors associated with mode of transport to acute care hospitals in rural communities. *Journal of Emergency Medicine*, 24(2), 189–98.

Zerhouni, E.A. (2005). Translational and clinical science – time for a new vision. *New England Journal of Medicine*, 353(15), 1621–3.

9.5

Interpersonal violence

Rachel Jewkes

Introduction to interpersonal violence

Violence is a violation of human rights that has a profound impact on the physical and mental health and social functioning of its victims and survivors. It impacts households, families, communities, businesses, and economies. It is recognized globally as an obstacle to development, and the ambition to reduce or eliminate violence is expressed in two of the Sustainable Development Goals (SDGs). SDG 5.2 is to eliminate all forms of violence against all women and girls in the public and private spheres, including trafficking and sexual and other types of exploitation and SDG 16 is to promote peaceful and inclusive societies for sustainable development, provide access to justice for all and build effective, accountable, and inclusive institutions at all levels, with SDG 16.1 seeking to significantly reduce all forms of violence and related death rates everywhere. The use of interpersonal violence is, however, a social construct that is inherently preventable.

The *World Health Report on Violence and Health* described interpersonal violence as encompassing child maltreatment, youth violence, intimate partner violence, sexual violence, and elder abuse (Krug et al. 2002). Other forms of violence are increasingly being recognized including peer violence and the abuse of disabled persons. This is categorization spans the life cycle. For many forms of violence the victims are mostly likely to be girls or women, notable exceptions being peer and youth violence. In settings where there is a pronounced culture of violence, adults of all ages may use violence to resolve conflict or assert dominance in a range of other settings including conflict between neighbours, in contexts of crime, in prisons, in the workplace, and between health professionals and patients.

Interpersonal violence is experienced and perpetrated by millions of people worldwide each year and the health and social impact is huge. In 2013, interpersonal violence was in the top 20 leading causes of death and the 22nd leading causes of years of life lost worldwide, and 404,000 deaths were attributed to it. In the age group of 20–44 years, however, interpersonal violence is the second most important cause of mortality for men and women. In 2013, interpersonal violence was among the top 10 causes of years of life lost in many countries and global regions, and is the second highest cause in the Latin America and Caribbean Region, and the fifth in sub-Saharan Africa. Although not high across Europe, it was in the top 10 in Albania and Russia. Evidence of the preventability of death from interpersonal violence is provided by the fact that the age standardized mortality rate, which was 5.6 per 100,000, had declined by 16% since 1990.

Commensurate with this scale of violence is a massive economic impact. A report published in 2004 estimated the cost of violence in the United States to be 3.3% of the gross domestic product, and in England and Wales, the total costs from violence were estimated to amount to $40.2 billion annually (Waters et al. 2004). A costing study of intimate partner violence estimated that in 2009 it was more than £15 billion in England and Wales alone (Walby 2010). Similarly a study of the economic costs of intimate partner violence in Viet Nam that included both direct costs due to lost earnings and out-of-pocket expenditure of affected women found that nearly 1.4% of the GDP was lost annually and amounted to 2.5 billion VDN in 2010. Women experiencing violence were estimated to earn 35% less than those who did not (Duvvury et al. 2012).

Magnitude of interpersonal violence

One of the most common forms of interpersonal violence, and in all global settings the most common form of violence against women, is intimate partner violence. This encompasses physical, sexual, emotional, and economic abuse, and a range of acts from threats of violence, social isolation, intimidation, insults, through to rape, beatings, and at the most extreme end of the spectrum, forms of torture. A recent systematic review concluded that globally 30% of women aged 15 and over have ever experienced physical or sexual intimate partner violence (Devries et al. 2013). There is considerable variation in prevalence between global regions

and settings. In the Eastern Mediterranean, Africa, and South East Asia the lifetime prevalence was 37%, compared to 23–25% in high-income countries, the Western Pacific, and European regions (World Health Organization (WHO) 2013). Greater diversity, however, is seen in the proportion of women who report having been exposed to intimate partner violence in the past year. In the United States, 2.9% of women experience physical intimate partner violence in the past year (Smith et al. 2018), compared to 6.1% in Cambodia, 29% in Ethiopia and 21.3% Bougainville, Papua New Guinea (Garcia-Moreno et al. 2005; Jewkes et al. 2017). Similarly the proportion who have experienced intimate partner sexual violence in the past year has been shown to range from, for example, 1.3% in Japan, to 22.2% in Bangladesh and 44.4% in Ethiopia (Garcia-Moreno et al. 2005).

Global estimates of intimate partner violence often do not include economic and emotional violence. The types of acts commonly included as emotional abuse include, but are not limited to, personal insults, being belittled, socially isolated, acts of coercive control, yelling, intimidation, and threats of physical violence. The economic and structural violence acts include having earnings taken and being prevented from earning. In higher income countries, as the prevalence of physical and sexual violence has declined over time acts of emotional and economic abuse has remained more common. There is also greater gender parity in experience and perpetration of these acts in these settings. Thus in the United States 36.4% of women and 34.2% of men had ever experienced psychological aggression, compared to 18.3% of women and 8.2% of men who had experienced contact sexual violence from an intimate partner (Smith et al. 2018).

In some countries, like South Africa, intimate partner violence is the most common cause of female homicide, accounting for 52% of cases (Abrahams et al. 2013), but globally intimate partner violence is estimated to account for 36% of female homicides (WHO 2013). In some global settings, intimate partner violence occurs as part of a broader experience of abuse of married women within the home by in laws, wider domestic violence. This is a particularly important problem for women in parts of central and south Asia, in highly patriarchal settings where women's movement is controlled. A key perpetrator is the mother-in-law. Exposure to such broader domestic violence has been shown to be associated with even more severe mental ill-health than exposure to partner violence alone (Jewkes et al. 2019).

The global estimate for the proportion of women who have experienced non-partner sexual violence is 7.2% (95% confidence interval = 5.3% to 9.1%), with the prevalence ranging from 5% in South East Asia to 12% in Africa (Devries et al. 2013). Research with men on perpetration has shown that in some settings, including South Africa and Papua New Guinea, 20–40% of men have ever raped a female non-partner (Abrahams et al. 2011; Jewkes et al. 2013). In some global settings the prevalence of rape is particularly high, including some conflict-affected regions and areas of high population crowding, mobility, and poverty, notably informal settlements, or slums. For example among displaced women in the United Nations' Protection of Civilians (POC) sites of Juba, South Sudan, 25% has ever been raped by a non-partner. Some of the men in Juba had also experienced rape, but it was much less prevalent than among women (1%) (Global Women's Institute and International Rescue Committee. 2017). Sexual workers across the global are at particularly high risk of rape from clients, intimate partners, acquaintances, and men in authority, notably the police. In a RDS sample of sex workers from Soweto, South Africa, 53.8% had experienced physical and/or sexual intimate partner violence (IPV) in the past year, 46.8% violence from clients, and 18.5% by police. In all 70.8% had been exposed to sexual and/or physical violence by any perpetrator in the last year and 62.4% had ever been raped (Coetzee et al. 2017)

Child abuse can be physical, sexual, or emotional, and may also encompass physical and emotional neglect. Research has often used very different definitions and these often reflect cultural differences in normative child-rearing practices. Comparability across settings can thus be hard, but the recent global initiative of Violence Against Children Surveys (VACS) has enabled the generation of comparable data from many low- and middle-income countries (LMICs). An one estimate from high-income countries is that each year 4–16% of children are physically abused and one in ten is neglected or psychologically abused (Gilbert et al. 2009). This contrasts with the VACS, where their initial surveys in several African countries showed between 53%–75% of children had experience physical violence, with this being more common for boys, and between 1 in 4 and 1 in 3 emotionally abused (National Population Commission of Nigeria et al. 2016). Further during the whole of childhood, between 5% and 10% of girls and up to 5% of boys are exposed to penetrative sexual abuse, and 15–40% are exposed to any type of sexual abuse (Reza et al. 2007; Gilbert et al. 2009).

An estimated 200,000 deaths occur each year among young people aged 10–29 and that for each death 20–40 require hospitalization (Krug et al. 2002; WHO 2015). Youth violence mortality predominantly occurs in LMICs, notwithstanding the recent rise in death from stabbing in the United Kingdom, and is highly gendered. The overwhelming majority of perpetrators and 83% of victims are male. Recent research for South Africa, for example, has shown that the rate of homicide (21.7/100,000) for teenage boys (15–17 years) was nearly five times the teenage girls' rate (4.6/100,000) (Mathews et al. 2013).

Peer violence is a part of the problem of youth violence, and commonly occurs at school, in communities and in families. Both girls and boys con be victims and perpetrators, but the prevalence tends to be higher among boys. Peer violence overlaps with the phenomenon of bullying, and generally is considered to include physical violence and threats emotional abuse and name calling, damage to property, spreading rumour and social exclusion. Its definition highlights that it is a repeated act and intended to cause harm. It has been said to affect 10–65% of children in schools globally, but prevalences as high as 90% have been reported in parts of the United States and in Pakistan (Karmaliani et al. 2017).

Elder abuse is a much less researched form of interpersonal violence. Its subcategories generally include physical, psychological, sexual, financial, and discriminatory abuse and neglect. It is generally measured as exposure to people aged over 65 or over 60 (Cooper et al. 2008). The category is often measured in a way that fails to distinguish between other forms of violence experienced by older people (e.g. IPV) from violence that is a result of the constrained power of older people (i.e. because a person is elderly). This is an area that is growing in recognition, especially for elderly people with cognitive impairment. The main perpetrators are family members and caregivers. In some parts of sub-Saharan Africa a particularly horrific dimension of elderly abuse is seen in witchcraft-related violence and murders (Meel 2009). Understanding and preventing abuse of the elderly is an area that requires considerable further research.

Violence against adults and children with disabilities has fairly recently been given much greater recognition as part of the overall global problem of interpersonal violence. A review on the prevalence of violence experienced by children with disabilities found that they were four times more likely than children without disabilities to experience violence. This also included

sexual violence, and overall they were 2.9 times more likely to experience sexual violence, but among those with mental or intellectual impairments the increased risk was 4.6 fold (Jones et al. 2012). Adults with disabilities are also at increased risk of interpersonal violence, including women with disabilities at elevated risk of IPV. A recent meta-analysis has found the risk of physical or sexual partner violence to be 2 times higher among women with disabilities (Dunkle et al. 2018). Disability also increases women's risk of non-partner sexual violence and the risk of this and IPV rise with increasing severity of the impairment.

Health consequences

The health impact of interpersonal violence follows two pathways: immediate and longer term. Immediate health impact is normally injury, which may be fatal, and may lead to complications in a pregnant victim. Medium and longer term impacts can be particularly profound. These may be physical, in the case of disabling injuries, but a very important area of impact is on mental health, resulting in depression and anxiety, post-traumatic stress disorder (PTSD), self-harm, and substance abuse. The mental health impact is particularly severe with emotional and economic abuse, sexual abuse, or rape and with exposure to life threatening violence. For example, South African research shows that adolescent women who have experienced emotional neglect in their lives are at greater risk of depression, suicidality and alcohol abuse, and adolescent men were at greater risk of depression and drug abuse (Jewkes et al. 2010a). The same study found sexual abuse associated with alcohol abuse and depression in the men and alcohol abuse in the women. Peer violence, including bullying, has been likewise shown to have an important negative impact on children's mental health elevating the risk of depression (Asad et al. 2017).

Mental health is also impacted by experience of IPV and non-partner rape. Women so exposed are at more than twice the risk of depression, and the latter group also have more than twice the risk of alcohol abuse (WHO 2013). Rape is also particularly likely to result in PTSD, and in fact is more likely to cause PTSD than any other violence exposure (Tolin and Foa 2006). Emerging research has highlighted the importance of emotional and economic partner violence in mental health impact, showing that greatest levels of depression and suicidal ideation are found in women experiencing

both emotional or economic violence combined with physical and/or sexual partner violence (Gibbs et al. 2018).

Exposure to interpersonal violence can also have complex impacts on health. One of the more complex mechanisms of impact is illustrated by the pathways through which IPV impacts HIV acquisition in women, and undermines the use of and adherence to antiretroviral therapy (Jewkes 2010; Jewkes et al. 2010b, Hatcher et al. 2015). Research from South Africa and Uganda shows that nearly one in four new HIV infections could be prevented if women were not subjected to IPV and relationship power inequalities and exposure to sexual abuse in childhood in one South African study resulted in a two-thirds greater risk of incident HIV infections (Jewkes et al. 2010b; Kouyoumdjian et al. 2013). The complex web of impact is illustrated by Figure 9.5.1. This shows that this is a very direct pathway, in that women may become directly infected with HIV through rape, but a much more important mechanism of risk is indirect, through partner violence and child sexual abuse exposed women having less control of their sexual encounters, as a result of violence and controlling behaviours of their partner, with the result that they have sex more frequently, use condoms less, more sexual transmitted infections and engaging more often in a range of risky sexual practices, including having more, and often concurrent partners, transactional sex, substance abuse-related sex, and even sex work. These increase the risk of HIV acquisition as well as rendering women vulnerable to further male violence. The mechanisms of increased risk are partly though the mental health consequences of violence, including insecure adult attachment which is associated with difficulties in emotional relationships and having more partners. It is also the case that men who are violent towards women often have more risky sexual practices and sexual transmitted infections and HIV (Decker et al. 2009; Jewkes et al. 2011b). Women who have more complex and chaotic as a result of exposure to men's violence are less able to adhere to antiretroviral therapy. Research has shown that in the absence of treatment, women with HIV who have been exposed to emotional abuse experience a more rapid decline in cellular immunity (Jewkes et al. 2015).

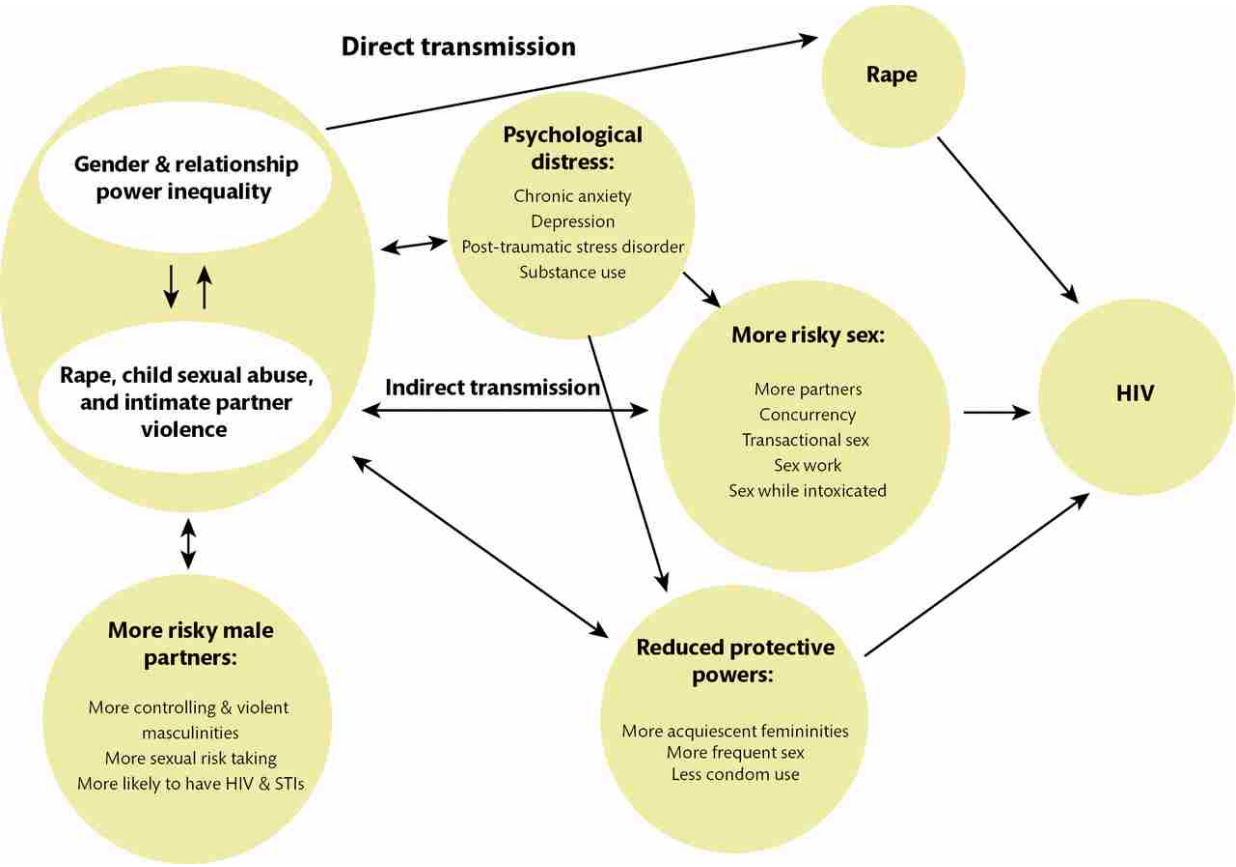


Fig. 9.5.1 Transmission.

Interpersonal violence has a major impact on women's reproductive health. IPV exposure is associated with a 30% increased risk of unintended pregnancy, also of infertility and subfertility, induced abortion, young teenage pregnancy as well as low birth weight and preterm birth (Hill et al. 2016; Stellar et al. 2016; Maxwell et al. 2018). Childhood sexual abuse and child marriage increase the risk of early pregnancy and childbearing with higher maternal and infant mortality.

Interpersonal violence also has a range of social and economic consequences apart from out-of-pocket costs for healthcare. These include, for women, a greater risk of school dropout due to pregnancy, which impacts on future employment and lifetime earnings. Violence results in more time off due to injury and its mental and other health effects, and even women not working experience social incapacity, impacting on childcare and homemaking. Women who experience violence who attend work have periods of lower productivity due to violence, presenteeism, and reduced productive engagement. A consequence of all of this is that it enhances

women’s dependency and lowers their power within the household, which in turn enhances their risk of experiencing IPV.

Understanding its roots

Interpersonal violence has different forms, is used in different settings and contexts, with diverse perpetrators and victims, that may be seen as only united by the common use of physical force to cause hurt or injury. But there is much more overlap in the roots and drivers of violence than first meets the eye. On closer examination its apparent that in many cases the same people perpetrate different forms of interpersonal violence at different times. For example, the aggressive young school bully may engage in violent fights with other youths, be sexually and physically violent to women, beat his children, and engage in be very violent in his job as a policeman. While people may be violent in one context and not in others, violence is strongly patterned and a propensity to the use of violence in one situation increases the likelihood of its use in others. This applies to both men and women.

Many perpetrators of violence have themselves been victims, particularly of emotional or physical maltreatment in childhood, of sexual abuse or bullying, and have witnessed their mother exposed to IPV. The complexity interrelationship was illustrated by a study of adolescents in the United States in which it was found that the connection (pathway) between experience of physical abuse in childhood and perpetration of sexual IPV was completely mediated by involvement in delinquent behaviour, which commonly includes use of other violence (Casey et al. 2009).

Victims of interpersonal violence do not bring the violence on themselves, except in gang fights, yet victimization is by no means random. Very commonly one exposure to violence, especially abuse in childhood, increases the risk of re-victimization. For example, women who have experienced sexual abuse as children are more likely to be raped as adults, and to experience IPV.

To explain the interconnectedness, it is important to understand that different forms of violence have overlapping root causes and risk factors. The use of violence to communicate dominance, punishment, or defend honour, is a learned, and it may be deployed when this is intended in multiple contexts. Social learning spans how violence is used, in what forms, who are appropriate victims, acceptable or required triggers for its use,

meanings that will be attached to its use, and patterns of reward or punishment that can be expected to follow. For example, Wood et al. (2007) researching in South Africa's Eastern Cape Province found use of violence by a man against how female partner to be widely regarded as a normal part of gender relations and considered unexceptional by the community, family, and providing the victim was not injured by the violence. Moreover it often viewed as deserved, good for teaching right from wrong, and when submission as the result of the violence it was socially rewarded. A further cultural dimension of interpersonal violence relates to the thresholds that trigger its use. Some societies, such as parts of Papua New Guinea, having very low thresholds for the use of many forms of interpersonal violence. In other cultural settings, thresholds are very much higher, and generally the use of violence in interpersonal relations is viewed as inappropriate. Social rewards for the use of interpersonal violence vary greatly, but in some settings 'men of violence' are revered, their physical power viewed with awe, and political power and economic rewards may flow from this. The use of violence is often coded into ideas of appropriate male behaviour, especially where demonstrating toughness, competition and honour are prominent features of masculinity and this further explains the cultural licence for violence use.

An obvious explanation for re-victimization is that victims of violence are often selected because they are vulnerable. This explains part of the higher risk of people with disability, poorer women, and children, and those working illegally as sex workers. Victimization increases such vulnerability, through its mental health impact and perceived disempowerment, furthermore negative coping strategies after some types of victimization may including problem drinking, create further vulnerability. In this way interpersonal violence creates cycles of vulnerability across different types of violence.

Risk factors for interpersonal violence

Risk factors for violence perpetration have been described as operating at a range of different ecological levels. This has been presented in an ecological model that highlights how some of these pertain and exert effects at an individual level, others at a family, peer, or relationship level, others at a community level and still others at a societal level.

Individual level

Genotype

There is evidence that genes are important in the development of aggression. Research points towards a general heritable tendency to antisocial behaviour, of which sexual coercion is one form. There may be important gene × environment interactions, particularly with the expression of aggression in genetically susceptible individuals after exposure to a harsh environment, or else a person with antisocial tendencies might be more likely to be drawn to situations and settings where antisocial behaviour is more common or more possible (Johansson et al. 2008). The evidence is predominantly focused on non-sexual aggression. However both sexual aggression and general aggression are heritable in mice (Canastar and Maxson 2003) and one study of twins concluded that over a quarter of sexual coercion was explained by genetic factors, but the genetic element in psychopathy was higher (Johansson et al. 2008).

The monoamine oxidase A gene (*MAOA*) has been described as impacting on a heritable tendency for antisocial behaviour. This encodes the MAOA enzyme, which metabolizes neurotransmitters such as norepinephrine (NE), serotonin (5-HT) and dopamine (DA) and thus inactivates them (Shih et al. 1999). It is very rare for the gene encoding MAOA to have been deleted, although this is found associated with heritable aggression in humans and mice (Rowe 2001). More common are functional polymorphisms that result in lower activity of the MAOA enzyme (Caspi et al. 2002). A prospective study conducted in New Zealand has shown that child maltreatment has a much stronger association with subsequent antisocial behaviour in males with low MAOA enzyme activity, than those with higher activity (Caspi et al. 2002). In this study 11% of violent convictions were attributable to the gene. Findings of research on exposure to physical punishment in early childhood, having the *MAOA* gene and likelihood of subsequent delinquency are inconsistent (Reif et al. 2007; Prichard et al. 2008; Edwards et al. 2010). The role of genetics in interpersonal violence, including sexual aggression, is clearly an important area for further work.

Other biological influences: androgens and neurological deficits and sexual offending

The role of androgens in interpersonal violence is questionable. A positive relationship between testosterone and self-reported aggressiveness has been reported for men and women (Christiansen and Knussmann 1987), but this has not been confirmed by all authors (Campbell et al. 1997). Several researchers have investigated links between blood testosterone levels and sexual or other aggression in prison populations, but found no evidence that testosterone levels influence the likelihood of sex offending, any more than general violent offending (Giotakis et al. 2003).

The role of brain damage in sexual offending has been the subject of research but the overall conclusion of a recent review of the literature is that the evidence is extremely weak (Fabian 2012). One of the few controlled studies with paedophiles found no differences on neuropsychological measures or positron emission tomography scans (Cohen et al. 2002). While brain trauma may be important in some individuals, there is currently no evidence that brain damage makes an important contribution at a population level.

Early years' exposures

Irrespective of genotype, there is evidence that an individual's risk of perpetrating interpersonal violence can be shaped in the first days, months, and years of life. Perpetrators of interpersonal violence, including sexual violence, are more likely to have failed to form secure attachment bonds in childhood. This is linked to exposure to emotional, physical, and sexual abuse as well as to domestic violence, removal from the family, family disruption, and parental loss due to death or divorce may be risk factors for perpetration (Maniglio 2010).

Attachment

One of the critical pathways through which the childhood home environment, particularly in early childhood, impacts upon psychological development is through attachment to primary caregivers, usually mothers (Bowlby 1969, 1973, 1980). If attachment bonds between the child and their primary caregiver are secure, children acquire the necessary skills to establish close relationships and grow to desire the intimacy of others. If these bonds are insecure, children do not acquire the necessary skills to establish close relationships, and may grow to fear intimacy with other

individuals, exhibit hostility or aggression within their relationships, or to seek intimacy in maladaptive ways (Bowlby 1969, 1973). The failure to form secure attachment bonds is due to poor socialization experiences, such as violent parenting, and disordered attachment has in particular been described as consequent of parenting that combines harsh punishment and expressions of affection (Marshall and Barbaree 1990; Fonagy and Target 2003). A failure to form secure attachment bonds in childhood results in a failure to learn interpersonal and other skills necessary to achieve intimacy as adults. Insecure attachments are marked by an expectation that what is needed will not be given, and as a result coercion, whether through physical violence, emotional manipulation, or sexual violence, is often used to get it (Bowlby 1969, 1973, 1980). In the arena of relationships, insecurity may result in more dismissive and hostile attitudes towards others and a tendency to perceive others negatively and distrustfully. This has been shown in sex offenders, but is also likely to increase propensity to IPV, youth violence and interpersonal violence in the community (Marshall 1989; Stirpe et al. 2006; Whitaker et al. 2008). Individuals with disorganized attachment are much more aggressive, distrustful, and less empathetic when older and have a propensity to general interpersonal violence (Fonagy and Target 2003).

Dysfunctional parenting and disturbed family histories, with unstable caregiving often due to maternal post-natal depression, or parental IPV, have been shown to be associated with interpersonal violence and various categories of offending among children, including sexual offending (Lightfoot and Evans 2000; Seto and Lalumiere 2010).

Child abuse victimization

Physical and emotional abuse and neglect in childhood are exposures that result in a failure to form secure attachments. Emotional abuse and neglect has been given less emphasis in research than physical abuse, but a six-country study on men's violence in Asia and the Pacific has shown that men who rape non-partners and those who have been violent to intimate partners are more likely to have been exposed to all of these in childhood, and these findings support the research from South Africa (Jewkes et al. 2011a; Fulu et al. 2013).

The evidence of childhood sexual abuse as a risk factor for subsequent perpetration is very extensive and probably has been the most comprehensively investigated of all risk factors for perpetration. In a meta-

analysis, a history of child sexual abuse was five times more common among adolescent sexual offenders than among adolescent non-sexual offenders (Seto and Lalumiere 2010). Further, sexual offenders against children are much more likely to have sexual abuse victimization histories than non-offenders or non-sexual offenders (Whitaker et al. 2008). Several large national studies of US adolescents and men in college have found that those who had sexually assaulted a partner were significantly more likely to have experienced sexual abuse themselves (Borowsky et al. 1997; Casey et al. 2009; Malamuth et al. 1991, 1995).

Witnessing IPV in the home is both extremely traumatic for children and may be part of social learning about intimate relations. Both of these may explain the consistent finding that men who have been so exposed in childhood are more likely to be physically or sexually violent towards their intimate partners (and girls so exposed to experience violence themselves) (Malamuth et al. 1991, 1995; Borowsky et al. 1997; Fulu et al. 2013). South African research has also shown that men who have been exposed in childhood are more likely to get involved in fights at work and in the community, and have been arrested for having an illegal firearm (Abrahams and Jewkes 2005). In Pakistan girls and boys who have witnessed violence against their mothers at home were more likely to perpetrate peer violence (Karmaliani et al. 2017).

Personality

These early years of exposure have a profound impact on personality when older. Psychopathy has been the most extensively discussed in relation to perpetration of interpersonal violence. Psychopathy is a constellation of personality traits and socially deviant behaviours, spanning affective, interpersonal, and behavioural traits. Hare has proposed that the characteristics of psychopathy include egocentricity, lack of remorse and empathy, pathological lying, manipulativeness, and the persistent violation of social norms (Hare 1996). These traits enable psychopaths to treat others as objects and feel no guilt or remorse. They are thus much more likely to perpetrate interpersonal violence, particularly severe physical violence, emotional violence, and rape.

In the US general population the prevalence of psychopathy is usually estimated at 1%, but the prevalence of psychopathic traits without a full

clinical diagnosis is higher (Lilienfield 2005). Individuals with heightened levels of some of the characteristics of psychopathy have been described as subclinical and non-clinical psychopaths (LeBreton et al. 2006). The prevalence of psychopathy as a diagnosis is dependent on assessment measure scores. Research on Canadian incarcerated sex offenders shows that a relatively small difference in cut point (25 versus 30) on the PCL-R (Hare Psychopathy Checklist—Revised) (Hare 2003) resulted in twice as many offenders being diagnosed as psychopathic (29% vs. 13%) (Olver and Wong 2006). This suggests that the presence of psychopathic traits may be important in understanding interpersonal violence at subclinical levels and indeed this has been suggested by research in South Africa where two traits were measured, Machiavellian egocentricity and blame externalization, and were significantly higher among men who had raped in both the past year and their lifetime (Jewkes et al. 2011a). There are similar findings from US college students, among whom subclinical scores on psychopathic traits have been associated with rape perpetration (Ouimette 1997; DeGue and DiLillo 2004; Abbey et al. 2011). Psychopaths have been described as having an inability to connect sexual behaviour with emotional feelings, such as warmth and tenderness, and a hedonistically orientated lack of impulse control and are thus more likely to be sexually as well as physically violent (Cleckley 1988).

Personality disorders of offenders

Research on personality disorders in populations of offenders shows that sex offenders against children and adults show significantly more Cluster A, Cluster B, and antisocial personality disorders than non-offenders (Whitaker et al. 2008). Psychopathy is common among more violent child sex offenders (Rosenberg et al. 2005). Schizoid personality features include social detachment and restricted emotionality, and avoidant features include social inhibition and inadequacy. Both of these interfere with the development of normal peer relationships, including intimate relationships with women, and are more prevalent among sex offenders, particularly against children (Fazel et al. 2002; Whitaker et al. 2008). Men who sexually molest children are socially and emotionally immature, have poor social skills, low levels of self-esteem, poor impulse control, and more signs of psychotic thinking (Finkelhor and Araji 1986). Relating to children provides

a sense of power and control and opportunity for dominance not available to them in adult relationships (Finkelhor and Araji 1986).

Substance abuse

Alcohol abuse is strongly associated with use of interpersonal violence (Machisa et al. 2019). There are physiological effects, but also complex cultural beliefs about and meanings given to the relationship between alcohol and violent, including sexually violent, behaviour. Alcohol has a physiological effect of reducing inhibitions, reasoning, and, in men who were more irritable, prone to anger, and have low levels of anger control, can result in ready use of violence (Finkelhor and Araji 1986). Research on IPV and rape in particular commonly finds that men who are violent are more likely to have heavy alcohol use (Koss and Dinero 1988; Jewkes et al. 2006, 2011a; Tsai et al. 2011; Fulu et al. 2013). Many cultures (or subcultures) link alcohol consumption to disinhibition, sexual desire, and performance, and perceive it to give courage to make advances that men may be reticent to make when sober, and as mitigating antisocial or aggressive behaviour (Lightfoot and Evans 2000; Seto and Lalumiere 2010). Alcohol abuse also has an economic impact on families, leading to greater financial strain, and may be a trigger for conflict over finances.

A further question is whether violence associated with alcohol more often results in serious injury or mortality. There is some evidence that shows that when a perpetrator is drunk and the victim sober there was a greater likelihood of rape completion and victim injury than when they had both been drinking (Brecklin and Ullman 2002). In the case of youth homicide, alcohol is estimated to be responsible for 26% and 16% of years of life lost by men and women, respectively, with the proportion lower in high-income countries and as high as 40% in low-income countries (WHO 2004).

Drug use

Drug use is also intimately connected with youth violence and homicide, most directly because the trade in drugs in many settings is associated with violence and turf wars between dealing gangs (Blumstein et al. 2000). Research on rape has shown that men who have raped are more likely to use drugs, indeed a prospective study from South Africa concluded that 24% of all rapes would have been prevented if drug use had not occurred (Jewkes et

al. 2012). The great majority of the drug use reported was marijuana use, but the authors argued that drug use was most likely an indicator of engagement in a subcultural peer context where both drug use and rape was seen as 'normal' (Jewkes et al. 2012). This conclusion is supported by findings of a multicountry study that drug use was associated with having ever engaged in multiple perpetrator rape (sometimes referred to as gang rape), but not single perpetrator rape (Jewkes et al. 2013).

Post-traumatic stress disorder (PTSD) and depression

PTSD symptoms have been associated with perpetration of several forms of interpersonal violence. PTSD symptoms of dysphoric arousal in particular have been were positively associated with violent behaviours. A recent systematic review of mental ill-health and recent IPV perpetration found that there was a positive association between this and having PTSD symptoms, but not with current depression (Machisa et al. 2019). This is not withstanding the observations that men who have used violence against intimate partners are often found to be more likely to be depressed (Nduna et al. 2010; Fulu et al. 2013). It is not clear whether depression causes men to become violent, or is an emotional response to having used violence and the negative impact that this has on a relationship.

Propensity to use violence

The use of interpersonal violence in one context is closely related to its use in others. This has particularly been highlighted with violence against women and children, for example rape of non-partners is more commonly perpetrated by men who are physically violent towards partners and men who beat women are more likely to beat their children (Malamuth et al. 1991, 1995; Jewkes et al. 2006, 2011a, 2013). It seems likely that a system of patriarchy and gender hierarchy that legitimates the use of physical violence by men to assert dominance in a relationship and 'correct' behaviour of wives and girlfriends would also find expression in sexual violence, as well as other forms of violence, such as verbal violence. There is also some evidence that verbal violence against women is more common among sexually coercive men in US colleges and among South African working men (Malamuth et al. 1991, 1995; Abrahams et al. 2004).

Relationship, peer, or microsocial level

Conflict

The dynamics of particular relationships are important influences of the propensity for use of violence. IPV is more common in relationships that are more conflictual, and experimental evidence suggests that interventions that build communications skills and reduce relationship conflict are impactful in reducing IPV (Jewkes et al. 2008). Poor communication skills of either partner may result in conflicts escalating readily to violence. Conflict escalating to violence is also more common in relationships where one partner is more irritable and prone to outbursts of anger. Perceived challenges to men's 'right' to what they perceive as traditional gender privileges and their non-fulfilment of expected provider roles are often causes of conflict.

Social learning around violence acceptability

In teenage peer relationships, it has been observed that friends are similar in respect to propensity for aggression (McPherson et al. 2001). There has been considerable debate about whether aggressive young people choose each other as friends or whether as friends they become more alike (Snyder et al. 1997; Steglich et al. 2010). Longitudinal research from Chile suggests that children do not select friends based on their levels of aggression, rather that similarity among friends is a product of the friendship, although it is also possible that aggressive male youth may be forced together if non-aggressive peers reject them (Steglich et al. 2010). This partly explains the strong connection between gang membership and violence perpetration that has been described in many settings (Bourgois 1996; Jewkes et al. 2006, 2011a). The importance of perceived social acceptability of violence was also highlighted by a study that found the parents of violence perpetrators, when compared with parents of non-perpetrators, were much more likely to have encouraged gang membership (Morris et al. 2002).

Social learning about sexual violence as normative, is demonstrated in research with incarcerated American adolescents that has shown that sexual assault perpetrators were nearly four times as likely as non-perpetrators to know a perpetrator of sexual assault (Morris et al. 2002). Among a community-based sample of men, perceived peer approval of coercing sex was much higher among men who rape (Abbey et al. 2007). A further aspect

of social learning is reflected in the perpetration of sexual abuse by those who have themselves been sexually abused in childhood. Research from New Zealand provides an additional perspective when sexually abusive adolescents and children were found to have been more likely to have had contact with adults who were known paedophiles, or suspected to have engaged in sexual abuse of other children, even when they did not have confirmed sexual abuse victimization themselves (Lightfoot and Evans 2000). These children may have experienced abnormal sexual socialization through exposure to inappropriate sexual experiences (e.g. pornography exposure or verbal experiences) (Lightfoot and Evans 2000).

There is also a social learning element in the use of violence in child-rearing practices, particularly physical punishment. Physical violence is generally used to punish or gain control over a child and is used more readily by parents raising children in contexts where beating is more common and where there is less role modelling of non-violent parenting and discipline practices.

Delinquent behaviour

The use of interpersonal violence, including rape, among adolescents is strongly associated with other antisocial behaviour (Ageton 1983; Malamuth et al. 1991; Calhoun et al. 1997; Knight and Sims-Knight 2003). Delinquency is an area where multiple risk factors and violent behaviour intersect. There may be genetic roots in some young men, as research has shown that men with genetically determined low MAOA enzyme activity and exposure to harsh punishment in early childhood are more likely to become delinquent (Edwards et al. 2010). Further it is strongly related to personality traits of psychopathy, which themselves may be genetically determined (Abbey et al. 2011). Delinquency is also socially modelled and much more common among youth who have themselves experienced physical aggression in childhood, and are more susceptible to peer pressure (Malamuth et al. 1995; Casey et al. 2009).

Multicountry research on both IPV perpetration and rape perpetration has shown that these are associated with having currently or previously been in a gang (Borowsky et al. 1997; Jewkes et al. 2011a; Fulu et al. 2013). They are also associated with having been involved in other interpersonal crime or theft, having a weapon, and having had an illegal gun (Jewkes et al. 2011a; Fulu et al. 2013). School bullying is one manifestation of delinquent

behaviour, and often a precursor of antisocial behaviour in adulthood (Bender and Lösel 2011). South African research has shown perpetration of school sexual bullying to be strongly associated with subsequent rape perpetration (Jewkes et al. 2011a).

One of the important avenues in the prevention of delinquency is health and education sector interventions for children with conduct disorders. Children with conduct disorders may go on to develop personality disorders as adults, particularly antisocial personality disorders. Conduct disorders are treatable in children, but they require parents to be trained in managing problem behaviour, or out-of-home placements with staff or foster parents appropriately trained. Treatment is most effective when provided to younger children (pre-adolescent); this has been part of the theoretical basis for the Incredible Years programme of Webster-Stratton and colleagues.

Masculinities and delinquent peers

Gang, or otherwise delinquent, peer contexts are also ones in which the masculinities of involved boys are shaped. Negative early childhood influences that generate feelings of insecurity greatly increase the likelihood that as boys grow into men they will adopt masculinities that emphasize strength, toughness, and demonstrate control over others. Such masculinities are familiar within a gang, or otherwise pugilistic peer context, where there is a readiness to show toughness among gang members, between them and potential rivals, dominance over women and to display power over others (e.g. in the context of crime).

Within any social setting there are multiple ways of, and aspirations for, being a man or boy. Some subcultural masculinities are particularly violent, and emphasize dominance over women, with a heightened sense of male sexual entitlement, which readily translates into rape (Bourgois 1996). This is particularly commonly found associated with gangs and among men positioned on the antisocial fringes.

Community level

Many of the structural drivers of interpersonal violence operate at a community level. These include poverty, patriarchy, and structural influences on these such as race or ethnicity. In a profound way these impact on the culture of the local context.

Poverty

Poverty, low educational attainment, and social disadvantage are structural drivers of interpersonal violence. Gangs, which are archetypically associated with many forms of violence, are very largely confined to life in low-income areas. In conditions of poverty, gangs provide young men access to experiences of power when other routes of attainment of power and status are unattainable, such as through working or fulfilling provider roles (Bourgois 1996; Wood and Jewkes 2001; Wood et al. 2007). Poverty increases the risk of IPV and child maltreatment both because it increases economic strains on families and relationships, and also because it generates feelings of inadequacy and powerlessness that may be expressed through the use of violence. Poverty further traps women within violent relationships and can reduce their options for taking action to protect themselves, or to get the violence to stop. Economic empowerment interventions for women have been shown to be a very fruitful avenue for enabling women to protect themselves from IPV (Pronyk et al. 2006).

The relationship between rape and poverty is complex. Multiple perpetrator (gang) rape is most clearly linked to poverty because this is important in the context of gangs (Jewkes et al. 2013). The role of social disadvantage in single perpetrator rape is less clear because perceptions of sexual entitlement are a very powerful part of the context of rape and these may stem from social advantage rather than disadvantage. South African research has shown that among the less well off, it is the slightly less poor or less disadvantaged who are more likely to have raped (Jewkes et al. 2006, 2011a). Authors have argued that a sense of entitlement stemming from relative advantage, within a context of general disadvantage, is played out in a perverse way in a context in which few or none would be able to attain objectively high levels of material ‘success’ (Jewkes et al. 2011a).

Social norms: the cultural scaffolding for interpersonal violence

For all forms of interpersonal violence, social learning around the acceptability of the use of violence, and for gender-based violence, the acceptability of directing it to women, within subcultural contexts is important in perpetration. The importance of social norms on the use of violence in partner violence perpetration was first made visible by anthropological work looking at factors associated with violence across a

wide range of cultural settings (Levinson 1989). More recently research on IPV has increasingly explored the importance of community level norms on the use of violence, for example indicated by community murder rates, as a risk factor for IPV and found it to be highly associated (Koenig et al. 2006).

Social norms on gender relations and violence are also very important. The phrase 'cultural scaffolding' was coined by Gavey to describe the protective social context that enables the reproduction of sexual violence through allowing men to reconfigure sexually coercive behaviour, especially with dates, girlfriends, or wives as 'not rape' (Gavey 2005). The scaffolding includes an array of rape myths that place the blame for rape at the feet of women and enable a cognitive justification for coercive behaviour, ideas of sexual entitlement within marriage (conjugal rights), and the positioning of women as worthy objects of male sexual conquest. This scaffolding is argued to stem from a framework of patriarchy whereby men are positioned as hierarchically superior to women and accorded greater social value. This ideological scaffolding provides the cultural space for men to perpetrate rape without embracing an identity as 'a rapist' or 'a perpetrator'. Similarly youth violence and bullying are often seen as being normal for the age group and is seen as 'not violence' but more the rough and tumble of growing up (Wood and Jewkes 2001).

In some settings the use of interpersonal violence is integrally related to culturally constructed notions of legitimate entertainment. This is very clearly seen in research on rape where men have been asked about their motivations. For example, in a six-country study on men's use of violence in Asia and the Pacific 59% of men who had raped a non-partner indicated that one of the reasons they did it was for entertainment (Jewkes et al. 2013). In South Africa, seeking 'fun' motivated nearly half of South African men who had raped (Jewkes et al. 2011a). Quite similar findings pertain in youth violence, where research in the United Kingdom has shown that fights in bars are very often started for fun (Graham and Wells 2003).

The cultural scaffolding around child abuse is such that physical violence used by caregivers against children is often described as being 'necessary' in order to teach children how to behave and instil discipline, and is thus seen as 'not violence'. In many cultures the use of violence in child-rearing is so ubiquitous that many parents and teachers cannot imagine that there is any other way to properly raise children. Further settings such as in sub-Saharan Africa, societies are characterized by a very pronounced age hierarchy (as

well as gender hierarchy) and children are raised to be obedient to adults in all ways, including to relative strangers. This can lead to children being placed in positions of vulnerability, especially to sexual abuse (Jewkes et al. 2005).

Masculinities

The most pervasive ways in which culture influences interpersonal violence is in the arena of gender. Gender attitudes held at an individual level are usually rooted in cultural models of gender relations, but within these there is considerable individual variation. Sexually aggressive men from diverse settings have greater hostility towards women, stronger sexual dominance desires, and more traditional attitudes towards gender roles and sexual relationships (Koss and Dinero 1988; Malamuth et al. 1991, 1995; Abbey et al. 2011; Jewkes et al. 2011a). IPV is much more common among men who hold more gender inequitable views as well as those who engage in more controlling behaviours (Fulu et al. 2013). However, individually held attitudes and social norms theory do not on their own explain the social structure of gender. R.W. Connell (1987), a highly influential gender theorist, describes the existence of multiple, mutable masculine positions and identities, that are produced against a backdrop of history, cultural, and material circumstances, arranged hierarchically with respect to each other (some are viewed as legitimate while others are censored), and superior to women. Recognizing the global dominance of patriarchy, Connell argues that all masculine positions are constructed within a gender value system that values men over women and within which men are more powerful. Connell emphasizes that gender identities are socially constructed and integrally linked to hierarchies of power. Her analysis is critical to understanding both gender relations (i.e. between men and women) as well as the interrelationships among men. Connell argues for the existence in a particular setting of a hegemonic masculinity, which is a cultural ideal of manhood that gains its legitimacy from acceptance that is shared between those who embody and benefit from the ideal and those subordinated through it (after Gramsci (1971)). Hegemony connotes the existence of agreed values and practices, not necessarily associated with repression and violence. However, in many contexts, being dominant and in control of women are important aspects of hegemonic masculinity, and violent

behaviours are justified in pursuit and demonstration of this (e.g. Wood and Jewkes 2001).

The different masculine positions that are adopted by men in a given setting vary, but there are global commonalities that have been observed. The first is that in any given setting there are normally masculinities that are notably violent and combine practices of violence between men with violent domination over others, particularly women. Men who adopt these masculine positions often are involved in illegal and antisocial behaviour. Further their dominance over women encompasses both the use of violence against women and demonstrations of power through (hetero)sexual success with or dominance over women, for example, in paying for sex, having many sexual partners, and so forth. These masculinities are sometimes referred to as hyper- or emphasized masculinities and are very closely associated with violence between men as well as sexual and physical violence against women.

The hegemonic masculinity within a given setting is much more moderate than this. It may not be numerically the most common position, but has the greatest cultural power, in that it is a masculinity that generally embodies a recognized cultural ideal of how men 'should' be, even if some men are unable to occupy the position due to economic circumstances. Hegemonic masculinities vary considerably globally, particularly in their relationship with men's violence. In some settings, such as South Africa, showing toughness and strength is absolutely integral to a hegemonic position. Men may be expected to be head of a household, a provider, and for women to be obedient. The cultural legitimacy accorded to a hegemonic masculine position results in men being able to exercise power over women without the use of violence, rather through women's acquiescence to men's dominance, but violence may be used. In some countries, such as Sweden, hegemonic masculinities are much more gender equitable and not expected to entail violence (Hearn et al. 2012). The relationship between hegemonic masculinity and the use of violence is extremely important in understanding the prevalence of interpersonal violence within a given cultural setting.

There are of course other masculinities and some are much more gender equitable and less violent, even in contexts of very patriarchal hegemonic masculinities. Some are constructed as resistance masculinities and may strive to be different, as is the case with gay masculine identities. Hegemonic masculinities often emphasize heterosexuality and are constructed very

consciously in opposition to the feminine or homosexual, and indeed men may be identified as victims of male violence because they are perceived as being ‘unmanly’.

The importance of masculinities in interpersonal violence should not mask the role of women. Women are responsible for perpetration of much violence against children, but for other forms of interpersonal violence the majority of perpetrators are men. Women do use violence against intimate partners, but much of it is retaliatory violence and where they initiate it their violence differs in meaning with the violence of men because it is not perpetrated within an overall context of gender power inequity. They also may engage in youth violence with other women. In a given setting women also adopt multiple feminine positions, some of which are highly acquiescent to dominant positions of men. These positions may render women vulnerable to violence but they may also provide some protection. Femininities which are constructed as resistance femininities may be viewed by men as threats and make women deliberate targets of violence, for example, in attacks on lesbians. There is an increasing literature from high-income countries and some settings with rapid social change, such as South Africa, on ‘modern’ femininities. These are characterized by women adopting many of the practices of men, without necessarily challenging the dominance of men in the gender order. They may be associated with particularly risky behaviour such as fighting among women, binge drinking, and having many sexual partners (Jewkes and Morrell 2010).

The importance of understanding gender identities in relation to interpersonal violence is that they provide a unifying framework for understanding both male dominance over women and the use of violence between men and between women. Other connected practices such as alcohol bingeing very often stem from links between alcohol excess and control and toughness, or alcohol and resistance (e.g. in conservative settings), which is highly gendered. It also connects interpersonal violence with some of its associated sexual practices and associated health risks. First, by linking gun or weapon ownership, or carrying, with manhood. Secondly, sexual assault and rape perpetrators from a range of countries, including the United States, South Africa, Botswana, and Swaziland, have more dating and consensual partners and more casual sex than other men, although they may not have more frequent sex as they have less emotionally engaged relationships with partners (Koss and Dinero 1988; Malamuth et al. 1991,

1995; Abbey et al. 2006; Jewkes et al. 2006, 2011a, 2011b; Kalichman et al. 2007; DeGue et al. 2010; Tsai et al. 2011). Changing the values upon which gender identities are constructed is an important avenue for intervention in interpersonal violence reduction.

Stigma

Gender values of a society are also expressed in perceptions of blame in cases of gender-based violence and in social responses to victims. Rape stigma is particularly powerful as it may be expressed in perceptions that women and girls who have been raped are sullied, and to blame for what happened, and this can lead to their ostracism. In some settings the rape is seen as a mark against the family's honour and not just a woman's own, and they may even be killed. Children born of rape may be rejected by families and communities too. Rape stigma is a very substantial barrier to access to services and legal redress after rape, and thus creates impunity for those who perpetrate.

Societal level

At a societal level it is important to understand that the dominant social values, especially gender values, are expressed in laws, policies, and through social institutions such as the police and courts. Thus in very patriarchal societies, except in settings with a strong commitment to building human rights (such as South Africa), laws very often restrict women's autonomy and may increase their vulnerability to violence and reduce their ability to bargain with, or leave, violent partners. Examples of such laws are those that make divorce hard to access, allow parents to marry very young girls, fail to criminalize rape in marriage, restrict married women's access to property and inheritance, or render women as legal minors.

Dominant values are also reflected in systems implementing laws and policies, even where laws are more progressive. The police may refuse to open cases if they believe an act not to be a crime, this has particularly been reported in cases of rape and domestic violence. Courts may have a low threshold for declining to prosecute and where women are viewed with suspicion their evidence in court may be insufficient for a conviction (Jewkes and Abrahams 2002). If women cannot get legal redress for acts of

IPV or rape or child abuse, perpetrators are sent a powerful message that they may anticipate impunity and that their acts are not viewed as serious.

At a societal level, laws and policies may influence access to risk and protective factors, including alcohol, and support networks. In some countries, alcohol is cheap, easily accessible at all hours, and poorly regulated. These factors result in considerably higher alcohol consumption patterns, especially by young people, and so to a stronger link between interpersonal violence and heavy drinking. Similarly firearms are very effectively controlled by laws in some settings. South Africa is an example of a country where tighter gun control translated into a substantial reduction in gun-related homicides, these accounted for 17% of female homicides in 2009, a reduction of nearly 50% from one decade earlier (Abrahams et al. 2013). In comparison in the United States, with its very liberal gun laws, 67.5% of homicide is perpetrated with a firearm. Some countries have considerable investment in service provision for victim/survivors of interpersonal violence, enabling them to get support, counselling, and assisting access to justice through legal channels. Where such support is provided through women's organizations, these often also play important roles as raising awareness and advocating for women's rights at a local level and can make a powerful impact on social norms on violence. In Nicaragua a substantial reduction in violence against women was found in surveys over a period of 20 years and has been largely attributed to the activities of the organized and empowered women's sector.

Summarizing the prevention challenge

A comprehensive review of risk factors for interpersonal violence, as provided earlier, includes many which are amenable to intervention in the short or medium term and others which require change across generations. We can draw from them seven key tasks for prevention of interpersonal violence: empowering women, changing gender relations and building gender equality; preventing the use of all forms of violence and changing violence-supportive social norms; reducing structural inequalities such as poverty and low levels of education; building relationships skills, especially communication skills; strengthening mental health and reducing substance abuse; elevating the status of children to strengthen child protection and rights and end corporal punishment; and to provide legal and social

frameworks of protection for women and girls and other vulnerable groups, and to enhance justice for victims of interpersonal violence.

Research shows the amenability of forms of interpersonal violence to prevention through interventions (Ellsberg et al. 2015; Lester et al. 2017). Knowledge of the field of prevention of violence against women and girls has been considerably advanced over the last decade by the DFID-funded *What Works to Prevent Violence Against Women and Girls? Programme* (https://www.whatworks.co.za). This has conducted research in 15 countries to understand drivers of violence, social and economic costs, and what works in prevention. Evidence from the Programme suggests that well-designed evidence-based multicomponent interventions, that address multiple drivers of violence, have considerable success in reducing violence. A critical feature of their design is the need for them to be built on a strong theory of change—rooted in an analysis of the local drivers of violence—with the preventive interventions using robust behaviour change and skills building methodologies. In the field of prevention of violence against women and girls, gender inequity is centrally important and so more effective interventions have a unifying characteristic of being based on theories of gender and power. In some settings small decreases in violence have been measured with single component interventions in the form of cash transfers, but these have also been shown to increase women's exposure to violence in other settings, as shown by research on the impact of short term cash transfers for conflict-affected women in Syria. Where violence prevention interventions engage individuals in journeys of behaviour change, successful interventions use participatory methods with single-sex groups, build relationship skills and seek to engage the individuals targeted at an emotional level through critical reflection and responding empathetically towards survivors. These need to be delivered by well-trained, experienced, and supported individuals who espouse the values of the programme.

Awareness arising is a key element of work on interpersonal violence in settings where awareness is low, but it does not change behaviour on its own. Effective interventions that have been implemented across communities that use awareness raising have this as a pillar of more complex interventions which also include providing support for survivors, constructively engaging, and often strengthening, local services (such as health sector and police responses) and engaging couples where the wife

experiences violence directly with counselling. Even in very conservative and fragile settings, gender attitudes and practices of gender relations including the use of violence, are amenable to change and interventions have been able to demonstrate this, but such interventions take time to deliver and to demonstrate impact.

Interventions that have been shown to be effective in reducing violence against women include those that seek to elevate the status of women through gender empowerment programming and some form of economic empowerment intervention. This modality was first demonstrated in the IMAGE study in South Africa with a brief gender empowerment intervention, microfinance, and mobilization of women into community action against violence (Pronyk et al. 2006). These interventions have subsequently been shown to be successful with multiple modalities of economic empowerment including cash transfers, microfinance, and assistance in establishing small enterprises. In highly conservative, patriarchal settings, such as Tajikistan and Nepal, family-based interventions have been shown to be valuable for protecting younger women. Elsewhere the evidence seems to suggest these interventions empower older women more consistently to protect themselves from violence, than younger women. Research with very poor and disempowered young men, suggests that their perpetration of violence may be reduced by an intervention combining livelihood strengthening programming and gender transformative elements. This is important as there are a key group at very high risk of perpetration of violence.

Community action driven interventions have been shown to effectively reduce violence where they have been delivered over multiple years in sufficient intensity to impact communities. The prototype was the SASA! Intervention in Uganda (Abramsky et al. 2014). A key feature of these are the training of members of the community, often nominees from communities, in gender equity and non-violence, and often counselling, and supporting them in outreach activities. The overall impact is shifting gender attitudes and challenging norms around the use of violence across communities, as well as individual couple and survivor impact through work engaging couples and supporting survivors directly. These interventions usually engage traditional and religious leadership in communities as well, and in some versions have worked in interpretation of scriptures to promote gender equity and non-violence.

Interventions that are designed as workshop series which seek to empower men to critically reflect on their masculinities, relations with women, and use of violence have also been shown to be effective in preventing violence. Early rigorous evaluations tested the South African adaptation of Stepping Stones, and showed impact in reducing men's perpetration of violence, a finding which has been confirmed in a second randomized controlled trial (Welbourn 1995; Jewkes et al. 2008). Other similar interventions have been effective in building more gender equitable attitudes among men such as Program H in Brazil (Ricardo et al. 2012). More recently couples' interventions have also been tested and shown to be effective in complex settings, notably Rwanda.

A particularly challenging dimension of partner violence is that where male perpetration of violence is substantially influenced by alcohol and drug use and poor mental health. A psychotherapeutic intervention has been developed and tested in Zambia using the Common Elements Transdiagnostic Approach (CETA) with an additional substance abuse module, and this has been shown to enable a substantial reduction in both substance abuse and violence, sustained for 2 years in couples.

Interventions with children has been designed to prevent different forms of violence. There is considerable evidence that peer violence can be prevented by socioemotional programmes, cognitive behaviour and life skills interventions delivered in schools, and recently a play-based life skills intervention was shown to be impactful in the challenging setting of Pakistan (Lester et al. 2017). There is very little evidence of effective intervention to reduce corporal punishment in LMICs where it is often illegal but widely practised. An exception is the Good Schools Programme in Uganda, which was shown to be effective in reducing this as well as violence among children in school (Devries et al. 2015). Dating violence has been an area in which interventions have been developed and tested in schools. Success has been mixed, and in LMICs there are still no examples of successful programmes. In Canada, the participatory 4th R intervention is a notable example of one that has been shown to be effective in reducing boys' perpetration of violence (Wolfe et al. 2009).

There have also been a series of key systematic reviews showing that interventions can be effective in reducing child maltreatment and strengthening positive parenting (McCloskey 2011; Knerr et al. 2013). More recent work has sought to focus on interventions with older children and

their parents, including harnessing the power of parents to protect teenage children from violence. Notable examples here are the Skhokho Programme and the Parenting for Lifelong Health (Sinovuyo) intervention in South Africa, which provide some evidence that these programmes can improve a range of parenting and adolescent outcomes.

A key element in interpersonal violence prevention is also the framework of laws and policies and their implementation, including support for survivors of violence. This includes interventions to strengthen child protection systems, supporting and assisting victim/survivors of violence to gain justice as well as in managing the health sequelae. It involves improving the functioning of police and courts in cases of interpersonal violence, rehabilitation of offenders and adopting and enforcing a range of other legislation and policies to reduce risk including those focused on gun control, illicit drug laws and prevention of trafficking, anticorruption measures, and the prevention of problem drinking. Support for women experiencing violence, who often do not have access to assistance from the State, or are unable to use legal avenues, through an active women's sector is incredibly important.

Over the last few years the INSPIRE (https://www.who.int/inspire) and RESPECT (https://www.who.int/topics/respect) frameworks have provided a practical basis for understanding the prevention challenge with respect to ending violence against children and violence against women. There is considerable overlap between the two frameworks.

Addressing violence globally is a daunting task, but available data on trends show that violence can be prevented. National data from several countries show declines over the last few decades due to a constellation of social change. However there are also settings where interpersonal violence has escalated, sometimes with declines reversed, notably in relation to drug-related activities. The best confirmatory evidence of preventability comes from the United States, where the prevalence of all forms of interpersonal violence has very substantially declined over the last 30 years. For example child maltreatment and victimization is believed to have declined by 40–70% (Finkelhor and Jones 2006), IPV has declined by 49%, and figures from the Federal Bureau of Investigation show that the homicide rate declined from 9.8 to 4.7 per 100,000 from 1990 to 2010, making the current homicide rate comparable to that of 1960. It is much less clear which interventions worked, if indeed it was a product of interventions, and why the impact has

been greater in some states than in others. Criminologists and public health researchers have extensively debated the causes of the decline, because of the importance of understanding this for its replication in other settings, but unfortunately they only agree that they do not know how to attribute it. The period of the decline was one in which there was much more public health intervention to raise awareness of, and the prevention of, child maltreatment. There was a 24% reduction in families living in poverty and there were increasing policing interventions and crime reduction activities, as well as a range of other social changes including changes to build gender equity (Finkelhor and Jones 2006). The clustering of reduction in all forms of interpersonal violence (except child neglect, but this may be an artefact of case identification) strongly supports an argument that they are all closely interconnected and the need for a multipronged approach to prevention.

Conclusions

Interpersonal violence is a problem of huge global magnitude, responsible for a very substantial global burden of disease and with considerable economic impact. It is largely preventable and this chapter has sought to argue that although forms of interpersonal violence differ, prevention can substantially be unified as there are a common set of risk factors. Further when interpersonal violence does decline, as in the United States, a decline across all forms is seen to occur. Social norms around gender and the use of violence are extremely important in all forms of interpersonal violence and need to be a key focus for intervention. The prevention agenda in interpersonal violence, especially gender-based violence and child maltreatment, has historically been dominated by secondary prevention, with responses from child protection and the criminal justice system and the provision of services to victims. This is very important, but it is only a small part of what is needed to effectively prevent violence. First prize for the future will go to developing robust multisectoral primary prevention interventions that have an impact on multiple risk factors and change the social context within which interpersonal violence occurs.

REFERENCES

Abbey, A., Jacques-Tiura, A.J., and Lebreton, J. (2011). Risk factors for sexual aggression in young men: an expansion of the confluence model. *Aggressive Behaviour*, **37**, 450–64.

Abbey, A., Parkhill, M.R., Beshears, R., Clinton-Sherrod, A.M., and Zawacki, T. (2006). Cross-sectional predictors of sexual assault perpetration in a community sample of single African American and Caucasian Men. *Aggressive Behaviour*, **32**, 54–67.

Abbey, A., Parkhill, M.R., Clinton-Sherrod, M., and Zawacki, T. (2007). A comparison of men who committed different types of sexual assualt in a community sample. *Journal of Interpersonal Violence*, **22**, 1567–80.

Abrahams, N. and Jewkes, R. (2005). Effects of South African men's having witnessed abuse of their mothers during childhood on their levels of violence in adulthood. *American Journal of Public Health*, **95**, 1811–16.

Abrahams, N., Jewkes, R., Hoffman, M., and Laubsher, R. (2004). Sexual violence against intimate partners in Cape Town: prevalence and risk factors reported by men. *Bulletin of the World Health Organization*, **82**, 330–7.

Abrahams, N., Jewkes, R., Martin, L.J., and Mathews, S. (2011). Forensic medicine in South Africa: associations between medical practice and legal case progression and outcomes in female murders. *PLoS One*, 6, e28620.

Abrahams, N., Mathews, S., Martin, L.J., Lombard, C., and Jewkes, R. (2013). Intimate partner femicide in South Africa in 1999 and 2009. *PLoS Medicine*, 10, e1001412.

Abramsky, T., et al. (2014). Findings from the SASA! Study: a cluster randomized controlled trial to assess the impact of a community mobilization intervention to prevent violence against women and reduce HIV risk in Kampala, Uganda. *BMC Medicine* 12, 122

Ageton, S. (1983). *Sexual Assault Among Adolescents*. Lexington, MA: Lexington Books.

Asad, N., et al. (2017). The intersection of adolescent depression and peer violence: baseline results from a randomized controlled trial of 1,752 youth in Pakistan. *Journal of Child and Adolescent Mental Health*, **22** (4), 232–41.

Bender, D. and Lösel, F. (2011). Bullying at school as a predictor of delinquency, violence and other anti-social behaviour in adulthood. *Criminal Behaviour and Mental Health*, 21(2), 99–106.

Blumstein, A., Rivara, F.P., and Rosenfeld, R. (2000). The rise and decline of homicide—and why. *Annual Review of Public Health*, **21**, 505–41.

Borowsky, I.W., Hogan, M., and Ireland, M. (1997). Adolescent sexual aggression: risk and protective factors. *Pediatrics*, 100, e7.

Bourgois, P. (1996). In search of masculinity—violence, respect and sexuality among Puerto Rican crack dealers in East Harlem. *British Journal of Criminology*, **36**, 412–27.

Bowlby, J. (1969). *Attachment and Loss. Vol. 1: Attachment*. New York: Basic Books.

Bowlby, J. (1973). *Attachment and Loss. Vol. 2: Separation*. New York: Basic Books.

Bowlby, J. (1980). *Attachment and Loss. Vol. 3: Loss, Sadness and Depression*. New York: Basic Books.

Brecklin, L.R. and Ullman, S.E. (2002). The roles of victim and offender alcohol use in sexual assaults: results of a National Violence Against Women Survey. *Journal of Studies on Alcohol*, **63**, 57–63.

Calhoun, K.S., Bernat, J.A., Clum, G.A., and Frame, C. (1997). Sexual coercion and attraction to sexual aggression in a community sample of young men. *Journal of Interpersonal Violence*, **12**, 392–406.

Campbell, A., Muncer, S., and Odber, J. (1997). Aggression and testosterone: testing a bio-social model. *Aggressive Behaviour*, **23**, 229–38.

Canastar, A. and Maxson, S. (2003). Sexual aggression in mice: effects of male strain and female estrous state. *Behavioural Genetics*, **33**, 521–8.

Casey, E.A., Beadnell, B., and Lindhorst, T.P. (2009). Predictors of sexually coercive behaviour in a nationally representative sample of adolescent males. *Journal of Interpersonal Violence*, **24**, 1129–47.

Caspi, A., et al. (2002). Role of genotype in the cycle of violence in maltreated children. *Science*, **297**, 851–4.
Christiansen, K. and Knussmann, R. (1987). Androgen levels and components of aggression in men. *Hormone Behaviour*, **21**, 170–80.
Cleckley, H. (1988). *The Mask of Sanity*. St Louis, MO: Mosby Press.
Coetzee, J., Gray, G.E., and Jewkes, R. (2017). Prevalence and patterns of victimization and polyvictimization among female sex workers in Soweto, a South African township: a cross-sectional, respondent-driven sampling study. *Global Health Action*, 10(1), 1403815.
Cohen, L.J., et al. (2002). Heterosexual male perpetrators of childhood sexual abuse: a preliminary neuropsychiatric model. *Psychiatric Quarterly*, **73**, 313–35.
Connell, R. (1987). *Gender and Power: Society, the Person and Sexual Politics*. Palo Alta, CA: University of California Press.
Cooper, C., Selwood, A., and Livingston, G. (2008). The prevalence of elder abuse and neglect: a systematic review. *Age and Ageing*, **37**, 151–60.
Decker, M.R., et al. (2009). Intimate partner violence functions as both a risk marker and risk factor for women's HIV infection: findings from Indian husband–wife dyads. *Journal of Acquired Immune Deficiency Syndromes*, **51**, 593–600.
Degue, S. and Dilillo, D. (2004). Understanding perpetrators of nonphysical sexual coercion: characteristics of those who cross the line. *Violence and Victims*, **19**, 673–88.
Degue, S., Dilillo, D., and Scalora, M. (2010). Are all perpetrators alike? Comparing risk factors for sexual coercion and aggression. *Sexual Abuse*, **22**, 402–26.
Devries, K.M., et al. (2013). Global health. The global prevalence of intimate partner violence against women. *Science*, **340**, 1527–8.
Devries, K.M., et al. (2015). The Good School Toolkit for reducing physical violence from school staff to primary school students: a cluster-randomised controlled trial in Uganda. *The Lancet Global Health*, 385, e78–86.
Dunkle, K.L., Jewkes, R.K., Brown, H.C., Gray, G.E., Mcintryre, J.A., and Harlow, S.D. (2004). Gender-based violence, relationship power, and risk of HIV infection in women attending antenatal clinics in South Africa. *The Lancet*, **363**, 1415–21.
Dunkle, K.L., Van Der Heijden, I., Stern, E., and Chirwa, E.D. (2018). *Disability and Violence Against Women and Girls. Emerging Evidence from the What Works to Prevent Violence Against Women and Girls Global Programme*. Pretoria: South African Medical Research Council.
Duvvury, N., Carney, P., and Nguyen, H.M. (2012). *Estimating the costs of domestic violence against women in Viet Nam*. Hanoi: UN Women.
Edwards, A.C., et al. (2010). MAOA-uVNTR and early physical discipline interact to influence delinquent behavior. *Journal of Child Psychology and Psychiatry*, **51**, 679–87.
Ellsberg, M., et al. (2015). Prevention of violence against women and girls: what does the evidence say? *The Lancet*, **385**, 1555–66.
Fabian, J. (2012). Neuropsychology, neuroscience, volitional impairment and sexually violent predators: a review of the literature and law and their application to civil commitment proceedings. *Aggression and Violent Behaviour*, **17**, 1–15.
Fazel, S., Hope, T., O'Donnell, I., and Jacoby, R. (2002). Psychiatric, demographic and personality characteristics of elderly sex offenders. *Psychological Medicine*, **32**, 219–26.
Finkelhor, D. and Araji, S. (1986). Explanations of pedophilia: a four factor model. *Journal of Sex Research*, **22**, 145–61.
Finkelhor, D. and Jones, L. (2006). Why have child maltreatment and child victimisation declined? *Journal of Social Issues*, **62**, 685–716.
Fonagy, P. and Target, M. (2003). *Psychoanalytic Theories: Perspectives from Developmental Psychopathology*. New York: Routledge.
Fulu, E., Jewkes, R., Roselli, T., and Garcia-Moreno, C. (2013). Prevalence and risk factors for male perpetration of intimate partner violence: findings from the UN Multi-country Study on

Men and Violence in Asia and the Pacific. *The Lancet Global Health*, 1(4), e187–207.
Garcia-Moreno, C., et al. (2005). *WHO Multi-country Study on Women's Health and Domestic Violence Against Women: Initial Results on Prevalence, Health Outcomes and Women's Responses*. Geneva: WHO.
Gavey, N. (2005). *Just Sex? The Cultural Scaffolding of Rape*. Hove: Routledge.
GBD 2013 Mortality and Causes of Death Collaborators (2015). Global, regional, and national age–sex specific all-cause and cause-specific mortality for 240 causes of death, 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013. *The Lancet*, **385**, 117–71.
Gibbs, A., Dunkle, K., and Jewkes, R. (2018). Emotional and economic intimate partner violence as key drivers of depression and suicidal ideation: a cross-sectional study among young women in informal settlements in South Africa. *PLoS One*, 13(4), e0194885.
Gilbert, R., Widom, C.S., Browne, K., Fergusson, D., Webb, E., and Janson, S. (2009). Burden and consequences of child maltreatment in high-income countries. *The Lancet*, **373**, 68–81.
Giotakis, O., Markianos, M., Vaidakis, N., and Christodoulou, G. (2003). Aggression, impulsivity, plasma sex hormones and biogenic amine turnover in a forensic population of rapists. *Journal of Sex & Marital Therapy*, **29**, 215–25.
Global Women's Institute and International Rescue Committee (2017). *No Safe Place: A Lifetime of Violence for Conflict-Affected Women and Girls in South Sudan*. Washington DC: George Washington University and London/IRC.
Graham, K. and Wells, S. (2003). 'Somebody's gonna get their head kicked in tonight!' aggression among young males in bars—a question of values. *British Journal of Criminology*, **43**, 546–66.
Gramsci, A. (1971). *Selections from a Prison Notebook*. London: Lawrence & Wishart.
Hare, R. (1996). Psychopathy: a clinical construct whose time has come. *Criminal Justice and Behaviour*, **23**, 25–54.
Hare, R. (2003). *Psychopathy Checklist—Revised*. Toronto: Multi-health Systems.
Hatcher, A.M., Smout, E.M., Turan, J.M., Christofides, N., and Stockl, H. (2015). Intimate partner violence and engagement in HIV care and treatment among women: a systematic review and meta-analysis. *AIDS*, 29(16), 2183–94.
Hearn, J., et al. (2012). Hegemonic masculinity and beyond: 40 years of research in Sweden. *Men and Masculinities*, **15**, 31–55.
Hill, A., Pallitto, C., McCleary-Sills, J., and Garcia-Moreno, C. (2016). A systematic review and meta-analysis of intimate partner violence during pregnancy and selected birth outcomes. *International Journal of Gynaecology & Obstetrics*, 133(3), 269–76.
Jama-Shai, N., Jewkes, R., Levin, J., Dunkle, K., and Nduna, M. (2010). Factors associated with consistent condom use among rural young women in South Africa. *AIDS Care*, **22**, 1379–85.
Jewkes, R. (2010). Gender inequities must be addressed in HIV prevention. *Science*, **329**, 145–7.
Jewkes, R. and Abrahams, N. (2002). The epidemiology of rape and sexual coercion in South Africa: an overview. *Social Science & Medicine*, **55**, 1231–44.
Jewkes, R., et al. (2006). Rape perpetration by young, rural South African men: prevalence, patterns and risk factors. *Social Science & Medicine*, **63**, 2949–61.
Jewkes, R., Dunkle, K., Nduna, M., Jama, N., and Puren, A. (2010a). Associations between childhood adversity and depression, substance abuse & HIV & HSV2 in rural South African youth. *Child Abuse and Neglect*, **34**, 833–41.
Jewkes, R., Dunkle, K., Nduna, M., and Shai, N. (2010b). Intimate partner violence, relationship gender power inequity, and incidence of HIV infection in young women in South Africa: a cohort study. *The Lancet*, **367**, 41–8.
Jewkes, R., Fulu, E., Roselli, T., and Garcia-Moreno, C. (2013). *Prevalence* and risk factors for non-partner rape perpetration: findings from the UN Multi-country Cross-sectional Study on Men and Violence in Asia and the Pacific. *The Lancet Global Health*, 1(4), e208–18.
Jewkes, R., Dunkle, K., Jama-Shai, N., and Gray, G. (2015). Impact of exposure to intimate partner violence on CD4+ and CD8+ T cell decay in HIV infected women: longitudinal study. *PLoS*

One, 10(3), e0122001.
Jewkes, R., et al. (2017). Women's and men's reports of past year prevalence of intimate partner violence and rape and women's risk factors for IPV: a multi-country cross-sectional study in Asia and the Pacific. *PLoS Medicine*, 14(9), e1002381.
Jewkes, R. and Morrell, R. (2010). Gender and sexuality: emerging perspectives from the heterosexual epidemic in South Africa and implications for HIV risk and prevention. *Journal of the International AIDS Society*, **13**, 6.
Jewkes, R. and Morrell, R. (2012). Sexuality and the limits of agency among South African teenage women: theorising femininities and their connections to HIV risk practices. *Social Science & Medicine*, **74**, 1729–37.
Jewkes, R., et al. (2008). Impact of Stepping Stones on incidence of HIV, HSV-2 and sexual behaviour in rural South Africa: cluster randomised controlled trial. *British Medical Journal* 337, a506.
Jewkes, R., Nduna, M., Jama-Shai, N., and Dunkle, K. (2012). Prospective study of rape perpetration by young South African men: incidence & risk factors for rape perpetration. *PLoS One*, 7, e38210.
Jewkes, R., Penn-Kekana, L., and Rose-Junius, H. (2005). 'If they rape me, I can't blame them': reflections on the social context of child sexual abuse in South Africa and Namibia. *Social Science & Medicine*, **61**, 1809–20.
Jewkes, R., Sikweyiya, Y., Morrell, R., and Dunkle, K. (2011a). Gender inequitable masculinity and sexual entitlement in rape perpetration South Africa: findings of a cross-sectional study. *PLoS One*, 6, e29590.
Jewkes, R., Sikweyiya, Y., Morrell, R., and Dunkle, K. (2011b). The relationship between intimate partner violence, rape and HIV amongst South African men: a cross-sectional study. *PLoS One*, 6, e24256.
Jewkes, R., Corboz, J. and Gibbs, A. (2019). Violence against Afghan women by husbands, mothers-in-law and siblings-in-law/siblings: Risk markers and health consequences in an analysis of the baseline of a randomised controlled trial. *PLoS One*, 14(2), e0211361.
Johansson, A., et al. (2008). Genetic effectics on male sexual coercion. *Aggressive Behaviour*, **34**, 190–202.
Jones, L., et al. (2012). Prevalence and risk of violence against children with disabilities: a systematic review and meta-analysis of observational studies. *The Lancet*, **380**, 899–907.
Kalichman, S., Simbayi, L.C., Cain, D., Cherry, C., Henda, N., and Cloete, A. (2007). Sexual assault, sexual risks and gender attitudes in a community sample of South African men. *AIDS Care*, **19**, 20–7.
Karmaliani, R., et al. (2017). Peer violence perpetration and victimization: prevalence, associated factors and pathways among 1752 sixth grade boys and girls in schools in Pakistan. *PLoS One*, 12(8), e0180833.
Knerr, W., Gardner, F., and Cluver, L. (2013). Improving positive parenting skills and reducing harsh and abusive parenting in low- and middle-income countries: a systematic review. *Prevention Science*, **14**, 352–63.
Knight, R.A. and Sims-Knight, J.E. (2003). The developmental antecedents of sexual coercion against women: testing alternative hypotheses with structural equation modelling. *Annals of the New York Academy of Sciences*, **989**, 72–85.
Koenig, M.A., Stephenson, R., Ahmed, S., Jejeebhoy, S.J., and Campbell, J. (2006). Individual and contextual determinants of domestic violence in North India. *American Journal of Public Health*, **96**, 132–8.
Koss, M.P. and Dinero, T.E. (1988). Predictors of sexual aggression among a national sample of male college students. *Annals of the New York Academy of Sciences*, **528**, 133–47.
Kouyoumdjian, F.G., et al. (2013). Intimate partner violence is associated with incident HIV infection in women in Uganda. *AIDS*, **27**, 1331–8.

Krug, E., Dahlsberg, L., Mercy, J.A., Zwi, A.B., and Lozano, R. (2002). *World Health Report on Violence and Health*. Geneva: WHO.

LeBreton, J.M., Binning, J.F., and Adorno, A. (2006). Subclinical psychopaths. In: Thomas, J.C. and Segal, D. (eds.) *Comprehensive Handbook of Personality and Psychopathology*, Vol. *1, Personality and Everyday Functioning*. New York: Wiley, pp. 388–411.

Lester, S., Lawrence, C., and Ward, C.L. (2017). What do we know about preventing school violence? A systematic review of systematic reviews. *Psychology, Health & Medicine*, 22(supp1), 187–223.

Levinson, D. (1989). *Family Violence in Cross-Cultural Perspective*. Newbury Park, CA: Sage Publications.

Lightfoot, S. and Evans, I. (2000). Risk factors for a New Zealand sample of sexually abusive children and adolescents. *Child Abuse and Neglect*, **24**, 1185–98.

Lilienfield, S. (2005). *Psychopathic Personality Inventory—Revised*. Lutz, FL: Psychological Assessment Resources, Inc.

Machisa, M., Hatcher, A., Christofides, N., and Jewkes, R. (2019). Systematic review and meta-analysis of the associations between alcohol abuse, depression, or PTSD symptoms and adult male perpetration of intimate partner violence (In Press).

Malamuth, N., Linz, D., Heavey, C.L., Barnes, G., and Acker, M. (1995). Using the confluence model of sexual aggression to predict men's conflict with women: a 10 year follow up study. *Journal of Personality & Social Psychology* 69, 353–69.

Malamuth, N.M., Sockloskie, R.J., Koss, M.P., and Tanaka, J.S. (1991). Characteristics of aggressors against women: testing a model using a national sample of college students. *Journal of Consulting and Clinical Psychology*, **59**, 670–81.

Maniglio, R. (2010). The role of deviant sexual fantasy in the ethiopathogenesis of sexual homicide: a systematic review. *Aggression and Violent Behaviour*, **15**, 294–302.

Marshall, W. (1989). Invited essay: intimacy, loneliness and sexual offenders. *Behaviour Research and Therapy*, **27**, 491–503.

Marshall, W.L. and Barbaree, H. (1990). An integrated theory of the etiology of sexual offending. In: Marshall, W., Laws, D., and Barbaree, H. (eds.) *Handbook of Sexual Assault: Issues, Theories and Treatment of Offenders*. New York: Plenum, pp. 257–75.

Mathews, S., Abrahams, N., Jewkes, R., Martin, L., and Lombard, C. (2013). The epidemiology of child homicides in South Africa. *Bulletin of the World Health Organization*, 91(8), 562–8.

Maxwell, L., Nandi, A., Benedetti, A., Devries, K., Wagman, J., and Garcia-Moreno, C. (2018). Intimate partner violence and pregnancy spacing: results from a meta-analysis of individual participant time-to-event data from 29 low-and-middle-income countries. *BMJ Global Health*, 3(1), e000304.

McCloskey, L. (2011). *Systematic Review of Parenting Interventions to Prevent Child Abuse Tested with Randomised Control Trial (RCT) Designs in High Income Countries*. Pretoria: Sexual Violence Research Initiative.

McPherson, M., Smith-Lovin, L., and Cook, J. (2001). Birds of a feather: homophily in social networks. *Annual Review of Sociology*, **27**, 415–44.

Meel, B.L. (2009). Witchcraft in Transkei region of South African: case report. *African Health Sciences*, **9**, 61–4.

Morris, R.E., Anderson, M.M., and Knox, G. (2002). Incarcerated adolescents' experiences as perpetrators of sexual assualt. *Achives of Pediatric and Adolescent Medicine*, **156**, 831–5.

National Population Commission of Nigeria, UNICEF Nigeria, and the U.S. Centers for Disease Control and Prevention (2016). *Violence Against Children in Nigeria: Findings from a National Survey*, 2014. Abuja, Nigeria: UNICEF.

Nduna, M., Jewkes, R.K., Dunkle, K.L., Shai, N.P., and Colman, I. (2010). Associations between depressive symptoms, sexual behaviour and relationship characteristics: a prospective cohort

study of young women and men in the Eastern Cape, South Africa. *Journal of the International AIDS Society*, **13**, 44.
Olver, M.E. and Wong, S. (2006). Psychopathy, sexual deviance and recidivism among sex offenders. *Sexual Abuse*, **18**, 65–82.
Ouimette, P. (1997). Psychopathology and sexual aggression in nonincarcerated men. *Violence and Victims*, **12**, 389–95.
Prichard, Z., Mackinnon, A., Jorm, A.F., and Easteal, S. (2008). No evidence for interaction between MAOA and childhood adversity for antisocial behavior. *American Journal of Medical Genetics Part B*, 147B, 228–32.
Pronyk, P., Hargreaves, J.R., Kim, J.C., and Al, E. (2006). Effect of a structural intervention for the prevention of intimate partner violence and HIV in rural South Africa: a cluster randomised trial. *The Lancet*, **368**, 1973–83.
Reif, A., et al. (2007). Nature and nurture predispose to violent behavior: serotonergic genes and adverse childhood environment. *Neuropsychopharmacology*, **32**, 2375–83.
Reza, A., et al. (2007). *Violence Against Children in Swaziland: Findings from a National Survey on Violence Against Children in Swaziland May 15–June 16, 2007*. UNICEF. [Online] Available at: http://www.unicef.org/swaziland/sz_publications_2007violenceagainstchildren.pdf
Ricardo, C., Eads, M., and Barker, G. (2012). *Engaging Boys and Young Men in the Prevention of Sexual Violence: A Systematic and Global Review of Evaluated Interventions*. Pretoria: Sexual Violence Research Initiative.
Rosenberg, A.D., Abell, S.C., and Kanitz Mackie, J. (2005). An examination of the relationship between child sexual offending and psychopathy. *Journal of Child Sexual Abuse*, **14**, 49–66.
Rowe, D. (2001). *Biology and Crime*. Los Angeles, CA: Roxbury.
Seto, M.C. and Lalumiere, M. (2010). What is so special about male adolescent sexual offending? A review and test of explanations through meta-analysis. *Psychological Bulletin*, **136**, 526–75.
Shih, J.C., Chen, M., and Ridd, J. (1999). Monoamine oxidase: from genes to behaviour. *Annual Review of Neurosciences*, **22**, 197–217.
Snyder, J., Horsch, E., and Childs, J. (1997). Peer relationships of young children: affiliative choice and the shaping of aggressive behaviour. *Journal of Clinical Child Psychology*, **26**, 145–56.
Steglich, C.E.G., Snijders, T.A.B., and Pearson, M. (2010). Dynamic networks and behaviour: separating selection from influence. *Sociological Methodology*, **40**, 329–93.
Stellar, C., Garcia-Moreno, C., Temmerman, M., and van der Poel,S. (2016). A systematic review and narrative report of the relationship between infertility, subfertility, and intimate partner violence. *International Journal of Obstetrics and Gynaecology*, **133**, 3–8.
Stirpe, T., Abracen, J., Stermac, L., and Wilson, R. (2006). Sexual offenders' state-of-mind regarding childhood attachment: a controlled investigation. *Sexual Abuse*, **18**, 289–302.
Tolin, D.F. and Foa, E.B. (2006). Sex differences in trauma and posttraumatic stress disorder: a quantitative review of 25 years of research. *Psychological Bulletin*, **132**, 959–92.
Tsai, A.C., et al. (2011). Prevalence and correlates of forced sex perpetrtaion and victimisation in Botswana and Swaziland. *American Journal of Public Health*, **101**, 1068–74.
Walby, S. (2010). *The Cost of Domestic Violence: Update 2009*. Lancaster: Lancaster University.
Waters, H., Hyder, A., Rajkotia, Y., Basu, S., Rehwinkel, J.A., and Butchardt, A. (2004). *Economic Dimensions of Interpersonal Violence*. Geneva: WHO.
Welbourn, A. (1995). *Stepping Stones*. Oxford: Strategies for Hope.
Whitaker, D.J., et al. (2008). Risk factors for perpetration of child sexual abuse: a review and meta-analysis. *Child Abuse and Neglect*, **32**, 529–48.
Wolfe, D.A., et al. (2009). A school-based program to prevent adolescent dating violence: a cluster randomized trial. *Archives of Pediatrics & Adolescent Medicine*, **163**, 692–9.
Wood, K. and Jewkes, R. (2001). 'Dangerous' love: reflections on violence among Xhosa township youth. In: Morrell, R. (ed.) *Changing Men in Southern Africa*, pp. 317–36. Pietermaritzburg: University of Natal Press, Zed Press.

Wood, K., Lambert, H., and Jewkes, R. (2007). ‘Showing roughness in a beautiful way’: talk about love, coercion, and rape in South African youth sexual culture. *Medical Anthropology Quarterly*, **21**, 277–300.

World Health Organization (WHO) (2004). *Global Status Report on Alcohol 2004*. Geneva: WHO.

World Health Organization (WHO) (2013). *Global and Regional Estimates of Violence Against Women: Prevalence and Health Effects of Intimate Partner Violence and Non-Partner Sexual Violence*. Geneva: WHO.

World Health Organization (WHO) (2015). *Preventing Youth Violence: An Overview of the Evidence*. Geneva: WHO.

World Health Organization/London School of Hygiene and Tropical Medicine (WHO) (2010). *Preventing Intimate Partner and Sexual Violence Against Women: Taking Action and Generating Evidence*. Geneva: WHO.

9.6

Collective violence

War

Barry S. Levy

Introduction

There are three categories of violence: self-directed violence, interpersonal violence, and collective violence. Collective violence is ‘the instrumental use of violence by people who identify themselves as members of a group —whether this group is transitory or has a more permanent identity— against another group or set of individuals in order to achieve political, economic, ideological, or social objectives’ (Krug et al. 2002, p. 215). Collective violence includes armed conflict (such as war), state-sponsored violence (such as genocide, repression, disappearances, and torture), and organized violent crime (such as gang warfare and banditry). This chapter focuses on war. (Chapter 8.7 focuses on self-directed violence. Chapter 9.5 focuses on interpersonal violence, including gang warfare. Chapter 8.19 covers *bioterrorism*.)

During 2019, there were 32 armed conflicts: 15 in sub-Saharan Africa, 7 in the Middle East and North Africa, 7 in Asia and Oceania, 2 in North American and South America, and 1 in Europe. Most of these were intrastate conflicts (mainly ‘civil wars’), occurring within a low- or middle-income country between government forces and one or more armed non-state groups. There were three wars with 10,000 or more battle-related deaths in 2019—in Afghanistan (approximately 41,900), Yemen (25,900), and Syria (15,300) (Stockholm International Peace Research Institute 2020).

There are multiple underlying and precipitating causes of war. Many wars are caused by disputes over government power, territory, or

environmental resources. Risk factors for war include: (a) demographic factors, such as high infant mortality and high population density; (b) social and community factors, such as inequality among groups, availability of small arms and light weapons, and racial, ethnic, or religious tensions; (c) political factors, such as inadequate public role in decision-making, rapid regime change, and unequal access to power; (d) economic factors, such unequal access to or distribution of resources; and (e) other factors, such as deterioration in public services and cycles of violent revenge (Krug et al. 2002, pp. 213–39; Carnegie Commission on Preventing Deadly Conflict 1998). Conflicts over access to freshwater, which are increasing (Gleick 2018), can lead to war (Levy and Sidel 2011). Climate change, by causing increased temperature and extremes of precipitation with associated consequences, may increase the risk of armed conflict (Levy and Patz 2015). Ethnic tensions, militarism, and the arms trade contribute to the risk of war.

The health and environmental consequences of war

The consequences of war include morbidity and mortality due to weapons; damage to the health-supporting infrastructure of society; contamination of air, water, and land; forced displacement; violation of human rights; diversion of financial and human resources; and promotion of additional violence (Levy and Sidel 2008, 2013). Each of these consequences is described in more detail next.

Morbidity and mortality due to weapons

Many people, including an increasing percentage of civilians, have been killed or injured by weapons during war. An estimated 191 million people died directly or indirectly as a result of conflict during the twentieth century (Ferguson 2006). More than half of those who died were civilians and tens of millions of people died as a result of mass murder and genocide (Rummel 1994; Power 2002). In addition to physical injuries and illnesses, wars cause—both among military personnel and civilians—mental health consequences, including anxiety, depression, posttraumatic stress disorder,

alcohol and drug abuse, violence inflicted on family and friends, and suicide (Blore et al. 2015; Bonde et al. 2016).

From the early twentieth century to the start of the Second World War, the vast majority of deaths during war were among uniformed combatants, usually members of national armed forces. Although non-combatant civilians suffered from the social, economic, and environmental consequences of war, they were usually not targeted. However, since 1937, when Nazi forces bombed Guernica, a city in Spain, military operations, in addition to causing 'collateral damage' to non-combatant civilians as a result of attacks on military targets, have increasingly targeted—and injured and killed—non-combatant civilians (Tirman 2011).

Damage to the health-supporting infrastructure of society

Much morbidity and mortality during war, especially among non-combatant civilians, results from damage to the health-supporting infrastructure of society. In some wars, such as the civil war in the Democratic Republic of the Congo (in which more than five million people died during the war and its aftermath), the vast majority of deaths among non-combatant civilians have been caused by infectious diseases, malnutrition, and pregnancy- and neonatal-related conditions—largely due to damage to the societal infrastructure (Coghlan et al. 2008). This damage includes adverse effects on medical care and public health services. Physicians, nurses, and other health workers are injured or killed or they flee; clinics and hospitals are physically damaged; public health services are reduced or eliminated; and there is inadequate or no access to essential medicines and vaccines. Damage to the societal infrastructure also includes adverse impacts on food and water supply systems, sewage disposal systems, power plants and electrical grids, and transportation and communication systems.

Contamination of air, water, and soil

The physical environment may be contaminated during war—or the preparation for war—by military exercises on land, in air, and in water and by the production, testing, use, and disposal of weapons. Hazardous wastes

from military operations can contaminate air, water, and soil. For example, groundwater contamination occurred due to disposal of petroleum products and other hazardous materials at the Otis Air Force Base in Massachusetts. Pesticides, solvents, other chemicals, and radioactive materials were dumped for over 30 years at the Rocky Mountain Arsenal in Colorado. Benzene, arsenic, chromium, and mercury have been detected at the McChord Air Force Base in Washington State (Westing 2008). In addition, release of hazardous and toxic substances from damaged industrial facilities can contaminate the environment. For example, during the Iraq War, environmental contamination occurred due to hazardous materials from a metal-plating facility, pesticides, and petrochemicals warehouses, a sulphur mining complex, and a military scrapyard site (United Nations Environmental Programme 2005).

Conventional weapons have accounted for the overwhelming majority of adverse environmental consequences in war. During the Second World War, for example, extensive carpet bombing of cities in Europe and Japan caused many deaths and injuries, as well as widespread devastation of urban environments. During the Gulf War, more than 600 oil fires in Kuwait caused widespread environmental devastation and respiratory disorders among people who were exposed to the smoke from these fires. Bombing during the Vietnam War destroyed mangrove forests; resultant bomb craters remained long afterward, often filling with stagnant water that enabled breeding of mosquitoes that transmit malaria and other diseases.

Forced displacement

War uproots people. Many of the approximately 26 million refugees globally were forced by war to leave their native countries; the five leading countries of origin of refugees in late 2019 were Syria, Venezuela, Afghanistan, South Sudan, and Myanmar, all of which had experienced war or internal conflict. Refugees often flee to neighbouring low- and middle-income countries, which are often already facing significant challenges in addressing the public health needs of their own populations. Seven of the eight countries sheltering the most refugees in late 2019—Turkey, Colombia, Pakistan, Uganda, Germany, Sudan, Iran, and Lebanon—are low- and middle-income countries already challenged in meeting the needs

of their own citizens. Many of the approximately 46 million internally displaced persons and approximately 8 million stateless people were also forced to leave their homes as a result of war. In 2019, there were approximately 11 million people newly displaced, mainly by conflict and violence; the leading countries in which they were internally displaced were Syria, Colombia, the Democratic Republic of the Congo, Yemen, Afghanistan, Somalia, and Nigeria—all of which had experienced significant conflict and violence. The vast majority of all these uprooted people are women, children, and elderly people, many of whom are at increased risk of injury and illness, threats to their safety and security, and death. Internally displaced persons are generally worse off than refugees because they frequently do not have security or access to safe food and water, healthcare and public health services, shelter, and other basic human needs.

Violation of human rights

Human rights have often been violated during war. These violations have included assaults on civilians, including sexual assaults on women and abduction of children into military forces; mass killing; torture of prisoners and other violations of the Geneva Conventions; and violations of medical neutrality. There are more than 250,000 children forcibly engaged in armed conflict. They have been abducted to join fighting forces or to serve as ‘sex slaves’. These children suffer physical injuries and illnesses, including mental health problems (Song and de Jong 2015).

The Geneva Convention IV mandates the protection of civilian populations during war. This convention states that civilian populations and individual civilians shall not be the objects of attacks. It prohibits acts or threats of violence whose primary purpose is to spread terror among civilian populations. It prohibits indiscriminate attacks on civilians, such as those that are not directed at a specific military objective or those that may be expected to cause incidental civilian deaths or injuries and/or damage to ‘civilian objects’. In addition, it bans reprisal attacks against civilian populations or civilians.

Diversion of resources

Former United States President Dwight D. Eisenhower said: ‘Every gun that is made, every warship launched, every rocket signifies, in the final sense a theft from those who hunger and are not fed, those who are cold and not clothed’ (Eisenhower 1953). Military expenditures, which often constitute a high percentage of national budgets, often divert financial—and human—resources away from health and human services and other beneficial activities. In many countries, the total amount of military expenditures approximates the amount spent on healthcare and public health. In 2019, the United States spent $732 billion on military expenditures, representing 38% of global military expenditures—more than the next 10 countries (China, India, Russia, Saudi Arabia, France, Germany, the United Kingdom, Japan, South Korea, and Brazil) combined (Stockholm International Peace Research Institute 2020, p. 226). As a reflection of diversion of resources, the United States ranks first among the nations of the world in arms exports, military expenditures, and military technology, but ranks 30th in life expectancy and 43rd in infant mortality rate.

As another example of the impact of the diversion of resources, the first $204 billion spent by the United States on the Iraq War could have achieved all of the following goals for three years: a 50% decrease in world hunger, a comprehensive global programme to address HIV/AIDS, complete immunization of all children in ‘less-developed countries’, and meeting all needs globally for clean water and sewage treatment (Bennis et al. 2005).

Promotion of additional violence

Violence often promotes further violence. In the aftermath of war, a number of factors can lead to additional violence, such as persistent hatred and unresolved disputes among warring countries or groups, militarism, and the international arms trade. There are many examples in which, in the aftermath of war, persistent hatred and unresolved disputes have led to further armed conflict, such as in the Middle East. Additional violence can result from militarism, which is defined as a philosophy or a system of beliefs and attitudes that a country should develop and maintain strong

military power and aggressively deploy it consistent with its values and national interests. The international trade in conventional weapons also heightens the risk of violence especially in low- and middle-income countries. The volume of international transfers of major arms increased 5.5% between the 2010–2014 and 2015–2019 periods. During the 2015–2019 period, the five largest suppliers of major weapons were the United States (36%), Russia (21%), France (7.9%), Germany (5.8%), and China (5.5%), and the Middle East region received 35% of imports of major weapons (Stockholm International Peace Research Institute 2020, p. 278 and p. 295).

The United Nations General Assembly, in 2013, took a major step towards regulating the international trade in conventional weapons by overwhelmingly approving the Arms Trade Treaty, which links purchases of arms to the human rights records of the purchasing countries. The treaty is designed to prevent exported weapons from contributing to human rights violations, terrorism, or organized crime, and to prohibit arms shipments that are deemed to be harmful to women or children.

Types of weapons

Conventional weapons

Conventional weapons consist of small arms and light weapons, heavy conventional weapons (such as tanks, artillery, combat aircraft, warships, and missiles), explosives, and incendiaries. Small arms and light weapons, which include pistols, rifles, machine guns, and other hand-held or easily transportable weapons, are the weapons that are most often used in war. They have occasionally been described as ‘weapons of mass destruction’ because the number of deaths that they cause ‘dwarf that of all other weapons systems—and in most years greatly exceed the toll of the atomic bombs that devastated Hiroshima and Nagasaki’ (Annan 2000).

Armed drones are a category of conventional weapons. Since 2004, the United States has been conducting armed drone strikes in Afghanistan, Pakistan, Somalia, and Yemen, in attempts to kill suspected terrorists. The Bureau of Investigative Journalism has reported that, through 2020, most of the people who had been killed in 14,040 armed drone strikes in these countries had been militants, but at least 910 of them were civilians.

Nuclear and radioactive weapons

The detonation of nuclear bombs over Hiroshima and Nagasaki in August 1945 during the Second World War immediately killed about 200,000 people, primarily non-combatant civilians, and later caused many more deaths from malignancies and other diseases. These attacks injured and disabled many more people, and caused massive devastation—and radioactive contamination—of the environment in these two cities (Sutton and Gould 2008).

As of 2020, there were 13,400 nuclear warheads possessed by nine nations—in order of decreasing number of warheads, Russia, the United States, China, France, the United Kingdom, Pakistan, India, Israel, and North Korea (Stockholm International Peace Research Institute 2020, p. 326). The United States (with 5,800 nuclear warheads, 1,750 of which are deployed) and Russia (with 6,375 nuclear warheads, 1,570 of which are deployed) possess the vast majority of nuclear weapons. Both countries have hundreds of nuclear warheads on alert on land-based intercontinental ballistic missiles and submarine-launched ballistic missiles. All of the nuclear-armed countries are planning major programmes to upgrade and modernize their nuclear weapons and delivery systems. In addition to the possible intended or accidental use of nuclear weapons by these countries, it is increasingly recognized that non-state actors, such as terrorist groups, could acquire and detonate nuclear weapons.

Atmospheric testing of nuclear weapons by the United States, the former Soviet Union, and other countries caused environmental contamination, with increased rates of leukaemia and other cancers among downwind populations. In addition, production of nuclear weapons caused environmental contamination. For example, the Techa River in the Chelyabinsk Region of Russia was heavily contaminated with radioactive materials from the nuclear weapons production facility there (Burmistrov et al. 2006). Leakage from stored radioactive wastes at the nuclear weapons production facility at Hanford, Washington, has also caused environmental contamination (Westing 2008).

There are several international treaties concerning nuclear weapons. The Partial Test Ban Treaty (PTBT) of 1963, promoted in part by concerns about radioactive contamination of the environment, banned nuclear tests in the atmosphere, under water, and in outer space. The expansion of the

PTBT, the Comprehensive Nuclear-Test-Ban-Treaty (CTBT)—a key step towards nuclear disarmament and preventing proliferation of nuclear weapons—was opened for signature in 1996 but has not yet entered into force. The CTBT bans explosions of nuclear weapons, for either military or civilian purposes, but it does not ban computer simulations and subcritical tests, which some nations use to maintain the option of developing new nuclear weapons. The CTBT has been ratified by 168 nations and signed, but not ratified, by another 16 nations. Entry into force requires ratification by the 44 nuclear-capable nations; by 2018, only 36 had ratified the CTBT and the following eight nuclear-capable nations had not: China, Egypt, India, Iran, Israel, North Korea, Pakistan, and the United States.

The Treaty on the Non-Proliferation of Nuclear Weapons (the ‘Nuclear Non-Proliferation Treaty’, or NPT) attempts to prevent the spread of nuclear weapons by restricting transfer of certain technologies. It relies on a control system operated by the International Atomic Energy Agency, which also promotes nuclear energy. In exchange for the non-nuclear-armed countries’ commitment not to develop or otherwise acquire nuclear weapons, the NPT commits the nuclear-armed countries to good-faith negotiations on nuclear disarmament. The NPT was opened for signature in 1968 and entered into force in 1970. A total of 191 countries have ratified the treaty. All of the five nuclear-armed countries recognized under the NPT —China, France, Russia, the United Kingdom, and the United States—have ratified it. Every 5 years since 1970, the states-parties to the treaty have held a review conference to assess implementation of the treaty. In 1996, the International Court of Justice (the World Court), in an advisory opinion, urged that the nations possessing nuclear weapons move expeditiously towards nuclear disarmament, as is required by the NPT.

All of the nuclear-armed countries can help stop the spread of nuclear weapons by actively supporting and adhering to these treaties, and can renounce the first use of nuclear weapons and the development of new nuclear weapons. The United States and Russia can collaborate in dismantling their nuclear warheads and can secure nuclear materials better so that they will not be acquired by non-state actors.

Attention is increasingly being focused on the total elimination of nuclear weapons. In 2017, the United Nations adopted the Treaty of the Prohibition of Nuclear Weapons (TPNW), which comprehensively bans nuclear weapons, including their possession. The TPNW is akin to the chemical and

biological weapons conventions. Under the treaty, it is illegal for its parties to develop (or finance to develop), test, produce, acquire, possess, store, transfer, transport, plan or threaten to use, or use nuclear weapons. Nuclear-armed countries that join the treaty will need to destroy their nuclear weapons and terminate their nuclear weapons programmes. The treaty also mandates that parties to the treaty assist victims of nuclear weapons testing and use and it requires environmental remediation of areas contaminated by nuclear weapons. By the end of 2020, the treaty had been signed by 86 countries and ratified by 51.

Possible deployment of ‘dirty bombs’, which are conventional explosive devices mixed with radioactive materials, and possible attacks on nuclear power plants represent other risks. So does depleted uranium (DU), from which the uranium isotope usable for nuclear weapons or for fuel rods in nuclear power plants has been removed. DU is used militarily as a casing for armour-penetrating shells. An extremely dense material, uranium increases the ability of shells to penetrate the armour of tanks; uranium also bursts into flames on impact. DU-encased shells were used by the United States during the Gulf War, the Iraq War, and the war in Kosovo; similar shells were used by the United Kingdom in the Iraq War. DU, which is both radioactive and toxic, can contaminate soil and groundwater (Bešić et al. 2018). Use of DU is considered legal by the nations that have used it, but its use is considered by other nations to be illegal under the Geneva Conventions and other international treaties.

Chemical weapons

Chemical weapons, which were first used during the First World War, have been used on several occasions since then. The government of Syria under President Bashar al-Assad has used chemical weapons during the civil war in Syria many times against civilians, causing many deaths (Arms Control Association 2020). A Japanese cult, Aum Shinrikyo, used sarin in the subway system of two Japanese cities in the mid-1990s, killing 19 people and injuring thousands.

The Chemical Weapons Convention (CWC), which came into force in 1997, prohibits development, production, acquisition, stockpiling, transfer, and use of chemical weapons. It requires each participating country to destroy its chemical weapons and chemical weapons production facilities,

as well as any chemical weapons it may have abandoned on the territory of another participating country. The CWC has been signed by 193 nations; one nation has signed, but not yet ratified it; and three nations have neither signed nor ratified it. The verification provisions of the CWC affect not only the military sector but also the civilian chemical industry with restrictions and obligations regarding production, processing, and consumption of chemicals that are considered relevant to the convention. These provisions are verified by reporting requirements, routine onsite inspection of declared sites, and short-notice challenge inspections. The CWC established the Organization for the Prohibition of Chemical Weapons (OPCW) in The Hague to ensure implementation of its provisions (Lee et al. 2012). Stockpiles of chemical weapons, which remain in a number of countries, pose a continuing threat to health and the environment.

Biological agents

Biological agents, which consist of bacteria, viruses, other microorganisms, and their toxins, can produce illness in humans and can be used against other animals or plants, thereby adversely affecting agricultural resources, food supplies, and human health. Biological agents have been used as weapons since ancient times (Barras and Greub 2014). In the sixth century BC, Persia, Greece, and Rome tried to contaminate drinking water sources with diseased corpses. In AD 1346, Mongols besieging the Crimean seaport of Kaffa placed cadavers of plague victims on hurling machines and threw them into the city. During the French and Indian War in the United States in the mid-eighteenth century, a British commander sent blankets infected with smallpox to Native Americans. During the First World War, Germany dropped bombs containing plague bacteria over British positions and used cholera in Italy. During the 1930s, Japan contaminated food and water supplies of several cities and sprayed the cities with cultures of microorganisms. In 1942, a test use of anthrax spores by the United Kingdom and the United States contaminated Gruinard Island, off the coast of Scotland (Harris and Paxman 1962). During the 1950s and 1960s, secret large-scale open-air tests at the US Army Dugway Proving Ground in Utah used the microorganisms that cause Q fever and Venezuelan equine encephalitis (Cole 1988). In 1979, the accidental release of anthrax spores

near Sverdlovsk in the Soviet Union resulted in at least 77 cases of inhalation anthrax, at least 66 of which were fatal (Meselson et al. 1994).

Biological agents could be used as terrorist weapons. In 2001, anthrax spores were disseminated through the US Postal Service, causing 23 cases of inhalational and skin anthrax, five of which were fatal. The Centers for Disease Control and Prevention (CDC) has identified three categories of diseases caused by biological agents, according to level of concern that they may be used as terrorist weapons. Category A consists of the agents that can be easily disseminated or transmitted from person to person, result in high mortality rates and have the potential for major public health impact, might cause public panic and social disruption, and require special action for public health preparedness. Category A agents include those that cause anthrax, botulism, plague, smallpox, tularaemia, and viral haemorrhagic fevers, such as the Ebola, Marburg, and Lassa viruses. Category B consists of the agents that are moderately easy to disseminate, result in moderate morbidity rates and low mortality rates, and require specific enhancements of CDC's diagnostic capacity and enhanced disease surveillance. Category B agents include those that cause brucellosis, glanders, melioidosis, psittacosis, Q fever, typhus fever, viral encephalitis, water safety threats, and food safety threats (such as *Salmonella* and *Shigella* species, and *Escherichia coli* O157:H7), as well as epsilon toxin of *Clostridium perfringens*, ricin toxin from castor beans, and staphylococcal enterotoxin B. Category C agents include emerging pathogens that could be engineered for mass dissemination in the future because of availability, ease of production and dissemination, and potential for high morbidity and mortality rates and for major health impact (Centers for Disease Control and Prevention 2018).

The 1975 Biological and Toxin Weapons Convention (BWC), prohibits the development, production, transfer, or use of biological weapons. The BWC has been signed and ratified by 183 countries; signed, but not ratified, by four countries; and neither signed nor ratified by 10 countries. The verification measures included in the BWC are weak and attempts to strengthen them have been unsuccessful. The United States has blocked attempts to strengthen the verification measures of the BWC, stating that such measures might lead to exposure of US industrial or military secrets.

Antipersonnel landmines and unexploded ordnance

Antipersonnel landmines, which are designed to be exploded by the presence, proximity, or contact of a person, can incapacitate, injure, or kill one or more people. Together with unexploded ordnance (UXO), they represent significant hazards in many countries and continue to account for much morbidity and mortality (Frost et al. 2017). Millions of antipersonnel landmines are still deployed in many countries. Antipersonnel landmines injure or kill thousands of people annually; the vast majority of landmine victims are civilians, primarily poor people, including children, living in rural areas. It is estimated that half of all landmine victims die of their injuries before they reach appropriate medical care. Although a mine may cost only a few dollars to produce, it may cost more than US$1,000 to remove, and its economic impact on human health and well-being can be substantially higher.

Antipersonnel landmines, in addition to maiming and killing people, make large areas of land uninhabitable. Remaining in place for many years, they pose long-term threats to people, including refugees and internally displaced persons returning to their homes in the aftermath of war. Since the Mine Ban Treaty entered into force in 1997, production of landmines has been markedly reduced and millions of landmines that had been implanted in the ground have been removed (International Campaign to Ban Landmines n.d.). However, about 36 countries continue to retain antipersonnel landmines and millions of landmines are still deployed in 58 countries; additional resources will be required to unearth and destroy them, tasks that pose inherent risks to demining personnel (Sirkin et al. 2008). A total of 164 nations are parties to the Mine Ban Treaty. Thirty-two nations have neither signed nor ratified it, including the United States, Russia, and China.

Prevention

Public health workers and other health professionals can play important roles in minimizing the health and environmental consequences of war and preventing war itself (Levy and Sidel 2008; Wiist and White 2017). Categories of preventive measures include: (1) documentation; (2) education and awareness-raising; (3) advocacy and support for policies and

programmes to minimize the consequences of war and to prevent war; and (4) provision of preventive services.

Documentation

It is important to document the health and environmental consequences of war in order to inform policymakers and the public about these consequences, provide necessary services to affected populations, document violations of international law, protect human rights, and reduce the likelihood of future war or other forms of violent conflict. Documentation can be performed by surveillance, epidemiological surveys, and rapid assessment (Levy and Sidel 2016).

Public health surveillance includes the systematic collection, analysis, and dissemination of data. Before, during, and after war, surveillance can provide information on morbidity and mortality as well as hazardous exposures and other factors that place non-combatant civilians at increased risk of injury, illness, or death. Even before a war has begun, surveillance may be able to identify population-based risk factors that may lead to interventions that may prevent the war from occurring.

Epidemiologic studies may provide critically important information on the nature, severity, and magnitude of morbidity and mortality, as well as on related risk factors. These studies are usually cross-sectional, but prospective and retrospective studies generally provide more accurate morbidity and mortality data.

Rapid assessment is often performed intensely during a short period, but can still be used to obtain preliminary information on a disastrous event or situation during war or its aftermath. It is also used to estimate the nature and magnitude of future health and related problems so that appropriate services can be provided.

Challenges to documenting accurately the adverse effects of war include inadequate or absent data systems, social breakdown, forced displacement, biases in reporting of data, problems in distinguishing combatants and non-combatant civilians, inadequate concern for the health of ‘enemy’ civilians, and ‘the fog of war’.

Education and awareness-raising

Much can also be accomplished by health professionals educating and raising the awareness of other health professionals, policymakers, and the general public about the health and environmental consequences of war. A multifaceted approach can include presentations and publications by professional organizations and citizens' groups, reports by the news media, and use of social media and other forms of communication. Health professionals who have witnessed war and its consequences and non-combatant and military personnel who have suffered from the consequences of war can be very effective in raising awareness about the impact of war.

Advocacy and support for policies and programmes

Health workers can advocate for and support policies and programmes to minimize the consequences of war, such as protecting the health-supporting infrastructure of society during war and its aftermath, preventing the exploitation of women and children, protecting the rights of displaced persons and prisoners of war, and maintaining safe zones for hospitals and other healthcare facilities. Advocacy is often most effective when conducted by professional associations, such as the American Public Health Association, or advocacy organizations, such as the International Physicians for the Prevention of Nuclear War and its national affiliate organizations, including Physicians for Social Responsibility and Medact.

Provision of preventive services

Health professionals can directly provide preventive services during war and its aftermath. These may include providing healthcare, including mental health services, for displaced individuals and populations; providing immunizations and other preventive services for children, women, and other civilians; and assisting in the restoration of healthcare and public health services.

Roles of the United Nations

Since its founding in 1945, the United Nations has had as its primary mission the maintenance of international peace and security. The United Nations works to accomplish its mission by preventing conflict, helping

nations and other groups involved in conflict achieve peace, providing peacekeeping activities, and promoting conditions in which peace can be sustainably maintained and flourish.

The Security Council of the United Nations determines the presence of a threat to peace or an act of aggression. It asks parties to a dispute or conflict to settle it by nonviolent means and it often recommends terms of settlement. The Security Council can implement a number of measures, such as economic sanctions and international military action, to restore or maintain peace and security among nations.

The United Nations General Assembly provides a forum for member nations to express their perspectives on issues and to try to find consensus on challenging matters. It makes decisions concerning issues of peace and security.

By using special envoys, political missions, and other forms of diplomacy, the United Nations tries to resolve conflicts before they become violent. In addition, it engages in peacekeeping activities to assist member countries in transitioning from conflict to peace. With troops and police contributed by member states, it maintains peace and security and also promotes peaceful political processes, protects civilian populations, and promotes disarmament, demobilization, and reintegration of former combatants. It also supports constitutional processes, helps to organize elections, protects and promotes human rights, and helps restore the rule of law. The United Nations has deployed 71 peacekeeping missions since 1948 and currently has 13 peacekeeping missions. The United Nations also engages in peacebuilding activities for countries that have recently been engaged in conflict, supports sustainable peace and development, coordinates efforts to counter terrorism, and works to eliminate weapons of mass destruction and regulate conventional arms.

Roles of non-governmental organizations

Non-governmental organizations (NGOs) can be important partners for health professionals in preventing the health and environmental consequences of war. These civil society organizations can help by intervening to mitigate the consequences of war (such as by providing humanitarian assistance to victims of war), performing research on the health and environmental impacts of war, educating the general public and

policymakers, and advocating for policies and programmes to prevent war and promote peace.

Levels of prevention

The public health framework of primary, secondary, and tertiary prevention can be applied to the prevention of war and its health and environmental consequences. In this context, *primary prevention* refers to preventing war before it begins, *secondary prevention* refers to prevention and minimizing the consequences of war after it has begun, and *tertiary prevention* refers to treating or ameliorating the health and environmental consequence of war.

Health professionals can contribute to the primary prevention of war in a number of ways. As previously described, they can educate policymakers and the general public about the health and environmental consequences of war on civilian populations. They can help identify the presence of risk factors for war, including socioeconomic inequalities and other forms of social injustice (Levy 2019). They can promote disarmament and the reduction and elimination of weapons of mass destruction. They can promote international cooperation through partnerships with health professionals in other countries. And they can promote a culture of peace, in which conflicts and disputes are settled by non-violent means.

When war occurs, its consequences can also minimized by secondary prevention measures to limit injuries, illnesses, and deaths (especially among non-combatant civilians), by preventing damage to the environment, and by supporting a prompt end to hostilities. Secondary prevention measures include strengthening and ensuring compliance with the Geneva Conventions.

During war and its aftermath, health professionals can promote and support:

Protection of non-combatant civilians.
Prosecution of targeting of civilians as a war crime.
Protection of the health-supporting infrastructure of society.
Evaluation of non-combatant civilians who are at risk of illness, injury, or death.
Maintenance of medical neutrality.
Transformation of the public mindset about civilians during war, including demonstrating that non-combatant civilians are not enemy supporters or members of the enemy's military and addressing social and psychological forces that permit and encourage military forces to injure and kill civilians.

In the aftermath of war, many tertiary prevention measures are critically important. For example, it is necessary to rehabilitate and reintegrate people who have incurred disabilities due to the war; repair the damage to the physical, social, cultural, and economic environments; and take measures to prevent further violence. After conflict is over, health workers have important roles to play, such as advocating for prompt return and restoration of communities and livelihoods and opposing forced relocation or return to insecure environments. They can also help to restore essential health services, which may include recruiting and training health workers. They can help to address mental health problems that may be otherwise overlooked. They can help restore food and water supply systems that were disrupted during war. And they can also promote and support programmes for truth, justice, and reconciliation. These programmes may include processes for determining the truth about crimes committed during war, ensuring trials and punishment for these crimes, compensating victims for their losses, and establishing memorials to commemorate the victims of these crimes.

Host-agent-environment model

The traditional host-agent-environment framework for prevention in public health (Fig. 9.6.1) can be adapted as an approach to the prevention of war, with people representing the 'host', weapons and the military representing 'agent', and the conditions in which people live representing 'environment' (Fig. 9.6.2). This framework can enable development of war-prevention strategies, such as the following:

Strategies aimed at people, including promoting understanding and tolerance, economic and social interdependence among countries, nonviolent resolution of disputes and conflicts, and developing the popular and political will to prevent war and promote peace
Strategies aimed at weapons and the military, including controlling the international arms trade, eliminating weapons of mass destruction, reducing military expenditures, and intervening in disputes to prevent war
Strategies to improve the conditions in which people live, including protecting human rights and civil liberties, reducing poverty and economic disparities, improving education and employment opportunities, and ensuring personal security and legal protections.

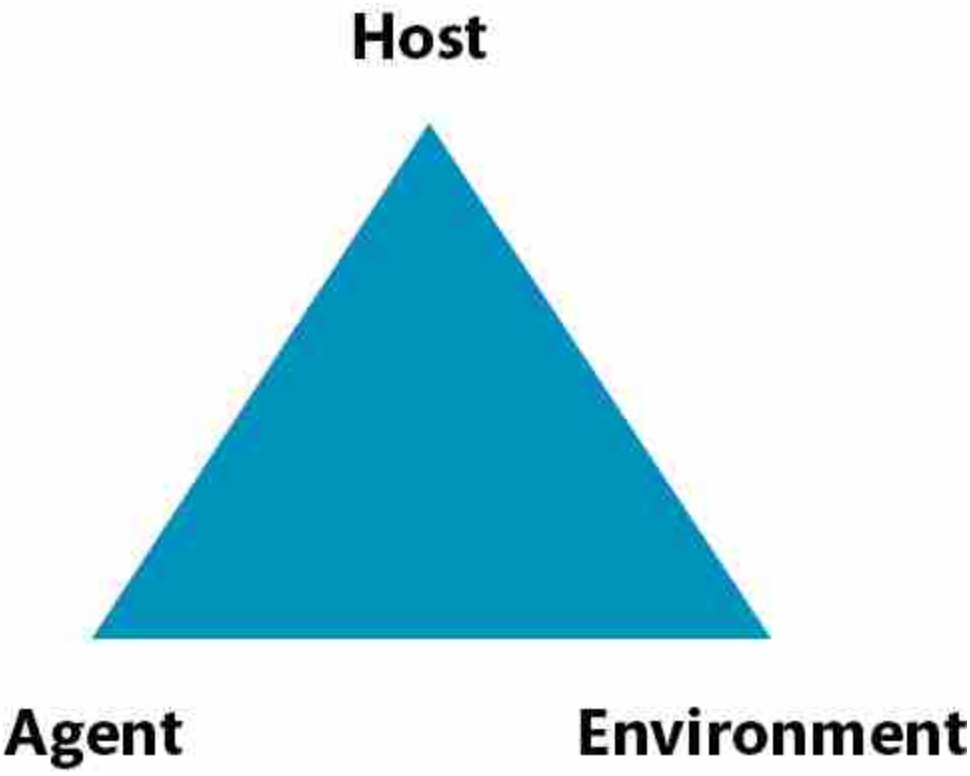


Fig. 9.6.1 Host-agent-environment model used to develop preventive measures in public health.

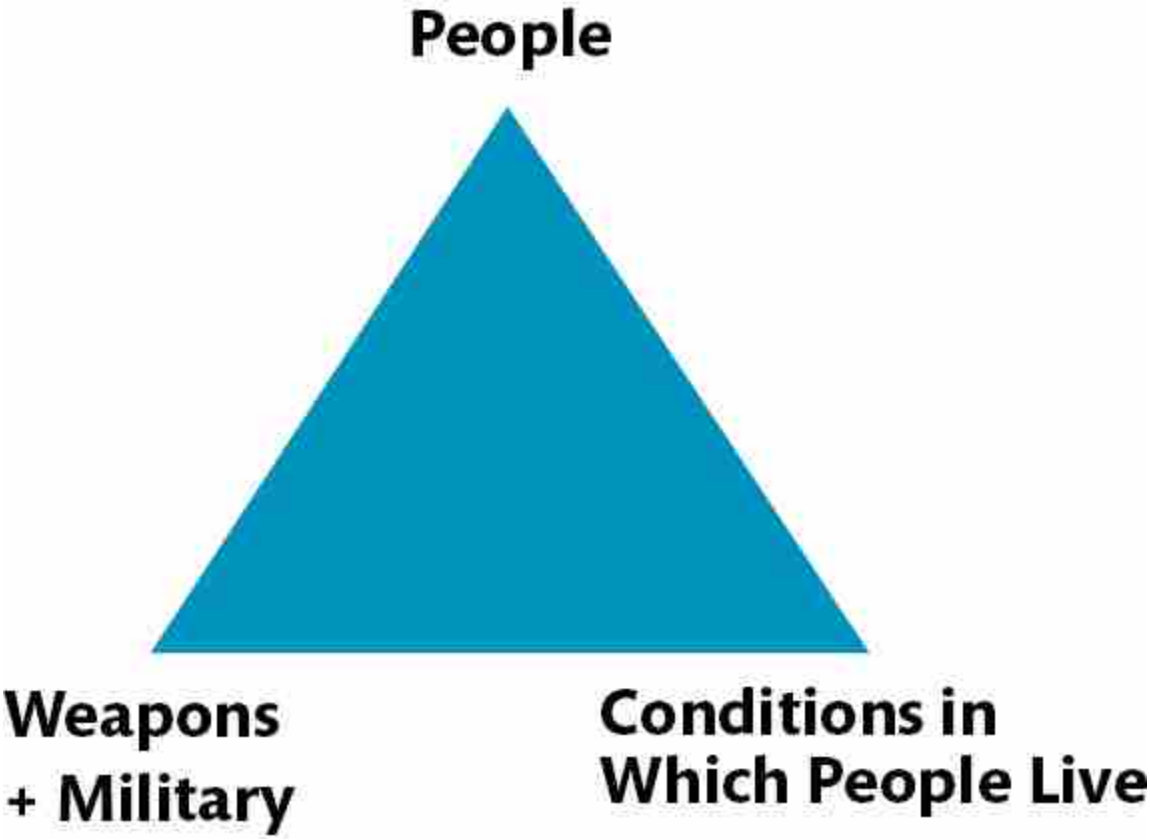


Fig. 9.6.2 Host-agent-environment model applied to the prevention of war.

The only way to prevent all of the health and environmental consequences of war, however, is to prevent war.

REFERENCES

Annan, K.A. (2000). *We the Peoples: The Role of the United Nations in the 21st Century*. New York: United Nations. [Online] Available at: http://www.un.org/en/events/pastevents/pdfs/We_The_Peoples.pdf

Arms Control Association. (2020). Timeline of Syrian chemical weapons activity, 2012-2020. [Online] Available at: https://www.armscontrol.org/factsheets/Timeline-of-Syrian-Chemical-Weapons-Activity.

Barras, V. and Greub, G. (2014). History of biological warfare and bioterrorism. *Clinical Microbiology and Infection*, **20**, 497–502.

Bennis, P., Leaver, E., and IPS Iraq Task Force. (2005). *The Iraq Quagmire: The Mounting Costs of War and the Case for Bringing Home the Troops.* Washington, DC: Foreign Policy in Focus.

Bešić L, Muhović I, Mrkulić F, Spahić, L., Omanović, A., and Kurtovic-Kozaric, A. (2018). Meta-analysis of depleted uranium levels in the Middle East region. *Journal of Environmental Radioactivity*, **192**, 67–74.

Blore, J.D., Sim, M.R., Forbes, A.B., Creamer, M.C., and Kelsall, H.L. (2015). Depression in Gulf War veterans: a systematic review and meta-analysis. *Psychological Medicine*, **45**, 1565–

80.
Bonde, J.P., et al. (2016). Risk of depressive disorder following disasters and military deployment: systematic review with meta-analysis. *The British Journal of Psychiatry*, **208**, 330–6.
Burmistrov D., Kossenko M., and Wilson R. (2006). *Radioactive Contamination of the Techa River and its Effects*. Cambridge, MA: Harvard University. [Online] Available at: http://wilsonweb.physics.harvard.edu/publications/pp747/techa_cor.htm
Carnegie Commission on Preventing Deadly Conflict (1998). *Preventing Deadly Conflict: Final Report*. New York: Carnegie Corporation of New York.
Centers for Disease Control and Prevention (2020). *Bioterrorism Agents/Diseases*. Centers for Disease Control and Prevention. [Online] Available at: https://emergency.cdc.gov/agent/agentlist-category.asp
Coghlan, B., et al. (2008). *Mortality in the Democratic Republic of Congo*. New York: International Rescue Committee and Melbourne; Australia: Burnet Institute.
Cole, L.A. (1988). *Clouds of Secrecy: The Army's Germ Warfare Tests Over Populated Areas*. Totowa, NJ: Rowman & Littlefield.
Eisenhower, D.D. (1953). '*The Chance for Peace*', *Address Delivered Before the American Society of Newspaper Editors*, 16 April. [Online] Available at: https://www.eisenhowerlibrary.gov/sites/default/files/file/chance_for_peace.pdf
Ferguson, N. (2006). *The War of the World: Twentieth-Century Conflict and the Descent of the West*. New York: Penguin Group.
Frost, A., Boyle, P., Autier, P., King, C., Zwijnenburg, W., Hewitson, D., and Sullivan, R. (2017). The effect of explosive remnants of war on global public health: a systematic mixed-studies review using narrative synthesis. *Lancet Public Health*, 2, e286–96.
Gleick, P.H. (2020). The water conflict chronology. [Online] Available at: https://www.worldwater.org/water-conflict/.
Harris, R. and Paxman, J. (1962). *A Higher Form of Killing: The Secret Story of Chemical and Biological Weapons*. New York: Hill and Wang.
International Campaign to Ban Landmines (n.d.). *Why the Ban.* [Online]. Available at: http://www.icbl.org
Krug, E.G., Dahlberg, L.L., Mercy, J.A., Zwi, A.B., and Lozano, R. (eds.) (2002). *World Report on Violence and Health*. Geneva: WHO. [Online] Available at: http://www.who.int/violence_injury_prevention/violence/world_report/en/full_en.pdf
Lee, E.C., Bleek, P.C., and Kales, S.N. (2012). Chemical weapons. In: Levy, B.S. and Sidel, V.W. (eds.) *Terrorism and Public Health: A Balanced Approach to Strengthening Systems and Protecting People* (2nd ed.). New York: Oxford University Press, pp. 183–202.
Levy, B.S. and Sidel, V.W. (eds.) (2008). *War and Public Health* (2nd ed.). New York: Oxford University Press.
Levy, B.S. and Sidel V.W. (2011). Water rights and water fights: preventing and resolving conflicts before they boil over. *American Journal of Public Health*, **101**, 778–80.
Levy, B.S. and Sidel, V.W. (2013). Adverse health consequences of the Iraq War. *The Lancet*, **381**, 949–58.
Levy, B.S. (ed.) (2019). Social Injustice and Public Health (3rd ed.). New York: Oxford University Press.
Levy, B.S. and Patz, J.A. (2015). Climate Change and Public Health. New York: Oxford University Press.
Meselson, M., et al. (1994). The Sverdlovsk anthrax outbreak of 1979. *Science*, **266**, 1202–8.
Power, S. (2002). '*A Problem from Hell': America and the Age of Genocide*. New York: Basic Books.

Rummel, R.J. (1994). *Death by Government: Genocide and Mass Murder Since 1900*. New Brunswick, NJ: Transaction Publications.

Sirkin, S., Cobey, J.C., and Stover, E. (2008). Landmines. In: Levy, B.S. and Sidel, V.W. (eds.) *War and Public Health* (2nd ed.). New York: Oxford University Press, pp. 102–16.

Song, S.J. and de Jong, Jr. (2015). Child soldiers: children associated with fighting forces. *Child and Adolescent Psychiatric Clinics of North America*, **24**, 765–75.

Stockholm International Peace Research Institute (2020). SIPRI Yearbook 2020: Armaments, Disarmament, and International Security. Oxford, UK: Oxford University Press.

Sutton, P.M. and Gould, R.M. (2008). Nuclear weapons. In: Levy, B.S. and Sidel, V.W. (eds.) *War and Public Health* (2nd ed.). New York: Oxford University Press, pp. 152–76.

Tirman, J. (2011). *The Death of Others: The Fate of Civilians in America's Wars*. New York: Oxford University Press.

United Nations Environmental Programme (2005). *Assessment of Environmental 'Hot Spots' in Iraq*. Nairobi, Kenya: UNEP.

Westing, A.H. (2008). The impact of war on the environment. In: Levy, B.S. and Sidel, V.W. (eds.) *War and Public Health* (2nd ed.). New York: Oxford University Press, pp. 69–84.

Wiist, W.H. and White, S.K. (2017). *Preventing War and Promoting Peace: A Guide for Health Professionals*. Cambridge: Cambridge University Press.

SECTION 10

Public health needs of population groups

10.1 The changing family
Ann Evans and Gavin W. Jones

10.2 Women, men, and health
Diane Cooper and Hanani Tabana

10.3 Child health
Tyler Vaivada, Amira Khan, Omar Irfan, and Zulfiqar A. Bhutta

10.4 Adolescent health
George Patton, Peter Azzopardi, Natasha Kaoma, Farnaz Sabet, and Susan Sawyer

10.5 Intersectional and social epidemiology approaches to understanding the influence of race, ethnicity, and caste on global public health
Jennifer Beard, Nafisa Halim, Salma M. Abdalla, and Sandro Galea

10.6 The health of Indigenous peoples
Papaarangi Reid, Donna Cormack, Sarah-Jane Paine, Rhys Jones, Elana Curtis, and Matire Harwood

10.7 People with disabilities
Anne Kavanagh, Marissa Shields, and Alex Devine

10.8 Health of older people
Samir K. Sinha and Brittany Ellis

10.9 Forced migrants and other displaced populations
Catherine R. Bateman Steel and Anthony B. Zwi

10.10 Prisoners: a wicked problem for public health
Tony G. Butler and Peter W. Schofield

10.1

The changing family

Ann Evans and Gavin W. Jones

Family change across the world

Kinship, family structure, and effect of socioeconomic and ideational change

> For the past thousand years, the Western family systems have been very different from those in China, India, Japan, and the Arab countries. … We cannot, then, view non-Western family systems as basically similar to the Western system at some undefined earlier historical phase just before industrialization. In fact, their differences extend for centuries back in time. (Goode 1963, p. 22)

Goode went on to propose that there would in future be a convergence towards some form of nuclear, or conjugal, family system. In this, he was mistaken (Cherlin and Furstenberg 1988, p. 578; McDonald 1992). His key point that family systems in different regions have to be studied in their own right, however, holds. ‘No useful analysis of African family systems can be undertaken using European concepts of marriage, household, parenthood, kinship’ (Oppong 1992, p. 82). There is also no doubt that almost everywhere, family systems are changing in response to pervasive and far-reaching technological, socioeconomic, and ideational forces. One of the best-known family sociologists in the United States recently wrote: ‘it is no exaggeration to say that the American, and more broadly the Western, family changed more dramatically in the latter half of the twentieth century than in any comparable span of time in our history’ (Furstenberg 2011, p. 192). Is the same generalization true for the rest of the world? Though conditions differ so greatly that it is difficult to generalize, Furstenberg‘s comment is certainly equally applicable to the East and Southeast Asian family, including China, and to the family in Latin America, though arguably not to the South Asian family.

What are some of the socioeconomic changes that have been impinging on family life? Firstly, the rapid pace of urbanization and rapid economic development in many parts of the world that were once characterized as ‘less developed countries’ has dramatically changed the options and the imperatives for family members. Latin America and East Asia are now heavily urbanized regions, and although urbanization is not as advanced in the rest of Asia or in Africa, even there some countries that were once predominantly rural, such as Indonesia, have passed the 50% urban population threshold. With more of the population drawn into the cities, and rising levels of education, particularly for women, the forces of globalization and neo-liberal economic structures have both drawn more women into the workforce and lessened job security. Then there have been remarkable technological developments in such fields as high-speed transportation, thus cutting travel times, and in social media, enabling a wider net of contacts to be maintained and providing instantaneous access to knowledge of all kinds. Technological developments in the medical field have widened the options in extending life and in avoiding conception and birth, options which bring with them not only benefits but also a range of ethical issues.

Across many developed countries in Europe, the United States and beyond these technological and economic changes have led to increasing social stratification, and the impact of disadvantage on family behaviour has changed. Cohabitation, once typified by increasing education and economic power of women (Lesthaeghe 2010), is becoming more widespread as young people navigate their way through young adulthood and settle on a marriage partner (Guzzo 2014; Sassler et al. 2018). There is now increasing evidence that marriage is not as accessible for the less educate and those with no or low skills (McLanahan 2004; Perelli-Harris and Lyons-Amos 2016; Manning et al. 2019). Indeed fertility within cohabiting relationships now signals a pattern of disadvantage where the least educated are more likely to have births within cohabitation than marriage (Perelli-Harris et al. 2010; Heard 2011).

In Asia, the world’s most populous continent, trends in family change differ greatly between subregions and countries. Common elements in most cases are rising levels of female education and labour force participation, fertility decline and a trend towards later age at childbearing (partly resulting from postponement of marriage), and rising rates of divorce. Both rising levels of education and economic (and advertising) pressures to maintain or

increase consumption worked to propel women into the labour force. Ideological changes led to increased demands for equality in the workplace and in the home. Improved contraception has both loosened the tight connection between sexual gratification and marriage and enabled married women to postpone childbearing.

Family structures differ markedly between Asian countries, but it is possible to detect general patterns which differentiate Southeast Asia from the other main regions of Asia. Rigidly patrilineal kinship systems characterize northwest India, China, and Korea, with patrilocal residence. In such systems, 'daughters were effectively lost to their parents when they married' (Das Gupta 2010, p. 128); women had no role as sister and daughter, but only as wife and mother. In the Indian system, there was resistance to enabling husbands and wives to be too emotionally attached, as this could threaten the unity of the patriarchal family (Caldwell et al. 1982). In contrast, the Southeast Asian region is characterized by bilateral kinship, which provides flexibility in post-marital residence, and continuing close relations between a woman and her natal family. In this region, there is a desire for children of both sexes; it is noteworthy that the only part of Southeast Asian that suffers from distorted sex ratios at birth due to misuse of pre-natal diagnostic techniques for non-medical purposes is northern Vietnam, which is the one area of Southeast Asia where most of the population does not follow a bilateral kinship system (Jones 2009, p. 13; Guilmoto 2012a). Southeast Asian kinship systems also facilitate the option of divorce if a marriage turns out unsatisfactorily.

These family systems have major health implications, for example through the distribution of resources within and across generations. How resources are allocated depends on relationships of authority within families. Those in the family whose status is relatively low (typically based on sex, age, or birth order) tend to receive fewer resources. In societies where there is a strong preference for male children, infant mortality is higher among girls than boys (for India, see Das Gupta 1987; for Egypt, see Yount 2003; for China, see Coale and Banister 1994; Croll 2000), though in recent years the son preference has been manifested more in sex-selective abortion, as a result of the ability to detect the sex of the foetus at an early stage (Guilmoto 2012b).

Accepted gender roles within families determine how much time women spend in the paid labour force as compared with childrearing and housework.

There are complex implications for resource allocation and health. In Western countries, working women's greater earnings and increased status have given them more control over resources, but have probably contributed to increases in divorce rates and in rates of non-marital childbearing, both of which have resulted in increased poverty among women and children. Also, the dual responsibilities of market work and childrearing and housework responsibilities (see Hochschild and Machung 1989) can take a toll on women's physical and mental health, and potentially have adverse effects on other members of the family.

Smaller families

The size and composition of households is a dynamic process, changing over the life cycle as children are born, grow up, and leave the family. Migration affects household composition, as does the decreasing prevalence of extended family forms where these were once more prominent. Smaller nuclear, or conjugal, families do not necessarily mean fewer family relationships; as in many parts of the world, including Western countries, families tend to be 'residentially nuclear but functionally extended', in the words of a Filipino sociologist (Castillo 1979), and engage in 'intimacy at a distance'. The web of contacts, in other words, can be quite dense although members of the extended family may be living far apart, even in different countries. Revolutionary developments in communications have made such contacts easier to maintain.

Household composition, of course, is not the same as family composition. Solitary living has increased enormously in Western societies. Whereas in 1950, just 10% of households in Europe were one-person households, five decades later, they accounted for 30% of British and 40% of Swedish households (Coontz 2004, p. 975). But solitary residence does not necessarily mean lack of family contact; nor are absent members necessarily uninvolved in family matters. Indeed, maintenance of the role of absent members—including those who live overseas for lengthy periods—both in household functioning and decision-making is increasingly observed and accepted as part of the reality of a globalizing world with mobile populations (see, for example, Parrenas 2008; Douglass 2011).

Whether we focus on the residential household or the family, however, the declines in fertility that have taken place almost universally except in parts

of sub-Saharan Africa, clearly make for smaller family size. As measured by the total fertility rate (TFR)—the number of children a woman would bear by the end of her reproductive period if she followed the currently prevailing age specific fertility rates throughout her childbearing ages – fertility over the past three decades remained very low in Europe, declining from 1.8 to 1.6; it declined from 2.6 to 1.6 in East Asia, from 4.6 to 2.4 in Southeast Asia, from 4.6 to 2.4 in South Asia and from 3.5 to 1.6 in Latin America. Only in sub-Saharan Africa did fertility remain really high, slightly below five children per woman in 2015. Fewer children being born into families directly lowers the number of family members living together, though more complex factors also affect family size. In particular, different family types —whether nuclear or some kind of extended family, have different implications for family size. Nuclear families are the norm in Western countries and in a wider range of non-Western countries than was often recognized.

Average household size has been falling almost everywhere. For example, in Indonesia, average household size decreased from 4.87 in 1971 to 4.27 in 1995 and 3.84 in 2010. In Thailand, it declined from 4.4 in 1980 to 3.8 in 1990, and to 3.1 in 2018 (Peek et al. 2015). Extended family contacts will, of course, widen the network of interaction, but not to anything like the extent in the past, when an average individual in Thailand had five siblings, six children, 30 nieces and nephews, 10 uncles and aunts, 60 first cousins, 120 first cousins of his parents and 720 second cousins (Mason et al. 1993, p. 70).

When fertility falls to very low levels, there are fewer close relatives for an individual to interact with. An extreme example was the one-child family in China, resulting from a policy that required over one-third of the population to have only one child, and over another half to have ‘1.5 children’ (i.e. those whose first child is a girl can have a second child); in other words, more than 60% of all Chinese couples were required to have only one child (Baochang 2009, pp. 76–80). Where each family has only one child, that child has no siblings, and while he/she may have a number of uncles and aunts, because the parent generation was not yet subject to the one-child family rule, there will be very few cousins. Once the second generation of one-child families emerges, the web of close relatives shrinks further. If two only children, who are themselves the children of only children, marry, between them they will have eight grandparents (living or

deceased) and four parents, but no cousins, aunts, or uncles at all. Hence extended family relationships disappear in this context. Not only this, but the family responsibility to care for the elderly falls squarely on the young couple.

Many countries have even lower fertility rates than China. Where fertility has fallen to really low levels—TFR below 1.5—half of all women have at most one child (Le Bras 1979). This does not necessarily mean that most children are growing up as only children. For one thing, children are born to a cohort of women whose average number of children is currently higher than suggested by the cross-sectional snapshot provided by period fertility rates. But this aside, many women are not having children at all. The proportion of women not marrying before reaching the end of their reproductive period is now 19% in Japan, around 14% in Singapore and Myanmar, well above 10% in South Korea and Thailand, and considerably higher than this for the tertiary educated in these and some other Asian countries. Having children outside of marriage is rare in these countries. Moreover, some of those who marry are not having children. In European countries, though having children outside of marriage is now common, an increasing proportion of adults are remaining childless; in Western Europe, around 10–20% of the cohort of women born in 1965 remained childless at the end of their reproductive period. However, the situation appears rather different when viewed from the perspective of the child; no children, of course, grow up in no-children families. Among women who have children, many are having just one, but many are having two or more. Taking into account childless women, this can still average out at just above one per woman. As Preston (1976, p. 105) and Le Bras (1979) noted, women with large families contribute disproportionately to the average family size of children.

Comparing the families of five or six children typical of earlier generations and those averaging two children, there are significant differences in the way children are raised and the social dynamics within the family. Typically, in large families, the older children play a role from a young age in assisting with the raising of their younger siblings, and perhaps later in financing their education. With low fertility, parents' attention is focused on just one or two children. This has positive implications for human capital; the investment of money and time by the parents will no doubt assist their one or two children achieve better results in school and in

career-related activities. The trade-off between quantity and quality of children has been resoundingly resolved in favour of quality in the East Asian countries (Montgomery et al. 2000), in Western countries and more recently in countries as varied in background as Brazil and Iran, and this is a major reason for the very low fertility they are now experiencing. The smaller cohorts of children should also enable government investment in education to achieve more per child.

What are the downsides to small family size? In China, both parents and the government fear the so-called 4-2-1 syndrome. This can be interpreted in two alternative ways: a great deal of attention focused on the one child by two parents and four grandparents, or as a weighty responsibility placed on the one child for the care of two parents and four grandparents as they age. When two only children marry, they will be responsible for four parents and eight grandparents. The young couple can only live with one set of parents, which means that the other half of their parent and grandparent families will be 'empty nesters'. What proportion of families is in this situation? It has certainly risen substantially, though exactly how much needs to be carefully studied. The issue is not unique to China. There are many single child families throughout East Asia, despite the absence of a one-child policy, though in the other East Asian countries ultra-low fertility is due to a much larger extent than in China to non-marriage, and therefore to a larger proportion of potential couples with no children at all (Frejka et al. 2010).

The issue of 'self-centredness' of only children is also frequently raised—the 'little emperor' syndrome. It is frequently speculated in China that divorce rates will rise with the clash of wills and personalities when two 'little emperors' marry each other. Inherently plausible as these speculations may be, they require evidence and research to back them up.

Delayed marriage and the rise of cohabitation

In Western countries, the 'golden age' of marriage from the 1950s to the 1970s (Festy 1980) has given way to an age of delayed marriage and increased cohabitation. Even though the rise of cohabitation has meant that delays in marriage do not accurately reflect the delay in forming relatively stable unions, overall, union formation and childbearing has been delayed (Mathews and Hamilton 2009). There are large differences between Western countries in the extent of the rise in cohabitation. Three broad groupings are

evident. Cohabitation is very common in the Nordic countries, France, the United Kingdom, and the Netherlands (where among those aged 20–34 who are either married or cohabiting, 40 to over 50% are cohabiting); there is an intermediate group of Germanic and Eastern European countries, the United States, and Australia, where the proportion is around 20–30%; and the Southern European countries and Poland, where cohabitation is less common (around 5 to 15%) (OECD Family database).

The United States differs from the countries of Europe in significant ways in matters affecting cohabitation patterns; for example, Americans are more religious than other Westerners, though this religiosity is adapted to a culture of expressive individualism (Cherlin 2009, pp. 103–4; p. 113). Americans begin to have partners (whether married or cohabiting) at a relatively younger age than most Europeans, and to become parents at a younger age. Marriages and cohabiting relationships in the United States, and to some extent in Great Britain (Kiernan 2003, p. 32; Beaujouan and Bhrolcháin 2011) are far more fragile than in most of Europe. American women are much more likely to spend time as lone parents in their teens or twenties than are women in Western Europe (Cherlin 2009, pp. 16–18). Overall, in the United States, about 41% of births were non-marital in 2010—in other words being borne by single mothers or those in cohabiting relationships (Martin et al. 2012). The widening interval between sexual debut and marriage has exposed many more teenagers to the risk of pregnancy and childbirth outside marriage (Furstenberg 1990, p. 150), although the rate of teenage childbearing has declined considerably over the past two decades. In developing countries, very early marriage has long characterized most of sub-Saharan Africa and South Asia, as well as parts of Southeast Asia, the Caribbean, and Central America. In recent decades, average age at marriage has risen almost universally, but substantial proportions of women are still marrying very early. The latest figures show that globally, 21% of young women married before their 18th birthday, the highest incidence being 38% in sub-Saharan Africa and 30% in South Asia (UNICEF 2018).

Unlike Western countries or Latin America, cohabitation does not feature importantly in Asian patterns of partnering, though it is becoming more common in Japan, China, and the Philippines and probably elsewhere. Asia does, however, show remarkable diversity in marriage systems and marriage patterns. While in general, average age at marriage is rising for both males and females throughout Asia, this rise is from a base that varies greatly. It is

in East and Southeast Asia that truly remarkable changes in marriage patterns have taken place. In recent decades, the traditional universal marriage norms have been abandoned and the average age at marriage and proportions not marrying at all have risen sharply (Fig. 10.1.1; for more detail, see Jones 2007; Jones and Gubhaju 2012). However, there are wide differences between countries, and even between different ethnic groups within one country, such as Malays, Chinese, and Indians in Malaysia.

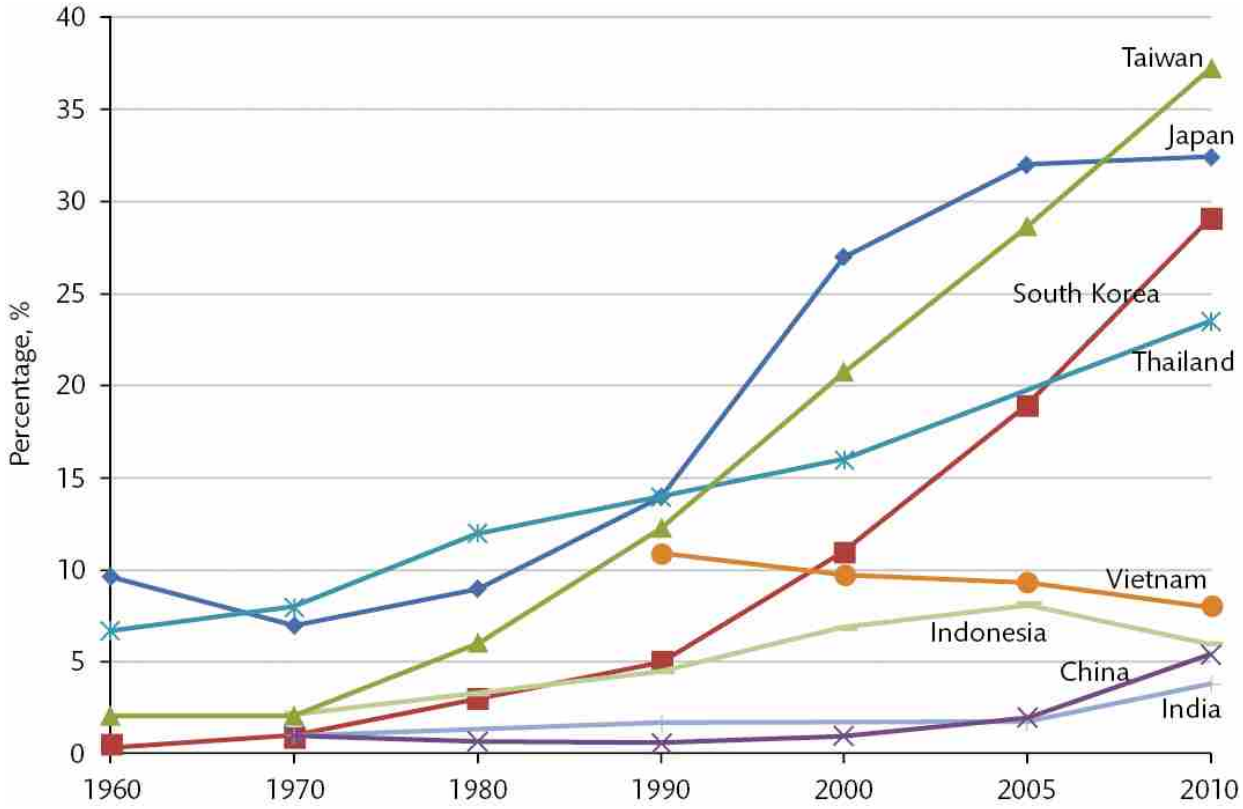


Fig. 10.1.1 Proportion of single among females aged 30–34, 1960–2010.
Source: data compiled by Gavin W. Jones from population census reports for various years for the countries included.

Broadly comparing East Asia, Southeast Asia, and South Asia, the changes in Southeast Asia as a whole are not as great as those in East Asia (excluding China), while most of South Asia remains firmly locked in an entirely different pattern of universal, arranged marriage. South Asian marriage systems have turned out to be remarkably enduring. While in East and Southeast Asia the location of most marriages on the continuum from

parent-arranged to self-arranged marriage has shifted decisively towards the self-arranged end, very little such change has occurred in South Asia.

Socioeconomic change is undoubtedly responsible for much of the change in marriage patterns, but it is hard to explain the very high singlehood in Myanmar and the lack of change in Vietnam in these terms. Within East Asia, Chinese marriage patterns differ widely from elsewhere; though average age at marriage for women is rising in China, marriage is 'concertinaed' at ages in the 20s, and by age 30, only a tiny proportion of Chinese women remain single. By contrast, in Japan, Korea, Thailand, Myanmar, Singapore, and among the Chinese population of Malaysia, delays in marriage have been such that substantial proportions of women remain single in their late 30s, an age beyond which not much childbearing takes place. Indonesia has seen gradually rising age at marriage but this has reversed according to the 2010 Population Census.

Turning attention to South Asia and South-West Asia, The Islamic Republic of Iran shows a strong shift towards a delayed marriage pattern, with the proportion of women single at ages 35–39 not much below that in the Republic of Korea. However, in the Indian subcontinent, the relatively tardy change in arranged marriage patterns in the face of very considerable socioeconomic change is striking. Evidence from a large survey in India (the India Human Development Survey 2005, IHDS) indicates that

> Marriage decisions remain within the purview of the family, and less than 5% of the IHDS respondents had a primary role in choosing their husbands. However, nearly 62% of the respondents felt that they were consulted in this decision. (Desai and Andrist 2010, p. 675)

'Consultation' clearly did not go to the extent of arranging for the spouses-to-be to meet; among the respondents who felt they had a choice in partner selection, around 55% met their husbands on or around the day of the wedding. The persistence of early and arranged marriage in India is seen by Desai and Andrist (2010, p. 681) as 'part of a script in which gender is performed by women through a symbolic display of segregation, modesty and chastity, and early marriage is part and parcel of culture in action'. However, rising education levels in South Asia have led to an increase in involvement of brides in the selection of their partners, with parents seeking appropriate matches with the involvement of and consent from the bride and groom (Yeung et al. 2018).

In Latin America, as in Europe, the key trend in union formation is the rise in cohabitation, which appears to have been just as marked as in Europe,

though from a more complex starting point. Latin America was marked by patterns of ‘old cohabitation’ (characterizing countries with mixed indio, mestizo, and European populations, and those with a history of slavery) and ‘new cohabitation’ (characterizing countries with more dominant European populations, notably Argentina, Chile, Brazil, Mexico, and Colombia). Cohabitation has increased sharply throughout Latin America in recent decades, though to varying degrees in different countries (Fussell and Palloni 2004; Esteve et al. 2012), and the long-standing educational gradient, whereby the better educated cohabited less and married more, has been attenuated. The factors lying behind these trends are complex, but ideational change in the direction of greater secularization and tolerance of non-conformist behaviour appears to be strongly implicated, except perhaps in Brazil (Esteve et al. 2012).

Late marriage, delayed transition to adulthood, and the role of singles

A trend affecting most of the world is the tendency for children to be dependents for a longer period of time, and to remain in the natal home until a later age. In agrarian and early industrial societies, children frequently began to work at quite a young age. Though age at marriage was young in many such societies, young marrieds frequently continued to live in the home of one or other set of parents until they had the resources to form their own household. The rise of universal schooling tended to delay both entry into the workforce and marriage, to a marked degree in countries where most children continue through secondary education and substantial proportions pursue tertiary education.

In the United States, 50 years ago, the vast majority of women and most men had completed schooling, left home, entered the workforce and begun families of their own by the time they reached their early 20s (Furstenberg 2010). These transitions have now been considerably delayed (Arnett 2004). Likewise, in much of Asia, extended education and delayed marriage are leading to later home leaving and in general a delay in reaching the milestones that signify adulthood and independence (Xenos et al. 2006). If adolescence is defined as the period between attaining sexual maturity and marriage, then it is a lengthening period in most young Asians’ lives, though

in South Asia, with the exception of Sri Lanka and to some extent Pakistan, early marriage remains the norm. In most of Asia, unlike in most Western countries (though the 'family-centric' Mediterranean countries share characteristics with Asian countries in this respect), there is a continued tradition of living at home until married, and in many countries, the need to do so is strongly reinforced by the high cost of owning or renting housing, and the convenience of living at home without paying board (on Japan, see Retherford and Ogawa 2006, p. 17–19; on Singapore, see Jones 2012). Therefore, young people are living longer at home, in an unmarried state.

Many issues have to be negotiated between parents and children in this evolving situation. In East and Southeast Asian countries, only in the last few decades has extended singlehood become really common, so those who are remaining single into their 30s are a pioneering group, whose roles in family and society are still being negotiated (Maeda 2008; Situmorang 2007, 2011; Tan 2010; Ibrahim and Hassan 2009). Japan, Thailand, Singapore, and Myanmar are somewhat different in that over a fairly long period, substantial proportions of men and women have been remaining single, and this has led to adjustments of societal expectations, such that most singles do not feel much pressure to marry from family and friends. Little is known about the role of such singles beyond their immediate family, for example, whether working singles are contributing financially to education of cousins, nephews, and nieces.

In societies where girls were formerly married off on reaching puberty, or shortly thereafter, the transition to adulthood was abrupt and, for many, traumatic. This remains the case in much of Africa and South Asia. But elsewhere in Asia, delayed marriage has led to an extended period of adolescence between puberty and marriage, raising issues about sexuality and relationships with the opposite sex that simply did not arise in times of early, parent-arranged marriage (Lloyd et al. 2005). In Indonesia, for example, extended education for girls, long past the age at which their grandmothers were already married and having children, and widened employment opportunities, has brought them out of the household and into daily contact, not only with a range of men, but also with the stimulus of sexually explicit material through movies, videos, magazines, books, and the internet. Premarital sexual activity is taken for granted in many of these information sources, but the material is designed to stimulate and titillate rather than to educate. Very little sex education is provided in schools or

other contexts (Utomo and McDonald 2009, p. 137). At the same time, influences deriving from religion and ethnicity are still strong, and in some ways growing stronger. ‘Faced with the conservative forces of idealized morality portrayed through religion, on the one hand, and Western influences promoting a more liberal approach on the other, it is little wonder that many teenagers are confused’ (Jones 2002, p. 232).

There are other aspects of delayed marriage to keep in mind. One has to do with the maturity of the marriage partners. Five decades ago, when half of Malay girls in Malaysia were married before they reached age 17, they were unlikely to be emotionally mature at the time they married. But at least co-residence with one or other set of parents—most commonly the bride’s parents—meant a period of ‘apprenticeship’ into many aspects of married life. Age at marriage has now increased dramatically, and the age at which Malay women are bearing their first child is four years older than for their mother’s generation and seven years older than their grandmother’s generation; some of them are having their first child in their mid to late 30s. There are a number of possible implications of later childbearing in both Asian and Western countries. Women are no doubt emotionally more mature when they are bearing their children, but they and their husbands may be more settled in ‘child-free’ ways, requiring a greater adjustment when the first child comes along than for couples who married at a younger age. The window of opportunity for having a second child is also very small when childbearing is postponed to these ages. Furthermore, there is a somewhat greater risk of birth abnormalities occurring when the mother and/or father are older. The greater availability of *in vitro* fertilization (IVF) and assisted reproductive technologies enables more older mothers to have babies, but the risks of multiple pregnancies and premature births could contribute to chronic health problems in the children.

Divorce

In the second half of the twentieth century, divorce rates were on the whole higher in Western countries, particularly the United States and Russia, than in most of Asia (Dommaraju and Jones 2012). In South Asian countries, divorce is a rare event; divorce is not a feasible way out of a disharmonious marriage, because the options of returning to the natal family, or of earning an independent income, are not generally open to a divorced woman.

However, the gap between Western and East Asian countries has narrowed because divorce rates stopped rising in many Western countries around the 1980s, while in East Asian countries, divorce rates have been rising steadily since 1980, particularly in the period since 1990 in the case of Japan, Republic of Korea, and Hong Kong. The rise in Republic of Korea was particularly sharp between 1995 and 2000—the period in which the Asian financial crisis occurred. In China, freeing up of divorce regulations in 2001 and 2003 may have contributed to the sharp rise in divorce rates in the 2000–2005 period. In Taiwan, as in other East Asian societies, increase in divorce has taken place in a context of decreased prevalence of marriage. The chance of re-marriage after divorce has declined (Yang and Tsai 2009), indicating a growing acceptance of and perhaps preference for singlehood.

Japanese divorce rates have risen to such an extent that Japan can no longer be considered as a society characterized by low levels of marital dissolution. About one-third of Japanese marriages are now likely to end in divorce (Raymo et al. 2004). Traditional marriage stability in Japan, while it did not necessarily have much to do with whether a marriage was happy, at least gave some assurance that children would be cared for, and that wives could expect economic security. With the awareness of increasing fragility of marriage, decisions about whether to marry and about whether to raise children when single motherhood is a possibility, come more into play (Bumpass et al. 2009, p. 223).

Trends in divorce in East Asia appear, then, to be going the way of the West, and partly for the same reasons (Dommaraju and Jones 2012). Substantial increases in divorce rates in East Asian countries signify a significant change in circumstances and attitudes to divorce, because in the past divorce carried a considerable stigma, and the pressure to remain in a disharmonious marriage for ‘the sake of the children’ and also for the sake of appearances and family honour, was very strong. It would appear that factors such as increasing economic independence of women, and the pressures (and widened range of personal contacts) in the big city environments in which an increasing proportion of East Asians live, are influencing divorce trends.

In Western countries, divorce rates no longer reflect very accurately trends in the breakdown of long-term relationships, because of the rise of cohabitation. It is impossible for couples who are not officially married to divorce. Understanding the relationship of cohabitation to breakdown of

unions requires analysis of the proportion of cohabiting relationships that are dissolved; the proportion that are converted to marriages; and the proportion of such marriages that are dissolved. In Europe, around one-third of cohabiting unions are dissolved within five years (Kiernan 2003), considerably higher than the proportion of marriages that are dissolved within the same length of time; of cohabiting relationships that are converted to marriages, divorce rates also tend to be higher than for marriages that did not commence with a cohabiting relationship, though 'when the pre-existing characteristics of those who cohabit are taken into account, the risk of marriage breakdown is no different from and may be lower among those cohabiting before marriage than among those who marry directly' (Beaujouan and Bhrolcháin 2011, p. 19). In any case, official figures on divorce trends clearly do not fully reflect the breakdown of long-term relationships, many of which, both marital and cohabiting relationships, involve children. It might be noted that in Japan, while divorce rates do not fully reflect relationship breakdown because cohabitation is increasing, the breakdown of cohabiting relationships rarely involves children, because very few children are born to cohabiting couples.

Non-marital childbearing, divorce, and effects on children

In Europe, 'sexual activity, childbearing and marriage have become disconnected so that increasingly it is socially acceptable to have one without either of the other two. But this has happened in different ways in different parts of Europe' (Pritchett and Viarengo 2012, p. 67). Notably, in northwest Europe, childbearing remains reasonably robust, much of it in non-marital relationships; in many countries of Eastern and southern Europe, though, marriage rates remain high and extramarital births are rare but fertility is very low.

Changing conventions about marriage and divorce in the West and in East Asian countries have multiple causes. 'Throughout Western Europe and North America family relations have become more discretionary; a greater premium is placed on emotional gratification and a lower value on obligation and authority' (Furstenberg 1990, p. 153). Such factors, stressed in 'second demographic transition' theory (van de Kaa 1987; Lesthaeghe 2010) have explanatory power for Western countries and perhaps for Japan; for East Asia more generally, however, more important than individualism

may be women's willingness to break from the shackles of patriarchal structures in a context of widened economic opportunities. Moreover, some weakening in these countries of community perceptions of divorce as a disgraceful event has undoubtedly lowered the bar to the decision to divorce.

In all these contexts, however, the effect of divorce on children needs to be considered. As Carlson and England (2011, p. 2) put it, 'The freedom to leave unhappy relationships might be counted as a victory for adults, but the same cannot be said for children'. Multiple studies suggest that 'children of divorce are at higher risk for experiencing a variety of developmental problems during late childhood and early adolescence, including school failure, delinquency, precocious sexual behaviour, and substance abuse' (Furstenberg 1990, p. 161–2). Nevertheless, the effect of divorce on children is a complex subject. Many studies suffer from methodological shortcomings, and assessments differ, particularly with regard to whether family conflict produces more adverse effects on children than divorce itself (Cherlin 2009, pp. 19–24; 192–7).

The effect of marital disruption (and more generally, in the era of cohabitation, relationship disruption) on children is a controversial and emotional issue. Many studies show that children raised by solo mothers or by a mother in unstable relationships are more likely to experience emotional and behavioural problems than children growing up in stable married families. For example, in Britain, 'children born into cohabiting families where the parents had separated and solo mothers who had not married the natural father … were around three times more likely than children in stable married families to be exhibiting behavioural problems' (Kiernan and Mensah 2009). However, careful control for factors lying behind the differences in the family situation in which children grow up is necessary before attributing the behaviour to the family situation itself. In the British study under discussion, 'Taking account of the mental well-being of mothers and family poverty attenuated the differences and the introduction of additional controls attenuated them still further. However, there are still indications that children born to cohabiting parents who separate and children of solo mothers who cohabit with the natural father or partner another man are more likely to be exhibiting behavioural problems.' (Kiernan and Mensah 2009). On the other hand, when the child's emotional well-being was the dependent variable, it appeared that the parental socioeconomic circumstance and mother's mental well-being may fully

account for differences between children living in different family trajectories.

Similarly, when considering the health and well-being of adults, family relationships have been found to play a role. Research has found that marriage offers health benefits to men and women through boosting economic resources (Waite and Gallagher 2002), social networks and support (Umberson et al. 2010; Umberson and Montez 2010). But some of this difference may in fact be due to different levels of access to partnerships and marriage due to health and socioeconomic circumstances evident through childhood (Perelli-Harris et al. 2018; Perelli-Harris et al. 2019).

In sub-Saharan Africa, patterns whereby many fathers do not continue to live with their children are common (Adepoju and Mbugua 1997; Therborn 2004, pp. 177–81). In parts of Africa characterized by polygyny, the mother–son bond tends to be strong. In South Africa, among the African population, in 2007 only 29% of children aged 17 years or younger were living with both parents, while 43% were living with mother only and 26% were living with neither parent (Republic of South Africa 2013). Almost half of children whose father was alive had no contact with him; unfortunately, in South Africa, '… family instability in childhood plays a major role in setting young people on pathways to adulthood that potentially compromise their life chances'.

Ageing and the family

The trend towards an older population is an inexorable one in most parts of the world. Japan has the oldest age structure of any country in the world (26% of the population aged 65+ in 2015), followed by Europe as a whole (17.6%). Some Asian countries—Republic of Korea, China, Singapore, and Thailand—are well along in the ageing process, with between 10% and 13% elderly. But many of the higher-fertility countries in Africa and South Asia have only around three or four% in the over-65 age groups. Everywhere, however, the percentage of elderly is rising, and by 2030 it is expected to reach 30% in Japan, 23% in Europe, 17% in China, and 19% in Thailand.

There is little doubt that issues related to population ageing will be central, not only to national policy reactions to ultra-low fertility, but also to family level adjustments needed to cope with radically altered population structures. At an aggregate level, the issues can be well illustrated by trends

in ageing support ratios—i.e. the ratio of the population ages 20–64 to the population aged 65+. Admittedly, the age limits used for this calculation do not accurately reflect the actual dependency situation, which depends on the ages at which people actually enter and withdraw from the labour force, and on the proportion of the elderly requiring various kinds of support (in the case of physical support, usually quite small). But as a general indicator, the ageing support ratio does give a sense of the trends which are so dramatic that modification of these ages will not alter the picture greatly.

The ageing support ratios were vastly different in different world regions in 1980 (see Fig. 10.1.2). Since then, they have fallen sharply in East Asia and South America, and this sharp decline is expected to continue until 2050. South Asia is 35 years behind East Asia in reaching an ageing support ratio as low as 8. In the 40 years between 2010 and 2050, the ageing support ratio will halve in Japan, and will fall to less than one-third its current level in countries such as China, Thailand, and Brazil. Leaving aside Africa, where the decline in the ageing support ratio will be very slow, and South Asia, where it will be relatively slow, there will be a marked convergence in ageing support ratios in the different regions at very low levels—around 2 to 3. Such declines will require major adjustments of policies with regard to income support programmes, orientation of health services to give much more weight to gerontology, and labour market policy, including retirement ages.

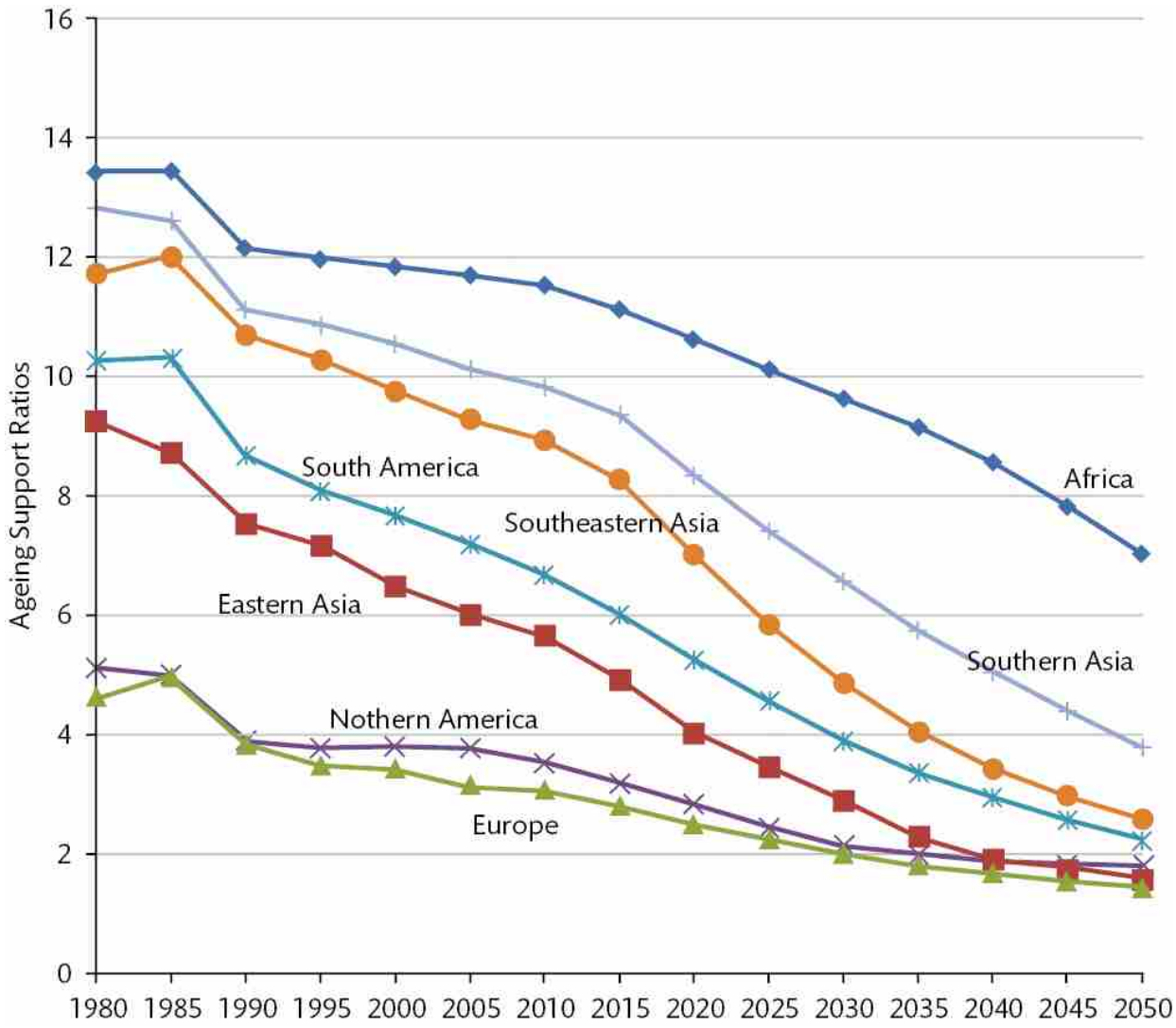


Fig. 10.1.2 Ageing support ratios (population aged 20–64/65+) by region.
Source: data from the detailed indicators in United Nations Population Division, World Population Prospects: The 2012 Revision, Copyright © 2014 United Nations, available from http://esa.un.org/unpd/wpp/unpp/panel_indicators.htm.

It should be noted that living arrangements of the elderly differ considerably in Asia from those in Western countries, with a much higher proportion living with a child or grandchild (United Nations 2005). And although the percent of elderly Japanese and Koreans living with a child or grandchild has been gradually declining, it does not appear likely to reach levels as low as in France or the United States. Demographic trends do, however, raise issues about whether the family can be relied on as in the past to provide care for those elderly who need it. There are a number of aspects to the problem. The first is the sharply declining ratio of potential carers to

numbers of elderly, as crudely indicated by trends in the support ratios. The number of living children of older persons has been declining through movement of smaller birth cohorts into the main 'caring ages'. And as documented by Schroder-Butterfill and Kreager (2005) in a study in East Java, *de facto* childlessness may differentially affect the poorest and most vulnerable groups. The second is that a high proportion of the potential carers (usually women) are now drawn into the paid workforce (consistent with government policy in these countries), and are therefore not available for full-time caring work. Moreover, many of these potential carers are in the so-called 'sandwich generation', required to support not only the grandparent generation but also their own children who are spending a longer time living with (and depending on) their parents. The third is that, despite the general improvement in health status of the elderly at any given age, the rising proportion of the elderly, and the rising proportion of the elderly who are 'old-old' (i.e. 75 years and above) will inevitably mean a rising load of frail and disabled elderly, who need full-time care. The ratio of elderly population who suffer from senile dementia to non-working women at various ages is projected to rise substantially over the coming two decades in Japan (Ogawa 2006). This suggests that Japan will not be able to rely on the traditional family support network to care for the infirm elderly; the rise in the burden placed on middle-aged Japanese women providing in-home nursing care will simply be too great.

There will be less possibility of the family providing intensive care for the increasing number of disabled or frail elderly in Asian countries, for three reasons: the lowered proportion of children per elderly person, the increased labour force participation of women (the traditional carers), and the movement of children of the elderly to other localities. These trends have major implications for policy on elder care. Despite the tradition in Asian families of caring for family members in various kinds of need, and official rhetoric about the key support role of the Asian family as a fundamental (and by implication, unchanging) characteristic of Asian societies, this traditional reliance on family will be subjected to intensive pressures. Governments will almost inevitably be forced to step in.

We need to bear in mind that co-residence is not the only indicator of close caring relationships. Assistance can notionally be divided into three categories: financial assistance, emotional support, and physical support. The first two of these do not require co-residence, particularly in the age of

the cell-phone. In Thailand, there has been an extraordinary rise in the proportion of the elderly living in houses with a telephone—from 15% in 1994 to 76% in 2007. Moreover, while the proportion of elderly living in the same house as a child is slowly declining in Thailand, the proportion is considerably higher if those living in close proximity are included (Knodel and Chayovan 2008, p. 41).

The care needs of the elderly should not be exaggerated. The great majority of the elderly do not need care; for example, in Thailand, more than 80% of those in their 70s and even 65% of those aged over 80, say that they can care for themselves (Knodel and Chayovan 2011). The positive contributions made by the elderly in society and family, for example through taking care of grandchildren and through serving as mentors and confidants of their grandchildren as they enter adolescence need to be recognized and supported (Tengku-Aizan and Nurizan 2008; Ananta and Ariffin 2009).

The challenge, of course, in harnessing the capacities of grandparents to develop a special relationship with their grandchildren, is to avoid burdening them with what are sometimes exploitative child-minding arrangements. In all societies, the role of grandparents in childcare appears to be greatly valued, as opposed to the alternatives of formal childcare arrangements or, in some countries, reliance on domestic servants. ‘Grandparent care is commonly viewed as the ‘next best thing’ to parental care, and is highly valued for its supposed intrinsic benefits: the fostering of closer intergenerational relationships, the quality of grandparent care in contrast with the formal market, and the satisfaction that grandparents apparently derive from the caregiving role’ (Jenkins 2010, p. 6; see also Wheelock and Jones 2002). However, grandparents are not always so positive about the role, which can be very stressful, especially when they feel that they lack the choice to determine the parameters of their caregiving.

The family, development, and health

Being born into the wrong family can almost guarantee a difficult life. ‘The family has long been a mechanism by which advantage—and disadvantage—is transferred across generations.’ (Carlson and England 2011, p. 9). In the American context, Lareau and Cox argue that class differences in parents’ cultural resources (e.g. knowledge necessary to help their children interact with institutions—such as gaining access to college-prep courses in high

school and navigating the college admission process) serve to exacerbate inequality in the life chances of youth from different classes, especially in getting a college education. (see Carlson and England 2011, p. 14). In Indonesia, only 2% of the children of the poorest 20% of the population continue to tertiary education, compared with 40% of children of the richest 20%.

From a public health perspective, issues abound in relation to conventional family structures as well as to rapid changes in these structures. Many examples could be given, but one will be chosen here: reproductive health issues arising from two contrasting situations—patterns of early marriage and childbearing in Africa and South Asia, and delays in marriage in other developing countries.

In many sub-Saharan African countries, as well as in Bangladesh and parts of India, data from the 1980s and 1990s showed that more than 10% of women in their early 20s had had a child by age 15, and between a quarter and a half had a child by age 18 (Singh 1998). In conditions of high mortality, 'socialization processes that support and encourage early marriage and childbearing are often deeply entrenched' (Singh 1998, p. 133), and these socialization processes do not change rapidly in the face of declining mortality. The dangers of early childbearing to the health of mother and child are well known. But other dangers emerge when marriage is delayed and young people are exposed to the risks of premarital intercourse for longer periods of time: a greater risk of unintended pregnancies, induced abortion and sexually transmitted infections (STIs). Increased levels of infection by STIs and HIV can affect young women's health and indeed survival, the outcome of their pregnancies, as well as their future fertility. Yet policies in many countries make it difficult for the unmarried to obtain access to contraception.

The family is seen in most countries as the key institution on which a stable society is built. Thus changes in family form and function are normally suspected of threatening the social order. One concern tends to be about ageing, and whether the family will continue to have the capacity to provide the necessary physical, social, and emotional support for the elderly as their share of the population increases. Delayed marriage also leads to disquiet, though not, in most countries, to specific policies to encourage marriage. But it raises important issues about the role of singles in society, as

well as practical issues such as their access to contraception, which proves to be divisive in many countries.

Family policy in many countries tends to be conservative in promoting the generally accepted family values. The model of universal heterosexual marriage in which children are raised in an intact family until they marry and leave home is the norm. In Asian countries, this of course remains overwhelmingly the main family form, and there is a great deal to be done to enhance the quality of life for mainstream families of this kind. One key need is to develop policies that will facilitate marriage and enable women and men to combine work with childrearing—to avoid the low fertility trap that East Asian countries have perhaps entered into, but also to enable women to realize the potential resulting from their higher levels of education, without succumbing to the exhaustion of the 'double shift'. At the same time, in countries such as India, Bangladesh, and Indonesia, stricter enforcement of legal minimum ages at marriage is needed, to lower the proportion of girls being married below this age.

But there are also other family forms in all societies, including childless couples, those cohabiting without entering formal marriage, blended families following re-marriage, the various kinds of lesbian/gay/bisexual/transgender (LGBT) pairings (Blasius 2001
), single parents post-divorce raising a child or children, and singles continuing to live with parents until into their 30s. The reality of considerable pluralism in family types and functions needs to be recognized. It is important that those who are not in mainstream family relationships are not made to feel ostracized and discriminated against by policy.

On elder care, increased strain will be placed on families whose characteristics and circumstances are changing in ways that will make it increasingly harder to deal with these strains. For example, in India, the percentage of elderly people is expected to double in 27 years after 2015, to 11.2%. In Thailand, this percentage is expected to increase by 2.5 times over the same period, to 26.7%. Although the percentage of elderly in India in 2042 will be less than half that in Thailand and less than a third that in Japan, even the smaller increase India faces poses significant challenges. The policy emphasis in all countries must be on active ageing, financial independence, and a positive attitude towards the contribution the elderly make within families. Nonetheless, an increasing role of government in elder care cannot be avoided. The retirement age in many countries also needs to

be increased, as it reflects the life expectancy situation decades ago, and will be increasingly untenable in terms of providing income maintenance for the rapidly increasing number of retirees. The extent to which government needs to replace the role of family in elder care, and its capability of doing so, will vary across regions.

REFERENCES

Adepoju, A. and Mbugua, W. (1997). The African family: an overview of changing forms. In: Adepoju, A. (ed.) *Family, Population and Development in Africa*. London: Zed Books.

Ananta, A. and Nurvidya Ariffin, E. (2009). *Older Persons in Southeast Asia: An Emerging Asset*. Singapore: Institute of Southeast Asian Studies.

Arnett, J.J. (2004). *Emerging Adulthood: The Winding Road from the Late Teens through the Twenties*. New York: Oxford University Press.

Baochang, G. (2009). The arrival of low fertility in China. In: Jones, G., Straughan, P.T., and Chan, A. (eds), *Ultra-low Fertility in Pacific Asia: Trends, Causes and Policy Issues*, London: Routledge.

Beaujouan, É. and Bhrolcháin, M.N. (2011). Cohabitation and marriage in Britain since the 1970s. *Population Trends* No. 145. London: Office for National Statistics.

Blasius, M. (2001). *Sexual Identities, Queer Politics: Lesbian, Gay, Bisexual and Transgender Politics*. Princeton, NJ: Princeton University Press.

Bumpass, L.L., Rindfuss, R.R., Choe, M.K., and Tsuya, N.O. (2009). The institutional context of low fertility: the case of Japan. *Asian Population Studies*, 5(3), 215–35.

Caldwell, J.C., Reddy, P.H., and Caldwell, P. (1982). The causes of demographic change in rural South India: a micro-approach. *Population and Development Review*, 8(4), 689–728.

Carlson, M.J. and England, P. (eds.) (2011). *Social Class and Changing Families in an Unequal America*. Stanford, CA: Stanford University Press.

Castillo, G. (1979). *Beyond Manila: Philippine Rural Problems in Perspective*. Ottawa: International Development Research Center.

Cherlin, A.J. (2009). *The Marriage-Go-Round: The State of Marriage and the Family in America Today*. New York: Alfred A. Knopf.

Cherlin, A. and Furstenberg, F.F. (1988). The changing European family: lessons for the American reader. *Journal of Family Issues*, 9(3), 291–7.

Coale, A.J. and Banister, J. (1994). Five decades of missing females in China. *Demography*, 31(3), 459–79.

Coontz, S. (2004). The world historical transformation of marriage. *Journal of Marriage and Family*, 66(4), 974–9.

Croll, E.J. (2000). *Endangered Daughters: Discrimination and Development in Asia*. London: Routledge.

Das Gupta, M. (1987). Selective discrimination against female children in rural Punjab, India. *Population and Development Review*, 13(1), 77–100.

Das Gupta, M. (2010). Family systems, political systems and Asia's 'missing girls': the construction of son preference and its unravelling. *Asian Population Studies*, 6(2), 123–152.

Desai, S. and Andrist, L. (2010). Gender scripts and age at marriage in India. *Demography*, 47(3), 667–87.

Dommaraju, P. and Jones, G. (2012). Divorce trends in Asia. *Asian Journal of Social Science*, 39(6), 725–750.

Douglass, M. (2011). *Global Householding and International Migration Research—Paradigms, Emerging Dynamics and Public Policy in East and Southeast Asia*. Paper presented at Workshop on Householding in Transition: Emerging Dynamics in 'Developing' East and Southeast Asia, National University of Singapore, 25–26 July.

Esteve, A., Lesthaeghe, R., and Lopez-Gay, A. (2012). The Latin American cohabitation boom. 1970–2007. *Population and Development Review*, 38(1), 55–81.

Festy, P. (1980). On the new context of marriage in Western Europe. *Population and Development Review*, 6(2), 311–15.

Frejka, T., Jones, G., and Sardon, J.-P. (2010). East Asian childbearing patterns and policy developments. *Population and Development Review*, 36(3), 579–606.

Furstenberg, F.F. (1990). Coming of age in a changing family system. In: Shirley Feldman, S. and Elliott, G.R. (eds.) *At the Threshold: The Developing Adolescent*. Cambridge: Harvard University Press.

Furstenberg, F. (2010). On a new schedule: transitions to adulthood and family change. *Future of the Child*, **20**, 67–87.

Furstenberg, F.F. (2011). The recent transformation of the American family: witnessing and exploring social change. In: England, P. and Carlson, M. (eds.) *Social Class and Changing Families in an Unequal America*. Stanford, CA: Stanford University Press, pp. 192–220.

Fussell, E. and Palloni, A. (2004). Persistent marriage regimes in changing times. *Journal of Marriage and Family*, 66(5), 1201–13.

Goode, W.J. (1963). *World Revolution and Family Patterns*. New York: The Free Press.

Guilmoto, C. (2012a), *Sex Imbalances at Birth: Trends, Differentials and Policy Implications*. Bangkok: UNFPA.

Guilmoto, C. (2012b). Son preference, sex selection and kinship in Vietnam. *Population and Development Review*, 38(1), 31–54.

Guzzo, K. (2014). Trends in cohabitation outcomes: compositional changes and engagement among never-married young adults. *Journal of Marriage and Family*, **76**, 826–42.

Heard, G. (2011). Socioeconomic marriage differentials in Australia and New Zealand. *Population and Development Review*, 37(1), 125–60.

Hochschild, A. and Machung, A. (1989). *The Second Shift: Working Parents and the Revolution at Home*. New York: Viking Penguin.

Ibrahim, R. and Hassan, Z. (2009). Understanding singlehood from the experiences of never-married Malay Muslim women in Malaysia: some preliminary findings. *European Journal of Social Sciences*, 8(3), 395–405.

Jenkins, B. (2010). Grandparent childcare in Australia: a literature review. *Elder Law Review*, **6**, 1–19.

Jones, G.W. (2002). The changing Indonesian household. In: Robinson, K. and Bessell, S. (eds.) *Women in Indonesia: Gender, Equity and Development*. Singapore: Institute of Southeast Asian Studies.

Jones, G.W. (2007). Delayed marriage and very low fertility in Pacific Asia. *Population and Development Review*, 33(3), 453–78.

Jones, G.W. (2009). Women, marriage and family in Southeast Asia. In: Devasahayam, T. (ed.) *Gender Trends in Southeast Asia: Women Now, Women in the Future*. Singapore: Institute of Southeast Asian Studies.

Jones, G.W. (2012). Population policy in a prosperous city-state: dilemmas for Singapore. *Population and Development Review*, 38(2), 311–36.

Jones, G.W. and Gubhaju, B. (2012). Marriage trends in Southeast Asia. In: Williams, L. and Guest, P. (eds.) *Southeast Asian Demography*. Ithaca, NY: Southeast Asia Program, Cornell University.

Kiernan, K. (2003). *Cohabitation and Divorce Across Nations and Generations*. CASE paper 65. London: Centre for Analysis of Social Exclusion, London School of Economics.

Kiernan, K.E. and Mensah, F.K. (2009). Unmarried parenthood, family trajectories, parent and child well being. In: Hansen, K., Joshi, H., and Dex, S. (eds.) *Children of the 21st Century: From Birth to Age 5*. Bristol: Policy Press.

Knodel, J. and Chayovan, N. (2008). Population ageing and the well-being of older persons in Thailand: past trends, current situation, and future challenges. Papers in Population Ageing Series, Number 5. Bangkok: UNFPA Thailand and Asia and the Pacific Regional Office.

Knodel, J. and Chayovan, N. (2011). *Intergenerational Family Care for and by Older People in Thailand*. Paper presented in Conference on Shifting Boundaries of Care Provision in Asia: Policy and Practice Changes, Asia Research Institute, National University of Singapore, 14–15 March.

Le Bras, H. (1979). *Child and Family*. Paris: OECD.

Lesthaeghe, R. (2010). The unfolding story of the second demographic transition. *Population and Development Review*, 36(2), 211–51.

Lloyd, C., et al. (2005), *The Changing Transitions to Adulthood in Developing Countries: Selected Studies*. Washington, DC: National Academies Press.

Maeda, E. (2008). Relational identities of always-single Japanese women. *Journal of Social and Personal Relationships*, 25(6), 967–87.

Manning, W., Smock, P., and Fettro, M. (2019). Cohabitation and marital expectation among single millennials in the U.S. *Population Research and Policy Review* 38, 327.

Martin, J.A. et al. (2012). Births: final data for 2010. *National Vital Statistics Reports*, 61(1), 1–72.

Mason, A., Phananiramai, M., and Poapongsakorn, N. (1993). Households and their characteristics. In: Campbell, B.O., Mason, A., and Pernia, E.M. (eds.) *The Economic Impact of Demographic Change in Thailand, 1980–2015*. Honolulu: University of Hawaii Press.

Mathews, T.J. and Hamilton, B.E. (2009). *Delayed Childbearing: More Women are Having Their First Child Later in Life*. *NCHS Data Brief*, No. 21. Atlanta, GA: Centers for Disease Control and Prevention.

McDonald, P. (1992). Convergence or compromise in historical family change?. In: Berquo, E. and Xenos, P. (eds.) *Family Systems and Cultural Change*, Oxford: Clarendon Press.

Mclanahan, S. (2004). Diverging destinies: how children are faring under the second demographic transition. *Demography*, **41**, 607.

Montgomery, M.R., Arends-Kuenning, M, and Mete, C. (2000). The quantity-quality transition in Asia. *Population and Development Review*, 26(suppl), 223–56.

OECD Family Database. [Online] Available at: https://www.oecd.org/els/social/family/database

Ogawa, N. (2006). Population aging and policy options for a sustainable future: the case of Japan. *Genus*, LX1(3–4), 369–410.

Oppong, C. (1992). Traditional family systems in rural settings in Africa. In: Berquo, E. and Xenos, P. (eds.) *Family Systems and Cultural Change*. Oxford: Clarendon Press.

Parrenas, R.S. (2008). *The Force of Domesticity: Filipina Migrants and Globalization*. New York: New York University Press.

Peek, C., Im-em, W., and Tangthanaseth, R. (2015). *The State of Thailand's Population 2015, Features of Thai Families in the Era of Low Fertility and Longevity*. Bangkok: UNFPA.

Perelli-Harris, B. et al. (2018). Do marriage and cohabitation provide benefits to health in mid-life? the role of childhood selection mechanisms and partnership characteristics across countries. *Population Research and Policy Review*, **37**, 703.

Perelli-Harris, B., and Lyons-Amos, M. (2016). Partnership patterns in the United States and across Europe: the role of education and country context. *Social Forces*, 95(1), 251–82.

Perelli-Harris, B., Sigle-Rushton, W., Kreyenfeld, M., Lappegård, T., Keizer, R., and Berghammer, C. (2010). The educational gradient of childbearing within cohabitation in Europe. *Population and Development Review*, 36(4), 775–801.

Perelli-Harris, B., Hoherz, S., Lappegård, T., and Evans, A. (2019). Mind the 'happiness' gap: the relationship between cohabitation, marriage, and subjective well-being in the United Kingdom,

Australia, Germany, and Norway. *Demography*, **56**, 1219–46.
Preston, S.H. (1976). Family sizes of children and family sizes of women. *Demography*, 13(1), 105–114.
Pritchett, L. and Viarengo, M. (2012). Why demographic suicide? The puzzles of European fertility. *Population and Development Review*, 38(Supplement), 55–71.
Raymo, J., Iwasawa, M., and Bumpass, L. (2004). Marital dissolution in Japan: recent trends and differentials. *Demographic Research*, 11(14), 395–419.
Republic of South Africa, Department of Social Development (2013). *White Paper on Families in South Africa, Pretoria*. [Online] Available at: https://www.westerncape.gov.za/assets/departments/social-development/white_paper_on_families_in_south_africa_2013.pdf
Retherford, R.D. and Ogawa, N. (2006). Japan's baby bust: causes, implications and policy responses. In: Harris, F.R. (ed.) *The Baby Bust: Who Will Do the Work? Who Will Pay the Taxes?*. Lanham, MD: Rowman and Littlefield Publishers, pp. 5–47.
Sassler, S., Michelmore, K., and Qian, Z. (2018). Transitions from sexual relationships into cohabitation and beyond. *Demography*, 55(2), 511–34.
Schroder-Butterfill, E. and Kreager P. (2005). Actual and de facto childlessness in old age: evidence from East Java, Indonesia. *Population and Development Review*, 31(1), 19–55.
Singh, S. (1998). Adolescent childbearing in developing countries: a global review. *Studies in Family Planning*, 29(2), 117–36.
Situmorang, A. (2007). Staying single in a married world: never-married women inn Yogyakarta and Medan. *Asian Population Studies*, 3(3), 287–304.
Situmorang, A. (2011). Delayed marriage among lower socio-economic groups in an Indonesian industrial city. In: Jones, G.W., Hull, T.H., and Mohamad, M. (eds.) *Changing Marriage Patterns in Southeast Asia: Economic and Socio-Cultural Dimensions*. London: Routledge.
Tan, J.-E. (2010). Social relationships in the modern age: never-married women in Bangkok, Jakarta and *Manila*. *Journal of Comparative Family Studies*, 41(5), 749–65.
Tengku-Aizan, H. and Nurizan, Y. (2008). National policy for the elderly in Malaysia: achievements and challenges. In: Lee, H.G. (ed.) *Ageing in Southeast and East Asia: Family, Social Protection and Policy Challenges*. Singapore: Institute of Southeast Asian Studies.
Therborn, G. (2004), *Between Sex and Power: Family in the World, 1900–2000*. London: Routledge.
Umberson, D., Crosnoe, R., and Reczek, C. (2010). Social relationships and health behavior across the life course. *Annual Review of Sociology*, 36(1), 139–57.
Umberson, D., and Montez, J.K. (2010). Social relationships and health: a flashpoint for health policy. *Journal of Health and Social Behavior*, 51(Suppl), S54–S66.
UNICEF (2018). *Data 2018. Percentage of Women Aged 20 to 24 Years Who Were First Married or In Union Before Ages 15 and 18*. [Online] Available at: https://data.unicef.org/topic/child-protection/child-marriage/
United Nations (2005). *Living Arrangements of Older Persons Around the World*. New York: United Nations.
Utomo, I.D. and McDonald, P. (2009). Adolescent reproductive health in Indonesia: contested values and policy inaction. *Studies in Family Planning*, 40(2), 133–46.
Van de Kaa, D.J. (1987). Europe's second demographic transition. *Population Bulletin*, 42(1), 1–59.
Waite, L. and Gallagher, M. (2002). *The Case for Marriage: Why Married People are Happier, Healthier and Better Off Financially*. New York: Broadway Books.
Wheelock, J. and Jones, K. (2002). Grandparents are the next best thing: informal childcare for working parents in urban Britain. *Journal of Social Policy*, **441**, 451–2.
Xenos, P., et al. (2006). Delayed Asian transitions to adulthood: a perspective from national youth surveys. *Asian Population Studies*, 2(2), 149–86.

Yang, C.-L. and Tsai H.-J. (2009). Changes in fertility and marriage rates in Taiwan. In: Ya-Chen, C. (ed.) *Women in Taiwan: Sociocultural Perspectives*. Indianapolis, IN: University of Indianapolis Press, pp. 61–82.

Yeung, W.-J.J., Desai, S., and Jones, G.W. (2018). Families in Southeast and South Asia. *Annual Review of Sociology*, **44**, 469–95.

Yount, K.M. (2003). Gender bias in the allocation of curative health care in Minia, Egypt. *Population Research and Policy Review*, 22(3), 267–99.

10.2

Women, men, and health

Diane Cooper and Hanani Tabana

Introduction

This chapter focuses on sex (biological) and gender-related influences on women and men's health. We pose the questions: what determines women and men's health status? What are the effects of sex and gender? What else influences health status, policies, programmes, and service provision? These questions are of critical significance in improving women and men's health status and promoting policies, programmes, and services that improve their well-being. A gender focus in health systems should aim at informing the development of a more 'strategic, effective and equitable health systems interventions, programs and policies' (Morgan et al. 2016, p. 1).

We highlight key health issues that are influenced by a mix of interlinking sex and gender factors. The categories 'women' and 'men' are not a binary categories and biological sex differences and gender-related factors do not influence health conditions homogeneously (Spence and Pilote 2015), but interact with a host of other factors.

Cross-cutting social determinants of health

We focus specifically on sex and gender factors impacting on women and men's health, with their influence needing to be understood within the context of other cross-cutting determinants of health. The social determinants of health (SDH) (CSDH 2008)—gender, class, race, or ethnicity, family structures, migration status, geographical location, age, sexuality, sexual orientation, and ability/disability combine in varying ways within and across time, influencing people's health lifespan. The SDH are

individually determined and underpinned by social, economic, and political systems embedded in power structures. Education significantly impacts positively on girls' and women's health. These are at personal, proximal, and distal levels (Krieger 2008). A socioecological model encompasses individual, interpersonal (relationships with partners, families, community), systems (such as the healthcare system) and structures (social, political, economic, cultural) (Bronfenbrenner 1979). Adopting a life-course approach, the contexts, experiences, and health status during childhood and adolescence has an impact on the health of adults (Dick et al. 2015; Chandra-Mouli et al. 2017). Within and between societies, these shape equitable or inequitable opportunities for people's good health enjoyment (CSDH 2008). Institutionalized through legislation, and culture and traditions (Christodoulou and Zobnina 2009), the health system can facilitate or hinder the attainment of good health. Health and the health system were influenced by the Millennium Development Goals (MDG) and now monitored by the Sustainable Development Goals (SDG). SDG 3 health goal consists of 13 targets with a number of implementation targets, that cross-cut each other. These include nutrition, environmental risk, reproductive and maternal health. SDG 3.1 includes reducing maternal mortality. This broader health vision links to health systems issues such as Universal Health coverage and biological and socially determined risk factors. SDH are often considered outside of the ambit of health, but are critical if we are to address health's root causes and promote intersectoral collaboration. In our chapter, we document health system and integrated interventions to enhance health (World Health Organization 2016a).

What does this chapter cover?

We begin by clarifying what we mean by the terms sex and gender. These terms are frequently confused and written interchangeably in health literature. We discuss various gendered approaches to health. Next we provide an overview of differences in life expectancy and the Burden of Disease in women and men and their determinants. We then examine more closely the manner in which sex and gender shape five specific examples of health issues: HIV, gender-based violence (GBV), cardiovascular disease (CVD), mental health problems, and maternal mortality. Each one is chosen to emphasize the varying manner in which gender and sex influence health.

We discuss biomedical interventions to address these examples of health problems and discuss gender transformative health interventions. We conclude by emphasizing conceptualizing health as a complex interplay between many determinants, of which gender is one key social determinant. These have key implications for we how we address them, approach research, develop health policy, and design services.

Sex and gender differences in health

Sex refers to the biological differences between females and males. Some health phenomena and experiences are biologically exclusive to males and females. These include male or female specific reproductive health events such as pregnancy and childbirth and illnesses such as female or male reproductive cancers. Other sex-related differences include conditions that appear to be commoner in women or men. For example, autoimmune conditions such as thyroid disease, rheumatoid arthritis, lupus, and multiple sclerosis is commoner in women (Whitacre 2001). Whereas, strongly related to lifestyle issues such as alcohol and tobacco use, liver and respiratory health conditions are commoner in men (Murray and Lopez 1997) (Box 10.2.1).

Box 10.2.1 What is sex/biology?

Sex-related health differences between women and men include anatomical, genetic, biochemical, physiological, and physical factors (Wizemann and Pardue 2001). These characteristics are universal and tend to be constant, differentiating humans as males and females. However there are variations and fluidity in the sex or biological characteristics of individuals, with some individuals possessing both, combined in different ways (Christodoulou and Zobnina 2009).

Cellular and molecular research is needed for a fuller understanding of biological explanations for women and men's health conditions. Many of the original biological explanations for an excess of certain health conditions in men or women are open to debate. Biological vulnerability does not act in isolation in placing women and men at more risk or advantage. It is frequently difficult to disentangle the mixed influences of biological and non-biological factors in women and men's health. Whether these biological

health risks lead to health problems depends on a host of non-biological factors.

Box 10.2.2 What is gender?

Gender is a social construction that embodies social attitudes, values, and norms and varying expectations of women and men's roles and responsibilities between and within societies, across time and space. It includes how individuals perceive themselves and others as female and male, and how they behave and engage with each other. It includes learned social attributes that comprise roles, activities, and responsibilities attributed to being female or male. It is therefore acquired through socialization and is changeable, varying across time and cultural contexts. It includes societies' perceptions of how women or men should act, what is valued in being a woman or a man and what behaviour is seen as permissible, respectively. In general these differences equate to inequalities between females and males in their responsibilities, activities, resource access and control and opportunities for decision-making (Christodoulou and Zobnina 2009; Sprague 2018).

Some frequently used key terms in gender analyses include gender identity, expression, and sexual orientation (Box 10.2.2 and 10.2.3).

Box 10.2.3 What is gender identity and sexual orientation?

Gender identity refers to an individual's personal and social identity regardless of their biological sex. Cisgender is a term used for an individual who identifies with their sex assigned at birth. Transgender includes identities of individuals who do not identify or do not wholly identify with their sex assigned at birth. Sexual orientation refers to sexual attraction and intimacy. For lesbian women, homosexual men, gay, queer, or bisexual individuals, attraction and intimacy is directed primarily or additionally to a person of the same sex (Christodoulou and Zobnina 2009).

While gender constructions are fluid (Sprague 2018), dominant socially stereotypical representations of masculinities are portrayed as idealized, heterosexual forms of masculinities. These are associated with characteristics such as self-reliance; competitiveness; containment of emotions; aggression; physical strength; primary economic providers; and prowess in sexual achievement (Jewkes et al. 2015). There are similarly stereotypically and idealized forms of emphasized femininities, associated with gentleness, empathy, sensitivity, care, and accommodating the interests and desires of men. Hegemonic forms of masculinities develop patterns of practices of masculinities that dominate and allow a global, ideologically

legitimate basis for subordination. This suggests that women and men position themselves in terms of power, in relation to these hegemonic or dominant patterns (Connell and Messerschmidt 2005, p. 832). These reflect systems of gender inequity. Inequity can be seen as not adequately addressing the different health needs of women and men (Fig. 10.2.1).

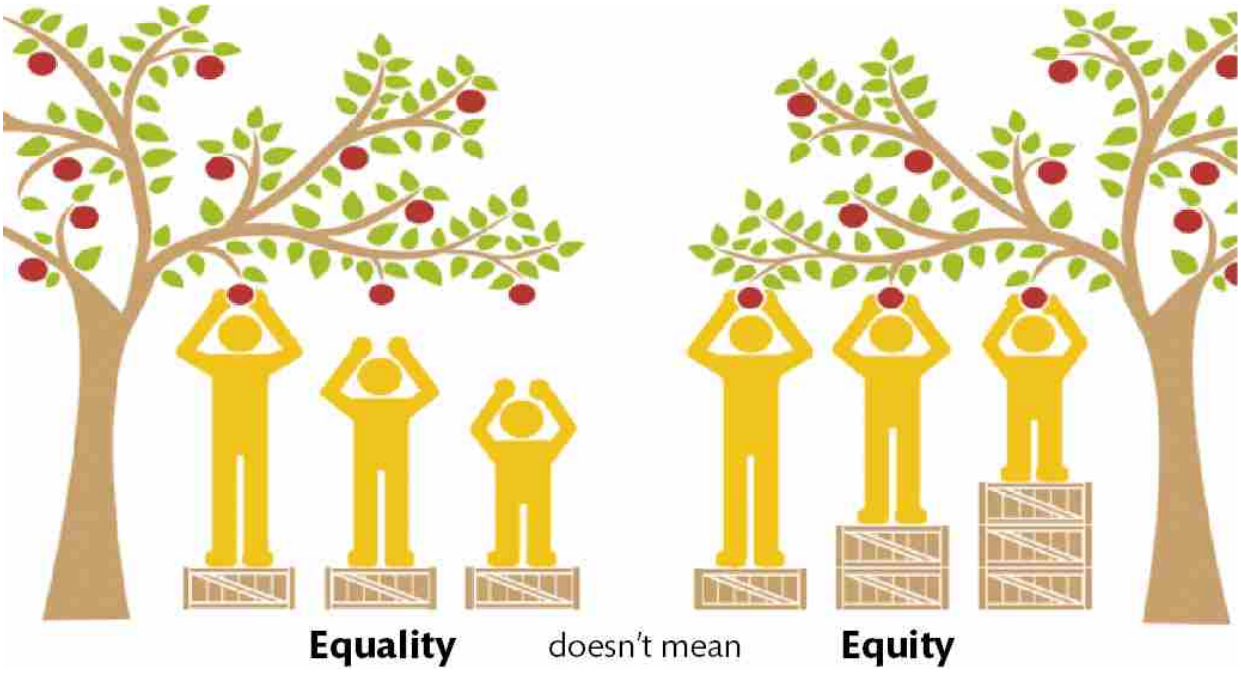


Fig. 10.2.1 The difference between equity and equality.

Power hierarchies include symbolic and actual power relations, acted upon at an individual, interpersonal, institutional, and societal level. Butler's (1999) concept of 'performative' gender recognizes gender as a set of social, economic, and political practices, *performed* by men and women. This concept of 'hegemonic' masculinities provides a framework for examining institutionalized hierarchies in gender relations within and between societies.

Despite the dominance of certain forms of masculinity and femininity, multiple forms of masculinities and femininities exist (Connell 2012; Jewkes and Morrell 2012; Morrell et al. 2016; Shefer 2016). Within the contextual nature of masculine and feminine norms gender socialization reflects a diverse range of female and male characteristics, social expectations, and identities (Hammarström 2018). 'For men, hegemonic masculinities emerge

in a contested process, among vying identities' (Connell and Messerschmidt (2005) in Sprague 2018, p. 53). Women and of men who adopt non-hegemonic patterns of practice are vulnerable to censure and marginalization. This makes for a complex analysis of how gender affects female and male power in achieving health status, health practices, and healthcare provision.

Gender hierarchies have changed historically, new forms replacing older ones (Connell and Messerschmidt 2005, p. 836). How individuals or groups of people 'perform' gender differs across time and social contexts. While there are multiple masculinities and femininities, there are socially central forms that predominate, associated with greater power for men (Box 10.2.4).

Box 10.2.4 Key messages

- There are important conceptual differences between sex and gender.
- Considering both is critical to understanding women and men's health status policies, programmes, and services and specific health problems.
- Gender is socially constructed, influencing roles, behaviours, activities, and attributes assigned to females and males and perceptions of their appropriate behaviour.
- Multiple masculinities and feminities exist.
- Hegemonic masculinities and feminities 'affects how people live, work and relate to each other at all levels, including in relation to the health system' (Morgan, et al. 2016, p1).

Intersectionality

The origins of intersectionality theories and approaches, emphasizing cross-cutting factors influencing health, are a blend of legal, critical feminist, race, anticolonial, and queer theories (Hankivsky 2014). They focus on analysing diversity in inequity, exclusion, and social injustice. Power, privilege, and marginality emerge from a multiplicity of interacting contextual social stratifiers, including class, race, or ethnicity, gender, geography, family structures, ability/disability, migration status, religion, and sexual orientation. These interlocking systems of power (Larson et al. 2016, p. 1) are reflected in laws, policies, and institutions (Larson et al. 2016, p. 2). As different and interdependent types of privilege, vulnerability, marginalization, and exclusion exist; they argue we should not a priori assign one component as a most significant health inequity factor. They

argue centralizing one factor encourages an add-on of other power influences as discrete factors, rather than in conjunction with each other (Hankivsky 2014, p. 252). Their analysis focuses on relative social privilege or disadvantage depending on people's life contexts (Hankivsky 2014). This influences understanding of poorer health and health service availability and access in vulnerable or disadvantaged groups (Larson et al. 2016, p. 3)

Gender mainstreaming

An approach that was popularized internationally from the mid-1990s is *gender mainstreaming*. The United Nations Economic and Social Council in 1997, the World Health Organization (WHO) and development organizations adopted it. It was in response to promoting gender as including women and men and addressing limitations in development-oriented activities. These included failing to integrate gender considerations into the work of all departments (Mannell 2010). It involved incorporating a gender perspective into policies, strategies, programmes, project activities, and administrative functions and into the organizational cultures—ways of thinking and doing.

While the contribution of gender mainstreaming is acknowledged in promoting the need to address gender inequality and power relations in the policy sphere, its health and development impact has been limited. One reason is its different interpretation in various implementation contexts. This led to vagueness and dilution of awareness and action on gender issues. Its orientation was primarily on equal female and male representation. A key and most serious critique is that it depoliticised gender power relations in its emphasis on integration and a largely technocratic approach (Tolhurst et al. 2012, p. 1,826). Issues of gender should not be mere add-ons or ticked boxes but should be integrated and analysed to illuminate how systems and institutions generate gender inequalities or perpetuate them and how to change these (Morgan et al. 2016).

Gender inequity remains important

Witter et al. (2017) argue that *not* focusing specifically on gender inequity leaves an unintended a gap in health and health system analysis. While an explicit focus on poor women, unemployed men, and female-headed households is needed to achieve health system equity (Witter et al. 2017), gender needs to be recognized as a critical division. It commonly impacts on

people's access to health systems resources and services. The predominance of reproductive health needs in low- or middle-income countries (LMIC) countries, disadvantages women as a group. They focus on the disadvantaging gendered effects of the out-of-pocket health expenditure on women, using an Indian case study. Their findings showed that while teenage pregnant women suffered most out-of-pocket cost-related decreased access to maternal health services, all women experienced reduced access. They found that untreated illness was higher in women than men, even among educated women with higher socioeconomic status and caste (Witter et al. 2017, p. v8). This is one specific case study, but draws attention to the continued value of considering gender inequity in designing healthcare policies and programmes. Hegemonic forms of masculinities are a useful concept to understand dominant patterns of gendered practices. Jewkes and Morrell (2012) and Shefer (2016), build on Connell's (2012) approach, in examining gender influences on women and men's health issues. It acknowledges that gender roles, relations, and identities are unstable, continuously undergoing active change (Tolhurst et al. 2012); that there are dynamic links between multiple gender identities and hierarchies of power that play out in complex ways. At the same time it highlights intersecting sex, age, social, and economic determinants, that impact differentially on different groups of women and men's health (Jewkes and Morrell 2012; Stern et al. 2014a) (Box 10.2.5).

Box 10.2.5 Key messages

- Theories and frameworks, such as intersectionality; approaches highlighting gender to avoid unintended negative health consequences; acknowledging the dynamic links between gender identities and hierarchies and their intersection with broader SDH, are helpful in analysing the roots of women and men's health problems.
- Gender inequity and power that favours dominant heterosexual forms of masculinities is key in understanding women and men's health differences.

Gender, sex, and key health issues

Our discussion of HIV highlights sex and gender determinants while GBV primarily reflects the effects of gender dynamics. The development of

cardiovascular disease (CVD) and mental health problems in women and men underscores biological and gendered influences. Their diagnosis and outcomes show the gendered effects of the health system, health professionals' decision-making, and treatment and healthcare. The effect of gender on health remains strongly mediated by other social factors. Maternal health foregrounds women's maternal mortality and morbidity differences. Similar divides between men exist.

Gender analysis recognizes that gender power relations are reflected in the division of labour, resources, and rewards (Gupta 2000). These are important in health-related interventions to improve men and women's health status, decision-making, behaviour, health service access, and management and treatment of health problems. We should aspire to change gender relations in health and health systems, but at the very least not further worsen them.

Life expectancy and Global Burden of Disease

A crude health measure is life expectancy. Estimates of life expectancy at birth show that women generally live longer than men. Average global life expectancy for children born in 2015 was approximately 4.7 years longer for females than males (73.8 years for females and 69.1 years for males) (WHO 2016b). This is due to biological, behavioural, and environmental factors, including sex-specific patterns in exposure to health risks and access to treatment.

In most countries, there is higher mortality risk for boys <5 years than for girls, contributing to lower male life expectancy. There is inconclusive evidence of whether there are biological differences. Sex-related theories for shorter male life expectancy have also been based on the possibility that the oestrogen may offer women premenopausal protection against cardiovascular disease (CVD), a key cause of earlier male mortality. However, most recent research does not appear to support this (Phillips 2006).

Gender differences affecting men's lower life expectancy are clearer. There are differences between women and men in accidental and non-accidental deaths. The gender division of labour has typically placed men in more hazardous jobs in construction, transport, and emergency services, with more dangerous work environmental conditions and accidents (Doyal 1995). Gender-related behavioural patterns, with young men more likely to engage

in greater risk-taking behaviours than women, combined with socioeconomic factors influence premature male mortality. Non-accidental deaths in young men (14–44 years) are due to road traffic accidents, homicides, and health-related conditions from higher male rates of smoking, alcohol, and substance abuse (Prince et al. 2007).

There are huge differences in overall in life expectancy between countries. In Japan, a country with the highest world life expectancy, females born in 2015 can expect an average lifespan of 86.8 years and males 80.5 years (Table 10.2.1). This contrasts with Sierra Leone, with the lowest life expectancy—females and males born in 2015 can expect a lifespan of 50.8 and 49.3 years, respectively (WHO 2017).

Table 10.2.1 Life expectancy at birth 2015 (top five and bottom five countries in years)

Country	Female	Male
Japan	86.8	80.5
Switzerland	85.3	81.5
Singapore	86.1	80.0
Australia	84.8	80.9
Spain	85.5	80.1
Cote d’ iVoire	54.4	52.3
Chad	54.5	51.7
Central African Republic	54.1	50.9
Angola	54	50.9
Sierra Leone	50.8	49.3

Source: data from World Health Organization 2017.

Lower overall life expectancies for women and men within and between countries, indicates inequity in health, between the wealthier and the poorer and marginalized. Life expectancy is much higher for women and men in high-income countries (HIC) than LMIC. Poverty-related factors: poor

nutrition, sanitation, low household income, and weak health systems underpin differences. Improvements in social, economic, and environmental conditions and access to good, quality healthcare have increased life expectancies in HIC. Even in HIC there are sharp differences in life expectancies, with poorer, more marginalized populations experiencing substantially shorter life expectancies.

Situations of war and conflict in certain regions and countries have led to lower life expectancies for populations overall, but particularly for men. High maternal mortality lowers female women's life expectancies in regions with poor maternal healthcare. Before antiretroviral therapy became universally available, AIDS-related mortality caused dramatic decreases in life expectancy in sub-Saharan African countries. In Eastern Europe and Russia, lower male life expectancy peaked in the period leading up to and after the dissolution of the USSR. These were attributed to gender-related male lifestyle practices such as excessive alcohol consumption, in response to the tremendous socioeconomic and political upheavals (Pietilä and Rytkönen 2008, p. 1,070).

The Burden of Disease (BOD) is a good indicator of the health quality of life people enjoy or lose. Amatya Sen captures this: 'we are not able to do much if we are disabled or ceaselessly bothered by illness' (Sen 2004, p. 28 in Sprague 2018, p. 126). The WHO adopted the concept of the Global Burden of Disease (GBD) to take account of illnesses and disabilities. It uses disability-adjusted-life years (DALY), quality-adjusted life years (QALY), and disability-adjusted life expectancy (DALE) as measures to quantify quality of health (Sprague 2018, p.128). This occurs throughout the life cycle with childhood and adolescence setting much of the scene for later adult health status. Perceptions of health vary between individuals and societies. Actions taken to decrease the BOD and promote enjoyment of health recognizes the special nature of health as a 'social good' (Sprague 2018).

LMIC countries especially in sub-Saharan Africa continue to struggle to contain major infectious or communicable health conditions such HIV/AIDS, tuberculosis, and Malaria. Interventions to address these are human resource and skills intensive. Many LMICs have limited infrastructure, human resources, and technical expertise to contain epidemics and devote resources to other health and social, economic, education, and development needs (Box 10.2.6).

Box 10.2.6 Key messages

- Overall, women enjoy a longer life expectancy than men.
- This varies within and between countries, reflecting the influence of a complex mix of sex, gender, age, behavioural, and social determinants of health.
- Lower female than male life expectancy signifies higher levels of societal gender discrimination.
- The WHO uses a number of measures for compromised health quality of life—the burden of disease.
- Child and adolescent health influence the health trajectory in adulthood.
- LMIC countries struggle with a dual epidemic of communicable and non-communicable diseases.

The impact of sex and gender on specific health problems

HIV/AIDS

HIV is one of the most significant threats to women and men's health in many parts of the world. The primary modes of HIV transmission are through unprotected vaginal or anal sex, use of contaminated needles during injecting drug use, and vertical transmission. We will on focus on the impact and variation in HIV's impact on women and men's health in LMIC in Southern Africa that bear the greatest burden of HIV infection (Abdool Karim et al. 2010; Ramjee and Daniels 2013).

Sprague's (2018) model in Figure 10.2.2 is valuable in showing graphically, the biological and structural factors impacting on HIV acquisition.

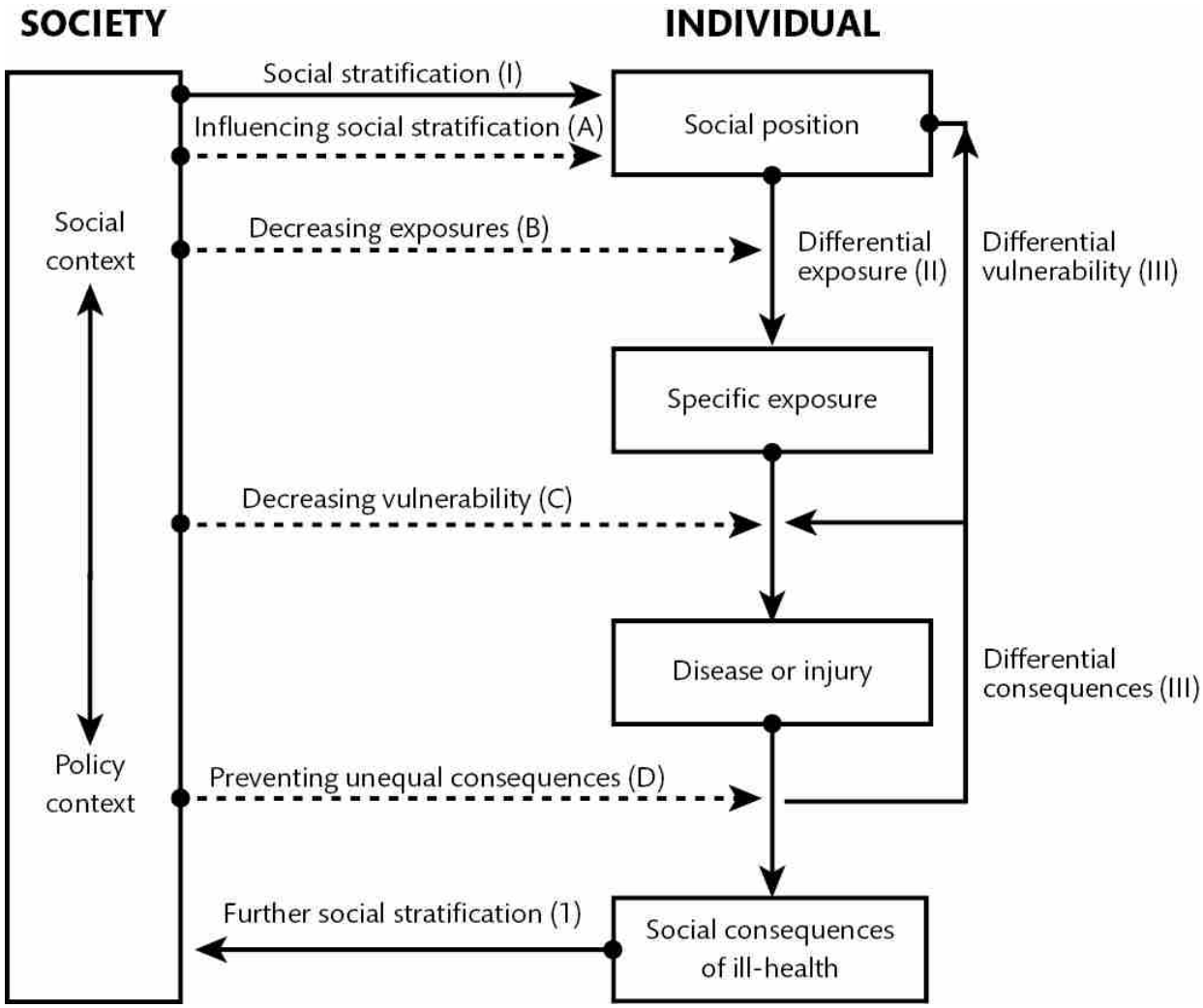


Fig. 10.2.2 Pathways in vulnerability and consequences of HIV acquisition.
Adapted with permission from Prof Finn Diderichsen from his and his colleagues framework that appears as figure 2.2. in Sprague (2008, p 47).

Biological HIV vulnerability

Women are biologically significantly more vulnerable than men to HIV infection. This is on a physical, genetic, and immunological level. The female vaginal tract (and the anus) has a larger mucosal surface exposed to the virus during condomless sex. It is also exposed to HIV for a longer period (Abdool Karim et al. 2010; Ramjee and Daniels 2013). The greater susceptibility of the vaginal, cervical (and anal) mucosal surface to microabrasions during sex facilitates HIV entry. This may be due to inflammations, ulcerations, lesions, and pre-existing sexually transmitted infections (STI). Undiagnosed STI are commoner in women, due to low

knowledge, greater difficulties in symptom diagnosis problems and access to STI healthcare is more stigmatized for women (Abdool Karim et al. 2010). The herpes simplex type 2 virus is linked to increased transmission risk ((Ramjee and Daniels 2013) and bacterial vaginosis (BV) and syphilis are also implicated. This is particularly significant for sub-Saharan Africa that has high levels of HSV-2, BV, and HIV.

Young women, aged 15–24 years, are up to 8 times more likely to be HIV-infected than their male peers (Dellar et al. 2015, p. 64). Gendered social and economic factors are combined with heightened biological risk of HIV transmission, due to young women's cervical immaturity and higher levels of genital inflammation (Abdool Karim et al. 2010; Ramjee and Daniels 2013). Hormones and immunological responses may also play a role. Increased levels of oestrogen after ovulation may increase HIV infectivity (Ramjee and Daniels 2013). Women's hormonal and physiological changes during pregnancy and lactation may be linked to higher rates of HIV acquisition (Ramjee and Daniels 2013). There are increased risks of maternal mortality as a result of untreated HIV-related infections (Kendall et al. 2014). There is potentially enhanced HIV risk for women using the injectable progesterone contraceptive Depo-Provera (Abdool Karim et al. 2010). This is currently under intensive study (Harper 2015; Tasker et al. 2018).

Gender and other social factors impact on HIV vulnerability

Biological factors interact with social, economic, and behavioural factors, to increase HIV infection vulnerability. Gender inequity is a key factor underscoring HIV infection vulnerability (Abdool Karim et al. 2010; Jewkes and Morrell 2012).

Hegemonic patriarchal cultures of masculinity underpin men frequently dominating in in heterosexual intimate relationship decision-making. Women have lesser say in whether to engage in condomless sex than men. Hegemonic forms of masculinities promote men having multiple concurrent sexual relationships and older men engaging in sex with younger women. This heightens young women's HIV vulnerability (Dellar et al. 2015, p. 67). Women may experience violence if they refuse sex or insist on condom use (Jewkes and Morrell 2010; Ramjee and Daniels 2013). Sexual violence can be a cause and a consequence of HIV (Jewkes 2014).

Young women's disadvantaged socioeconomic status is associated with earlier and coerced sexual debut, less condom use, multiple partners, and having transactional sex with more likely to be infected older men (Abdool Karim et al. 2010; Ramjee and Daniels 2013).

Social constructions of masculinity may lead to increased male risks of HIV infection. Social expectations that 'real' (heterosexual) men show their masculinity through condomless sex and multiple partners places pressure men to engage in riskier sexual behaviour. Traditional social norms combine with a health system with poor integration of sexual and reproductive health (SRH) and HIV. Men's involvement in SRH, apart from for STI, is not encouraged. As Gutmann (2007, p. 11) states men are 'the missing (f)actors in reproductive health'. Social context and barriers within health services hinder Lesbian, Gay, Bisexual, Transgender, Intersex, Questioning, and other gender identities (LGBTQI+) individuals and groups from seeking SRH care.

There are marked differences in women and men's use of HIV care. Gender norms make it more acceptable for women to seek healthcare. Women seek care for reproductive health and child health needs. In parts of the world where the HIV epidemic is most severe, prevention of infant HIV transmission programmes has added to women's greater health-seeking behaviour. This has created its own set of problems for women. Because they are frequently the first partner in an intimate relationship to test for HIV, if infected, they could face stigma and blame. However, women's access to testing and diagnosis for HIV and access to ART has led to greatly improved healthcare access and outcomes for women living with HIV. Gender norms, that perceive seeking healthcare as a display of 'weakness', may discourage men from seeking healthcare until very ill (Stern et al. 2014b; Jewkes et al. 2015). Men are less likely to test for and be diagnosed with HIV infection, initiate ART, be retained in care and more likely to die of AIDS-related illnesses (Cornell et al. 2017). Men are less likely to participate in social support structures for people living with HIV (Madiba and Canti-Sigaqa 2012) (Box 10.2.7).

Box 10.2.7 Key messages

- Biological and gender-related factors make women, particularly younger women, more vulnerable to HIV infection.
- They are more likely to be tested, diagnosed with HIV, and access and remain on ART due to greater health service access. However, this can lead to stigma and blame.
- Hegemonic masculinities promote riskier sexual behaviour. This and less health service contact deter men from health-seeking and testing, being diagnosed with HIV, and accessing and remaining on ART.

Cardiovascular disease

Non-communicable diseases (NCD) represent a significant BOD in the twenty-first century (Mendis and Chestnov 2014). Of the estimated 36 million deaths globally approximately 63% in 2008 were caused by NCD. Cardiovascular diseases (CVD) has the largest individual proportion of NCD deaths (48%) (Mendis and Chestnov 2014). Coronary Heart Disease and Stroke are the leading causes of CVD death and disability. Key risk factors include high blood pressure, overweight or obesity, and higher cholesterol levels. These appear to have similar effects on CVD outcomes for women and men.

The Gendered nature of CVD

Despite CVD perceived for decades as a ‘male’ disease, it is commoner in women (Appelman et al. 2015), likely due to stroke being commoner and due to women’s longevity. Women present with CVD on average 7–10 years later than men. This has led to speculation that oestrogen plays a protective role against CVD for women. However, recent research fails to support this (Riska 2010) and has shown no benefit in prevention of CVD through hormonal replacement therapy (Spence and Pilote 2015, p. 209).

Younger women have specific sex-related risk factors for CVD associated with pregnancy-related health problems such as gestational hypertension, diabetes and preeclampsia and endocrine disorders (Appelman et al. 2015). Risk factors such as migraine, commoner in women than men, appears to be associated with stroke in women. While smoking is commoner in men than women, multiple studies show that prolonged heavy smoking in women increases their relative risk of CVD compared to men (Appelman et al. 2015, p. 213).

NCD in LMIC

NCD have become increasingly common in LMIC, influenced by lifestyle changes due to the health transition (Kengne et al. 2013). CVD shows a prevalence of pre-transitional diseases such as rheumatic fever and idiopathic cardiomyopathies, with high levels of hypertensive disease and stroke and health transition diseases. Common individual risk factors include tobacco use, physical inactivity, and unhealthy diets. Risk factors vary by sex, with tobacco use commoner in men and obesity in women (Mayosi et al. 2009). The African region has the highest age standardized death rate from NCD, 779 per 100,000 (Mendis and Chestnov 2014). The double burden of infectious diseases and NCD is referred to as 'colliding epidemics' (Mayosi et al. 2012, p. 5). NCD are the second leading cause of death in Sub-Saharan Africa comprising 30% of deaths in 2011 (Airhihenbuwa et al. 2016). Hypertension is a major underlying cause of CVD. The numbers of people with hypertension is expected to increase by 68% from 7.5 million in 2008 to 126 million in sub-Saharan Africa by the year 2025 (Airhihenbuwa et al. 2016). An analysis of country variation in strokes show that the poorest countries have the highest BOD (Mayosi et al. 2009, p. 936). The WHO estimates the burden of NCD in South Africa is two to three times higher than that in HIC and deaths due to CVD are likely to double by 2040. This is similar to that in other sub-Saharan Africa countries and central European countries, increasing in rural and poor urban communities. This is 'driving a rise in the demand for chronic disease care' (Mayosi et al. 2009, p. 935). Prevention and treatment of NCD in LMIC including CVD, is neglected due to the overwhelming burden of infectious illnesses (Mayosi et al. 2009, p. 934). HIV has exacerbated the burden of NCD. This demonstrates an important role for improved health services in reducing health inequities that underpin these conditions (Mayosi et al. 2009, p. 936).

CVD and healthcare

Health systems factors play an important role in the gender differences in women's perceptions of and presentation with possible symptoms of heart disease. Women present more frequently with symptoms of tiredness, weakness, light-headedness, palpitations, or loss of appetite rather than chest pain (Aggarwal et al. 2018, p. 3). When women report with symptoms of chest pain, this is more often attributed to unexplained chest pain (Crea et al.

2015, p. 159), because these do not conform to typical male symptoms (Crea et al. 2015; Aggarwal et al. 2018). Diagnostic tests for women and men with potential heart disease have differed. The Euro Heart Survey, in which 42% of patients were female, found physicians are more likely to conduct exercise ECG and coronary angiography on men with reported chest pain or discomfort than women. It found gender biases in the use of evidence-based medical treatment for women, leading to under or inappropriate treatment (Crea et al. 2015, p. 160).

The risk of heart disease has been underestimated in younger women in whom, in-hospital post-acute myocardial infarction mortality is higher than among their male peers (Crea et al. 2015, p. 157). Factors related to younger women and ethnicity in the United States (black, Latino and south Asian women) create subgroups of women who are more disadvantaged and marginalized in recognition of heart attack symptom recognition and treatment (Aggarwal et al. 2018, p. 5). Angina and heart attacks are perceived as 'male diseases', with poorer recognition of symptoms and quality of care in women. Previously a greater proportion of men reportedly died from sudden, pre-hospital cardiac attacks. This is no longer the trend (Crea et al. 2015, p. 161). This is viewed as due primarily to gender biases in perceptions of cardiac attack symptoms.

Less post-cardiac rehabilitation occurs in women. This may be due to health service and socioeconomic related factors, such as poorer promotion and referral by healthcare providers, cost and transport problems, inadequate social support systems, and the burden of family responsibilities (Aggarwal et al. 2018, p. 5). Socioeconomic factors are similarly likely to affect women's prevention activities and health-seeking behaviour with respect to CVD (Spence and Pilote 2015, p. 209).

Further research into potential biological risk factors for CVD in women and men can improve our understanding of the aetiology of CVD and assist in tailoring specific interventions for detection, treatment, and management (Appelman et al. 2015) (Box 10.2.8).

Box 10.2.8 Key messages

- Contrary to perceptions that CVD is a ‘male disease’ due to higher levels of hypertension and strokes, it is commoner in women.
- Younger women and those with HIV have specific CVD risks.
- Difference risk factors including socioeconomic influences lead to different CVD outcomes in women and men.
- Healthcare professionals’ gender bias in detection, diagnosis, testing and treatment has had negative outcomes for women.
- LMIC have ‘colliding epidemics’ of infectious diseases and NCD.

Gender-based violence

GBV covers a wide range of categories of violence (Box 10.2.9).

Box 10.2.9 Gender-based violence

The United Nations defines Gender-based violence against women (VAW) as: Any act of GBV that harms of may result in harm (physical, sexual, or emotional/mental) or suffering to women (Sprague 2018, p.63). This may include threats or acts which coerce or deprive women of freedom. Physical violence includes being slapped, punched, choked, having objects thrown at one, pushed or shoved, beaten, intentionally burnt, or threatened with a weapon. This can be in either the private or public sphere. Rape or sexual violence includes any form of forced sexual act, including those due fear of harm or intimidation (WHO 2013, p. 6) It can be perpetrated regardless of the relationship of the perpetrator to the survivor, in any place. It includes penetration with a body part or object (Sprague 2018, p. 63). GBV includes femicide, homicide, and intimate partner (domestic) violence. These are most commonly perpetrated by men against women.

Violence against women

There is pervasive violence against women internationally. Thirty-five percent of women report physical or sexual violence by a male intimate or non-intimate partner (WHO 2013). Women are most likely to be exposed to violence in the home or its surroundings. Most physical and sexual violence against women is by an intimate partner (WHO 2013). While physical and sexual violence is more visible, intimate partner emotional abuse or violence, especially from male intimate partners, including ‘verbal abuse, threats of violence, engendering fear, humiliation, destruction of property, enforcement of social isolation, taking or withholding earnings, and flaunting other sexual partners’ (Jewkes 2010, p. 851) is less recognized and addressed (Matheson et al. 2015). Emotional abuse’s reported prevalence is

between 9 and 70% (Jewkes 2010, p. 851). People experiencing abuse frequently face unsympathetic criminal justice services (Meyer 2016) and unprepared health services (Global Burden of Disease Study 2013 Collaborators et al. 2015).

Violence, including femicide against lesbian or gender non-conforming women, those living and not living with HIV, is a specific form of violent hate crime (Cooper et al. 2016). In some regions there is violent abduction of women for sex and/or marriage. For example in the Eastern Cape province of South Africa, a specific form of sexual violence, known as *ukuthwala* occurs—this includes abduction, forced marriage, and rape of young girls. It is a distortion of a past custom for consenting, eloping young adults (Monyane 2013).

Male-on-male violence

Almost all violence against men is perpetrated by other men. It is estimated that for every violence male death, there are 29 male-on-male assaults serious enough to warrant healthcare (Krug et al. 2002).

Gender-related factors promoting notions of toughness, competitiveness, and a desire for dominance over other men, as positive attributes of 'being a man' (Jewkes et al. 2015) are most prominent in underpinning male-on-male violence. These occur against a backdrop of other socioeconomic risk factors for male-on-male violence. Like male-on-female violence, these include socioeconomic deprivation, urban concentrations of poverty, and substance abuse (Global Burden of Disease Study 2013 Collaborators et al. 2015).

Interpersonal violence

The global prevalence of interpersonal violence (IPV) against women is estimated as 15 to 71% (Sprague 2018, p. 63). IPV was responsible for 8.5% of deaths in 2013 (Global Burden of Disease Study 2013 Collaborators et al. 2015). GBV-related DALY differs by sex, age, and region. Injuries are higher among men than women, except in the older age group and apart from road accident injuries are due to male-on-male assault and homicide. Homicide in adult males aged 15 years to 49 years, in DALY are 2,651 per 100,000 population in western Europe and 10 780 in Eastern Europe (Global Burden of Disease Study 2013 Collaborators et al. 2015, p. 6). Interpersonal violence is responsible for 317/100,000 DALY. This is commoner in LMIC,

with boys under the age of 15 years having a low rate of 468.4 per 100,000 DALY in Western Europe and a high rate of 6,471.4 in central Africa. Male-on-male interpersonal violence has reached epidemic proportions, among young men in a country such as South Africa, that is not at war (Norman et al. 2010).

No convincing biological factors underpin interpersonal violence. Testosterone in men doesn't explain why all men are not violent. Greater physical strength in men than in most women enables them to engage in more physical violence. GBV against women and in male-on-male violence is highly gendered.

While several other factors cross-cut gender in influencing male violence against women, a key driver is gender power inequity in interpersonal relationships and society (Jewkes 2002). This manifests in male controlling behaviour, notions of male entitlement and power to 'punish' women that is condoned by individuals, within communities, the criminal justice system, and among policy makers and service providers (Seedat et al. 2009). There is an interaction between women's experiences of violence and mental health problems. Violence against women has important negative mental health consequences such as depression and anxiety (Jewkes 2010). Physical health effects of violence include injury or death and limited sexual and reproductive control and health-seeking behaviour. This can lead to unintended pregnancies, with poorer maternal and perinatal health outcomes and risk of STI.

Same-sex IPV has been under researched. However, research among gay men link emotional, physical, and sexual abuse in intimate partnerships to dominant forms of masculinities in which one partner exercises controlling behaviour and power over another (Coles 2009; Creighton and Oliffe 2010). Given that men are the main perpetrators of IPV, gay men may suffer disproportionately among men (Oliffe et al. 2014). Like men generally, they frequently view physical and mental injuries as 'part of being a man' (Oliffe et al. 2014, p. 574). This has health system and services implications due to reluctance to seek medical or other assistance (Oliffe et al. 2014) (Box 10.2.10).

Box 10.2.10 Key messages

- Violence against women and male-on-male violence among young men is of epidemic proportions. Males are generally the perpetrators.
- Interpersonal violence is commonest for women and men, particularly from non-strangers.
- Several socioeconomic factors influence GBV, but gender power inequity in interpersonal relationships and society, with males more powerful, is the key GBV driver.
- GBV has severe negative physical and mental consequences.
- Unsympathetic criminal justice systems and unprepared health service present major problems in detecting and addressing GBV.

Mental health

The burden of mental health and gendered aspects—commonalities and differences

The burden of mental, neurological, and substance (MNS) use problems have increased by between 1990 and 2010. It currently accounts for one out of 10 healthy years of life lost globally. This causes excess mortality, impacting socially and economically on those affected, caregivers, and society. Historically, mental health research has focused on women's issues (Smith et al. 2016). This focus may have led to a neglect of men's mental health problems. Gender differences in mental health problems can be viewed through a gender conception and practices lens. During the industrial revolution, men left home for work, while women assumed child caring responsibilities in the home.

This gendered labour division influenced on how men and women were viewed and expected to behave in society. Productive work in the public sphere became associated with men and masculinity, while socioemotional and domestic work became associated with women and femininity. Women were normatively viewed as fragile and emotional beings, with children requiring their mothers' special care for moral and psychological development (Rosenfield and Smith 2010). While the current global shifts in ideological norms empower women striving for equality especially in the workplace, gender power divisions between men and women, in responsibilities, remain relevant for women and men's mental health. Given dominant feminine role expectations, women tend to put others' interests first more often, promoting internalizing problems (Doyal 1995). Childcare and household responsibilities often take a psychological toll on women, ultimately resulting in high stress levels. Where childcare is secure and home responsibilities are shared with spouses, women's depressive and anxiety

symptoms resemble the lower levels diagnosed among men. Gender norms encourage women to maintain social ties that are emotionally intimate. They enjoy higher levels of feeling that others care deeply about them (Taylar and Turner 2001). However, close ties that bring support can also be sources of negative interactions with conflicting demands (Taylar and Turner 2001). Negative interactions can offset the benefits of supportive interactions, leaving women with higher distress levels. Neurological differences in women and men's mental health are contested and social factors, including gender play a dominant role in mental health problems.

Large-scale epidemiological studies assessing mental health problems in the United States (Kessler et al. 2005) and elsewhere (Andrews et al.1999; Seedat et al. 2009), found no differences in men and women's overall rates of psychopathology. The differences lay in the diagnoses (Rosenfield and Smith 2010). Men exhibiting problems of substance abuse and antisocial behaviour, are likelier to be destructive and problematic to others as well as themselves. Alcohol and drug addiction can lead to serious psychological or physical consequences. Research shows that men are likelier to take their lives in violent ways, hence suicide death rates are higher for all ages of men than women (Zarit and Zarit 2011). Antisocial personality disorder is more commonly diagnosed in men, as are schizoid and paranoid personality disorders. In contrast, histrionic, dependent, and avoidant personality disorders (female stereotyped), are more commonly diagnosed in women (Torgersen et al. 2001). Eating disorders are diagnosed as 10 times commoner in women than men (Byrne and McLean 2001). These occur mostly in late adolescence and early adulthood. Rates of dementia, perhaps related to longevity, are higher for older women, leading to older women having a greater need for mental health services.

For women and men mental health problems are commonly comorbid with physical disorders such as CVD, cancers, metabolic disorders, and HIV (De Hert et al. 2009; Murthy 2004).

Notions of hegemonic masculinities encourage males to suppress emotions (feelings of helplessness, worry, and insecurity) defined these as feminine and weak. These expectations ultimately prescribe how men and women express their feelings. Men and women develop coping mechanisms, which for men may lead to hiding their feelings, often turning to substance abuse. Some describe substance abuse as a male version of depression (Rosenfield et al. 2005). Men may experience negative mental health

consequences through greater gendered difficulties in forging close long lasting social ties (Castillo et al. 2007).

The health system and gendered mental health practices

Health system-related, health services and health practitioners' behaviour, and clinician measurement bias has likely influenced much of the scholarship and current findings on gender and mental health. This has led to problematic assumptions about gendered mental health symptoms (Currin et al. 2011). Health professionals tend to perpetuate gender stereotypes about men and depression, by being less likely to correctly diagnose men presenting with symptoms of depression or anxiety (Borowsky et al. 2000). Men's mental problems tend to be underestimated. Clinicians are less likely to use standard psychological measurement tools to assess men's symptoms (Martin et al. 2013). Research comparing men's and women's mental health has tended to focus on different outcomes of stress instead of similarities and differences in men and women's mental health experiences (Elliott 2013) (Box 10.2.11).

Box 10.2.11 Key messages

- Four conventional explanations for gender differences in mental health are gender socialization, help seeking, coping, and symptom stratification (Currin et al., 2011).
- Understanding factors underlying mental health problems among women and men involves adopting a holistic consideration of the biological social, cultural, economic, and personal contexts of their lives (Judd et al. 2009).
- Researchers, policymakers, and clinicians should reassess the validity and reliability of current measurement tools for detecting symptoms for men and women, and focus on what factors might be particularly relevant for men's experiences (Smith and Hemler 2014). This can more effectively address mental health in women and men.

Maternal health

In 2015, the maternal mortality ratio (MMR) (the number of female deaths during pregnancy, childbirth and immediately after childbirth per 100,000 live births) was an estimated 216. The overall MMR declined by 44% and

2.3% annually from 1990 to 2015. It would need to decrease to a minimum of 7.3%, to reach the SDG target of 70 by 2030 (WHO 2016a).

Differences between women and women: region, class, and ethnicity

The overall decline in global MMR, belies huge differences in the position of pregnant women based on class, race, age, and within and between world regions. High levels of maternal morbidity and mortality are key indicators of global inequity reflecting structural inequity between women in different parts of the world. Women's higher maternal morbidity and mortality in LMIC are linked to poverty. The World Bank's 2018 Health Equity database found that people in 35 and 34 countries faced catastrophic and impoverishing health expenditure, respectively (Wagstaff et al. 2018). Poor maternal health and related problems remain important causes of women's death and disability in LMIC. In HIC the risk of maternal death is 1 in 3,700 compared to 1 in 38 in in sub-Saharan African countries (United Nations Development Programme 2015). LMIC accounted for roughly 99% of global maternal deaths in 2015, with sub-Saharan Africa accounting for approximately 66%, followed by Southern Asia (WHO et al. 2015, p. xi). Eighteen sub-Saharan African countries had a MMR ranging between 500 and 1,000 deaths per 100,000 live births (WHO et al. 2015). India, Nigeria, Pakistan, Afghanistan, Ethiopia, and the DRC together had more than half of the global number of maternal deaths (Hogan et al. 2010). MMR in LMIC regions such as in sub-Saharan Africa and south Asia are similar (351 and 336 per 100,000 live births, respectively).

The health system and healthcare

Gender-related factors hinder women's maternal healthcare in some regions, where women need permission from or be accompanied by a male partner, to attend health services. However, most differences in maternal health outcomes are due to preventable conditions related to health system and service issues, such as skilled birth attendance (WHO 2016a). While overall skilled birth attendance was estimated to be 73% by 2013, in Africa and Southeast Asia, more than 40% of births took place without skilled healthcare. Poor healthcare provider attitudes in maternal care, low healthcare provider knowledge and inadequate training underpin problems with access to skilled maternal healthcare (Sprague 2018, p. 98).

Younger, poorer women and those in lower income or rural areas have poorer access to or are less likely to use effective methods of contraception. Globally, more than 10% of all women do not have access to or are not using an effective method of contraception, with much higher proportions in LMIC countries. Less than half of those in the WHO African Region had their contraceptive needs met (WHO 2016a). It is estimated if unmet need for contraception was met, it could decrease maternal deaths by nearly a third (United Nations Development Programme 2015).

Most maternal health complications are preventable if women enjoy sufficient economic resources to satisfy educational and informational and health needs and are able to delay age at first pregnancy. The main direct causes of maternal morbidity and mortality include obstetric haemorrhage, hypertensive disorders, and sepsis during pregnancy (WHO 2016a). Indirect causes include the following: Women with untreated AIDS-related health conditions have an eightfold increased risk of maternal death than HIV negative women (Kendall et al. 2014, p. S251). In the absence of HIV approximately 281,500 global maternal deaths could have been avoided in 2008 (Hogan et al. 2010). Inadequate access to skilled maternal healthcare, especially emergency obstetric care is a major problem. This is due to unavailability of adequate maternal health services, distances, costs in travelling to maternal health services, and inadequately skilled healthcare providers.

Unsafe abortion's impact on maternal health

Unsafe abortion is a 'preventable pandemic' (Grimes et al. 2006, p. 1). Unsafe abortion is on a continuum unsafe care, with abortions conducted either by unskilled individuals or in an unsafe medical environment (Sedgh and Adewole 2016). Unsafe abortion contributes significantly to maternal morbidity and mortality. An estimated 19–20 million unsafe abortions occur annually, with 97% in LMIC. An estimated 68,000 women die annually from unsafe abortion and millions have complications, often with long-term, poor maternal health consequences. Key outcomes from unsafe abortion are haemorrhage, infection, and poisoning from substances used in induced abortions (Grimes et al. 2006, p. 1). There has been a decline in abortions, in Eastern Europe, but no significant decreases in LMIC (Singh et al. 2017, p. 5). An estimated 55% were safe; 31% less safe and 14% least safe. Countries

with the most restrictive abortion laws were responsible for the majority of unsafe abortions (Singh et al. 2017, p. 7).

Men and maternal health

While maternal mortality and morbidity affects only women physically and emotionally, it should not be regarded as only a women's health issue. The importance of men as partners in maternal health (attending antenatal, childbirth, and postnatal care) cannot be overemphasized (Brittain et al. 2015; Sibanda et al. 2017; Konje et al. 2018) (Box 10.2.12).

Box 10.2.12 Key messages

- Maternal mortality highlights health inequity between women in different world regions.
- Most maternal deaths are preventable through good quality, skilled maternal care, and availability of emergency care.
- Improved access to contraception and safe abortion would decrease maternal deaths.
- Men as partners in maternal health should be promoted.

Interventions and initiatives to address women and men's health

Biomedical interventions

HIV

Biomedical interventions to decrease HIV infection have been the focus of much research. These include promotion of male and female condoms as barrier methods to protect against HIV acquisition.

Biologically, uncircumcised men have higher female-to-male HIV infection risk than circumcised men. Randomized controlled trials in Kenya, Uganda, and South Africa show that medical male circumcision can provide partial protection against risk of female-male HIV transmission by up to 60% as well as to penetrative partners in same-sex male intimate relationships. The WHO and UNAIDS have recommended medical male circumcision as part of an HIV prevention strategy for men, particularly in

countries with a high HIV prevalence and low circumcision rates (WHO/UNAIDS 2008). ART's critical role in HIV treatment and in 'treatment as prevention', through pre-exposure prophylaxis, reducing HIV infectivity and transmissibility. People living with HIV (PLWH) with undetectable viral loads, have minimal risks of transmission (McNairy and El-Sadr 2014). Research for a microbicide, applied to the vaginal or anal tract to reduce HIV transmission, has had limited success in women, over the past 30 years. However there is active ongoing research (Karim et al. 2010; Doggett et al. 2015). This has highlighted HIV prevention's gender-related factors—gender hierarchies in sexual decision-making and gender norms influencing younger women's lack of adherence to microbicides (Doggett et al. 2015). There is also research for a vaccine to prevent or ameliorate the effects of HIV transmission (Bekker and Gray 2017).

CVD

Key changes that have occurred to address gender issues in CVD. These include addressing knowledge among individuals and in communities and important health system changes in the delivery of care by health professions. In the United States in 2014, overall male mortality from CVD overtook female mortality for the first time since 1984. This is attributable to increased awareness of women's symptoms, presentation, and appropriate treatment for possible heart attacks (Aggarwal et al. 2018, p. 2). Primary and secondary prevention has been largely responsible for a decline in CVD over the past decades (Appelman et al. 2015). Since 2014, an American Heart Association consensus statement has recommended the full battery of non-invasive diagnostic tests for women with suspected heart attack symptoms (Aggarwal et al. 2018, p. 4).

Maternal health

Measures addressing poor maternal health have been introduced at an individual, community, and health systems level. These include improving knowledge of skilled care antenatally, during birth and postnatally at an individual, partner, and community level; and training of health professionals, particularly in rural areas to deal with complications and emergency care and ensuring emergency care facilities are available and well-functioning. Programmes like *Momconnect* in South Africa (Barron et

al. 2016); and Maternal Waiting houses for pregnant women, operating in a variety of LMIC settings such as Lesotho, Cambodia, Malawi, Nigeria, Papua New Guinea, Timor, and Zambia (WHO 1996; Wild et al. 2012; Siv et al. 2014) are examples of health systems initiatives to improve maternal healthcare. Given that apart from HIV and respiratory illnesses, health-system-related causes are the commonest preventable maternal deaths causes, health system initiatives need to be prioritized.

Mental health

Numerous medical and pharmaceutical interventions are effective in treating MNS problems. However psychological and social interventions are important to prevent and treat MNS disorders (Patel et al. 2016).

Transformative gendered initiatives to improve women and men's health

Multiple theories and interventions of change to address the impact of gendered factors on health have emerged. Gupta's (2000) work on HIV introduced interventions along a continuum, to address the gender aspects of the epidemic. These include interventions that are: (1) gender exploitative; (2) gender neutral; (3) gender sensitive, or accommodating; or (4) gender transformative (Rolleri 2014). An example of gender exploitative programming in HIV prevention is one that uses an aggressive image of masculinity to promote condoms, taking advantage of patriarchal masculinity norms, reinforcing gender stereotypes (Rolleri 2014). Gender-neutral programmes take account of specific needs of women or men but miss opportunities to address gender determinants of health. An example is where fidelity messages are promoted to prevent HIV, without reference to gendered power relations in many women and men's intimate sexual relationships (Gupta 2000). Gender-sensitive approaches address specific needs, such as introducing male-friendly reproductive healthcare services at convenient times for men. They recognize and respond to gender norms and inequities, but do not work structurally changing harmful norms or power imbalances (Rolleri 2014). Gender transformative programmes aim to increase awareness of inequitable gender norms, question these and promote, individually and structurally, equitable gender norms, and relations (Rolleri 2014). An example is changing inequitable norms in which men hold power

in decision-making in sexual relationships to those in which women and men engage in equitable decision-making (Greene and Levack 2010).

Individual agency

Some women living in contexts in which hegemonic patriarchal patterns prevail, engage in a variety of individual actions to assert their agency (Jewkes and Morrell 2012; Stern et al. 2016). Jewkes and Morrell's (2012) research in rural South Africa shows that women resist male control and potential male IPV. Some women avoid initiating such relationships, while others accommodate them but simultaneously seek respect and non-violence. Some women in countries such as South Africa and Botswana avoid formal marriage, in part to avoid male control over their lives (Cooper et al. 2013).

While many men conform to dominant masculinity associated with high risk sexual behaviour, some have adopted health-seeking behaviours protective towards their own and their partners' health (Stern and Cooper 2014; Morrell et al. 2016). As not all masculinities are equally powerful, some men go against prevailing gender norms even if they face sanction. Some men act out their masculinity individually through reflecting discomfort with harmful hegemonic forms through caring fatherhood (Morrell et al. 2016). A study among 18–49-year-old men in South Africa found 84% reported helping their children with school homework; only just over half (54%) reported discussing personal matters with their children. Men who were more gender egalitarian in their attitudes were more likely to be caring fathers (Morrell et al., 2016, p. 81).

Structural changes

Sexual and Reproductive Rights are a fundamental human right. Less restrictive abortion laws decrease unsafe abortions. The South African Choice of Termination Act of 1996 is a good example of this, with abortion-related morbidity and mortality decreasing by 91% between 1997 and 2002 (Cooper et al. 2016). Legalizing abortion is a first step, with social stigma, political will for implementation and healthcare provider unwillingness to provide abortions, remaining major problems (Harries et al. 2014). As the US and South African experience show, it is important even when abortion is legal to prevent rolling back gains made (Cooper et al. 2016). In all countries, but particularly those where abortion is illegal, post-abortion care

is crucial in protecting women's health (Singh et al. 2017). A 16-country study showed that factors most influencing access to safe abortion services were having abortion services integrated into comprehensive SRH care, strong country-based women's rights organizations, supported by international agencies and professional medical associations; involvement of non-governmental organizations in provision of services; task sharing among providers so that provision of abortion services is more 'normalized'; and availability of new and less complicated techniques (such as the use of misoprostol and misoprostol-mifepristone in medical abortion). Political will was a key factor for legalization, provision, and ensuring access to abortion services (Chavkin et al. 2018). Although the need for abortion will always remain, better knowledge and access to contraception is critical in preventing pregnancy in the first place.

A number of social interventions prevent and treat MNS problems. Initiatives with good results include legislation measures to restrict access to means of self-harm/suicide and reduce the availability of and demand for alcohol. At health service and community levels, parenting programmes and life skills training in schools can be effective in building social and emotional competencies. At the health service level, self-management, web-based psychological therapy for depression and anxiety disorders; and primary care and community outreach programmes are positive initiatives. This may include psychological and pharmacological management for specific conditions.

Hospitals play a role in providing specialist services for MNS disorders, providing for severe emergencies and for integrating mental healthcare into broader other healthcare services. At a specialist care level, the management of complex and severe cases of MNS disorders, and the integration of primary care level interventions in other medical care platforms, such as maternal and HIV care have shown positive results. The PRIME programme has focused on a district health systems level. It involved partnerships with stakeholders in Ethiopia, India, Kenya, Nepal, South Africa, and Uganda. It developed a district health framework for dealing with mental health problems. Research to test implementation of pilot projects and estimate costs and evaluate their outcomes and processes was conducted. It was aimed at health ministries' scale-up of mental health services at district level in order to improve individuals' and communities' psychological well-being. It has prioritized LMICs, where there is a significant gap in addressing

prevention, treatment, and care for mental health problems (Lund et al. 2016, p. S1). These would be cost-effective interventions, if scaled up. Less than 1% of development assistance for health and government spending on health is allocated to MNS problems, estimated at US$1–2 per capita per year in LMIC, and US$ 3–5 per capita in upper-middle-income countries. National Health Insurance or Universal coverage policies should include addressing these issues through allocation of public health resources (Patel et al. 2016, p. 1).

Advocacy and group/community initiatives

HIV and GBV

A number of examples of gender transformative programmes have been introduced to address HIV and violence against women.

A gender transformative HIV and GBV prevention programme, the *Stepping Stones* programme (Welbourn 1995), was developed in Uganda and adapted and used in 40 countries. Targeting women and men, it sought to improve sexual health through building stronger, more gender-equitable relationships (Jewkes et al. 2007). A randomized controlled trial evaluating the *Stepping Stones* programme in rural South Africa among young men and women (15–26 years old), showed of short term success in reducing sexual risk-taking and violence perpetration by young male participants. Similar changes did not occur among young female participants. This speaks to the difficulties women face in implementing individual behaviour change, in the absence of community or structural changes in gender norms and relationships (Jewkes et al. 2007).

The International AIDS Alliance strategy (2016–2020) on HIV, health and rights conducts a programme based on a ‘people centred’ approach. It works through community action in partnership with PLWH, based on the premise that they are best able to assess their own needs. It promotes equity through focusing on those systematically excluded or marginalized; activities that are criminalized (such as women who use drugs in Kenya and sex workers); and people with diverse sexual orientations or gender identities. It analyses the social determinants, including gender’s, effect on peoples’ lives, and health (International AIDS Alliance 2017, p. 1). Its Vihaan project in India, promotes community-based organizations and PLWH taking a leadership role in care, treatment, and support programmes (International AIDS

Alliance 2017, p. 3). In Kenya work with women using drugs involves issues of concern for young women as lovers, mothers, keeping safe, HIV, SRH and rights and childcare (International AIDS Alliance 2017, p. 2).

Recent transformative research interventions aimed at girls involve cash transfer programmes to address one of the root economic HIV vulnerabilities in young women. These were implemented and evaluated in South Africa (Cluver et al. 2013) and other southern and eastern African countries affected by HIV (Samson 2009). They showed some success in encouraging younger women to complete schooling, be economically independent and engage in more positive sexual health behaviours.

Over the past few decades, a substantial number of male-oriented initiatives have addressed the gender effects of HIV and GBV. These focused on changing patriarchal masculinities among men that are harmful to men and women's health.

Examples of transformative individual, community, and structural initiatives that engage primarily with men on GBV, HIV, and fatherhood include the *Sonke Gender Justice Network* and *Primundo* (Dworkin et al. 2012; Ricardo et al. 2011). Sonke Gender Justice Network's the 'One Man Can' campaign, is a rights-based gender equality health initiative, recognized as best practice by the WHO. Its programme has been translated into over 10 languages and subsequently implemented across the African continent (Van den Berg et al. 2013). It organized men's community-based workshops to reflect on human rights, women's rights, how masculinities are defined and practised and can be more caring (Dworkin et al. 2013). It assisted establish community action teams to address GBV and HIV (Dworkin et al. 2012). Evaluations showed increased actions among male participants against GBV, engaging in safer sex and accessing HIV testing and ART (Dworkin et al. 2013).

Pharmaceuticals and patent laws—structural initiatives

In India, people living with HIV struggled and won changes in India's medicine patent laws to allow access to cheaper antiretroviral medication (People's Health Movement 2018). The Treatment Action Campaign in South African launched and won a similar struggle as well as pressurizing the government into providing antiretroviral treatment in public sector health facilities (Colvin 2014). These initiatives confronted structural determinants

of health and involved the support of health professionals and healthcare providers.

Mental health

The ‘Emerging mental health systems in LMICs’ (Emerald) programme has aimed at improving health outcomes of people with MNS problems in Ethiopia, India, Nepal, Nigeria, South Africa, and Uganda. It generated evidence that was used to boost health system performance in providing improved mental healthcare and enhance the capacity of policymakers, and planners. It focused on strengthening health service use and caregiver capacity to support efforts within mental health systems. The programme sought to deal with stigma and discrimination as key barriers to health-seeking behaviour and provision of mental health services (Semrau et al 2015).

Other informative initiatives that could inform women and men’s health

The *Genuino Clandestino* is an Italian network of collectives, associations, and individuals engaged in dealing with the struggles for food and control over land. Founded in 2009, with origins in small-scale producers of food, it links its activities to the specific needs of local communities, promotes small-scale organic food markets, certification of products, self-organized kitchens, and land access. Its strong grassroots community base has enabled it to take on local struggles important to female and male nutrition and land issues and to confront the interests of large multinational food and land companies (People’s Health Movement 2018).

Changing gender power relations

There has been ongoing tension on place of men and boys in changing gender power relations (Tolhurst et al. 2012). Initial approaches for male involvement stressed their strategic importance for women’s health. This raised questions about whether gender was only about women’s interests. Later approaches emphasized male involvement as important in its own right and to improve female health (Tolhurst et al. 2012). These approaches argued that transforming gender relations acts in the interests of both women and men’s health. Yet others have drawn critical attention to women and men

not being equally vulnerable or privileged. Hence addressing gender inequity in power relations likely involves the relinquishment of personal and structural hegemonic patriarchal forms of power and privilege, vested in institutions such as the legal, political, and economic system (Box 10.2.13).

Box 10.2.13 Key messages

- Biomedical interventions have been important in HIV, CVD, mental health, and maternal health improvements.
- There are multiple theories of gender change, with gender transformation being the most far-reaching.
- Individual agency, structural changes and advocacy and community initiatives are important in transforming gender power relations that can improve women and men's health and healthcare.
- Several evaluated transformative initiatives have shown promise in improving women and men's health.

In conclusion

Researchers, practitioners, policymakers, and policy implementers should recognize health as a complex interplay between many determinants, of which gender is one key social determinant. These operate at individual, interpersonal, systems (such as the healthcare system) and structural levels. This has key implications for how we conceptualize health problems, how we address them, how we approach research and design a health system and services. It involves work intersectorally to promote facilitators of health and engage in advocacy activities to transform the legal, policy, social, political, and economic structures, in which health is embedded.

ACKNOWLEDGEMENTS

Huge appreciation goes to Tanya Jacobs and Courtney Sprague who previewed this chapter at different stages and provided invaluable informal reviews and comments. We cannot thank you enough for your insights and encouragement.

Our thanks to:

Ntobeko Nywagi who assisted in searching for data on Life expectancy and Cardiovascular disease;
To Prof. Finn Diderichsen for permission to adapt his and his colleagues framework that appears as figure 2.2 in Sprague (2018, p 67), for use in Figure 10.2.2 in this Chapter;
To Hillevi Sellen for her patience in extending deadlines for the chapter.

The first author wishes:

To thank to Justine Evans for beginning to teach me when and how to say 'no' or 'I need a deadline extension'!

I express my huge appreciation to Jon, Tara, and Mikey for encouraging me to believe 'I can do this' in writing this chapter.

REFERENCES

Abdool Karim, Q., Sibeko, S., and Baxter, C. (2010). Preventing HIV infection in women: a global health imperative. *Clinical Infectious Diseases*, 50(s3), S122–9.

Aggarwal, N.R., et al. (2018). Sex differences in ischemic heart disease. *Circulation: Cardiovascular Quality and Outcomes*, 11(2), e004437.

Airhihenbuwa, C.O., et al. (2016). Claim your space: leadership development as a research capacity building goal in global health. *Health Education & Behavior*, 43(1_suppl), 17S–24S.

Andrews, G., Hall, W., Teesson, M., and Henderson, S. (1999). *The Mental Health of Australians*. Canberra: Mental Health Branch, Commonwealth Department of Health and Aged Care.

Appelman, Y. et al. (2015). Sex differences in cardiovascular risk factors and disease prevention. *Atherosclerosis*, 241(1), 211–18.

Barron, P., Pillay, Y., Fernandes, A., Sebidi, J. and Allen, R. (2016). The MomConnect mHealth initiative in South Africa: early impact on the supply side of MCH services. *Journal of Public Health Policy*, 37(2), 201–12.

Bekker, L.G. and Gray, G.E. (2017). Hope for HIV control in southern Africa: the continued quest for a vaccine. *PLoS Medicine*, 14(2), e1002241.

Van den Berg, W., Hendricks, L., Hatcher, A., Peacock, D., Godana, P. and Dworkin, S. (2013). 2013. 'One Man Can': shifts in fatherhood beliefs and parenting practices following a gender-transformative programme in Eastern Cape, South Africa. *Gender & Development*, 21(1), 111–25.

Borowsky, S.J., Rubenstein, L.V., Meredith, L.S., Camp, P., Jackson-Triche, M. and Wells, K.B. (2000). Who is at risk of nondetection of mental health problems in primary care? *Journal of General Internal Medicine*, 15(6), 381–8.

Brittain, K., et al. (2015). Pregnant women's experiences of male partner involvement in the context of prevention of mother-to-child transmission in Khayelitsha, South Africa. *AIDS Care—Psychological and Socio-Medical Aspects of AIDS/HIV*, 27(8), 1020–4.

Bronfenbrenner, U. (1979). *The Ecology of Human Development: Experiments by Nature and Design.* Boston, MA: Harvard University Press.

Byrne, S. and McLean, N. (2001). Eating disorders in athletes: a review of the literature. *Journal of Science and Medicine in Sport*, 4(2), 145–59.
Castillo, R.J., et al. (2007). *Diagnostic and Statistical Manual of Mental Disorders*. Washington, DC: American Psychiatric Association Press.
Chandra-Mouli, V. et al. (2017). A never-before opportunity to strengthen investment and action on adolescent contraception, and what we must do to make full use of it. *Reproductive Health*, 14(1), 85.
Chavkin, W. et al. (2018). Implementing and expanding safe abortion care: an international comparative case study of six countries. *International Journal of Gynecology and Obstetrics*, Supplement, 3–11.
Christodoulou, J. and Zobnina, A. (2009). Glossary of gender-related terms. *Mediterranean Institute of Gender Studies (MIGS)*. [Online] Available at: https://eige.europa.eu/library/resource/aleph_eige000000498
Cluver, L., et al. (2013). Child-focused state cash transfers and adolescent risk of HIV infection in South Africa: a propensity-score-matched case-control study. *The Lancet Global Health*, 1(6), e362–70.
Coles, T. (2009). Negotiating the field of masculinity: the production and reproduction of multiple dominant masculinities. *Men and Masculinities*, 12(1), 30–44.
Colvin, C.J. (2014). Evidence and AIDS activism: HIV scale-up and the contemporary politics of knowledge in global public health. *Global Public Health*, 9(1–2), 52–72.
Connell, R. (2012). Conceptualizing the issue, in local and world perspective. *Social Science & Medicine*, 74(11), 1675–83.
Connell, R.W. and Messerschmidt, J.W. (2005). Hegemonic masculinity rethinking the concept. *Gender and Society*, 19(6), 829–59.
Cooper, D., et al. (2016). Coming of age? Women's sexual and reproductive health after twenty-one years of democracy in South Africa. *Reproductive Health Matters*, 24(48), 79–89.
Cooper, D., Moore, E., and Mantell, J.E. (2013). Renegotiating intimate relationships with men: how HIV shapes attitudes and experiences of marriage for South African women living with HIV: 'Now in my life, everything I do, looking at my health'. *Acta Jurid (Cape Town)*, **2013**, 218–38.
Cornell, M., et al. (2017). Twelve-year mortality in adults initiating antiretroviral therapy in South Africa. *Journal of the International AIDS Society*, 20(1), 21902.
Crea, F., Battipaglia, I., and Andreotti, F. (2015). Sex differences in mechanisms, presentation and management of ischaemic heart disease. *Atherosclerosis*, 241(1), 157–68.
Creighton, G. and Oliffe, J.L. (2010). Theorizing masculinities and men's health: a brief history with a view to practice. *Health Sociology Review*, 19(4), 409–18.
CSDH (2008). *Closing the Gap in a Generation. Closing the Gap in a Generation. Health Equity Through Action on the Social Determinants of Health. Final Report of the Commission on Social Determinants of Health*, p. 246. [Online] Available at: https://www.who.int/social_determinants/final_report/csdh_finalreport_2008.pdf
Currin, J.B., Hayslip, B., and Temple, J.R. (2011). The relationship between age, gender, historical change, and adults' perceptions of mental health and mental health services. *The International Journal of Aging and Human Development*, 72(4), 317–41.
Dellar, R.C., Dlamini, S., and Karim, Q.A. (2015). Adolescent girls and young women: key populations for HIV epidemic control. *Journal of the International AIDS Society*, 18(2), 64–70.
Dick, B., et al. (2015). Health for the world' s adolescents: a second chance in the second decade. *Journal of Adolescent Health*, **56**, 4–7.
Doggett, E.G., Lanham, M., Wilcher, R., Gafos, M., Karim, Q.A. and Heise, L. (2015). Optimizing HIV prevention for women: a review of evidence from microbicide studies and considerations for gender-sensitive microbicide introduction. *Journal of the International AIDS Society*, 18(1), 20536.

Doyal, L. (1995). *What Makes Women Sick: Gender and the Political Economy of Health*. New Brunswick, NJ: Rutgers University Press.
Dworkin, S.L., Colvin, C., Hatcher, A. and Peacock, D. (2012). Men's perceptions of women's rights and changing gender relations in South Africa: lessons for working with men and boys in HIV and antiviolence programs. *Gender & Society*, 26(1), 97–120.
Dworkin, S.L., Treves-Kagan, S. and Lippman, S.A. (2013). Gender-transformative interventions to reduce HIV risks and violence with heterosexually-active men: a review of the global evidence. *AIDS and Behavior*, 17(9), 2845–63.
Elliott, M. (2013). Gender differences in the determinants of distress, alcohol misuse, and related psychiatric disorders. *Society and Mental Health*, 3(2), 96–113.
Global Burden of Disease Study 2013 Collaborators, et al. (2015). Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013. *The Lancet*, 386(9995), 743–800.
Greene, M., and Levack, A. (2010). *Gender Synchronizing Strategies: A Cooperative Model for Improving Reproductive Health and Transforming Gender Relations.* Washington DC: Population Reference Bureau.
Grimes, D.A. et al. (2006). Unsafe abortion: the preventable pandemic. *The Lancet Sexual and Reproductive Health Series*. [Online] Available at: https://www.who.int/reproductivehealth/publications/general/lancet_4.pdf?ua=1
Gupta, G.R. (2000). Gender, sexuality, and HIV/AIDS: the what, the why, and the how. *Canadian HIV/AIDS Policy & Law Review/Canadian HIV/AIDS Legal Network*, 5(4), 86–93.
Gutmann, M. (2007). *Fixing Men. Sex, Birth Control, and AIDs in Mexico*, Berkeley, CA: University of California Press.
Hammarström, A. (2018). How gender theories are used in contemporary public health research. *International Journal for Equity in Health*, 17(1), 34.
Hankivsky, O. (2014). Rethinking care ethics: on the promise and potential of an intersectional analysis. *American Political Science Review*, 108(2), 252–64.
Harper, K.N. (2015). Two recent studies find Depo-Provera use increases HIV risk. *AIDS*, 29(9), N1.
Harries, J., et al. (2014). Conscientious objection and its impact on abortion service provision in South Africa: a qualitative study. *Reproductive Health*, 11(1), 16.
De Hert, M., et al. (2009). Cardiovascular disease and diabetes in people with severe mental illness. *European Psychiatry*, 24(6), 412–24.
Hogan, M.C., et al. (2010). Maternal mortality for 181 countries, 1980–2008: a systematic analysis of progress towards Millennium Development Goal 5. *The Lancet*, 375(9726), 1609–23.
International AIDS Alliance (2017). *Putting People at the Heart of the HIV Response*. [Online] Available at: https://frontlineaids.org/wp-content/uploads/2019/02/The-Alliances-person-centred-approach.pdf
Jewkes, R., Nduna, M., Levin, J., Jama, N. and Dunkle, K. (2007). *Evaluation of Stepping Stones: A Gender Transformative HIV Prevention Intervention.* [Online] Available at: https://steppingstonesfeedback.org/wp-content/uploads/2016/10/SS_MRC_GenderTransformativeEvaluation_SouthAfrica_2007.pdf
Jewkes, R. (2014). (How) can we reduce violence against women by 50% over the next 30 years? *PLoS Medicine*, 11(11), e1001761.
Jewkes, R. (2010). Emotional abuse: a neglected dimension of partner violence. *The Lancet*, 376(9744), 851–2.
Jewkes, R., et al. (2015). Hegemonic masculinity: combining theory and practice in gender interventions. *Culture, Health and Sexuality*, **17**, 96–111.
Jewkes, R. (2002). Preventing sexual violence: a rights-based approach. *The Lancet*, 360(9339), 1092–3.

Jewkes, R. and Morrell, R. (2010). Gender and sexuality: emerging perspectives from the heterosexual epidemic in South Africa and implications for HIV risk and prevention. *Journal of the International AIDS Society*, **13**, 6.
Jewkes, R. and Morrell, R. (2012). Sexuality and the limits of agency among South African teenage women: theorising femininities and their connections to HIV risk practises. *Social Science and Medicine*, 74(11), 1729–37.
Judd, F., Armstrong, S., and Kulkarni, J. (2009). Gender-sensitive mental health care. *Australasian Psychiatry*, 17(2), 105–11.
Karim, Q.A., et al. (2010). Effectiveness and safety of tenofovir gel, an antiretroviral microbicide, for the prevention of HIV infection in women. *Science*, 329(5996), 1168–74.
Kendall, T., et al. (2014). Eliminating preventable HIV-related maternal mortality in sub-Saharan Africa: what do we need to know? *Journal of Acquired Immune Deficiency Syndromes (1999)*, 67 Suppl 4, S250–8.
Kengne, A.P., et al. (2013). Cardiovascular diseases and diabetes as economic and developmental challenges in Africa. *Progress in Cardiovascular Diseases*, 56(3), 302–13.
Kessler, R.C., et al. (2005). Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the National Comorbidity Survey Replication. *Archives of General Psychiatry*, 62(6), 617–27.
Konje, E.T., et al. (2018). Missed opportunities in antenatal care for improving the health of pregnant women and newborns in Geita district, Northwest Tanzania. *BMC Pregnancy and Childbirth*, 18(1), 394.
Krieger, N. (2008). Proximal, distal, and the politics of causation: what's level got to do with it? *American Journal of Public Health*, 98(2), 221–30.
Krug, E.G., et al. (2002). The world report on violence and health. *The Lancet*, 360(9339), 1083–8.
Larson, E., et al. (2016). 10 Best resources on. . . intersectionality with an emphasis on low- and middle-income countries. *Health Policy and Planning*, 31(8), 964–9.
Lund, C., Tomlinson, M., and Patel, V. (2016). Integration of mental health into primary care in low-and middle-income countries: the PRIME mental healthcare plans. *The British Journal of Psychiatry*, 208(s56), s1–s3.
Madiba, S. and Canti-Sigaqa, V. (2012). Barriers to participate in support groups for people living with HIV: a qualitative study with men receiving antiretroviral treatment in a HIV clinic in Mthatha, South Africa. *Global Journal of Health Science*, 4(6), 119–28.
Mannell, J. (2010). Gender mainstreaming practice: considerations for HIV/AIDS community organisations. *AIDS Care—Psychological and Socio-Medical Aspects of AIDS/HIV*, 22(Suppl. 2), 1613–19.
Martin, L.A., Neighbors, H.W. and Griffith, D.M. (2013). The experience of symptoms of depression in men vs. women: analysis of the National Comorbidity Survey Replication. *JAMA Psychiatry*, 70(10), 1100–6.
Matheson, F.I., et al. (2015). Where did she go? The transformation of self-esteem, self-identity, and mental well-being among women who have experienced intimate partner violence. *Women's Health Issues*, 25(5), 561–9.
Mayosi, B.M., et al. (2012). Health in South Africa: changes and challenges since 2009. *The Lancet*, 380(9858), 2029–43.
Mayosi, B.M., et al. (2009). The burden of non-communicable diseases in South Africa. *The Lancet*, 374(9693), 934–47.
McNairy, M.L. and El-Sadr, W.M. (2014). A paradigm shift: focus on the HIV prevention continuum. *Clinical Infectious Diseases*, 59(suppl_1), S12–S15.
Mendis, S. and Chestnov, O. (2014). The global burden of cardiovascular diseases: a challenge to improve. *Current Cardiology Reports*, 16(5), 486.
Meyer, S. (2016). Still blaming the victim of intimate partner violence? Women's narratives of victim desistance and redemption when seeking support. *Theoretical Criminology*, 20(1), 75–90.

Monyane, C. (2013). Is ukuthwala another form of 'forced marriage'? *South African Review of Sociology*, **44**, 64–82.
Morgan, R., et al. (2016). How to do (or not to do) gender analysis in health systems research. *Health Policy and Planning*, 31(8), 1069–78.
Morrell, R., et al. (2016). Fathers who care and those that don't: men and childcare in South Africa. *South African Review of Sociology*, 47(4), 80–105.
Murray, C.J. and Lopez, A.D. (1997). Alternative projections of mortality and disability by cause 1990–2020: Global Burden of Disease Study. *The Lancet*, 349(9064), 1498–504.
Murthy, R.S. (2004). Psychiatric comorbidity presents special challenges in developing countries. *World Psychiatry*, 3(1), 28.
Norman, R., et al. (2010). Interpersonal violence: an important risk factor for disease and injury in South Africa. *Population Health Metrics*, 8(1), 32.
Oliffe, J.L., et al. (2014). Gay men and intimate partner violence: a gender analysis. *Sociology of Health and Illness*, 36(4), 564–79.
Patel, V., et al. (2016). Addressing the burden of mental, neurological, and substance use disorders: key messages from Disease Control Priorities. *The Lancet*, 387(10028), 1672–85.
People's Health Movement (2018). *Global Health Watch. An Alternative Health Report*. London: Zed books.
Phillips, S.P. (2006). Risky business: explaining the gender gap in longevity. *Journal of Men's Health & Gender*, 3(1), 43–6.
Pietilä, I. and Rytkönen, M. (2008). 'Health is not a man's domain': lay accounts of gender difference in life-expectancy in Russia. *Sociology of Health and Illness*, 30(7), 1070–85.
Power, J. (2015). Microbicides and HIV prevention in women: the state of research. *HIV Australia*, 13(2), 41.
Prince, M., Patel, V., Saxena, S., Maj, M., Maselko, J., Phillips, M.R. and Rahman, A. (2007). No health without mental health. *The Lancet*, 370(9590), 859–77.
Ramjee, G. and Daniels, B. (2013). Women and HIV in sub-Saharan Africa. 1–9.
Ricardo, C., Eads, M., and Barker, G. (2011). *Engaging Boys and Young Men in the Prevention of Sexual Violence*. [Online] Available at: https://svri.org/sites/default/files/attachments/2016-03-21/menandboys.pdf
Riska, E. (2010). Coronary heart disease: gendered public health discourses. In: Kuhlmann, E. and Annandale, E. (eds.) *The Palgrave Handbook of Gender and Healthcare*. London: Palgrave Macmillan, pp. 158–71.
Rolleri, L.A. (2014). *Gender Transformative Programming in Adolescent Reproductive and Sexual Health: Definitions, Strategies, and Resources Applying a Gender Transformative Approach.* Ithaca: ACT for Youth Center for Excellence. Cornell University, University of Rochester and New York School Safety.
Rosenfield, S., Lennon, M.C., and White, H.R. (2005). The self and mental health: self-salience and the emergence of internalizing and externalizing problems. *Journal of Health and Social Behavior*, 46(4), 323–40.
Rosenfield, S. and Smith, D. (2010). Gender and mental health: do men and women have different amounts or types of problems? In: Scheid, T.L. and Brown, T.N. (eds.) *A Handbook for the Study of Mental Health: Social Contexts, Theories, and Systems* (2nd ed.). Cambridge: Cambridge University Press.
Samson, M. (2009). The impact of social transfers on growth, development, poverty and inequality in developing countries. In: Townsend, P. (ed.) *Building Decent Societies*. London: Palgrave Macmillan, pp. 122–50.
Sedgh, G. and Adewole, I. (2016). Insights from an expert group meeting on the definition and measurement of unsafe abortion. *International Journal of Gynecology & Obstetrics*, 134(1), 104–6.

Seedat, M. et al. (2009). Violence and injuries in South Africa: prioritising an agenda for prevention. *The Lancet*, 374(9694), 1011–22.
Semrau, M., et al. (2015). Strengthening mental health systems in low-and middle-income countries: the Emerald programme. *BMC Medicine*, 13(1), 79.
Shefer, T. (2016). Resisting the binarism of victim and agent: critical reflections on 20 years of scholarship on young women and heterosexual practices in South African contexts. *Global Public Health*, 11(1–2), 211–23.
Sibanda, E. et al. (2017). Knowledge, attitudes and practices of health professionals in public health institutions on emergency contraception in Pietermaritzburg, KwaZulu-Natal Province, South Africa. *South African Journal of Obstetrics & Gynaecology*, 23(1), 7–11.
Singh, S., Remez, L., Sedgh, G., Kwok, L., and Onda, T. (2017). *Abortion Worldwide 2017. Uneven Progress and Unequal Access*. New York. [Online] Available at: https://www.guttmacher.org/report/abortion-worldwide-2017
Siv, S., Teap, S., Houy, C., Ray, S., Edvaredsen, O., and Ingstad, B. (2014). *Safe Motherhood: A Qualitative Study on the Impact of Maternal Waiting Houses in Rural Areas of Battambang and Pailin Provinces, Cambodia. End-of-Project Report*. Battambang: Trauma Care Foundation.
Smith, D.T. and Hemler, J. (2014). *Constructing Order. Social Issues in Diagnosis: An Introduction for Students and Clinicians.* Baltimore, MD: JHU Press.
Smith, D.T., Mouzon, D.M., and Elliott, M. (2016). Reviewing the assumptions about men's mental health: an exploration of the gender binary. *American Journal of Men's Health*, 12(1), 78–89.
Spence, J.D. and Pilote, L. (2015). Importance of sex and gender in atherosclerosis and cardiovascular disease. *Atherosclerosis*, 241(1), 208–10.
Sprague, C. (2018). *Gender and HIV in South Africa. Advancing Women's Capabilities.* New York: Palgrave Macmillan.
Stern, E., Buikema, R., and Cooper, D. (2016). South African women's conceptualisations of and responses to sexual coercion in relation to hegemonic masculinities. *Global Public Health*, 11(1–2), 135–52.
Stern, E. and Cooper, D. (2014a). Experiences and conceptualizations of sexual debut from the narratives of South African men and women in the context of HIV/AIDS. *African Journal of AIDS Research*, 13(2), 121–31.
Stern, E., Rau, A., and Cooper, D. (2014b). Sexual and reproductive health perceptions and practices as revealed in the sexual history narratives of South African men living in a time of HIV/AIDS. *Journal of Social Aspects of HIV/AIDS*, 11(1), 233–44.
Tasker, C., Davidow, A., Roche, N.E., and Chang, T.L.Y. (2018). Depo-Provera alters cervical immune responses and increases HIV infectivity ex vivo: a longitudinal study.
Taylar, J. and Turner, R.J. (2001). A longitudinal study of the role and significance of materring to others for depressive symptoms. *Journal of Health and Social Behavior*, 42(3), 310–25.
Tolhurst, R., et al. (2012). Intersectionality and gender mainstreaming in international health: using a feminist participatory action research process to analyse voices and debates from the global south and north. *Social Science and Medicine*, 74(11), 1825–32.
Torgersen, S., Kringlen, E., and Cramer, V. (2001). The prevalence of personality disorders in a community sample. *Archives of General Psychiatry*, 58(6), 590–6.
United Nations Development Programme (UNDP) (2015). *The Millennium Development Goals Report 2015*. [Online] Available at: https://www.un.org/millenniumgoals/2015_MDG_Report/pdf/MDG%202015%20rev%20%28July%201%29.pdf
Wagstaff, A. et al. (2018). The 2018 health equity and financial protection indicators database overview and insights. *The Lancet*, (18), 1–62.
Welbourn, A. (1995). *Stepping Stones: A Training Package on HIV. AIDS, Communication and Relationship Skills.* London: ActionAid.

Whitacre, C.C. (2001). Sex differences in autoimmune disease. *Nature Immunology*, 2(9), 777–80.
Wild, K., Barclay, L., Kelly, P. and Martins, N. (2012). The tyranny of distance: maternity waiting homes and access to birthing facilities in rural Timor-Leste. *Bulletin of the World Health Organization*, **90**, 97–103.
Witter, S. et al. (2017). Minding the gaps: health financing, universal health coverage and gender. *Health Policy and Planning*, 32(5), 4–12.
Wizemann, T. and Pardue, M.-L. (2001). *Exploring the Biological Contributions to Human Health: Does Sex Matter?* Washington, DC: National Academies Press.
World Health Organization (WHO) (1996). *World Health Organization, 1996. Maternity Waiting Homes: A Review of Experiences*. Geneva: WHO.
WHO/UNAIDS (2008). *Male Circumcision: Global Trends and Determinants of Prevalence, Safety and Acceptability*. Geneva: WHO, p. 7.
World Health Organization (WHO) (2013). *Global and Regional Estimates of Violence Against Women: Prevalence and Health Effects of Intimate Partner Violence and Non-Partner Sexual Violence*. Geneva: WHO, p. 57.
World Health Organization (WHO) (2014). *Trends in Maternal Mortality: 1990–2013. Estimates by WHO, UNICEF, UNFPA, The World Bank and the United Nations Population Division*. Geneva: WHO, p. 56. Available at: http://apps.who.int/iris/bitstream/10665/112682/2/9789241507226_eng.pdf?ua=1
World Health Organization (WHO) (2015). *Trends in Maternal Mortality: 1990 to 2015: estimates by WHO, UNICEF, UNFPA, World Bank Group and the United Nations Population Division*. Geneva: WHO. pp. 1–38. [Online] Available at: http://whqlibdoc.who.int/publications/2010/9789241500265_eng.pdf
World Health Organization (WHO) (2016a). *Health in 2015: from MDGs to SDGs.*, Geneva: WHO.
World Health Organization (WHO) (2016b). *World Health Statistics. Monitoring Health for the SDGs*. Geneva: WHO.
World Health Organization (WHO) (2017). *Country Profiles*. Geneva: WHO.
Zarit, S.H. and Zarit, J. (2011). *Mental Disorders in Older Adults: Fundamentals of Assessment and Treatment.* Guilford: Guilford Press.

10.3

Child health

Tyler Vaivada, Amira Khan, Omar Irfan, and Zulfiqar A. Bhutta

Current situation of child health globally

Within the life course, the period of life before attaining adulthood is divided into three age subgroups based on epidemiology and healthcare needs: the first 5 years (*under-5 children*), the next 5 years (*older children*), and the second decade of life (*adolescents*). The first 5 years of life are further subdivided into the neonatal period (the first 28 days of life), infancy (the first year of life), and preschool years (from 1 to 5 years). This chapter focuses on the health of under-5 children, in addition to older children from 5 to 10 years.

The most important indicators of health status of children relate to their mortality, morbidity, nutrition, growth, and development. Poor health and survival are linked to social and economic development and child mortality is one of the most sensitive barometers of this relationship. An overview of the health situation and inequalities among children in the world is presented in this section.

Global burden and epidemiology of mortality in children under 5

There has been remarkable progress in the reduction of the under-5 mortality burden, with the under-5 mortality rate (U5MR) decreasing from 93 per 1,000 livebirths in 1990 to 39 per 1,000 livebirths in 2017; a reduction of nearly 56% (UN IGME 2018). Much of this success in reducing under-5 deaths is attributable to falling infectious disease mortality. Despite progress made, 5.4 million children died in 2017—a staggering 2.5 million of them were neonates. Nearly half of under-5 deaths occur within the first month of

life, and the risk of death is the highest on the day of birth; it then decreases over the subsequent days, months, and years. Unfortunately, as compared to the 60% reduction in the 1–59-month mortality between 2000 to 2017, mortality in the neonatal period has only decreased by 41% (UN IGME 2018).

More than 80% of all under-5 deaths are clustered in just two regions of the world: sub-Saharan Africa and South Asia. Sub-Saharan Africa leads with the highest global U5MR of 76 deaths per 1,000 live births in 2017 with the chances of an under-5 child dying being 1 in 13. It is noteworthy that globally 50% of all under-5 deaths are concentrated in six countries: India, Nigeria, Ethiopia, Democratic Republic of Congo, Pakistan, and China.

Most children eventually die from only a small number of diseases and conditions. Globally, the leading causes of under-5 mortality are those occurring in the neonatal period, such as preterm birth complications (18%), intrapartum-related events (12%), and neonatal sepsis or meningitis (7%). Although global data on preterm birth rates are sparse, in 2014, an estimated 10.6% (14.84 million) of live births occurred before 37 weeks' gestation, of which 80% were in sub-Saharan Africa and South Asia (Chawanpaiboon et al. 2018). Birthweight is a key indicator of maternal and foetal health during pregnancy, and in 2015, an estimated 14.6% (20.5 million) of newborns had low birthweight (LBW), defined as less than 2500g (Blencowe et al. 2019). LBW can be the result of intrauterine growth restriction, preterm birth, or both. Children born too soon or too small are at considerable increased risk of mortality and morbidity. Pneumonia (16%), diarrhoea (8%), and malaria (5%) are the leading causes of infection-derived post-neonatal under-5 mortality (UN IGME 2018).

The so-called epidemiological transition has a complex, dynamic course which reflects changes in demographic, socioeconomic, cultural, biological, technological, and environmental domains. As a result, important modifications are seen in the patterns of causes of death that usually shift from acute infectious diseases and deficiency-related conditions to non-communicable diseases (NCDs), leading to a relative increase in conditions such as perinatal causes (prematurity, birth asphyxia, and trauma), congenital abnormalities, and injuries. The World Health Organization (WHO) estimated that, between 2000 and 2012, childhood measles, diarrhoea, and

pneumonia mortality reduced by 80, 50, and 40%, respectively (WHO 2018a).

Poverty, maternal health and education, gender, race, and poor-quality healthcare are some of the underlying determinants of under-5 deaths. Levels, trends, and progress in both infant and under-5 mortality are unequally distributed between and within communities, regions, and countries. The likelihood of a child dying in the poorest household, before their fifth birthday, is nearly double that of the most affluent. Similarly, those in rural areas have a 1.5 times greater chance of under-5 death than children in urban settings. More than one-third of under-5 mortality is concentrated in the least developed countries. Unless this inequity is addressed, it is estimated that between 2016 and 2030, 69 million under-5 children will die with nearly half of them in sub-Saharan Africa (Watkins et al. 2016). Figure 10.3.1 illustrates the distribution of under-5 deaths by cause in 2016. Figure 10.3.2 shows trends in child mortality rates.

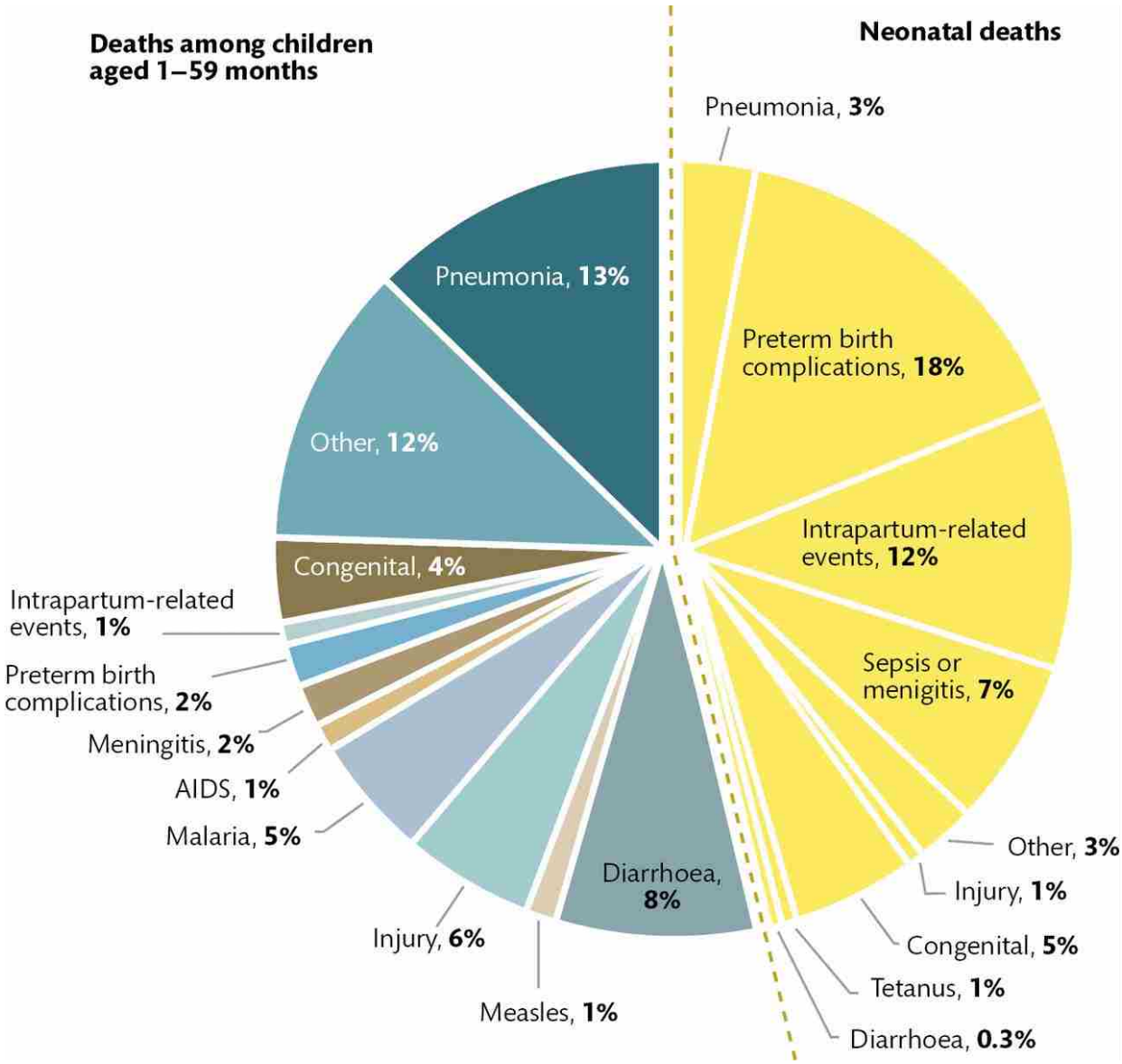


Fig. 10.3.1 Global distribution of deaths among children under-5, by cause, 2016.

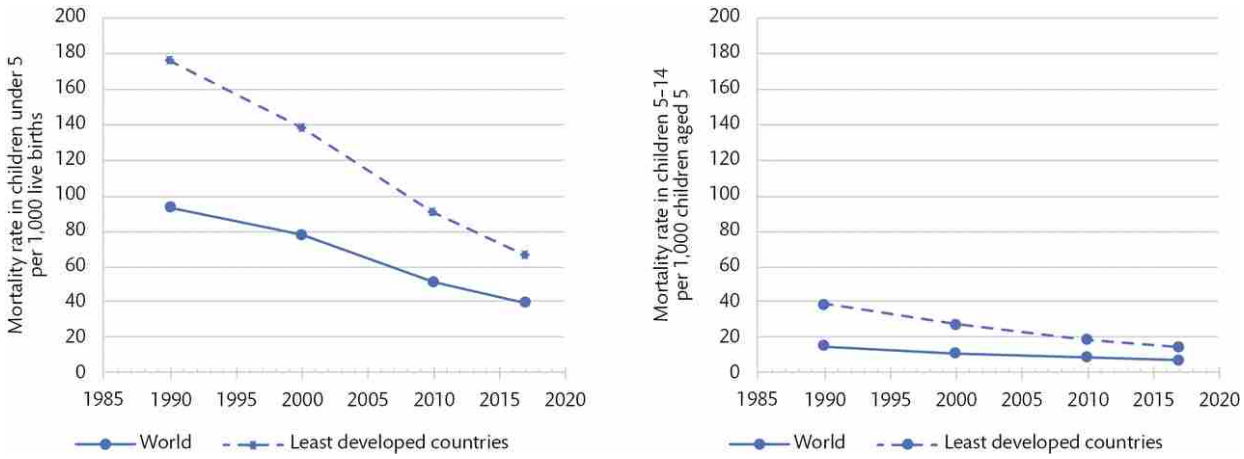


Fig. 10.3.2 Trends in child mortality rates, 1990–2017.
Source: data from Levels and Trends in Child Mortality Report 2017. (2017). [online] United Nations Inter-agency Group for Child Mortality Estimation (UN IGME). Available at: https://www.unicef.org/publications/files/Child_Mortality_Report_2017.pdf [Accessed 14 May 2019].

Global burden and epidemiology of morbidity in children under 5

Measurement of morbidity is more complex than that of mortality, thus data on the distribution of morbidity burden among the under-5 population is scarce. Despite data limitations, it is known that the main causes of morbidity highly correlate with the major causes of death in under-5 children (Table 10.3.1). Pneumonia, the second-leading cause of under-5 mortality is the leading cause of under-5 deaths in sub-Saharan Africa—nearly 490,000 annually. The global incidence of pneumonia in the 0–4 age group is nearly 0.19 episodes per child per year, with the highest incidence in Southeast Asia (0.26) and Africa (0.27) (Walker et al. 2013). In 2010, 120 million episodes of pneumonia occurred in children less than 5 years of age, of which 14 million progressed to severe illness (Fischer Walker et al. 2013).

Recent estimates of diarrheal disease burden show that diarrhoea incidence and mortality decreased by nearly 6 and 21%, respectively, between 2005 and 2015 (GBD Diarrhoeal Diseases Collaborators 2017). Of the 1.3 million deaths caused by diarrhoea in 2015, nearly 0.5 million were in children under-5, accounting for 8% of the total under-5 mortality. Nearly 90% of these diarrhoeal deaths in children occurred in South Asia and sub-Saharan Africa. Persistently high rates of diarrhoea morbidity have long-term detrimental effects on child growth and cognitive development.

Incidence of both pneumonia and diarrhoea are closely associated with poor home environments, undernutrition, and lack of access to health services. The prevalence of pneumonia and diarrhoea can be more than twice as high among the poorest children as among the richest ones. Malnutrition is an underlying determinant of nearly 45% of all under-5 deaths. With changing environments and lifestyles, rates of childhood overweight and obesity are rising. In many instances, overweight and obesity coexist with undernutrition and stunting, thus giving rise to the double burden of malnutrition.

Table 10.3.1 Estimated numbers of deaths caused by infections in neonates and children aged 1–59 months, at the global level in 2000–2015

	2000	2015*
Neonates		
Pneumonia	0.296 million	0.161 million
Sepsis/meningitis	0.532 million	0.402 million
Tetanus	0.161 million	0.034 million
Diarrhoea	0.04 million	0.018 million
Children (1–59 months)		
Pneumonia	1.4 million	0.76 million
Meningitis	0.284 million	0.115 million
Diarrhoea	1.2 million	0.509 million
Measles	0.485 million	0.074 million
Malaria	0.725 million	0.306 million
AIDS	0.22 million	0.086 million
Pertussis	0.056 million	0.054 million

* Neonatal sepsis/meningitis, neonatal pneumonia, neonatal tetanus, childhood pneumonia, childhood diarrhoea, meningitis, diarrhoea, measles, malaria and AIDS all showed >30% decline in mortality rates from 2000 to 2015

Source: data from: Liu, L., Oza, S., Hogan, D., et al. (2016). Global, regional, and national causes of under-5 mortality in 2000–15: an updated systematic analysis with implications for the Sustainable Development Goals. *The Lancet*, *388*(10063), 3027–35.

Global burden and epidemiology of mortality and morbidity in school-age children

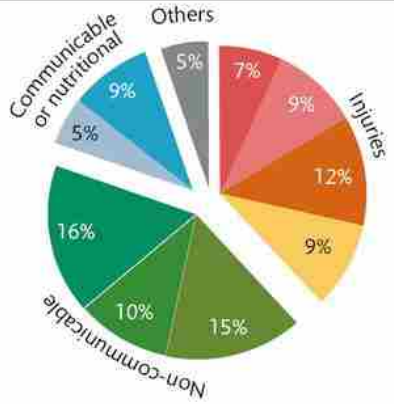
Although a great deal of progress has been made since 1990, there are still considerable differences in the risk of mortality between younger and older children (Fig. 10.3.2). The disparities in child mortality between the least developed countries and the global average have also narrowed, but a substantial gap remains that can be addressed by evidence-based interventions targeted at highly marginalized groups. Between 1990 and

2016, the global mortality rate for children between 5 and 10 years decreased from 10.3 to 4.5 deaths per thousand children (Masquelier et al. 2018). Despite this global trend, the mortality rates of older children remained high in 2016 in sub-Saharan Africa at 11.9 deaths per thousand. In 2016, there were still an estimated 585,000 deaths globally among children 5–9 years, with approximately 351,000 and 192,000 occurring in sub-Saharan Africa and South Asia, respectively.

Mortality rates are substantially lower, and the breakdown of causes of mortality in children over the age of 5 are quite different from that of younger children, with NCDs and injury-related causes increasing in importance to different degrees between countries (Fig. 10.3.3). The relative proportion of total mortality due to these causes increases in adolescence, and in higher income countries. A high proportion of infection-related mortality, characteristic of early childhood, extends into later childhood for those countries with higher levels of under-5 infectious morbidity. For example, the nutritional and communicable disease-related death rate of children 5–14 in India in 2016 was 32.6 and 26.2 per 100,000 for girls and boys, respectively, but the same rates for China were only 1.7 and 1.5 (Fadel et al. 2019). Furthermore, while NCDs and injuries in this age group were the principal causes of death in China, Brazil, and Mexico, India still experienced higher mortality rates due to these causes. Here, the typical epidemiological transition has not yet occurred, resulting in a substantial double burden of communicable and NCDs in older children, putting significant strain on health systems.

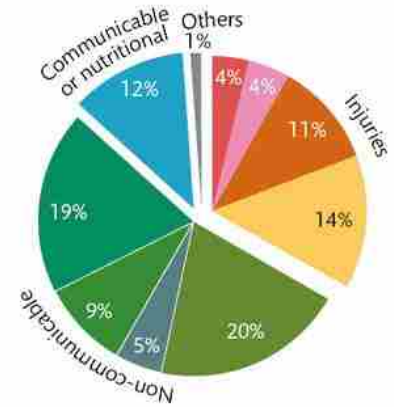
While their mortality risk is lower, older children bear the consequences of ill health and adversities that were experienced in their earlier years. Early and continued nutritional deficits and recurrent infections can lead to poor school performance, in addition to predisposing children to deleterious metabolic changes. The main causes of morbidity and mortality in children under 5, including respiratory infections, diarrhoea, and malaria, remain a significant burden for school-age children in the poorest low- and middle-income countries (LMICs). NCDs that become increasingly prominent and cause significant morbidity and disability include hemoglobinopathies, asthma, rheumatic heart disease, and sensory and neurological disorders. Additionally, school-age children experience a higher burden of injuries caused by drowning, road traffic accidents, poisoning, burns, and maltreatment.

Brazil



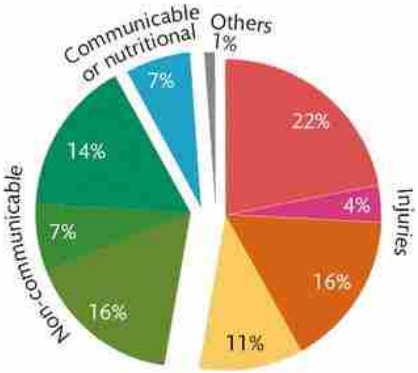
Girls	%	Boys	
4	8	Drowning
6	11	Homicide
11	13	Transport accidents
8	10	Other injuries
16	14	Cancer
11	9	Neurological
23	18	Other non-communicable
5	4	Pneumonia
11	8	Other communicable or nutritional
5	5	Others

Mexico



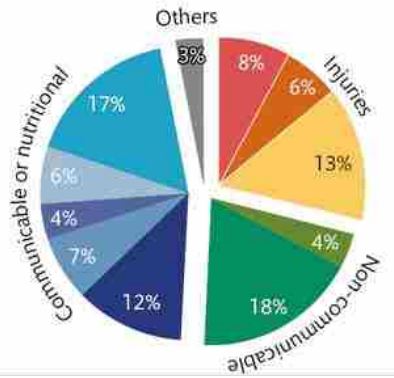
Girls	%	Boys	
2	5	Drowning
4	4	Suicide
9	12	Transport accidents
11	15	Other injuries
21	19	Cancer
6	5	Cardiovascular
9	9	Neurological
22	18	Other non-communicable
14	11	Communicable or nutritional
0	1	Others

China



Girls	%	Boys	
15	27	Drowning
4	4	Falls
16	16	Transport accidents
11	11	Other injuries
19	14	Cancer
8	7	Neurological
18	14	Other non-communicable
8	6	Communicable or nutritional
1	1	Others

India



Girls	%	Boys	
6	10	Drowning
4	8	Transport accidents
15	16	Other injuries
4	5	Cancer
17	17	Other non-communicable
14	10	Diarrhoea
7	6	Malaria
5	4	Meningitis or encephalitis
7	5	Pneumonia
18	16	Other communicable or nutritional
4	3	Others

Fig. 10.3.3 Breakdown of the causes of mortality among children 5–14 in Brazil, China, India, and Mexico from 2005 to 2016.

Source: data from Fadel, S.A. et al (2019). Trends in cause-specific mortality among children aged 5–14 years from 2005 to 2016 in India, China, Brazil, and Mexico: an analysis of nationally representative mortality studies. *The Lancet*, 393(10176), pp.1119–27.

Determinants and indicators of child health and well-being

The determinants of child survival, growth, and development can be broadly categorized as the social and physical environment in which a child resides, or their biological characteristics. Social and environmental determinants constitute the underlying risk factors further up the causal chain that can influence child health status. Biological determinants are often directly related to these underlying risks and include maternal risk factors (which influence a child's health even before they are born), birth conditions, and the health and nutritional status of child. The key adversities that have the highest impact on child health vary along the life course but are most critical in infancy and early childhood when most preventable child deaths occur. It is important to note that these risk factors affect entire populations of children to varying degrees, and those that contribute to a proportion of child deaths also contribute to a spectrum of morbidities, disabilities, and suboptimal growth and development trajectories in a much larger proportion of children.

Social and environmental determinants of child mortality and morbidity

Economic marginalization and poverty are the underlying determinants of much of the inequity in child health globally. Household food insecurity and reduced access to a diverse range of nutritious foods is strongly associated with poverty and predisposes children to malnutrition. When parents lack the means to provide a nurturing environment for their children, the risk of child mortality rises, child health suffers, and the consequences extend through adulthood and into the next generation. Parental educational achievement, particularly of the mother, is strongly linked to child health through multiple pathways. Community and household living conditions are also important determinants of childhood exposure to infectious diseases, chiefly water quality, sanitation infrastructure, and hygiene practices. Children raised in a

safe, secure, and stimulating home environment, with responsive caregivers, early educational opportunities, and good access to health services, are best placed to survive to adulthood and achieve their developmental potential. Supporting optimal child health is necessary to optimize human capital development within a population. When a child cannot reach their full intellectual and social potential, lifetime educational achievement and earnings are reduced, and the intergenerational cycle of poverty is perpetuated.

Gender and women's empowerment also play a key role in child health, the impacts of which begin even before birth. Adolescent girls who enter marriage and become pregnant at an early age are at higher risk for malnutrition, and subsequent foetal growth restriction and poor birth outcomes. Although female infants have a biological advantage over male infants in early life (who generally experience higher mortality rates than female infants in most regions of the world), inequitable gender preferences can put infant girls at a disadvantage with regards to infant feeding and care practices. These discriminatory parenting practices reflect societal gender inequality that manifests itself at the earliest stages of life.

Geographical disparities in child health stem, in part, from inequitable access to food, health services, and education—particularly in rural regions. Geography also dictates the physical environment in which children and their families live, and to which pathogens and pollutants they are exposed. Climactic conditions and agricultural productivity can affect household food security, especially in regions susceptible to the effects of climate change. Children are especially vulnerable to environmental shocks that disrupt the normal functioning of infrastructure, health systems, and food distribution. Migration, political instability, conflict, and natural disasters create adversities in the living conditions of children that increase the risk for malnutrition and disease.

Biological risk factors for child mortality and morbidity

There are several key maternal risk factors in the context of pregnancy that influence child health. Adolescent pregnancy is associated with poorer birth outcomes, and untreated infectious diseases in expectant mothers can cause serious birth complications or child death. Adequate maternal nutrition status, both before conception and during pregnancy, are required for

optimal growth and development of the foetus. Preterm birth and intrapartum-related complications such as birth asphyxia, in addition to severe neonatal infections, cause the majority of neonatal deaths. Those infants that survive are at substantially increased risk for developmental delay and disability (Mwaniki et al. 2012). Early childhood malnutrition and exposure to pathogenic organisms are the immediate causes of the majority of preventable child mortality, morbidity, and poor growth and development beyond the neonatal period. The subclinical condition known as environmental enteric dysfunction, or environmental enteropathy, is characterized by intestinal inflammation and changes to the intestinal lining. It is linked to poor sanitation conditions and chronic pathogen exposure, primarily through the repeated ingestion of faecal matter, and is proposed to represent an important link on the causal pathway to child stunting and suboptimal cognitive development, though the exact mechanism by which this occurs is still being investigated (Keusch et al. 2014).

Nutrition and growth in childhood

The nutritional status of under-5 children is usually assessed through three standard indicators: stunting, wasting, and underweight. In the case of stunting, a child's linear growth falls below minus two standard deviations (−2 SD) from the median height-for-age of the WHO Child Growth Standards (WHO 2006a). Stunting is usually a result of chronic undernutrition over a lengthy period of time, often in the context of a high burden of infectious pathogens that are associated with enteric dysfunction and nutrient malabsorption. A child is considered wasted when they fall below −2 SD from the median weight-for-height of the WHO Standards. Wasting usually reflects an acute nutritional deficiency, due either to reduced food consumption or acute weight loss during an illness. A child is said to be underweight if his/her weight falls below −2 SD from the median weight-for-age of the WHO Standards, which can be a result of stunting, wasting, or both. Conversely, overweight is defined as 2 SD above the median weight-for-height of the WHO Child Growth Standards.

While childhood malnutrition is a consequence of risk factors related to the social determinants of health, poor nutritional status itself is strongly associated with vulnerability to disease, delayed physical and mental development, and increased mortality risk. Undernutrition is usually

associated with inadequate maternal education and reduced child well-being. Despite declines in recent decades, in 2018, 21.9% (149 million) of children under 5 years globally experienced stunted linear growth, down from 32.5% (198.2 million) in 2000. In 2018, 5.9% (40.1 million) were overweight, up from 4.9% (30.1 million) in 2000. In 2018, an estimated 7.3% (49.5 million) of children under 5 experienced wasting (UNICEF/WHO/World Bank 2019). These global metrics mask striking inequities between regions; in 2018, 14.6% of children under 5 in Southern Asia experienced wasting, representing more than half of all wasted children globally. In Eastern Africa, 35.2% (24 million) of children under 5 were stunted (UNICEF/WHO/World Bank 2019).

An increasing number of countries are experiencing a ‘double burden’ of malnutrition, in which high rates of undernutrition, and overweight and obesity coexist. While Asia and Africa account, respectively, for 55 and 39% of all stunted children under 5, they also account for 47 and 24% of all overweight children under 5. Obese children are at increased risk for adult obesity, diabetes, liver disease, and poor social and economic performance in later life.

Micronutrient deficiencies are also a concern in children, sometimes causing serious deficiency-related morbidity, and affecting growth, development, and immune function. Deficiencies in iron (one cause of anaemia), vitamin A, iodine, and zinc are particularly harmful to child health, and are common in LMICs.

The nutritional status of children over 5 years globally is insufficiently measured and studied compared to the under-5 group. In addition to malnutrition and growth deficits that are carried forward from early childhood, micronutrient deficiencies such as iron deficiency anaemia can contribute to poorer school performance in school-age children.

Early child development

The first years of life provide the greatest opportunities for optimizing developmental trajectories. Young children have maximum plasticity of neuronal systems and are highly vulnerable to the negative effects of poor environments, and thus can benefit the most from interventions aimed at addressing risk factors in the critical period of early childhood. Estimates using stunting and poverty as proxy measures suggest that 43% (250

million) children under 5 in LMIC are at risk of not achieving their developmental potential (Lu et al. 2016; Black et al. 2017). Large regional disparities exist, with 66% (94.8 million) of children in the sub-Saharan African region at risk, versus 18% (9.7 million) in Latin America and the Caribbean.

A child's environment and activities while growing shape their cognitive, motor, and socioemotional development. Optimal child development can only occur with adequate nutrition, prevention and timely management of illness, responsive caregiving, and adequate psychosocial stimulation. Chronic malnutrition, poverty, illness, disability, and severe psychosocial deprivation represent key adversities that hamper a child's optimal development. Children experiencing these adversities have more difficulty learning, are less active, and may be less able to receive the attention needed from those who care for them. Later in childhood, school performance may suffer, and subsequent educational attainment and future earnings may be reduced.

Effective interventions that impact child development are available along the life course, and aim to improve maternal health and birth outcomes, caregiver–child interactions, responsiveness, and stimulation through play and communication. Through stimulation and interactions, these interventions intend to enhance the child's ability to learn socially and intellectually and grow physically. The frequent contact of mothers, their babies, and young infants with health services provides a unique opportunity for these services to actively promote and support early child development.

Key supportive actions for child health

Actions to improve child health do not exist solely within the health sector. This section briefly outlines a package of intersectoral and cross-cutting actions, rights frameworks, and policy improvements that support child health at the structural level. These actions address the most basic and distal social and environmental determinants of child health and are completely essential if actions within the health sector are to be effective.

Cross-cutting actions and supportive interventions to address the social determinants of health

There are a host of actions that must be taken at the national and subnational level to help to address the basic and underlying risk factors for child mortality and morbidity, and suboptimal development across the life course (Box 10.3.1). These include policies and programmes that aim to address poverty, food insecurity, illiteracy, gender inequality, poor infrastructure, unsafe home and community environments, suboptimal caregiving practices, and barriers to health service access. When specifically tailored and targeted with marginalized populations in mind, these actions have high potential to impact child health at the population level and should be the foundation of national strategies for improving child health and well-being across the life course.

Box 10.3.1 Key cross-cutting actions, programmes, and policies supporting child health, and health throughout the life course

- Poverty alleviation and cash transfer programmes
- Social protection services and safety nets
- Health insurance schemes
- Infrastructure and education for improved water, sanitation, and hygiene
- Improving living conditions and housing quality
- Reducing indoor and outdoor air pollution and exposure to toxins (e.g. lead, arsenic, aflatoxin)
- Food fortification with micronutrients
- Agricultural policies supporting food security
- Education and empowerment for women and girls
- Improved access and quality of health services
- Improved access to parenting programmes and comprehensive early child development services
- Universal primary education

Recognizing the rights of the child

Efforts to ensure effective accountability for children's health are further enhanced by the recognition that availability of and access to health information and services for children and their caregivers are basic human rights. Recognition of need for legal entitlements of every child, regardless of his/her status or situation, elevates associated responsibilities and political commitments to legal obligations. This paradigm shift thus enables the

establishment of not merely a normative but indeed legal foundation for addressing the health needs of children.

The UN Convention on the Rights of the Child (CRC), adopted in 1989 and ratified by all but two UN Member States, is the principal legal instrument for the protection of children's well-being, and outlines the core obligations of States in relation to children's health. Article 24 directly addresses the child's right to health and healthcare, while other provisions refer to the many underlying determinants of children's health. The CRC thus provides a valuable holistic blueprint of all essential needs and legal entitlements of children needed to ensure their survival, health, and optimal development.

Governance

Supportive laws and policies are prerequisites for strengthening the health sector's capacity to provide quality healthcare to a population. Governance in the context of a functional health system involves developing strategic policy frameworks, regulation, and system design and accountability. Examples of evidence-based policies that play a role in child health and that are currently being promoted in countries to tackle the burden of ill health and excess mortality are (WHO 2012b):

Maternity protection in accordance with the International Labour Organization Convention 183.
Policy on midwives authorized to administer a core set of life-saving interventions.
Policy on postnatal home visits in the first week of life.
The International Code of Marketing of Breast-milk Substitutes.
Policy on community treatment of pneumonia with antibiotics.
Policy on low osmolarity oral rehydration salts and zinc supplementation.
Policy on rotavirus and pneumococcal vaccines.

In order to facilitate the effective implementation of these policies in countries, advocacy and policy dialogue with a wide range of stakeholders is needed. Both public funding of healthcare and the quality of governance matter for the reduction of childhood mortality but remain inadequately monitored and evaluated. Some policies gain wide adherence rapidly while others lag behind. For example, among the 75 countries with the highest

burden of maternal and child mortality, the number of those with a policy allowing community health workers (CHWs) to treat pneumonia has more than doubled from 18 to 38 between 2008 and 2012. CHWs with appropriate training and supervision, and access to essential drugs and supplies have contributed significantly to improved pneumonia treatment (Kalyango et al. 2013). In contrast, by 2017 less than a third of the 81 Countdown countries had fully adopted the legislation on International Code of Marketing of Breast-milk Substitutes, which was endorsed in 1981 with consensus by the member states of the WHO.

Key child health issues, critical periods, and age-specific health sector interventions for prevention and management

Children living in resource-poor environments require protective interventions to prevent the major causes of mortality and morbidity summarized in prior sections. There are two main types of interventions: first, health promotion for morbidity reduction and prevention of specific disease conditions; second, case management interventions to treat disease, help reduce illness severity, and improve child health and survival. Key interventions are reviewed in the following text for each of the major causes of child mortality and morbidity, and summarized in Table 10.3.2

Table 10.3.2 Summary of key interventions to address risk factors for the major causes of child mortality, morbidity, and suboptimal growth and development

Important causes of child mortality and morbidity	Preventive interventions	Case management interventions
Maternal and Newborn (<1 Month)		
Newborn conditions: • Preterm birth • Perinatal asphyxia • Neonatal sepsis • Low birth weight • Congenital conditions	*Preconception:* • Maternal iron-folic acid or multiple micronutrient supplementation • Nutritional supplementation with balanced protein-energy in mothers at risk for undernutrition • Counselling on cessation of smoking, alcohol, and other substances • Delaying age at first pregnancy, and promoting optimal birth spacing *Antenatal care:* • Routine antenatal care visits • Maternal vaccination • Intermittent preventive treatment of malaria in pregnancy (IPTp) • Infectious disease screening and treatment • Prevention of Rhesus disease *Labour and childbirth care:* • Skilled care during labour and birth • Care at the time of birth: warmth, delayed cord clamping, hygienic cord care, and early initiation of breastfeeding *Postnatal care:* • Care during the first days and weeks after birth: exclusive breastfeeding, warmth, hygienic skin and cord care, vaccination, and prompt care-seeking for illness	• Detection and management of hypertension, diabetes, and foetal growth restriction during pregnancy • Tocolytics and antenatal corticosteroids for preterm labour (in appropriate populations) • Antibiotics for preterm premature rupture of membranes • Detection and management of maternal group B streptococcus (GBS), helminthic, malaria, tuberculosis, and HIV infections • Kangaroo Mother Care, and additional care of preterm and low birthweight infants (addressing feeding, warmth, and breathing problems) • Caesarean section or assisted vaginal delivery in case of foetal distress • Resuscitation if unable to breathe after birth • Diagnosis and treatment of neonatal sepsis and local infections
Infancy (<1 Year) and Early Childhood (1–5 Years)		

Important causes of child mortality and morbidity	Preventive interventions	Case management interventions
Pneumonia	• Immunization against *Pneumococcus*, *Haemophilus influenzae*, pertussis, diphtheria, and measles • Exclusive breastfeeding up to 6 months; continued breastfeeding with safe and appropriate complementary feeding at least up to 2 years • Reduction in environmental risk factors—e.g. indoor air pollution, overcrowding	• Treatment of pneumonia with amoxicillin • Treatment of severe pneumonia with parenteral antibiotics, oxygen, and other supportive treatment
Diarrhoea and enteric infections	• Early and exclusive breastfeeding up to 6 months; continued breastfeeding with safe and appropriate complementary feeding at least up to 2 years • Preventive vitamin A and zinc supplementation • Immunization against measles, rotavirus, cholera, and typhoid • Improved water and sanitation; hand washing with soap	• Oral rehydration therapy to prevent dehydration • Zinc treatment for 10–14 days during diarrhoea • Antibiotics for dysentery and typhoid
Malaria	• Vector control • Use of long-lasting insecticide-treated bed nets • Intermittent preventive treatment in infants (IPTi)	• Treatment of non-severe malaria with oral antimalarial drugs recommended for the area • Treatment of severe malaria with parenteral antimalarial drugs

Important causes of child mortality and morbidity	Preventive interventions	Case management interventions
HIV/AIDS	• Primary prevention of HIV and of unintended pregnancies in HIV-infected women • Antiretrovirals (ARVs)—treatment of pregnant women when indicated, prophylaxis for others for the prevention of mother-to-child transmission of HIV • Exclusive breastfeeding with ARVs or replacement feeding for HIV-positive mothers • Co-trimoxazole prophylaxis for HIV-exposed and HIV-infected children	• Antiretroviral therapy (ART) for children who are eligible for it
Malnutrition	• Early initiation of breastfeeding within 1 hour of birth • Exclusive breastfeeding up to 6 months of age • Continued breastfeeding with safe and appropriate complementary foods for at least 2 years • Preventing vitamin A, iron, iodine, and zinc deficiency through supplementation, food fortification, and dietary diversification • Community-based screening for severe acute malnutrition using mid-upper arm circumference (MUAC) measurements, and growth monitoring and promotion strategies	• Community-based management of children with severe malnutrition without complications • Hospital management of severely malnourished children with complications
School Age (5–10 Years)		
Overweight and obesity	• Education on healthy eating habits • Promotion of physical activity • Improving the local food environment	• Anthropometric screening, management, and treatment of comorbidities

Important causes of child mortality and morbidity	Preventive interventions	Case management interventions
Non-communicable diseases Sickle cell disease Asthma Rheumatic heart disease Epilepsy	• Improving air quality • Prompt detection and treatment of streptococcal pharyngitis and rheumatic fever	• Treatment of sickle cell disease with hydroxyurea • Screening, ongoing treatment, and supportive care for congenital and chronic conditions
Communicable diseases	• Vaccination for HPV and tetanus • Vaccination for typhoid, cholera, dengue where indicated • Catch-up on missed routine vaccines • Improved water and sanitation; hand washing with soap • Prevention of other infectious diseases from early childhood described earlier	• Deworming • Detection and case management of other invasive infections from early childhood described earlier
Unintentional injuries Drowning Road injuries Burns	• Swimming lessons • Helmet use • Improved cookstoves	• Emergency ambulance services • Trauma care

Maternal and newborn care interventions

Although neonatal mortality rates have reduced by 51% between 1990 and 2017, accelerated efforts are still needed to achieve Sustainable Development Goal targets for 2030 (Hug et al. 2019). Improved coverage and quality care along the continuum from pregnancy, childbirth, and the newborn period, into childhood and adolescence are critical for optimizing newborn health and survival. Such goals must focus on the most effective evidence-based interventions and intentional investment and implementation, which has so far been insufficient. Newborn health

interventions can be classified into those relevant for all mothers and newborns, and those relevant for those newborns with conditions that require additional care (Table 10.3.2).

Routine care for mothers and newborns

Improving newborn health and survival necessitates addressing social determinants, the education and empowerment of women in society, and especially the health and well-being of girls. This includes education and nutrition, prevention of harmful practices such as female genital mutilation, delaying age at first pregnancy, prevention and treatment of sexually transmitted infections, enhancement of pre-pregnancy nutrition status and optimization of interpregnancy intervals. Routine care during pregnancy holds particular importance for newborn survival and well-being. Evidence-based interventions for newborn health (Bhutta et al. 2014) provided during antenatal care include maternal vaccination (*Haemophilus influenzae* type b, influenza, pneumococcal), infection prevention, screening, and treatment, prevention of Rhesus disease, and maternal nutrition enrichment in populations with inadequate dietary intakes.

Antenatal care provides the platform for continued care during and after childbirth by planning for skilled birth attendance, preparing for any unforeseen complications, and by helping the family prepare for immediate newborn care practices, such as early initiation of exclusive breastfeeding.

The presence of skilled birth attendants and the provision of immediate obstetric care represent the cornerstone of modern obstetrics. A Delphi process involving expert consultation suggested that basic emergency obstetric care could reduce neonatal deaths during the intrapartum period by 40% and comprehensive emergency obstetric care could do so by 40% (Lee et al. 2011).

Skilled childbirth care in first-level health facilities, with the backup of a hospital that can manage complications, should be available and accessible to expectant mothers 24 hours a day, every day, as the complications that may occur cannot be predicted and may very rapidly become fatal. It is crucial to have access to a skilled attendant and a companion who can stay with the mother and child for the first 12–24 hours to ensure a successful home birth. Knowledge and recognition of maternal and newborn danger signs is crucial for women and families who cannot access a skilled

attendant, and a plan needs to be in place for a clean delivery and early care-seeking.

Vaccines are highly effective and cost-effective prevention tools for the prevention of morbidity and mortality in early life. Given their proven benefits, availability and accessibility of routine neonatal vaccines should be a public health priority. Several key vaccines are mentioned in Table 10.3.2.

Care for newborns at the time of birth and during the period that follows includes several relatively simple and effective interventions that do not require advanced technology (Table 10.3.2). The most critical time to deliver these interventions is immediately after birth and in the first 24 hours of life as a large proportion of maternal and neonatal mortality is reported during this window. Hence, at least one additional contact with a health provider within the first week, is required.

Care of at-risk and sick newborns

Fifteen million premature babies are born every year of which nearly 1 million die due to complications of preterm birth. About 75% of these deaths could be prevented if current cost-effective measures are made available to all (Liu et al. 2016). Such key interventions include tocolytics to delay birth, antibiotics for preterm premature rupture of membranes, and corticosteroids given to eligible mothers during preterm labour to promote lung maturation of the baby. Additional care of preterm and LBW infants is critical for improving newborn survival. Basic newborn care packages should involve identification of LBW infants within the first hours after birth. In settings where measuring birth weight is not feasible, such as for home births, all newborns perceived to be ‘small’ should receive additional care, much of which can be provided at home and in first-level health facilities. However, care for babies weighing less than 1,500 g, who have difficulty feeding or breathing due to lung immaturity, is best provided in a hospital.

In addition to basic newborn care, LBW infants need more attention, including immediate drying, adequate nutrition, provision of warmth, hygiene, early detection and management of infections, vitamin K administration, and growth monitoring. Kangaroo mother care or skin-to-skin contact is supported by strong evidence to reduce mortality and improve breastfeeding. Mothers who are unable to breastfeed directly should be taught to express breast milk and to feed from a cup.

Perinatal asphyxia is responsible for 10% of all under-5 deaths. Monitoring of progress and feto-maternal well-being in labour using a partograph has been proven to help reduce morbidity and mortality due to perinatal asphyxia. When foetal distress is suspected, assisted vaginal or Caesarean birth may be needed, which requires functional referral mechanisms. The second key intervention for prevention of asphyxia-related mortality and morbidity is neonatal resuscitation. All birth attendants should be skilled and equipped to perform neonatal resuscitation, which can include positive pressure ventilation, use of endotracheal intubation and laryngeal mark, in addition to thorough drying and stimulation.

The timely detection and optimum treatment of suspected neonatal bacterial infections, such as pneumonia and sepsis can substantially reduce neonatal mortality. It is therefore important to have community- and health-facility-based activities which focus on screening and management. Newborns with severe infections should receive parenteral antibiotics, along with intravenous fluids or alternative feeding methods and other supportive therapy such as oxygen in a hospital. Although the evidence for the use of injectable and muscular antibiotics in community settings has shown significant reductions in mortality in the infected newborns, implementation at scale has been scarce.

Care during Infancy and early childhood

The first 5 years of life is a critical period of rapid growth and development when children are particularly sensitive to adversity. The focus during this period is on supporting optimal nutrition, minimizing the burden of infectious diseases, and providing a nurturing, responsive, stimulating environment to optimize growth and development.

Malnutrition

The World Health Assembly (WHA) has declared 2016–25 as the ‘decade of nutrition’ and the Sustainable Development Goal (SDG) 2 aspires to end hunger, achieve food security, and eliminate all forms of malnutrition by 2030. Interventions to prevent malnutrition include promotion of optimal feeding of infants and young children, prevention and treatment of illness, and amelioration of micronutrient deficiencies in the diet. Early initiation of

breastfeeding, exclusive breastfeeding up to 6 months of age, and continued breastfeeding along with complementary foods up to at least 2 years of age are key interventions to combat undernutrition. Early initiation of breastfeeding, ideally within the first hour after birth, is associated with a reduced risk of neonatal mortality (Khan et al. 2014) and exclusive breastfeeding during the first 6 months of life is associated with a significantly lower risk of all-cause and infection-related mortality compared to those who were predominantly, partially, or not breastfed (Sankar et al. 2015). Safe and appropriate complementary feeding started at 6 months of age can reduce the prevalence of malnutrition and contribute to reduced mortality. Promotion of safe and appropriate complementary feeding has been shown to significantly improve weight-for-age and height-for-age scores (Lassi et al. 2013).

Prevention and management of acute malnutrition

Malnutrition, as a term, covers both being undernourished or overweight. Acute malnutrition is characterized by an acute loss of weight as a result of inadequate food intake or disease. However, there are many underlying and basic causes that contribute to the development of undernutrition. Several social factors such as decreased access to care, poor living conditions, inadequate care and feeding practices, and household food insecurity exacerbate this issue. In low-resource settings, children experiencing wasting are particularly vulnerable to infections, such as diarrhoea, pneumonia, and malaria, which can worsen malnutrition and further increase risk of disease, creating a vicious cycle of deteriorating health. The risk of mortality in acutely malnourished children is elevated, and nutrition-related factors directly or indirectly contribute to 45% of deaths in children under 5 years of age (Black et al. 2013).

To improve the nutritional status of children and prevent them from becoming severely malnourished, the dietary management of moderate acute malnutrition (MAM) (weight-for-height between –3 and –2 SD of the median of the WHO Child Growth Standards or 11.5 cm $\leq$ mid-upper-arm circumference [MUAC] < 12.5 cm) should be based on optimal use of locally available nutrient-dense foods. Inadequate nutrient content of habitual diets can be augmented through a number of approaches, including dietary diversification and fortification of certain staple foods with vitamins

and minerals. Supplementary feeding in children with MAM should be complemented with essential nutrition actions such as breastfeeding, nutrition counselling, and addressing the underlying social determinants of undernutrition.

Severe acute malnutrition (SAM) in children 6–59 months of age is defined as weight-for-height criterion less than −3 SD below median WHO Growth Standards, or MUAC less than 11.5 cm, or the presence of bilateral oedema. For infants less than 6 months, there is no agreed cut-off value for MUAC. Children with SAM should be clinically assessed for presence of medical complications and for their appetite. Children who pass the WHO-defined appetite test and are clinically well and alert are treated as outpatients while those with medical complications, danger signs, or who fail the appetite test should be treated as inpatients. Young infants less than 6 months with SAM are always treated as inpatient. Outpatient management involves providing the child with therapeutic feeding, micronutrient supplementation, antibiotics, and regular monitoring.

Inpatient management is generally organized in two phases: the stabilization phase (usually 3–7 days) includes careful evaluation and management of hypoglycaemia, hypothermia, dehydration, infections, electrolyte imbalance, and feeding initiation; and the rehabilitation phase (usually 4–6 weeks) which consists of the provision of nutritional and psychosocial support to achieve full nutritional recovery. Therapeutic feeding should start early in the stabilization phase and include giving a low-protein milk-based formula diet with additional vitamins and minerals such as the WHO-recommended F-75 (75 kcal/100 mL). In the rehabilitation phase, a milk formula with higher protein and energy content such as F-100 should be given and additional iron is needed to address deficiencies. Ready-to-use therapeutic foods (RUTFs) have replaced liquid F-100 in a variety of settings because they do not contain free water and can therefore be used without risk of bacterial contamination. Most RUTFs are lipid-based pastes combining milk powder, electrolytes, and micronutrients to meet recommended nutritional values and offer the malnourished child the same nutrient intake as F-100. RUTFs also contain 10–14 mg/100 g of iron, therefore no additional intake of iron is necessary.

In combination with community-based screening, RUTFs are highly effective in the treatment of SAM in children 6–59 months of age without complications (WHO 2013c). These children can be managed in the

community without the risks of nosocomial infections or any other disruptions regarding hospital care to the family. Often ignored is the sensory stimulation component of management where it is recommended that children with SAM be provided with structured play therapy and tender loving care in the rehabilitation phase.

In children less than 6 months of age who have SAM, care should focus on managing sepsis and other underlying complications and on re-establishing breastfeeding where this has been ineffective.

Prevention and management of chronic malnutrition

Chronic malnutrition in children is a global public health concern that is strongly linked to poverty. The process of nutritionally associated linear growth faltering begins *in utero* in the presence of maternal malnutrition, contributing to intrauterine growth restriction and low birthweight. During infancy, chronically suboptimal feeding practices, deficits in macronutrient and micronutrient intake, and a high burden of infectious diseases contribute to poor growth. A high prevalence of linear growth stunting within a population of children is a visible and easily measurable indicator of widespread chronic malnutrition. Children who experience this growth delay also suffer both higher mortality and suboptimal cognitive and motor development. Chronically malnourished children are at substantially higher risk of dying from common childhood infections such as diarrhoea, pneumonia, and malaria. Nutrition-related factors directly or indirectly contribute to 45% of deaths in children under 5 years of age, with an estimated 17% of mortality in children under 5 years attributable to stunting (Black et al. 2013).

The exact causes of stunting within a population are multifactorial, and still not well-understood. As such, a broad spectrum of interventions to prevent and address both undernutrition and high infectious disease burden is necessary to minimize the risk of growth faltering. This includes the promotion of optimal infant and young child breastfeeding and complementary feeding practices described earlier, ensuring adequate micronutrient intake in those at risk of deficiency, and high coverage of the infection prevention strategies described in the following sections.

Analyses of cohort data have shown that every unit increase in height-for-age z-score in children ≤2 years was associated with a 0.22 SD improvement

in measures of cognitive function at 5 to 11 years (Sudfeld et al. 2015), demonstrating the lingering effects chronic malnutrition in young children. This has serious implications not only for population health in countries with a high burden of chronic childhood malnutrition, but also for the fulfilment of their intellectual and economic potential. Countries that prioritize the implementation and scale-up of nutrition-sensitive and nutrition-specific policies and programmes stand to make the greatest improvements in human capital development and economic productivity, as these actions generally have very high benefit-cost ratios (Hoddinott et al. 2013).

Micronutrient deficiencies and hidden hunger

Sometimes described as 'hidden hunger', common micronutrient deficiencies such as those of vitamin A, iron, iodine, and zinc, both at clinical and subclinical levels, contribute to illness, suboptimal growth and development, and mortality in children. Strategies to prevent these deficiencies include supplementation (WHO 2011a, 2011b, 2016c), point-of-use fortification (WHO 2016d), mass food fortification of staple foods (WHO 2016e, 2018c), and dietary diversification (Pan American Health Organization and WHO 2003; WHO 2005b). See Table 10.3.3 for a summary of nutrition-specific and nutrition-sensitive interventions for the prevention and treatment of micronutrient deficiency in mothers and children, as recommended in the 2013 *Lancet Series on Maternal and Child Nutrition* (Bhutta et al. 2013; Ruel et al. 2013).

Table 10.3.3 Summary of interventions to prevent and treat micronutrient deficiencies in mothers and children.

Individual level	Population level
• Education and support for optimal infant and child breastfeeding and complementary feeding practices • Promotion of dietary diversification • Point-of-use fortification of complementary foods (e.g. micronutrient Sprinkles) • Single micronutrient supplementation (e.g. vitamin A capsules) • Multiple micronutrient supplementation (MMN) • Balanced energy protein supplementation (BEP) • Lipid-based nutrient supplements (LNS) • Delayed cord clamping • Deworming • Improved water, sanitation, and hygiene	• National nutrition-sensitive policies and programmes • Mass food fortification with micronutrients (e.g. salt iodization) • Biofortification • Cash transfer programmes • Social safety nets

Acute respiratory infections

Acute respiratory infections (ARIs), particularly pneumonia, are the second-leading cause of under-5 mortality, causing 16% of all child deaths worldwide. However, in children beyond the neonatal period, pneumonia continues to be the leading killer, responsible for nearly 13% of these deaths (Liu et al. 2016). Half of all childhood pneumonia deaths are attributable to *Pneumococcus* and *Haemophilus influenzae* type b (Hib). Other common causative organisms include viruses, most prominently respiratory syncytial virus, in addition to bacterial infections with *Staphylococcus aureus* and *Klebsiella pneumoniae*. *Pneumocystis jirovecii* is a significant cause in areas with high HIV burden. Interventions to control ARIs can be divided into preventive interventions such as immunization against specific pathogens, improvements in nutrition, reduction in household air pollution, and case management interventions.

Widespread use of vaccines against *Pneumococcus*, *Haemophilus influenzae*, pertussis, diphtheria, influenza, and measles has the potential to substantially reduce the incidence of ARIs in children in LMICs. Two particularly important vaccines are those against *Pneumococcus* and Hib pathogens, as they account for the largest burden of disease. *Pneumococcus*

and Hib were estimated to cause more than 300,000 and 30,000 under-5 deaths, respectively, in 2015 (Wahl et al. 2018).

Two types of pneumococcal vaccines are currently being used internationally: a polysaccharide vaccine effective against 23 different strains of *Pneumococcus* (primarily recommended for adults) and different formulations of polysaccharide-protein conjugate vaccine (PCV), of which WHO recommends the use of the 10- and 13-valent types (PCV10 and PCV13, respectively), for all infants. The serotypes included in these vaccines account for the majority of serious pneumococcal disease in children worldwide. PCVs are increasingly available in low-income, high-mortality countries, but gaps in vaccine uptake within countries could reduce their impact. A total of 141 countries have included the vaccine in their national immunization programmes; these include 7 of the 10 countries with the highest burden of pneumococcal disease.

The use of Hib conjugate vaccines in developed countries has resulted in the virtual elimination of invasive Hib disease (meningitis and bacteraemic pneumonia) because of immunity in vaccinated children and a herd effect in those not vaccinated. There are several Hib conjugate vaccines available, all of which are effective when given in early infancy and have virtually no side effects. The success of these vaccines is seen in most low-income countries which have introduced Hib vaccines. Recent research has estimated that the use of PCV has prevented 250,000 pneumococcal deaths from 2000 to 2015 (90% of these after 2010), while Hib vaccine prevented 1.2 million deaths in the same period (Wahl et al. 2018).

Vaccination should be seen as complementary to other strategies for pneumonia control, such as improved care-seeking and case management, better nutrition (including exclusive breastfeeding for the first 6 months of life), hand washing promotion, and measures to reduce indoor air pollution. Many studies have shown that young children exposed to smoke from the use of household biomass fuel had twice the risk of developing acute lower respiratory infection than unexposed children. A systematic review on household air pollution control strategies for child survival concluded that reducing exposure could reduce the risk of fatal pneumonia, especially if WHO guidelines are followed (Bruce et al. 2013).

Diagnosis and clinical guidelines for ARI case management rely on two simple clinical signs: fast breathing and lower chest wall indrawing. Chest indrawing is the inward movement of the lower chest wall when a child

breathes in. As per the revised WHO classification, children 2–59 months, with cough and/or difficult breathing plus fast breathing and/or chest indrawing are now classified as non-severe pneumonia that can be treated at home with oral amoxicillin (WHO 2014a). Severe pneumonia can be recognized by the presence of one or more danger signs (such as central cyanosis, severe respiratory distress, inability to drink, lethargy, or convulsions), and such cases must be managed in a hospital with parenteral antibiotics and supportive care including oxygen therapy. The updated guidelines to manage 'chest indrawing' pneumonia (previously labelled 'severe pneumonia') at home is significant, especially in settings when referral is difficult and also in reducing the use of unnecessary parenteral antibiotics. The role of CHWs in identifying and treating non-severe pneumonia is critical, especially in low-resource settings with poor access to facilities (Bhutta et al. 2013).

Diarrhoea

Diarrhoea remains one of the leading causes of under-5 child mortality in LMICs that lack sufficient resources and a strong infrastructure. Strategies for the control of diarrhoeal diseases have remained substantially unchanged since the early 1980s. A package of proven prevention and treatment measures is available that, if taken to scale, will lead to a significant reduction in the morbidity of diarrhoeal disease. It consists of a seven-point plan for diarrhoea control and includes promotion of early and exclusive breastfeeding, vitamin A supplementation, rotavirus and measles vaccinations, promotion of hand washing with soap and improved water supply quantity and quality, community-wide sanitation promotion, fluid replacement to prevent dehydration, and zinc treatment (UNICEF and WHO 2009). The package should be implemented as part of a concerted process alongside targeted and integrated behaviour change strategies which aim improve uptake by families and communities.

Exclusive breastfeeding for the first 6 months of life reduces the risk of infection-related neonatal mortality when compared to partial or no breastfeeding (Gülmezoglu et al. 2016). Components of breast milk such as antibodies, hormones, and growth factors provide optimal protection to the growing child. It is recommended that appropriate complementary feeding be introduced at 6 months of age alongside continued breastfeeding until at

least 24 months. This improves the nutritional status of infants and young children, optimizing physical and cognitive development and immune function.

Rotavirus is the leading cause of diarrhoea mortality among children younger than 5 years. It is estimated that introduction of the rotavirus vaccine was responsible for a 2.6% decrease in diarrhoea-related mortality in children younger than 5 years between 2000 and 2016 (Troeger et al. 2018). The international market has two rotavirus vaccines: the monovalent (RV1) and the pentavalent (RV5), which show 80 to 90% efficacy against severe infection in countries with very low or low mortality, and 40 to 60% efficacy in countries with high mortality in children and high or very high mortality in adults. WHO recommends the first dose of rotavirus vaccine be administered as soon as possible after 6 weeks of age, along with diphtheria-tetanus-pertussis (DTP) vaccine. RV1 and RV5 are administered in 2 dose and 3 dose schedules respectively, with 4 weeks each between the doses (WHO 2013a, b).

An estimated 800,000 children die each year from preventable diseases caused by lack of quality water, sanitation, and proper hygiene (Mills and Cumming 2016). Human faeces are the primary source of diarrhoeal pathogens. Water filtration and disinfection, and hand hygiene can significantly reduce diarrheal disease risk (Darvesh et al. 2017). However, behaviour change components are complex and often difficult to sustain. Poor-quality and quantity of water, sanitation, and hygiene (WASH) and related measures can hinder linear growth in children under 5 years of age (Dangour et al. 2013). Thus, resources should also be directed towards extending and sustaining water, sanitation, and hygiene services at the community level. SDG 6 includes targets for universal access to drinking water, sanitation, and hygiene.

Most diarrhoea-related deaths are due to dehydration, and dysentery (bloody diarrhoea) and electrolyte imbalance in severe cases. Management of diarrhoea includes:

Oral rehydration salts (ORS) solution administration to replenish fluid loss, or an intravenous electrolyte solution in cases of severe dehydration.
Mothers and health workers should supplement children with 20 mg per day of zinc for 10 to 14 days (10 mg per day for children under the age of 6 months).
Continuing or increasing breastfeeding during, and increasing feeding after, the diarrhoeal episode.
Using antibiotics only during episodes of bloody diarrhoea.

The use of oral rehydration solution has shown to reduce diarrhoea-specific mortality by 93% (Munos et al. 2010). Two noteworthy advances in the management of acute diarrhoea are: zinc supplementation, and newly formulated ORS containing lower concentrations of glucose and salts. Over the years, ORS formulation therapy has proven safe and effective in the prevention and treatment of dehydration. However, acceptance of standard ORS by mothers and health workers has been suboptimal because watery stools persist, and the duration of diarrhoea is not reduced. Efforts over the past decades have led to the development of a new ORS formulation with reduced osmolarity (low sodium and glucose). This new formulation reduces stool output, vomiting, and the need for any intravenous therapy. The new ORS is safe and effective in both non-cholera diarrhoea and cholera in children and adults. WHO and UNICEF now recommend the use of ORS containing 75 mEq of sodium and 75 mmol of glucose per litre (total osmolarity of 245 milliosmol per litre) everywhere (WHO 2005a). In addition to fluid replacement, optimal nutrition status can enhance fluid absorption and strengthen immune function.

A recent review of relevant clinical trials from areas with prevalent zinc deficiency showed benefit of zinc supplementation in children aged 6 months or more by reducing the duration of diarrhoea (Lazzerini and Wanzira 2016). Moreover, studies from India show that supplementing zinc in the management of acute diarrhoea can reduce the rate of hospital admissions for diarrhoeal disease (Mazumder et al. 2010). The addition of zinc to ORS in the treatment of diarrhoea has also increased the compliance of ORS and has reduced unwarranted use of prescription drugs of unknown identity and antibiotics (Bhandari et al. 2008).

Dysentery is an inflammatory disorder of the lower intestinal tract with a clinical picture of abdominal cramping, fever, and severe diarrhoea, often accompanied by the passage of bloody and mucoid stools. *Shigella* is the

leading and most severe cause of bloody diarrhoea. The primary treatment for shigellosis is the use of antibiotics, which can decrease dysentery-related mortality by more than 99% (Traa et al. 2010).

Antibiotic resistance is an emerging challenge in the management of shigellosis, and with ciprofloxacin-resistant strains already emerging, this increases the urgency for the development of a *Shigella* vaccine. Ciprofloxacin is the current treatment drug of choice due to its effectiveness, safety, ease of oral administration, short course, and low cost. For severe cholera cases, WHO recommends the use of oral antibiotics to which strains of *Vibrio cholerae* in the area are known to be sensitive; possible choices include tetracycline, erythromycin, doxycycline, co-trimoxazole, and chloramphenicol. Antibiotic treatment of cholera has shown a 63% reduction in rates of clinical failure and a 75% reduction in rates of bacteriological failure (Das et al. 2013).

Enteric infections and neglected tropical diseases

Neglected tropical diseases (NTDs) are a group of bacterial, parasitic, and viral infections highly endemic to the poorest regions of the world, where NTDs caused over 500,000 deaths per year (Naghavi et al. 2017). Several conditions disproportionately affect children compared to adults, who are often infected with multiple pathogens. Some of these infections are chronic, and contribute to substantial and prolonged illness and disability.

The WHO recognizes 17 core NTDs: rabies, trachoma, dengue, Chagas disease, Buruli ulcer, endemic treponematoses, human African trypanosomiasis (HAT), leprosy, echinococcosis, leishmaniasis, taeniasis/cysticercosis, dracunculiasis, trematodiases, lymphatic filariasis, schistosomiasis, onchocerciasis, and soil-transmitted helminthiases. From a programmatic point of view, WHO classifies NTDs into two groups: preventive chemotherapy and transmission control (PCT) NTDs, and innovative and intensified disease management (IDM) NTDs. The most prominent examples of PCT group NTDs are lymphatic filariasis, onchocerciasis, schistosomiasis, and soil-transmitted helminthiasis; they are mainly controlled by the administration of drugs to entire at-risk populations. IDM focuses on NTDs that lack appropriate tools for large-scale use. These include Buruli ulcer, Chagas disease, HAT, and leishmaniasis. Climate change is contributing to the intensifying spread of

many vector-borne infections, including malaria, dengue fever, Chagas disease, leishmaniasis, onchocerciasis to name a few (Mitra and Mawson 2017).

Disease control strategies, in line with WHO's 5 general strategies may involve: (i) prevention, including combined and concurrent preventive chemotherapy; (ii) improved access to clean water and sanitation; and (iii) control of insect vectors by use of environmentally safe insecticides, treated bed nets, and other personal protective measures, vegetation clearance, and improved sanitation to disrupt breeding sites (Mackey et al. 2014).

Food-borne enteric infections are also major causes of morbidity and mortality in LMICs, with outbreaks of *Salmonella*, *Escherichia coli*, and spongiform encephalitis occurring due to suboptimal sanitation, hygiene, and food processing practices. Typhoid and paratyphoid fevers are caused by infection with *Salmonella typhi* and *paratyphi*, respectively. After malaria and diarrhoea, typhoid fever is the most common cause of death in older children. An estimated 14.3 million cases of typhoid and paratyphoid fevers occurred globally in 2017, with 12.6% of cases reported among children younger than 5 years, and 55.9% among children less than 15 years. The highest mortality rates were reported in South Asia, accounting for 69.6% of worldwide mortality from typhoidal illness in 2017 (Stanaway et al. 2019). Antibiotics such as cephalosporins and azithromycin are primarily used in the affected regions. In December 2017, WHO endorsed the first conjugate vaccine for typhoid. This new vaccine has longer-lasting immunity than older vaccines, requires fewer doses and can be given to younger children. Provision of safe water and adequate sanitation, good hygiene among food handlers, and vaccination are all effective in preventing typhoid fever.

E. coli are normal inhabitants of the human intestine. Virulent strains are responsible for diarrhoea, neonatal meningitis, sepsis, and urinary tract infections. Transmission occurs via contaminated food and water or direct contact with an infected individual. Outbreaks have been associated with ground beef, exposure to animals in public settings, and contamination of water in recreational areas. Children who are suspected of having a systemic infection with *E. coli* should receive intravenous antibiotic therapy after diagnostic culture of the organism. Aminoglycoside or a third-generation cephalosporin is recommended as the empirical therapy. Preventive measures for *E. coli* infections involves hand hygiene, food safety, and contact isolation of infected individuals.

Dengue virus is estimated to infect 50 to 100 million people worldwide and is a significant cause of childhood morbidity and mortality (WHO 2013d). It is endemic to more than 125 countries across all WHO regions (excluding Europe) and has increased in incidence about 30-fold in the past 50 years. The factors that facilitate local outbreaks are shared with other NTDs, and include a lack of vector control, rapid urbanization, climate change, global travel to endemic regions, and social infrastructure weaknesses. Its arthropod vector is the *Aedes aegypti* mosquito. Public health interventions for dengue focus on vector and environmental management. Complications can lead to dengue haemorrhagic fever (DHF), which in some severe cases, such as in those who suffer from dengue shock syndrome (DSS), can be fatal.

Soil-transmitted helminths (STH) are a group of worms transmitted through soil contaminated with infected human faeces. Of particular importance are the hookworm *Necator americanus*, the roundworm *Ascaris lumbricoides*, and the whipworm *Trichuris trichiura*. Most STH infections result from ingestion of contaminated soil or through skin penetration. Endemic to tropical regions worldwide, STH infections are particularly burdensome in poorer areas of Asia, Africa, and Latin America, with most cases occurring among the extreme poor living in some of the large LMICs of Asia, such as India, Bangladesh, China, and Indonesia. Children usually suffer from *Ascaris* and *Trichuris* infections, whereas hookworms can be seen among both children and adults. In sub-Saharan Africa, an estimated 50 million children are infected with hookworm alone, and in Southeast Asia nearly 120 million preschool and school-age children require treatment for STH infections (Hotez et al. 2015). The predisposing factors include poor personal hygiene and sanitation, poverty, and overcrowding. Infections can result in anaemia, growth retardation, and impaired cognitive development, and are also responsible for school absenteeism and a substantial burden of disability-adjusted life years (Vandemark et al. 2010). A recent study reported deficits in children's cognitive development as a result of untreated STH infections (Pabalan et al. 2018). According to the WHO, administration of drugs like albendazole and mebendazole, education, and adequate sanitation are required for STH infection management.

Schistosomiasis is caused by schistosomes that penetrate the skin of those that have frequent contact with contaminated water bodies, which are often children and adolescents. More than 90% of schistosomiasis cases occur in

Africa alone. *Schistosoma* infection and non-treatment is significantly associated with educational, learning, and memory deficits (Ezeamama et al. 2018).

Malaria

Malaria accounts for 5% of child deaths worldwide. More than 300,000 children under 5 (1–59 months) were killed by malaria in 2015 (Liu et al. 2016), accounting for nearly two-thirds of all malaria deaths. The African region bears 90% of the global malaria burden, however from 2000 to 2015 the leading cause of under-5 mortality in sub-Saharan Africa shifted from malaria to pneumonia. Chapter 8.15 describes malaria interventions in greater detail.

Owing to effective preventive measures, the global malaria death rate has decreased by 60% since 2000. Vector control remains the most effective measure to prevent malaria transmission in all ages. Insecticide-treated bed nets (ITNs), and more specifically long-lasting insecticidal bed nets (LLINs), are particularly effective in reducing uncomplicated malaria incidence and parasite infection prevalence in pregnant women and young children. The WHO's Global Technical Strategy for Malaria 2016–2030 aims for universal coverage with effective vector control using LLINs or indoor residual spraying (IRS) of all households at risk for malaria (WHO 2015).

Prompt parasitological confirmation by microscopy or alternatively by rapid diagnostic tests in all patients suspected of malaria should happen prior to treatment. Treatment solely on the basis of clinical suspicion should only be considered when a parasitological diagnosis is inaccessible. Once diagnosed, the child should be treated early with a safe and effective antimalarial medicine as delaying treatment could result in progression to severe disease, which is associated with a high case-fatality rate.

Malaria case management has been greatly affected by the emergence and spread of chloroquine and sulfadoxine/pyrimethamine resistance. To counter the threat of drug resistance of *Plasmodium falciparum* to monotherapies and to improve treatment outcomes, the use of artemisinin-based combination antimalarials is recommended for treatment of uncomplicated malaria. In areas with poor access to health facilities, early recognition, prompt, and appropriate treatment require treatment in the home or

community. The development of new chemical formulations to target mosquitoes and the upcoming first global malaria vaccine (WHO 2018b) hold potential to reduce the burden of malaria further.

HIV and AIDS

In 2017, an estimated 1.8 million children under 15 globally were living with HIV, with just over half receiving antiretroviral (ARV) treatment. In 2015, HIV/AIDS accounted for about 1.4% of under-5 deaths globally (Liu et al. 2016), the vast majority of which occurred in sub-Saharan Africa. Although the proportional mortality due to HIV/AIDS among children under 5 remains unacceptably high in some African countries, it has declined significantly in the last decade due to scaling-up of interventions to prevent infections in children. In 2015, an estimated 45% of new HIV infections among children occurred in Central and West Africa. In children less than 18 months HIV infection diagnosis is established by the presence of HIV DNA through HIV nucleic acid-based testing (NAT) and beyond this age by detecting antibodies to HIV through an enzyme-linked immunosorbent assay. This is because maternal antibodies to HIV can pass through the placenta and remain present in HIV-exposed but uninfected children until the 18 months of age.

About 90% of HIV-infected children are infected through the mother-to-child transmission (MTCT), which may occur during pregnancy, birth, or through breastfeeding. Without any interventions, there is risk of transmission between 20 to 45%, but this can be reduced to less than 2% with a combination of evidence-based interventions (WHO 2010a). Antiretroviral (ARV) drug interventions given to pregnant HIV-infected women are very effective at preventing peripartum MTCT and also protects the health of the mothers (WHO 2010a). In 2013, about 49% of women continued to take antiretrovirals (ARVs) while breastfeeding, compared to 62% of women who took them during pregnancy. This demonstrates the urgent need to increase awareness about the importance of continuing treatment after birth.

Postnatal ARV interventions are now recommended to reduce the risk of maternal HIV transmission through breastfeeding. ARVs can be given either to the mother (Shapiro et al. 2010; Kesho Bora Study Group 2011) or to the infant (Coovadia et al. 2012). Based on the strength of this evidence and

programmatic considerations, national authorities are advised to support one infant feeding practice for HIV-infected mothers, namely, either breastfeeding while on ARVs or to avoid breastfeeding altogether and use replacement feeds (WHO 2010b). In settings where diarrhoea, pneumonia, and malnutrition continue to be significant causes of child mortality, breastfeeding is likely to give HIV-exposed infants the greatest chance of HIV-free survival (Kagaayi et al. 2008; Peltier et al. 2009). Mothers should exclusively breastfeed their infants for the first 6 months of life, introducing appropriate complementary foods thereafter, and continue breastfeeding for at least the first 12 months of life, and up 24 months and beyond while being actively supported for ARV adherence (WHO 2016a). It has been challenging to achieve high coverage of these interventions in resource-poor countries. Moreover, CD4 count testing to establish which pregnant women are eligible for lifelong ARV treatment or prophylaxis has hindered scale-up in many settings. There is recognition that women in high-burden countries are likely to have repeated pregnancies that would each require ARV interventions, either prophylaxis or treatment. To this end, and to protect mothers, several national programmes have opted to start all HIV-infected pregnant women on lifelong treatment (e.g. Centers for Disease Control and Prevention 2013). The questions of whether all affected pregnant women will want to start lifelong regimen and how best to support women to remain in care and achieve high rates of adherence remain unanswered.

In countries where infectious diseases and malnutrition do not contribute significantly to child mortality, mothers known to be HIV-infected may be advised to give their infants replacement feeds and avoid all breastfeeding. Commercial infant formula milk should be given as the replacement feed to the HIV-uninfected infants only when the following specific conditions are fulfilled:

Clean water and sanitation are assured at the household level and in the community, and
The mother or other caregiver can reliably provide sufficient infant formula milk to support normal growth and development of the infant; and
The mother or caregiver can prepare it cleanly and frequently enough so that it is safe and carries a low risk of diarrhoea and malnutrition; and
The mother or caregiver can, in the first 6 months, exclusively give infant formula milk; and
The family is supportive of this practice; and
The mother or caregiver can access healthcare that offers comprehensive child health services.

Pneumocystis jirovecii (previously known as *carinii*) infection often occurs before HIV status has been confirmed, and can cause life threatening pneumonia in HIV-infected or exposed individuals. Co-trimoxazole is a fixed-dose antimicrobial combination of sulfamethoxazole and trimethoprim targeting a broad range of aerobic Gram-positive and Gram-negative organisms, fungi, and protozoa including *Pneumocystis*. The drug is widely available in both syrup and solid preparations at a low cost, even in resource-limited settings. Regular prophylaxis with trimethoprim-sulfamethoxazole (co-trimoxazole) has been shown to reduce mortality of HIV-infected children by up to 40% even in the absence of ART (Chintu et al. 2004). Co-trimoxazole prophylaxis may also reduce the risk of malaria-related mortality in both HIV-infected and exposed children (Sandison et al. 2011), though its role in HIV unaffected children is questionable. Because of difficulty in diagnosing HIV infection in infants, co-trimoxazole prophylaxis is recommended for all HIV-exposed children around 4 to 6 weeks after birth and continuing until HIV infection has been cleared and the infant is no longer at risk of acquiring HIV through breastfeeding (WHO 2014b).

HIV-exposed and HIV-infected children should receive all childhood vaccines as early in life as possible, except Bacillus Calmette–Guérin (BCG) and yellow fever vaccines. In asymptomatic children, the decision to give late BCG should be based on the local risk of tuberculosis. Infants with symptomatic HIV infection should not receive yellow fever vaccines.

Antiretroviral therapy (ART) can substantially improve survival of HIV-infected children. The availability of ARV drugs with other formulations such as ‘sprinkles’ make earlier initiation and adherence to treatment by children more convenient. The particular needs of adolescents infected with

HIV such as access to adolescent-friendly sexual and reproductive health services are increasingly recognized as priorities.

The Global Health Sector Strategy on HIV for the period 2016–2021 (WHO 2016b) suggests the health sector response to HIV as critical for achieving universal health coverage, which is a key target of the SDGs. The strategy proposes a set of global targets aligned with the multisectoral UNAIDS strategy. It also identifies five strategic directions: strengthening national HIV programmes; defining a package of essential interventions; adapting services for underserved populations for equitable coverage; implementing funding mechanisms to reduce costs to patients; and embracing innovation.

Prevention and management of suboptimal development and disability

Experiencing repeated adversities between conception and early childhood disrupts normal brain development, social functioning, and learning. Developmental delay or disability can result from repeated exposure to infection or malnutrition, and from adversities related to insufficient stimulation, trauma, and maltreatment. Global data on direct measures of child development are still very sparse, although it is known that many children are at risk for poor development due to the conditions of poverty and chronic malnutrition.

The Nurturing Care Framework (Black et al. 2017) outlines five key domains for intervention that must be addressed in order to ensure children reach their full developmental potential: good health, adequate nutrition, responsive caregiving, safety and security, and early educational opportunities. The efforts of parents, families, and communities, within an enabling environment, make this possible. Many of the evidence-based interventions supporting child health and nutrition discussed in prior sections have the potential to support healthy development through a reduction in risk factors for suboptimal development (Britto et al. 2017; Vaivada et al. 2017). Interventions in multiple domains of the Framework can also be packaged together for delivery, such as combined stimulation and nutrition interventions to support optimal development.

It is imperative to provide extra support for children with motor, developmental, and learning disabilities, especially in LMICs, where these

children are particularly vulnerable (WHO 2012a; 2010b). Primary care providers are well placed to support child development in LMICs, as they are often the only health service providers that have contact with young children and families. There is a notable lack of research examining the effectiveness of interventions for prevention of developmental delay in early childhood.

Strengthening integrated service delivery

Despite the fact that effective interventions exist against all major conditions from which children die, the coverage of these interventions generally remains low. The key to achievement of substantial child mortality reduction is quality universal coverage of the interventions previously described.

Before the mid-1990s, child health programmes in LMICs focused on immunization, control of diarrhoeal diseases, ARIs, and nutrition. Their implementation led to reduction in under-5 mortality in many countries, but this approach encountered limitations. The narrow focus on single interventions has failed to consider the child in a holistic manner, leading to many missed opportunities for care in contacts with a health practitioner. It was also difficult to resolve upstream health-system constraints such as management or human resource policies using this approach. Solving these problems involves ensuring access to an integrated package of interventions that are organized around the continuum of care. It also means that care should span across the home, community, health centre, and hospital. The advantages of delivering interventions as packages include the following:

Many interventions go naturally together because they are delivered by the same person at the same time.
Packaging is more cost-effective in terms of training, implementation, and supervision.
Packaging meets the needs of the individual caregiver and the child much better than isolated and uncoordinated single-intervention delivery, thus reinforcing the continuum of care.

Potential obstacles to packaging of interventions may be related to policy, programmes, and the organizational structure of ministries and health services. For example, managers responsible for single issue-focused programmes may feel that packaging their interventions with other maternal

and child health services lessens their delivery efficiency. Another concern about packaging interventions is that vulnerable groups may not get any of the interventions if insufficient attention is given to equitable delivery of the package.

Delivery platforms relevant to child health ideally target the specific health needs of children at different stages of life with tailored packages of interventions for health promotion, prevention, screening, and treatment. Community delivery platforms aim to address the health needs of mothers and children where health facilities are sparse, often through the use of CHWs to facilitate outreach services, initiatives such as women's groups, and campaigns such as Child Health Days. Financial platforms such as conditional cash transfers can incentivize health seeking behaviours and school attendance, and user fee removal can increase health service utilization. School-based platforms target the needs of school-age children with vaccinations, nutrition interventions, and health education. Digital and technology-based platforms leverage the broad reach of mobile phones and the Internet to deliver both health messaging and healthcare. A tailored and context-specific mixture of these delivery strategies is necessary to achieve optimal coverage of essential interventions for child health.

Integrated management of neonatal and childhood illness (IMNCI) and integrated community case management (iCCM)

Integrated management of childhood illness (IMCI) was developed as an integrated delivery approach to managing common childhood diseases based on the principles of primary healthcare (PHC) as set out in the Alma Ata Declaration. Focusing on countries with a high under-5 mortality burden (>40/1,000 live births), the approach aims to enhance health workers' skills in treatment of common childhood diseases, strengthen health systems for appropriate management of childhood illnesses, and improve family and community practices to promote child health.

Over the years the approach has been expanded to include newborn care —IMNCI and integrated community case management (iCCM). Recognizing the key role CHWs can play when appropriately trained, equipped, and supervised, the equity-focused iCCM strategy involves training these workers to identify and treat children in remote areas for pneumonia, diarrhoea and malaria using appropriate treatment methods such

as ORS and zinc, oral antibiotics or artemisinin-based combination therapy (ACT) (WHO 2017b). A Cochrane review on the impact of IMCI concluded that the implementation of IMCI, especially with the neonatal care component, led to a decrease in infant and under-5 mortality rates (Gera et al. 2016).

Health system building blocks

It is crucial to address the challenges that affect health systems in many low-income countries in order to effectively scale-up child health interventions. The programmes which address newborn and child health are implemented through 'complex public and private organizations that rely on systems to provide medicines, finance health services, assure quality, and efficiency of care, manage the health workforce, and generate information needed for effective operational decisions' (Countdown Working Group on Health Policy and Health Systems et al. 2008). Key input components to the health system include an adequate and skilled health workforce, evidence-based policies, and sufficient health funding—critical determinants of access, quality, and demand for effective interventions. Improving child health services globally relies heavily on decisions made using health information systems (HIS). However, in most resource-limited settings, HIS are prone to lack of coverage, incomplete record-keeping, poor data quality, limited analysis, and fragmented systems that can hamper delivery of quality health services. Service delivery, together with medical products and technologies, reflect the immediate outputs of the health system: the availability and distribution of care (WHO 2010c).

Over the past two decades, increasing economic hardships and financial crises have disrupted healthcare delivery in many countries. Due to deteriorating infrastructure and a resource-limited public health sector, subject to structural adjustment and macroeconomic ceilings, challenges faced by healthcare sectors persist. As a result, human resources in the health sector have been destabilized and undermined. In many countries, households spend considerable amounts on healthcare and out-of-pocket expenditures for health services, sometimes two to three times greater than the total health expenditure by governments and donors. To establish an efficient, functional health system, it is important to address inequality in access to health services and provide financial risk protection. As part of the

SDGs and universal health coverage (UHC) agendas, many developing countries have implemented health programmes to improve access to basic health services for all and provide financial risk protection to reduce out-of-pocket (OOP) health expenses.

The higher the proportion of health-related user payments, the greater the relative financial burden on the poor, rendering many unable to afford necessary healthcare services. Household expenditure surveys show that more than 150 million individuals globally face severe financial crisis each year due to healthcare costs. The full assessment of such financial burden needs to evaluate a multidimensional framework to examine the broader impacts of illness or injury on the household economy, such as loss of work (Ruger 2012). Prepayment systems involve collection of funds in advance through tax-based insurance or social health insurance schemes. Both provide financial risk protection and promote equity through prepayment of healthcare costs and pooling of health risks. OOP expenditures should be gradually added into prepayment schemes, including community finance programmes.

Health workers are ‘all people engaged in actions whose primary intent is to enhance health’. This includes health service providers such as doctors, midwives, nurses, pharmacists, and CHWs. It also includes health management and support workers who dedicate all or part of their time to improving health (WHO 2006b). The availability of a competent workforce is vital to successfully delivering healthcare services. Critical shortages, inadequate skill mix, and uneven geographical distribution of the health workforce are major barriers to achieving global goals for child health. The minimum threshold necessary to deliver essential maternal and child health services is estimated as 23 doctors, nurses, and midwives per 10,000 population. Countries that fall below this threshold struggle to provide adequate skilled care at birth to pregnant women, as well as emergency and specialized services for newborns and young children, influencing the burden of women and children mortality directly. Among countries categorized as low-income economies by the World Bank, only five met this minimum threshold in 2010.

In almost all countries, a poor mix of the types of health workers and their range of individual skills, combined with a shortage of qualified health workers in remote and rural areas, impedes access by women and children to life-saving interventions and services. To address these challenges, countries

are exploring and applying different solutions. Community-based health workers can provide a number of life-saving child health interventions, such as management of non-severe pneumonia and immunizations. As a result, governments in various countries are instituting new cadres of CHWs or upgrading the skills of existing community resource persons to deliver an essential package of health services that include child health interventions. Similarly, midwifery personnel are enabled to perform life-saving interventions such as resuscitation of the asphyxiated baby. The process of sharing tasks at various levels of service delivery, sometimes referred to as task-shifting, is being investigated and applied in a number of countries.

Supporting children in humanitarian emergencies and conflict settings

The total number of people affected by humanitarian emergencies has been on the rise. In 2017 alone, there were an estimated 65 million forcibly displaced people, over half of whom were less than 18 years old (UNHCR 2018). A recent analysis commissioned by Save the Children was conducted by the Peace Research Institute in Oslo (PRIO) which mapped the numbers of children living in areas affected by armed conflict globally (Østby et al. 2018). The research reported that in 2017, 420 million children were living in areas defined as ‘conflict zones’ or ‘conflict-affected areas’—nearly one in every five children worldwide. Armed conflicts violate children’s rights, as established in the United Nations Convention on the Rights of the Child (UNCRC) (Hodgkin and Newell 2007).

Environmental factors, including climate change, can encourage conflict and intensify the severity of natural calamities (IPCC 2014). Large-scale disasters and warfare can increase the probability of family separation dismantling any existing support systems, putting children at risk for abuse, violence, and neglect. Such emergencies can create economic hardship and create further vulnerabilities for children. In Lebanon, over one million Syrian refugees have been registered with the United Nations High Commissioner for Refugees (UNHCR 2017), and child labour and marriage have been reported as families face financial insecurity (Bartels et al. 2018). Moreover, children in situations of economic and physical instability are also at increased risk of sexual abuse, human trafficking, and recruitment by extremist groups or armed forces.

The consequences of armed conflict on children can be both direct and indirect. Direct effects include physical injury or disability, developmental delay, mental health sequelae, or death. Armed violence has the intended and unintended consequence of killing and maiming children. Indirect effects are associated with the destruction of resources required by children for their optimal survival and growth, in addition to effects on the social determinants of health, such as living conditions, education, nutrition and health of caregivers (Guha-Sapir and D'Aoust 2011). Such conflicts can also displace children, and about 68.5 million people globally were forcibly displaced in 2017, with 25 million being refugees; over half of these refugees were children (UNHCR 2018).

The UN Security Council identifies six grave violations against children in situations of armed conflict (United Nations 2013):

killing and maiming of children
recruitment and use of children as soldiers
sexual violence against children
abduction of children
attacks on schools and hospitals
denial of humanitarian access

Given the prevalence of armed conflict globally and its deleterious effects on children, it is critically significant that child health experts respond both domestically and globally. Child health professionals should recognize children's traumatic life experiences and respond to the consequences of such trauma on children's health and well-being by providing holistic, multidisciplinary, culturally acceptable, and age-appropriate services that reflect children's need for physical, mental, emotional, and sexual and reproductive healthcare (Marsac et al. 2016).

Governments should develop systems to register all children at birth, so that no child remains stateless, and children separated from family have opportunities to reunite. A plan needs to be in place to regularly review plans for children in custody of the State. Doing so overcomes the barriers to accessing health services, education, and employment. There is a need to facilitate dialogue between states, the UN, NGOs, and non-state armed groups to protect children. It is important for the governing bodies to support international mechanisms to prosecute cases of children's rights violations in conflict.

Educational opportunities for children affected by armed conflict should be provided ensuring their swift reintegration into society. There is strong evidence supporting the role of education for boys and girls at all levels in reducing political violence (Østby and Urdal 2010). Interventions for optimal feeding including breastfeeding become even more critical for survival in humanitarian emergencies. During such displacements children may face risks from diarrhoeal diseases, pneumonia, and undernutrition. Moreover, unhygienic water and sanitation, overcrowding, food insecurity, and overburdened health facilities can further exacerbate the worsening situation for children.

Children should be protected from landmines, small arms, and light weapons through strict control on their ownership, sale, and safe storage. Environmental hazards, such as landmines, injure and kill children long after conflicts have ended. As a complement, initiatives for child protection may also take the form of direct implementation, such as the establishment of 'Child Friendly Spaces (CFS)' that allow children safe zones to play, parent trainings that emphasize alternatives to physical punishment and reunification for separated children. Domains for further research include the physical, developmental, and psychosocial health consequences of armed conflict exposure on children; the associations if any, between social determinants and these effects; the burden of child morbidity and mortality attributable to armed conflict; interventions; and child health assessments for humanitarian response.

Global initiatives to accelerate progress on child survival, good health, and optimal development

Several global initiatives related to specific causes of child mortality have been launched to help accelerate progress in closing the gap in child health outcomes between high-income countries and LMICs. Countries that have rapidly increased coverage for multiple interventions across the continuum of care have accelerated child survival. Their success can mainly be attributed to improvements in service quality—bringing them closer to home and expanding access to essential care.

The UN has launched the updated Every Woman Every Child's Global Strategy for Women's and Children's Health for the period 2016–2030

(WHO 2017c). The aims of this ‘Global Strategy’, which are universal in scope, are to end all preventable mortality for mothers, newborns, and children, and promote this vision as part of a plan for accelerating progress in achieving the SDGs. The strategy has helped catalyse new attention and investment to some of the most frequent causes of women’s and children’s mortality. Other recent global initiatives and events are tackling different aspects of implementing the ‘Global Strategy’, including the Child Survival Call to Action, The Family Planning Summit, and the UN Commission on Life-Saving Commodities for Women and Children, Every Newborn Action Plan, and the Global Action Plan for Pneumonia and Diarrhea.

To accelerate progress and achieve improved health outcomes for all children, it is important to ensure universal access to good-quality basic care, clean water, sanitation, and other social sector services such as education. This also includes creating the national legal, policy, and budget frameworks to promote the realization of children’s rights to these services.

School-age children

Children between 5 and 10 years of age normally attend primary school, and thus can be targeted by school-based health programmes. School feeding programmes and targeted micronutrient supplementation are effective strategies to prevent malnutrition. Education and messaging regarding the lifelong benefits of healthy eating habits and an active lifestyle can begin in this age group, coupled with supporting a healthy food environment. The school setting also provides an opportunity for the promotion of good oral hygiene practices and oral health screening. Furthermore, vision screening and the provision of corrective lenses for those children that require them help them learn more effectively. The WHO Health Promoting Schools initiative provides a framework of preventive actions that aim to promote healthy living habits and better learning conditions, and provide a safe learning environment with access to health services for schoolchildren.

Overweight and obesity

An estimated 38.3 million children under 5 worldwide were overweight or obese in 2017. Nearly half of these children live in Asia and about a quarter in Africa (WHO/UNICEF 2018). Past the age of 5, obesity rates climb even

further: 340 million children and adolescents between 5 and 19 years experienced overweight or obesity in 2016. Overweight and obesity increases a child's risk of developing type 2 diabetes, hypertension, asthma, and liver disorders. Moreover, they have a higher likelihood of suffering from depression and low self-esteem.

To assess, prevent, and manage childhood obesity, the WHO recommends primary care providers take anthropometric measurements for all young children to determine weight-for-height and assess the presence of wasting or overweight (WHO 2017a). They should also counsel caregivers on optimal infant and young child feeding, promote intake of healthy food and adequate physical activity, and develop a management plan for children identified to be obese. As childhood stunting raises the risk of becoming overweight or obese, it is not recommended to provide formulated supplementary foods to children who are only moderately wasted or stunted, unless there is household or community food insecurity or other medical considerations.

Non-communicable diseases and injuries

With the global epidemiological transition and change in lifestyles, NCDs are on the rise globally, especially in industrialized countries. Contrary to common perception, NCDs and their risk factors have a significant impact on the health of children and adolescents. Congenital disorders, diabetes, cancer, sickle cell disease, epilepsy, chronic respiratory diseases such as asthma, and rheumatic heart disease are some key NCDs that affect children and require a functioning health system providing quality care to effectively detect and manage. Childhood overweight and obesity, unhealthy diet, exposure to pollution, certain infectious diseases, and epigenetics can all increase the risk of developing an NCD in the short- and long-term.

While many NCDs are chronic and require ongoing management, some preventive actions do exist. Preventing obesity is a key strategy to prevent the development of several NCDs, both during childhood and later in life. Improving indoor and outdoor air quality can help prevent asthma exacerbations. Prompt detection and treatment of rheumatic fever can help prevent progression to rheumatic heart disease.

Supporting positive developmental progress, resilience, and self-efficacy is vital for children to achieve their full potential. As mental disorders can

begin to surface in childhood, routine screening and counselling for conditions such as anxiety, depression, and conduct disorder should be provided.

As rates of unintentional injuries rise as children become older, injury prevention is key to reducing injury-related mortality and disability. This includes requiring the use of helmets on bicycles and mopeds, and seatbelts in automobiles. Swimming lessons and education on water safety provided to children can reduce the risk of drowning. Burns can be prevented through the use of safer types of improved cookstoves, compared to open flames. Improving access to emergency services and trauma care in the event of injury is necessary to minimize injury-related mortality and morbidity.

Communicable diseases

Ensuring complete and up-to-date vaccinations status in school-age children is the most effective prevention strategy for infectious diseases. For children 5–10 years, these include universal vaccination for human papilloma virus (HPV) (age 9+) and tetanus, targeted vaccination for typhoid, cholera, dengue where indicated, and catch-up and booster vaccines as needed. Routine deworming treatment in endemic regions is also recommended for reducing the risk of anaemia. Improving WASH conditions in the school setting as well as the household also remains crucial for preventing infection transmission. School-age children also still require many key interventions that are carried over from the under-5 period for specific high-burden conditions, including ARIs, diarrhoea, and malaria. Typhoid fever in particular becomes increasingly prevalent in older children and requires prompt screening and treatment.

Child health in the SDGs era

In 2015, the 2030 Agenda for Sustainable Development was adopted by all United Nations Member States (UNGA 2015), representing a comprehensive plan for peace and prosperity, not only for people, but also for the planet. As part of this agenda, 17 SDGs (Box 10.3.2) were agreed upon, building on the legacy of the prior Millennium Development Goals (MDGs), whose purview ended in 2015. The MDGs contained specific targets for reductions in maternal and under-5 mortality, emphasizing massive inequities between low

and high-income regions, and were successful in bringing the global health and development community together with a common goal. Countries that have rapidly increased coverage for multiple interventions across the continuum of care have accelerated child survival. The most pronounced falls in under-5 mortality occurred in four regions: Latin America and the Caribbean; East Asia and the Pacific; Central and Eastern Europe and the Commonwealth of Independent States (CEE/CIS); and the Middle East and North Africa. All have more than halved their regional rates of under-5 mortality since 1990. Their success was attributed mainly to improvements in service quality—bringing them closer to home and expanding access to essential care.

While all of the SDGs are relevant to child health from a societal and environmental perspective, two specific goals contain ambitious and explicit targets for key indicators of child health. These include achieving global targets for stunting and wasting reduction by 2025, ending all forms of malnutrition by 2030 (Goal 2.2), and ending preventable newborn and under-5 deaths by 2030 (Goal 3.2), with targets for neonatal and under-5 mortality of 12 and 25 per 1,000 live births, respectively.

Box 10.3.2 The Sustainable Development Goals

- Goal 1. End poverty in all its forms everywhere
- Goal 2. End hunger, achieve food security and improved nutrition and promote sustainable agriculture
- Goal 3. Ensure healthy lives and promote well-being for all at all ages
- Goal 4. Ensure inclusive and equitable quality education and promote lifelong learning opportunities for all
- Goal 5. Achieve gender equality and empower all women and girls
- Goal 6. Ensure availability and sustainable management of water and sanitation for all
- Goal 7. Ensure access to affordable, reliable, sustainable, and modern energy for all
- Goal 8. Promote sustained, inclusive, and sustainable economic growth, full and productive employment, and decent work for all
- Goal 9. Build resilient infrastructure, promote inclusive and sustainable industrialization and foster innovation
- Goal 10. Reduce inequality within and among countries
- Goal 11. Make cities and human settlements inclusive, safe, resilient, and sustainable
- Goal 12. Ensure sustainable consumption and production patterns
- Goal 13. Take urgent action to combat climate change and its impacts*
- Goal 14. Conserve and sustainably use the oceans, seas, and marine resources for sustainable development
- Goal 15. Protect, restore, and promote sustainable use of terrestrial ecosystems, sustainably manage forests, combat desertification, and halt and reverse land degradation and halt biodiversity loss
- Goal 16. Promote peaceful and inclusive societies for sustainable development, provide access to justice for all and build effective, accountable, and inclusive institutions at all levels
- Goal 17. Strengthen the means of implementation and revitalize the global partnership for sustainable development

Measuring progress in child health

The Countdown to 2030 initiative aims to provide a framework through which to improve the monitoring and measurement of child, women, and adolescent health. The main focus is on measuring and reporting, at the national level, on the coverage of many of the key interventions described in this chapter, in addition to highlighting geographic and socioeconomic inequities in this coverage within countries.

There are 81 countries tracked as part of the Countdown initiative, and despite progress in recent years towards universal coverage, there is still a large gap in coverage of essential interventions along the continuum of care (Boerma et al. 2018). In order to achieve the ambitious targets set out by the SDGs, a great deal of investment is still required. Countdown aims to provide accurate data on coverage, quality, and inequalities so as to guide these investments towards those countries, regions, and populations who

have the greatest need. To this end, efforts to strengthen the collection of high-quality vital statistics data are required to improve accountability to commitments made to improve child health. Furthermore, developing a deeper understanding of the drivers of this coverage change, through quantitative analysis of coverage data coupled with assessment of national policy impacts, is also needed to develop evidence-based recommendations tailored to each Countdown country context. Ideally, this involves strengthening individual countries' capacity for conducting these analyses.

Conclusion

ARIs, diarrhoea, malaria, and injuries are the main causes of illness and death in children 1–59 months. Among newborn babies, most deaths are the result of preterm birth complications, intrapartum-related events such as birth asphyxia, and infections. Undernutrition is associated with 45% of all under-5 deaths. Despite the substantial progress made in reducing under-5 mortality, 5.4 million children under 5 died in 2017, 2.5 million of these being neonates.

Poor health and survival are linked to social and economic disadvantages. Child deaths are unequally distributed in the world. Within countries, child morbidity and mortality tend to be higher in the rural areas and within the poorer and least educated families. Poorly nourished, LBW, sick, and disabled children all are at particular risk of adverse outcomes and have special needs for care. Global progress in child health is attributable to a broad range of factors that include high coverage of good-quality, effective interventions, development programming, and strategic integrated delivery of packaged health services, along with improvements in education, child protection, respect for human rights, and economic gains in developing countries.

Key points

Approximately two-thirds of under-5 deaths could be prevented by universal coverage of a small set of existing interventions that do not require expensive technology. Among these interventions, the promotion of breastfeeding, management and prevention of neonatal conditions, ARIs, diarrhoea, malaria, and HIV infection, in addition to improved feeding, have the highest impact in reducing child mortality.

Reducing stillbirths and neonatal mortality requires high-quality coverage of key interventions delivered during pregnancy, childbirth, and in the postnatal period.

Many more children today survive infancy yet remain in impoverished conditions. Infections and chronic malnutrition during infancy and childhood delay growth and development, and thus promoting interventions for survival alone is inadequate.

Supporting early child development during critical time periods facilitates optimal developmental trajectories in which children grow into adults that can realize their full potential.

Coverage of key interventions remains unacceptably low. Integrating interventions into packages of care organized along a continuum through pregnancy, birth, the neonatal period, and childhood, and delivered at all levels—from the home and community to first-level facilities and hospitals—would contribute to improved efficiency in intervention delivery and to the achievement of high coverage.

Strengthening health systems is a prerequisite for increasing the coverage and quality of health interventions.

The development of an operational plan to which programmes can be held accountable, along with a good financing and human resource strategy, will yield long-term benefits for the child and the health system at large and is an essential element to follow the development of a child survival strategy.

REFERENCES

Bartels, S.A., et al. (2018). Making sense of child, early and forced marriage among Syrian refugee girls: a mixed methods study in Lebanon. *BMJ Global Health*, 3(1), e000509.

Bhandari, N., et al. (2008). Effectiveness of zinc supplementation plus oral rehydration salts compared with oral rehydration salts alone as a treatment for acute diarrhea in a primary care setting: a cluster randomized trial. *Pediatrics*, 121, e1279–85.

Bhutta, Z.A., et al. (2013). Interventions to address deaths from childhood pneumonia and diarrhoea equitably: what works and at what cost? *The Lancet*, **381**, 1417–29.

Bhutta, Z.A., et al. (2014). Can available interventions end preventable deaths in mothers, newborn babies, and stillbirths, and at what cost? *The Lancet*, **384**, 347–70.
Black, M.M., et al. (2017). Early childhood development coming of age: science through the life course. *The Lancet*, **389**, 77–90.
Black, R.E., et al. (2013). Maternal and child undernutrition and overweight in low-income and middle-income countries. *The Lancet*, **382**, 427–51.
Blencowe, H., et al.(2019). National, regional, and worldwide estimates of low birthweight in 2015, with trends from 2000: a systematic analysis. *The Lancet Global Health*, 7(7), e849–60.
Boerma, T., et al. (2018). Countdown to 2030: tracking progress towards universal coverage for reproductive, maternal, newborn, and child health. *The Lancet*, 391(10129), 1538–48.
Britto, P.R. et al. (2017). Nurturing care: promoting early childhood development. *The Lancet*, **389**, 91–102.
Bruce, N., et al. (2013). Control of household air pollution for child survival: estimates for intervention impacts. *BMC Public Health*, 13(3). https://doi.org/10.1186/1471-2458-13-S3-S8.
Chawanpaiboon, S., et al. (2018). Global, regional, and national estimates of levels of preterm birth in 2014: a systematic review and modelling analysis. *The Lancet Global Health*, 7(1), e37–e46.
Chintu, C., et al. (2004). Co-trimoxazole as prophylaxis against opportunistic infections in HIV-infected Zambian children (CHAP): a double-blind randomized placebo-controlled trial. *The Lancet*, **364**, 1865–71.
Coovadia, H.M., et al. (2012). Efficacy and safety of an extended nevirapine regimen in infant children of breastfeeding mothers with HIV-1 infection for prevention of postnatal HIV-1 transmission (HPTN 046): a randomised, double-blind, placebo-controlled trial. *The Lancet*, **379**, 221–8.
Countdown Working Group on Health Policy and Health Systems, et al. (2008). Assessment of the health system and policy environment as a critical complement to tracking intervention coverage for maternal newborn and child health. *The Lancet*, **371**, 1284–93.
Dangour, A.D., et al. (2013). Interventions to improve water quality and supply, sanitation and hygiene practices, and their effects on the nutritional status of children. *Cochrane Database of Systematic Reviews*, (8), CD009382.
Darvesh, N., Das, J.K., Vaivada, T., Gaffey, M.F., Rasanathan, K., and Bhutta, Z.A., (2017). Water, sanitation and hygiene interventions for acute childhood diarrhea: a systematic review to provide estimates for the Lives Saved Tool. *BMC Public Health*, 17(4), 776.
Das, J.K., Ali, A., Salam, R.A., and Bhutta, Z.A. (2013). Antibiotics for the treatment of *Cholera*, *Shigella* and *Cryptosporidium* in children. *BMC Public Health*, 13(Suppl. 3), S10.
Ezeamama, A.E., et al. (2018). Cognitive deficits and educational loss in children with schistosome infection—a systematic review and meta-analysis. *PLoS Neglected Tropical Diseases*, 12(1), e0005524.
Fadel, S.A., et al. (2019). Trends in cause-specific mortality among children aged 5–14 years from 2005 to 2016 in India, China, Brazil, and Mexico: an analysis of nationally representative mortality studies. *The Lancet*, 393(10176), 1119–27.
GBD Diarrhoeal Diseases Collaborators (2017). Estimates of global, regional, and national morbidity, mortality, and aetiologies of diarrhoeal diseases: a systematic analysis for the Global Burden of Disease Study 2015. *The Lancet Infectious Diseases*, 17(9), 909–48.
Gera, T., Shah, D., Garner, P., Richardson, M., and Sachdev, H.S. (2016). Integrated management of childhood illness (IMCI) strategy for children under five. *Cochrane Database of Systematic Reviews*, (6), CD010123.
Guha-Sapir, D. and D'Aoust, O. (2011). *Demographic and Health Consequences of Civil Conflict*. Washington, DC: World Bank. [Online] Available at: https://openknowledge.worldbank.org/handle/10986/9083
Gülmezoglu, A.M., et al. (2016). Interventions to reduce maternal and newborn morbidity and mortality. In: Black, R.E., Laxminarayan, R., Temmerman, M., and Walker, N. (eds.)

Reproductive, Maternal, Newborn, and Child Health: Disease Control Priorities (3rd ed.) (*Volume 2*). Washington, DC: The World Bank.

Hoddinott, J., Alderman, H., Behrman, J. R., Haddad, L., and Horton, S. (2013). The economic rationale for investing in stunting reduction. *Maternal & Child Nutrition*, **9**, 69–82.

Hodgkin, R. and Newell, P. (2007). *Implementation Handbook for the Convention on the Rights of the Child* (3rd ed.). New York: UNICEF.

Hotez, P.J., et al. (2015). Neglected tropical diseases among the Association of Southeast Asian Nations (ASEAN): overview and update. *PLoS Neglected Tropical Diseases*, 9(4), e0003575.

Hug, L., et al. (2019). National, regional, and global levels and trends in neonatal mortality between 1990 and 2017, with scenario-based projections to 2030: a systematic analysis. *The Lancet Global Health*, 7(6), e710–20.

Intergovernmental Panel on Climate Change (IPCC) (2014). *Climate Change 2014–Impacts, Adaptation and Vulnerability: Regional Aspects*. Cambridge: Cambridge University Press.

Kagaayi, J., et al. (2008). Survival of infants born to HIV-positive mothers, by feeding modality, in Rakai, Uganda. *PLoS One*, 3(12), e3877.

Kalyango, J.N., Alfven, T., Peterson, S., Mugenyi, K., Karamagi, C., and Rutebemberwa, E., (2013). Integrated community case management of malaria and pneumonia increases prompt and appropriate treatment for pneumonia symptoms in children under five years in Eastern Uganda. *Malaria Journal*, 12(1), 340.

Kesho Bora Study Group (2011). Triple antiretroviral compared with zidovudine and single-dose nevirapine prophylaxis during pregnancy and breastfeeding for prevention of mother-to-child transmission of HIV-1 (Kesho Bora study): a randomised controlled trial. *The Lancet Infectious Diseases*, **11**, 171–80.

Keusch, G.T., et al. (2014). Environmental enteric dysfunction: pathogenesis, diagnosis, and clinical consequences. *Clinical Infectious Diseases*, 59(suppl_4), S207–S212.

Khan, J., Vesel, L., Bahl, R., and Martines, J. (2014). Timing of breastfeeding initiation and exclusivity of breastfeeding during the first month of life: effects on neonatal mortality and morbidity—a systematic review and meta-analysis. *Maternal and Child Health Journal*, 19(3), 468–79.

Lassi, Z., Das, J., Zahid, G., Imdad, A., and Bhutta, Z. (2013). Impact of education and provision of complementary feeding on growth and morbidity in children less than 2 years of age in developing countries: a systematic review. *BMC Public Health*, 13(Suppl 3), S13.

Lazzerini, M. and Wanzira, H. (2016). Oral zinc for treating diarrhoea in children. *Cochrane Database of Systematic Reviews*, (12), CD005436.

Lee, A.C., et al. (2011). Care during labor and birth for the prevention of intrapartum-related neonatal deaths: a systematic review and Delphi estimation of mortality effect. *BMC Public Health*, 11(3), S10.

Liu, L., et al. (2016). Global, regional, and national causes of under-5 mortality in 2000–15: an updated systematic analysis with implications for the Sustainable Development Goals. *The Lancet*, 388(10063), 3027–35.

Lu, C., Black, M.M., and Richter, L.M. (2016). Risk of poor development in young children in low-income and middle-income countries: an estimation and analysis at the global, regional, and country level. *The Lancet Global Health*, 4(12), e916–22.

Mackey, T.K., Liang, B.A., Cuomo, R., Hafen, R., Brouwer, K.C., and Lee, D.E. (2014). Emerging and reemerging neglected tropical diseases: a review of key characteristics, risk factors, and the policy and innovation environment. *Clinical Microbiology Reviews*, 27(4), 949–79.

Marsac, M.L., et al. (2016). Implementing a trauma-informed approach in pediatric health care networks. *JAMA Pediatrics*, 170(1), 70–7.

Masquelier, B., et al. (2018). Global, regional, and national mortality trends in older children and young adolescents (5–14 years) from 1990 to 2016: an analysis of empirical data. *The Lancet Global Health*, 6(10), e1087–99.

Mazumder, S., et al. (2010). Effectiveness of zinc supplementation plus oral rehydration salts compared with oral rehydration salts alone for diarrhoea in infants aged less than 6 months in Haryana state, India. *Bulletin of the World Health Organization*, **88**, 754–60.

Mills, J.E. and Cumming, O. (2016). *The Impact of Water, Sanitation and Hygiene on Key Health and Social Outcomes*. New York: UNICEF. [Online] Available at: https://www.unicef.org/wash/files/The_Impact_of_WASH_on_Key_Social_and_Health_Outcomes_Review_of_Evidence.pdf

Mitra, A. and Mawson, A. (2017). Neglected tropical diseases: epidemiology and global burden. *Tropical Medicine and Infectious Disease*, 2(3), 36.

Munos, M.K., Walker, C.L., and Black, R.E. (2010). The effect of oral rehydration solution and recommended home fluids on diarrhoea mortality. *International Journal of Epidemiology*, 39(Suppl. 1), i75–87.

Mwaniki, M.K., Atieno, M., Lawn, J.E., and Newton, C.R. (2012). Long-term neurodevelopmental outcomes after intrauterine and neonatal insults: a systematic review. *The Lancet*, 379(9814), 445–52.

Naghavi, M., et al. (2017). Global, regional, and national age-sex specific mortality for 264 causes of death, 1980–2016: a systematic analysis for the Global Burden of Disease Study 2016. *The Lancet*, 390(10100), 1151–210.

Østby, G. and Urdal, H. (2010). *Education and Civil Conflict: A Review of the Quantitative, Empirical Literature*. Paris, France: UNESCO. [Online] Available at: https://unesdoc.unesco.org/ark:/48223/pf0000190777

Østby, G., Rustad, S.A., and Tollefsen, A.F. (2018) *Children Affected by Armed Conflict, 1990–2017, Conflict Trends, 10*. Oslo: PRIO.

Pabalan, N., Singian, E., Tabangay, L., Jarjanazi, H., Boivin, M.J., and Ezeamama, A.E. (2018). Soil-transmitted helminth infection, loss of education and cognitive impairment in school-aged children: a systematic review and meta-analysis. *PLoS Neglected Tropical Diseases*, 12(1), e0005523.

Pan American Health Organization and World Health Organization (2003). *Guiding Principles for Complementary Feeding of the Breastfed Child*. Washington, DC: Pan American Health Organization.

Peltier, C.A., et al. (2009). Breastfeeding with maternal antiretroviral therapy or formula feeding to prevent HIV postnatal mother-to-child transmission in Rwanda. *AIDS*, **23**, 2415–23.

Ruel, M.T., Alderman, H., and Maternal and Child Nutrition Study Group (2013). Nutrition-sensitive interventions and programmes: how can they help to accelerate progress in improving maternal and child nutrition?. *The Lancet*, 382(9891), 536–51.

Ruger, J.P. (2012). An alternative framework for analyzing financial protection in health. *PLoS Medicine*, 9(8), e1001294.

Sandison, T.G., et al. (2011). Protective efficacy of co-trimoxazole prophylaxis against malaria in HIV exposed children in rural Uganda: a randomised clinical trial. *British Medical Journal*, 342, d1617.

Sankar, M., et al. (2015). Optimal breastfeeding practices and infant and child mortality: a systematic review and meta-analysis. *Acta Paediatrica*, **104**, 3–13.

Shapiro, R.L., et al. (2010). Antiretroviral regimens in pregnancy and breast-feeding in Botswana. *The New England Journal of Medicine*, **362**, 2282–94.

Stanaway, J.D., et al. (2019). The global burden of typhoid and paratyphoid fevers: a systematic analysis for the Global Burden of Disease Study 2017. *The Lancet Infectious Diseases* 19(4), 369–81.

Sudfeld, C.R., et al. (2015). Linear growth and child development in low- and middle-income countries: a meta-analysis. *Pediatrics*, 135, e1266–75.

Traa, B.S., Walker, C.L., Munos, M., and Black, R.E. (2010). Antibiotics for the treatment of dysentery in children. *International Journal of Epidemiology*, 39(Suppl. 1), i70–4.

Troeger, C., et al. (2018). Estimates of the global, regional, and national morbidity, mortality, and aetiologies of diarrhoea in 195 countries: a systematic analysis for the Global Burden of Disease Study 2016. *The Lancet Infectious Diseases*, 18(11), 1211–28.

United Nations; Office of the Special Representative of the Secretary-General for Children and Armed Conflict (2013). *Working Paper No 1: The Six Grave Violations Against Children During Armed Conflict: The Legal Foundation*. New York: United Nations. [Online] Available at: https://childrenandarmedconflict.un.org/publications/WorkingPaper-1_SixGraveViolationsLegalFoundation.pdf

United Nations High Commissioner for Refugees (UNHCR) (2017). *Global Trends: Forced Displacement in 2017*. Geneva: UNHCR. [Online] Available at: https://www.unhcr.org/5b27be547.pdf

UNHCR (2018). *Figures at a Glance: Statistical Yearbooks 2016*. [Online] Available from: http://www.unhcr.org/en-us/figures-at-a-glance.html

United Nations Children's Fund and World Health Organization (2009). *Diarrhoea: Why Children Are Still Dying and What Can Be Done*. Geneva: WHO; New York: UNICEF. [Online] Available at: http://www.who.int/maternal_child_adolescent/documents/9789241598415/en/index.html.

United Nations Inter-agency Group for Child Mortality Estimation (UN IGME) (2017). *Levels & Trends in Child Mortality: Report 2017, Estimates Developed by the United Nations Inter-agency Group for Child Mortality Estimation*. New York: UNICEF. [Online] Available at: https://www.unicef.org/publications/index_101071.html

United Nations Inter-agency Group for Child Mortality Estimation (UN IGME) (2018). *Levels & Trends in Child Mortality: Report 2018, Estimates Developed by the United Nations Inter-agency Group for Child Mortality Estimation*. New York: UNICEF. [Online] Available at: https://www.unicef.org/publications/index_103264.html

United Nations General Assembly (UNGA) (2015). *Resolution 70/1: Transforming our world: the 2030 Agenda for Sustainable Development (25 September 2015)*. A/RES/70/1. [Online]. Available at: http://www.un.org/ga/search/view_doc.asp?symbol=A/RES/70/1

UNICEF/WHO/World Bank (2019). *Joint Child Malnutrition Estimates, March 2019 edition*. Available at: https://www.who.int/nutgrowthdb/estimates2018/en/

Vaivada, T., Gaffey, M.F., and Bhutta, Z.A. (2017). Promoting early child development with interventions in health and nutrition: a systematic review. *Pediatrics*, 140, e20164308.

Vandemark, L.M., Jia, T.W., and Zhou, X.N. (2010). Social science implications for control of helminth infections in Southeast Asia. In: *Advances in Parasitology* (Vol. 73). Cambridge, MA: Academic Press, pp. 137–70.

Wahl, B., et al. (2018). Burden of *Streptococcus pneumoniae* and *Haemophilus influenzae* type b disease in children in the era of conjugate vaccines: global, regional, and national estimates for 2000–15. *Lancet Global Health*, 6, e744–57.

Walker, C., et al. (2013). Global burden of childhood pneumonia and diarrhoea. *The Lancet*, 381(9875), 1405–16.

Watkins, K., et al (2016). The State of the World's Children 2016: a fair chance for every child. [online] New York: *United Nations Children's Fund (UNICEF)*. [Online] Available at: https://www.unicef.org/publications/files/UNICEF_SOWC_2016.pdf

World Health Organization (WHO) (2005a). *Diarrhoea Treatment Guidelines for Clinic Based Health Care Workers*. Geneva: WHO. [Online] Available at: http://www.who.int/maternal_child_adolescent/documents/a85500/en/index.html

World Health Organization (WHO) (2005b). *Guiding Principles for Feeding Non-Breastfed Children 6–24 Months of Age*. Geneva: WHO.

World Health Organization (WHO) (2006a). *Multicentre Growth Reference Study Group. WHO Child Growth Standards: Length/Height-for-Age, Weight-for-Age, Weight-for-Length, Weight-for-Height and Body Mass Index-for-Age: Methods and Development*. Geneva: WHO.

World Health Organization (WHO) (2006b). *World Health Report 2006: Working Together for Health*. Geneva: WHO.
World Health Organization (WHO) (2010a). *Antiretroviral Drugs for Treating Pregnant Women and Preventing HIV Infection in Infants*. Geneva: WHO. [Online] Available at: http://www.who.int/hiv/pub/mtct/antiretroviral2010/en/index.html
World Health Organization (WHO) (2010b). *Community-based Rehabilitation: CBR Guidelines*. Geneva: WHO. [Online] Available at: https://www.who.int/disabilities/cbr/guidelines/en/
World Health Organization (WHO) (2010c). *Monitoring the Building Blocks of the Health System: A Handbook of Indicators and their Measurement Strategies*. Geneva: WHO. [Online] Available at: http://www.who.int/healthinfo/systems/monitoring/en/index.html
World Health Organization (WHO) (2011a). *Guideline: Intermittent Iron Supplementation in Preschool and School-Age Children*. Geneva: WHO. [Online] Available at: http://whqlibdoc.who.int/publications/2011/9789241502009_eng.pdf
World Health Organization (WHO) (2011b). *Guideline: Vitamin A Supplementation in Infants and Children 6–59 Months of Age*. Geneva: WHO. [Online] Available at: http://whqlibdoc.who.int/publications/2011/9789241501767_eng.pdf
World Health Organization (WHO) (2012a). *Developmental Difficulties in Early Childhood: Prevention, Early Identification, Assessment and Intervention in Low- And Middle-Income Countries: A Review*. Geneva: WHO. [Online] Available at: https://www.who.int/maternal_child_adolescent/documents/development_difficulties_early_childhood/en/
World Health Organization (WHO) (2012b). *Countdown to 2015. Accountability for Maternal, Newborn & Child Survival. An Update on Progress in Priority Countries*. Geneva: WHO. [Online] Available at: http://countdown2015mnch.org/documents/countdown-news/count_complete_small.pdf
World Health Organization (WHO) (2013a). *Global Health Observatory Data Repository* [online database]. Geneva, World Health Organization. [Online] Available at: http://apps.who.int/gho/data/?theme=main
World Health Organization (WHO) (2013b). WHO position paper on rotavirus vaccines, 2013. *Weekly Epidemiological Record*, **88**, 49–64. [Online] Available at: https://www.who.int/wer/2013/wer8805.pdf?ua=1
World Health Organization (WHO) (2013c). *Guideline: Updates on the Management of Severe Acute Malnutrition in Infants and Children*. Geneva: WHO.
World Health Organization (WHO) (2013d). *Sustaining the Drive to Overcome the Global Impact of Neglected Tropical Diseases: Second WHO Report on Neglected Diseases* (No. WHO/HTM/NTD/2013.1). Geneva: WHO.
World Health Organization (WHO) (2014a). *Revised WHO Classification and Treatment of Childhood Pneumonia at Health Facilities*. Geneva: WHO.
World Health Organization (WHO) (2014b). *Guidelines on Post-exposure Prophylaxis for HIV and the Use of Co-trimoxazole Prophylaxis for HIV-related Infections Among Adults, Adolescents and Children*. Geneva: WHO. [Online] Available at: https://www.who.int/hiv/pub/guidelines/arv2013/arvs2013upplement_dec2014/en/
World Health Organization (WHO) (2015). *Global Technical Strategy for Malaria 2016–2030*. Geneva: WHO. [Online] Available at: http://apps.who.int/iris/bitstream/handle/10665/176712/9789241564991_eng.pdf
World Health Organization (WHO) (2016a). *Guideline: Updates on HIV and Infant Feeding*. Geneva: WHO. [Online] Available at: https://www.who.int/maternal_child_adolescent/documents/hiv-infant-feeding-2016/en/
World Health Organization (WHO) (2016b). *Global Health Sector Strategy on HIV, 2016–2021*. Geneva: WHO. [Online] Available at: https://www.who.int/hiv/strategy2016-2021/ghss-hiv/en/

World Health Organization (WHO) (2016c). *Guideline: Daily Iron Supplementation in Infants and Children.* Geneva: WHO. [Online] Available at: https://www.who.int/nutrition/publications/micronutrients/guidelines/daily_iron_supp_childrens/en/

World Health Organization (WHO) (2016d). *Guideline: Use of Multiple Micronutrient Powders for Point-of-Use Fortification of Foods Consumed by Infants and Young Children Aged 6–23 Months and Children Aged 2–12 Years.* Geneva: WHO. [Online] Available at: https://www.who.int/nutrition/publications/micronutrients/guidelines/mmpowders-infant6to23mons-children2to12yrs/en/

World Health Organization (WHO) (2016e). *Guideline: Fortification of Maize Flour and Corn Meal with Vitamins and Minerals*. Geneva: WHO. [Online] Available at: https://www.who.int/nutrition/publications/micronutrients/guidelines/maize-corn-fortification/en/

World Health Organization (WHO) (2017a). *Guideline: Assessing and Managing Children at Primary Health-Care Facilities to Prevent Overweight and Obesity in the Context of the Double Burden of Malnutrition. Updates for the Integrated Management of Childhood Illness (IMCI).* Geneva: WHO. [Online] Available at: https://www.who.int/nutrition/publications/guidelines/children-primaryhealthcare-obesity-dbm/en/

World Health Organization (WHO) (2017b). *Integrated Management of Childhood Illness (IMCI) Global Survey Report*. Geneva: WHO. [Online] Available at: https://www.who.int/maternal_child_adolescent/documents/imci-global-survey-report/en/

World Health Organization (WHO) (2017c). *Global Strategy for Women's, Children's and Adolescents' Health (2016–2030): Adolescents' Health. Report by the Secretariat: Seventieth World Health Assembly*. Geneva: WHO. [Online] Available at: http://apps.who.int/gb/ebwha/pdf_files/WHA70/A70_37-en.pdf?ua=1

World Health Organization (WHO) (2018a). *The Top 10 Causes of Death*. Geneva: WHO. [Online] Available at: http://www.who.int/en/news-room/fact-sheets/detail/the-top-10-causes-of-death

World Health Organization (WHO) (2018b). *Malaria Vaccine Implementation Programme (MVIP) —Progress Update*. Geneva: WHO. [Online] Available at: http://www.who.int/immunization/sage/meetings/2018/april/2_WHO_MalariaMVIPupdate_SAGE_Apr2018.pdf

World Health Organization (WHO) (2018c). *Guideline: Fortification of Rice with Vitamins and Minerals as a Public Health Strategy*. Geneva: WHO. [Online] Available at: https://www.who.int/nutrition/publications/guidelines/rice-fortification/en/

World Health Organization/UNICEF (WHO/UNICEF) (2018). *Levels and Trends in Child Malnutrition—UNICEF/WHO/World Bank Group Joint Child Malnutrition Estimates—Key findings of the 2018 edition*. UNICEF/WHO/World Bank. [Online] Available at: https://www.who.int/nutgrowthdb/2018-jme-brochure.pdf?ua=1

10.4

Adolescent health

George Patton, Peter Azzopardi, Natasha Kaoma, Farnaz Sabet, and Susan Sawyer

Introduction

The current generation of 10- to 24-year-olds is the largest in human history. Their health and well-being are increasingly shaped by the global forces of urbanization, migration, global communications, economic development, and the sustainability of planetary ecosystems. Over the past two decades we have come to new understandings of adolescence as a critical developmental phase in life for achieving human potential (Patton et al. 2016). It is characterized by maturation of all physiological systems including reproductive, musculoskeletal, cardiometabolic, and neurocognitive. Health during adolescence shapes patterns of growth and the acquisition of social and emotional capabilities, sets trajectories for health across the life-course and, because this is the next generation to parent, lays a foundation for the healthy development of the next generation (Patton et al. 2016). Thus investments in adolescence have the potential to bring a 'triple dividend', an understanding that has brought a fresh prominence to adolescent health in global policy (Every Woman Every Child 2015). However, the challenge for governments in investing in adolescent health is a relative lack of capacity extending from the health workforce to healthcare systems, monitoring frameworks and evidence base for policy and programming, and processes relative to other fields of health (Patton et al. 2016). This chapter outlines current understandings of adolescent health, growth, and development, the available evidence for effective interventions and priorities for research, and technical investment.

Concepts of adolescence

Adolescence is generally conceptualized as the phase of life that spans childhood and adulthood. It is a phase of major growth with neurocognitive maturation extending well into the third decade. Through history, adolescence has been considered to begin with puberty and end with transitions into marriage and parenthood (Worthman and Trang 2018). In today's contexts, the endpoint is often less clear-cut, particularly in higher income and urban settings where the adoption of other adult roles such as the transition to employment, financial independence, as well as the formation of life partnerships commonly mark the end of adolescence. These events occur at different ages in different parts of the world influencing local cultural concepts of adolescence, which vary greatly.

Sawyer et al. recently reviewed the range of different English terms have been used to describe the adolescent phase of life or parts of it (Fig. 10.4.1) (Sawyer et al. 2018). Within the UN system, adolescence has been generally defined as the second decade from 10 to 19 years and youth commonly refers to 15 to 24 years. The terms 'adolescents and young adults' and 'young people' are also used to refer to the 10–24-year-old age group (World Health 1989; Sawyer et al. 2012).

The Convention on the Rights of the Child defines a child as below the age of 18 years unless, under the laws applicable to the child, majority is attained earlier (Assembly 1989). Eighteen is commonly viewed as the legal age of majority, the point at which an individual is considered an adult in many countries, although not universally. In law there is no single definition of adulthood but rather a collection of laws that at differing ages, for different activities, bestow the status of adulthood. These include laws that have major consequences for health and health risks including those related to the age of consent to sexual intercourse or medical treatment, the minimum age that young people can legally work, leave school, drive, buy alcohol, marry, and be held accountable for criminal actions. In addition to 'young adults', 'emerging adulthood' is a term that has been used to describe the phase of life from the late teens to the late twenties when an individual acquires some of characteristics of adulthood without having necessarily reached the milestones that historically define fully-fledged adulthood.

This chapter focuses on this age range from 10 to 24 years, coinciding with the major years of physical, cognitive, and social development and the key social role transitions that mark adulthood. When we use the terms

‘adolescence and young adulthood’, or simply ‘adolescence’, we are referring to these years.

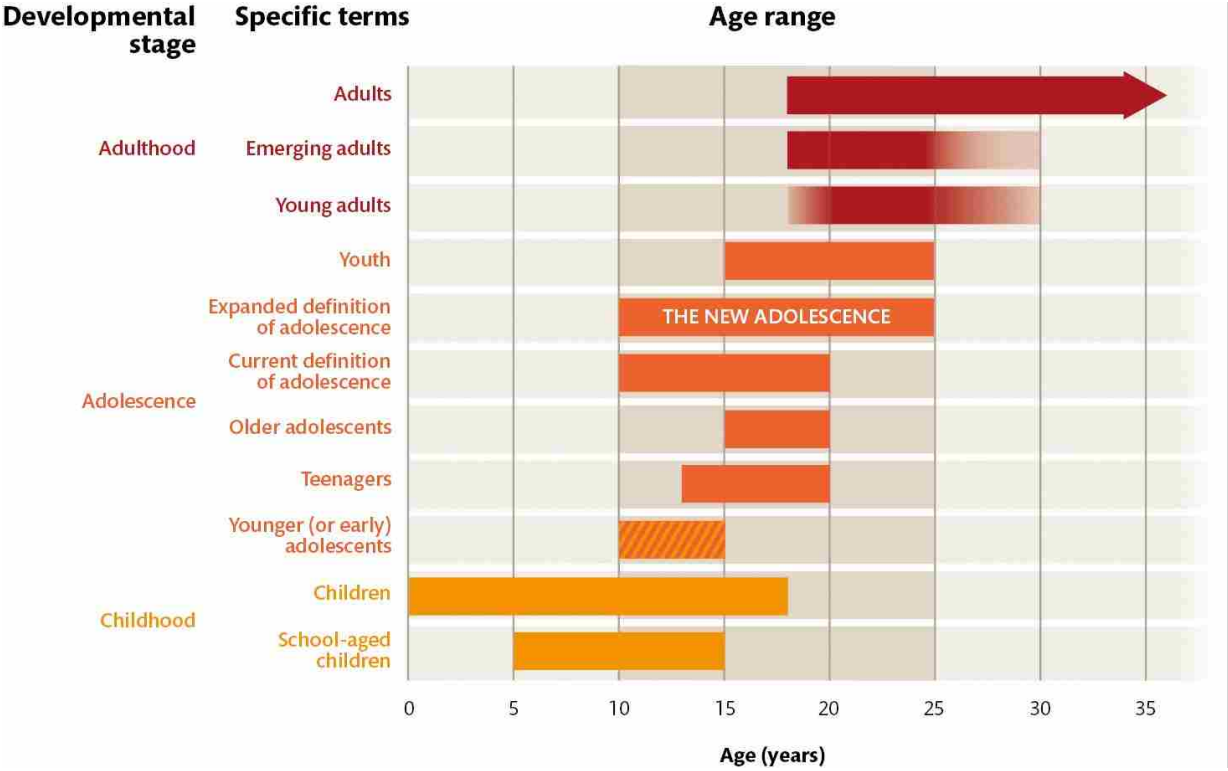


Fig. 10.4.1 Different terms used to describe the adolescent and young adult years.
Reprinted from Sawyer, S. M., et al. (2018), ‘The age of adolescence’, *Lancet Child and Adolescent Health*, 2(3):223–8, with permission from Elsevier.

Adolescent growth and development

Timing of adolescent growth and development

Later childhood and adolescence are increasingly viewed as sensitive developmental phases where an individual’s interaction with the physical, nutritional, social, and cultural environment shapes patterns of physical, social, and emotional growth (Patton et al. 2016). These aspects of the environment affect the timing and form of puberty. The effects of this changing environment are evident in secular trends in the age of menarche in girls. Over the past 150 years there has been a fall in the age of menarche of

around 4 to 5 years in Western Europe and North America, which is thought to arise from better childhood health from improved nutrition and reduced infectious disease (Worthman and Trang 2018). Occurring over a shorter period, similar trends have been observed in countries undergoing rapid economic development or in children whose families have moved from low-resource to high-resource settings (Houghton et al. 2014). In high resource settings the mean age for menarche is now 12–13 years, although the onset of puberty occurs much earlier with adrenarche (adrenal maturation) on average occurring around the age of 7–8 years and thelarche (breast budding) around 10 years (Patton and Viner 2007).

The timing of puberty predicts later life health as early puberty increases risks for premature all-cause mortality, cardiovascular deaths, type II diabetes and some mental disorders (Day et al. 2015). During adolescence, early puberty is also associated with higher rates of common mental disorders in girls; a mismatch between the social and sexual roles that accompany physical maturity and an individual's underlying cognitive–emotional capabilities is one possible explanation (Brooks-Gunn and Warren 1989).

In pre-industrial societies, the adolescent transition from puberty to adult roles, as defined by the transition to parenthood, ranged from around 2 years in girls to 4 years in boys (Schlegel and Barry 1991). In today's developed economies, longer periods of education and the availability of effective contraception means that adolescence commonly persists for well over a decade. The lengthening of adolescence means that it occupies a greater proportion of the life-course than ever before. It is also associated with a more diverse set of influences on adolescent health, growth, and development in many places. These include a longer duration of education and training, with the acquisition of professional capabilities and independence increasingly being hallmarks of adult status (Patton and Viner 2007; United Nations Children's Fund 2012). So too relationships with an adolescent's family of origin have shifted in many places with an extension of time living at home into the third and even fourth decade in many places. Opportunities for engagement with the peer group and cultural influences beyond the local neighbourhood have also changed with the growth of mass media, and more recently, digital and social media. For these reasons, influences on adolescent health differ markedly today from those in previous

generations. Equally the opportunities for using new platforms for health interventions have also expanded.

Psychosocial development

Early adolescence (10–14 years) is dominated by puberty and the effects of the rapid rise in pubertal hormones on body morphology, and sexual and brain development. It is accompanied by a remodelling of the brain's reward system, a sensitivity to peer influences, low levels of future orientation and altered risk perceptions, sometimes leading to risk-taking behaviour. It is a time of identity formation and development of new interests including sexual and romantic relationships. School and family environments have major effects on growth and development and are therefore important settings for health intervention.

Mid-adolescence (15–19 years) is characterized by continued pubertal maturation, particularly prominent for males, whose growth spurt starts later and continues for longer. Brain development continues with maturation of the prefrontal cortex and increased connectivity between brain networks enhancing executive and self-regulatory skills. It is a phase of greater autonomy and, in poor resource settings where there are limited opportunities for educational engagement, a common point of transition into marriage and parenthood for girls. In higher resource settings, this phase of development is accompanied by a greater future orientation and an ability to weigh up the short- and long-term implications of decisions. Education settings are influential for those who remain in school: for those who leave school, work settings have a growing influence on health, particularly in hazardous work environments in low-income settings.

Later adolescence or young adulthood (20–24 years) is accompanied by maturation of the prefrontal cortex and greater planning and self-regulation. It marks the end of a period of high brain plasticity. It often corresponds to the adoption of adult roles and responsibilities, including entering the workforce or tertiary education, marriage, childbearing, and economic independence.

Adolescent health in the global context

Demographic and epidemiological transitions

The demographic transition from high birth and early-life death rates to low fertility and longer life expectancy has reshaped the relative proportion of adolescents in high-income countries since the post-war years. The demographic transition is now proceeding rapidly in most other countries with different countries at very different points in that transition (World Bank 2007). This acceleration of the demographic transition has resulted in today's adolescent generation being the largest in history relative to other age groups. With a population of 1.8 billion, 10–24 year olds comprise over a quarter of the global population. Nearly 90% live in low- and middle-income countries where they may comprise more than a third of the population (Fig. 10.4.2). In contrast, in a growing number of high-income countries the proportion of 10- to 24-year-olds is now under 15%. Projecting forward to the end of the century, around a half of the world's adolescents will be growing up in sub-Saharan Africa.

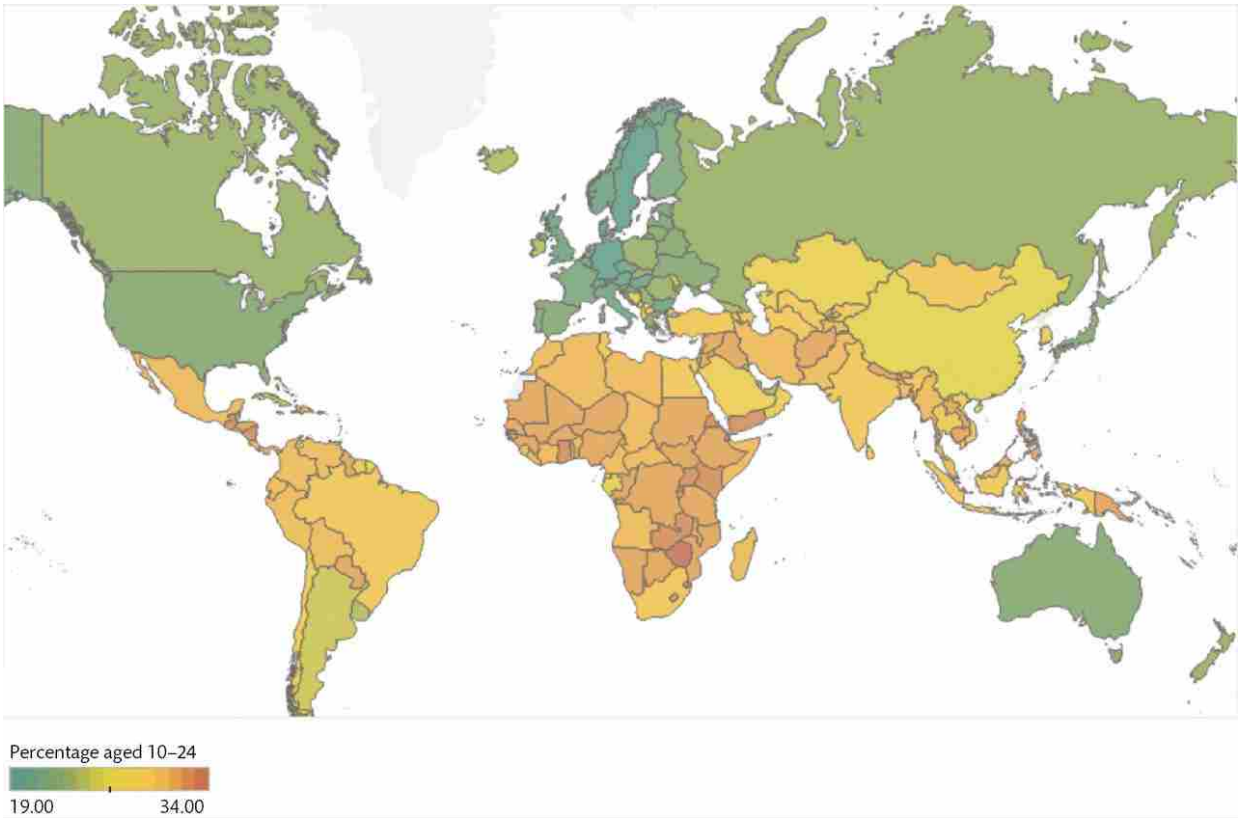


Fig. 10.4.2 Percentage of the population aged 10 to 24 years across countries in 2020.

Better childhood health and nutrition, extended time in education, delays in family formation and access to new technologies theoretically offer a prospect of this being the healthiest generation of adolescents ever. Indeed, many more adolescents are now growing up in non-communicable disease (NCD)-predominant countries with lower levels of reproductive health problems, infectious disease injury, and violence. Yet uneven demographic change means that across Asia and Africa many more young people are growing up with high risks for multiple health problems in the adolescent years (Azzopardi et al. 2019).

The demographic transition has been accompanied by epidemiological and nutritional transitions with profound implications for adolescent health and health risks. The nutritional switch towards greater consumption of foods that are high in added sugars, salt, and unhealthy fats, and low in important micronutrients is fuelling a global rise in obesity but with continuing high levels of micronutrient deficits. Further global trends in youth unemployment, less stable families, environmental degradation, armed conflict, and mass migration also pose new threats to adolescent health.

The epidemiological transition, that commonly accompanies the demographic transition and economic development, has changed health profiles for all age groups. Many adolescents have benefited from better control of infectious diseases including diarrhoeal disease, lower respiratory tract infections, tuberculosis, and malaria. Adolescent girls and young women in many settings have benefited from gains in maternal health, particularly in settings where early pregnancy has fallen markedly. However, the epidemiological transition has brought an increase in other health problems for adolescents, including road traffic injuries and suicide. In some countries and contexts, such as Latin America and Eastern Europe, these problems have extended to homicide, mostly in young males (Patton et al. 2009). A link to broader economic conditions was evident in the rise of mortality and alcohol-related disease burden in Eastern European countries following the fall of the Soviet Union in 1991 and is likely to be a factor underlying more recent civil unrest and armed conflict across the Middle East, North Africa, and sub-Saharan Africa (Yair and Miodownik 2014).

Other adolescent health problems have become more prominent with the epidemiological transition, including more persistent mental and physical health problems. For these reasons, self-harm and suicide have assumed greater significance in adolescent health in recent decades, particularly in high resource settings (Collishaw 2015). Mental disorders have a particular significance in adolescent health as they commonly have their onset in late childhood and adolescence. Because many persist into adulthood, adolescent mental disorders make a greater contribution to adult disease than ever before (Patton et al. 2014).

HIV is a further reason why adolescents and young adults, particularly in southern and eastern sub-Saharan Africa, have to date seen fewer benefits from the epidemiological transition. To a large extent, the burden of HIV has fallen on the young; new infections are concentrated in adolescent girls and young adults; the effectiveness of new antiretroviral medication has also brought the survival into adolescence and adulthood of a generation affected by maternal–child transmission. In this context of continuing high prevalence, adolescents and young adults have not had the same mortality declines as other age groups (Murray et al. 2014).

With ageing populations and a convergence to a disease burden dominated by NCDs in later life, adolescent-prevalent NCDs and adolescent-onset NCD risks assume greater significance (Jamison et al. 2013). In part, the relatively

high prevalence of adolescent NCDs reflect greater survival into adolescence of diseases that were previously fatal in early childhood (e.g. cystic fibrosis and childhood cancers): in part it reflects a changing epidemiology of mental and physical disorders with higher incidence rates in the adolescent years in developed settings, e.g. allergy and autoimmune conditions (Mayer-Davis et al. 2017). NCD risks that commonly emerge during adolescence include tobacco use, physical inactivity, obesity, alcohol, and illicit substance use and poor diets. Without specific policy interventions, the combination of increasing wealth and unrestrained marketing will see adolescent health risks continue to rise in prevalence.

Adolescent health by region

We have adopted the conceptual model from the 2016 Lancet Commission on Adolescent Health and Well-being to describe global and national patterns of adolescent health (Fig. 10.4.3). The model considers:

Current causes of disease burden and death across the adolescent years;
Health risks that emerge during adolescence with consequences for health during these years, for later adult health and/or a healthy start to life for the next generation; and
Important and structural determinants of health that affect health growth and development across the adolescent and young adult years (Sawyer et al. 2012). These include education and employment; marriage and parenthood; marketing and digital media; and the quality of universal health coverage.

The model also depicts the corresponding responses to adolescent health problems and their determinants. The most effective actions for adolescent health and well-being are intersectoral, occur at multiple levels, and with multiple components. They may include structural, media, community, online, and school-based elements as well as the provision of preventive and treatment health services. Intersectoral coordination has, for example, been prominent in countries that have made progress in reducing road traffic injury. In addition to the specific interventions outlined earlier, essential underpinning elements have included strong information systems, clear governance, civil society advocacy and a capacity to implement effectively within different sectors (Jamison et al. 2013).

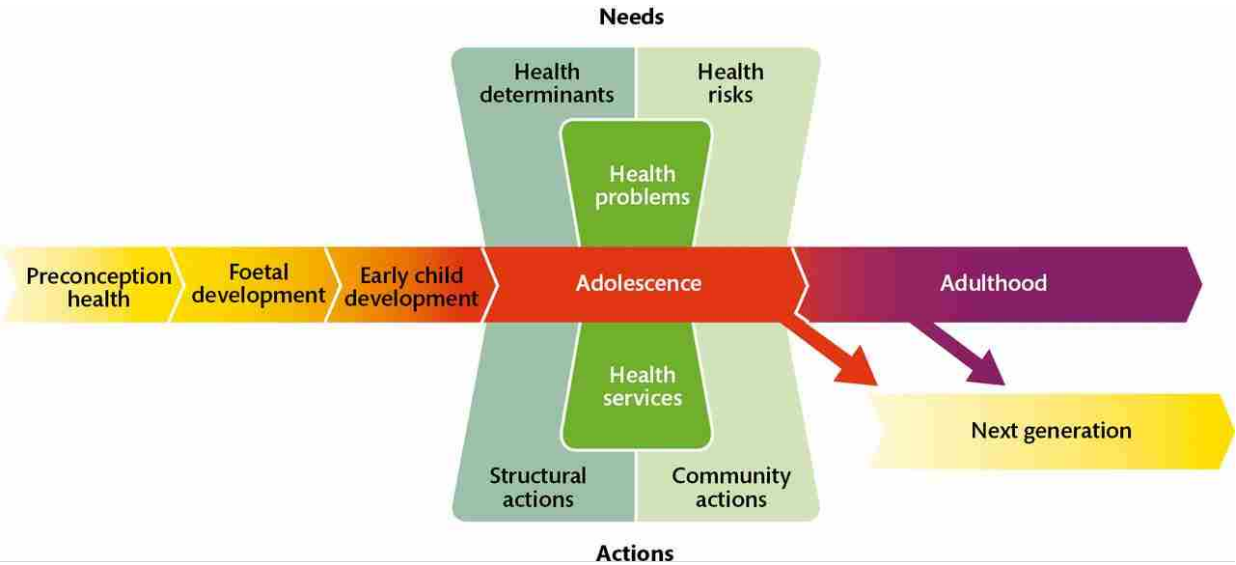


Fig. 10.4.3 A conceptual framework for adolescent health.
Reprinted from Patton, G, et al. (2016), 'Our Future: A Lancet Commission on Adolescent Health and Wellbeing', *Lancet*, **387**, 2423–78, with permission from Elsevier.

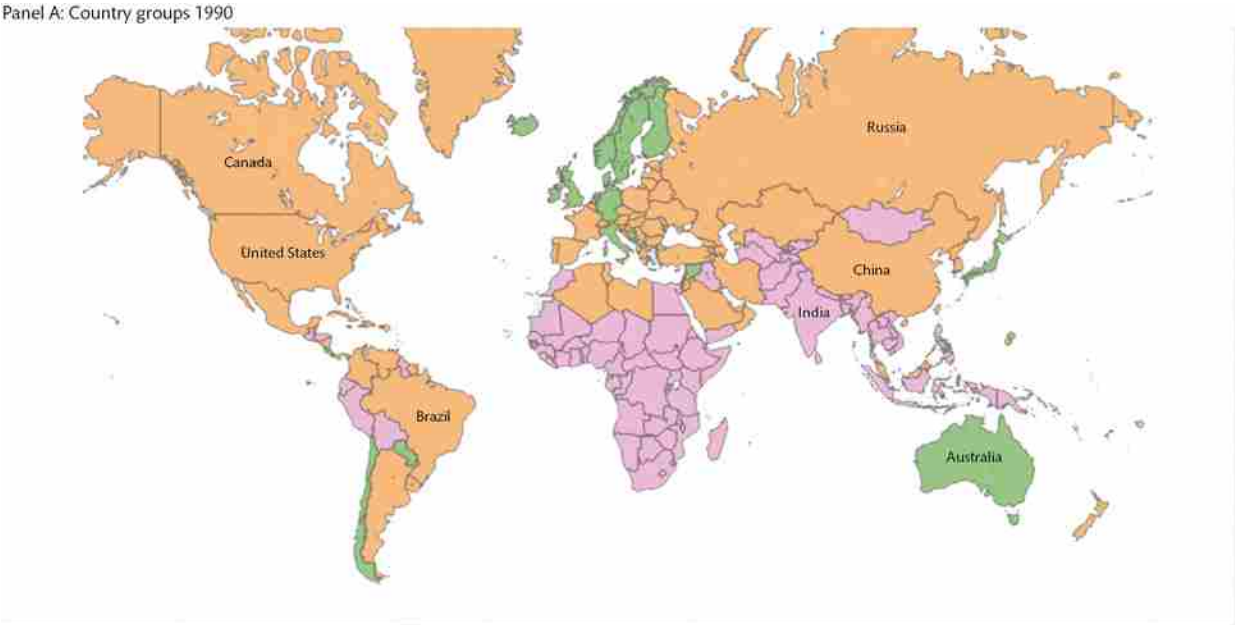
Disease burden across adolescence

Disease burden and mortality rates generally increase across the adolescent years into young adulthood, with those increases particularly marked for sexual and reproductive health problems and common mental disorders in females, and for injuries, violence, and substance use disorders in males.

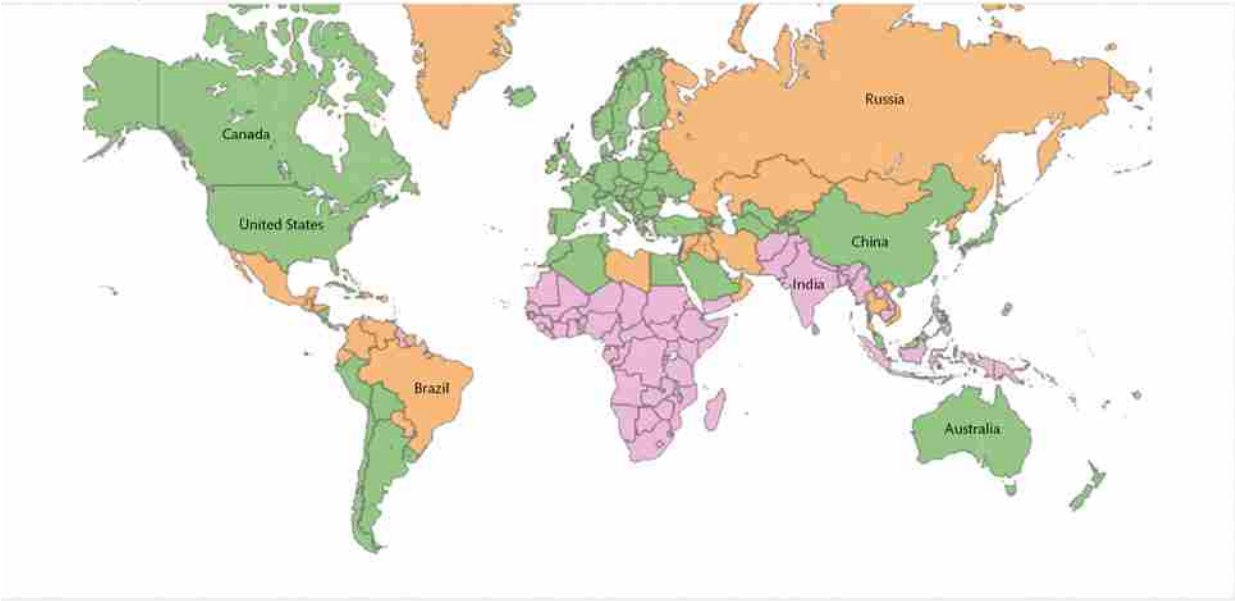
This is in turn reflected in increases in mortality across the adolescent years with particularly marked increases in deaths due to injury, violence, suicide in males and maternal causes for females in later adolescence and young adulthood (Patton et al. 2016). In comparison to younger children, adolescents have had lesser reductions in mortality in recent decades. In a study of 50 countries with longitudinal data, the mortality of children had declined by over 80% in the last 50 years (Grazzi 2004). In contrast, mortality among adolescents had improved only marginally. Indeed, there has been a reversal of historical mortality patterns in which early childhood mortality exceeds adolescent deaths. Over that time causes of death in adolescents have changed with a fall in infectious disease deaths in most high- and middle-income countries and a rise in deaths due to injuries, that now account for around 40% of all adolescent deaths globally.

Azzopardi et al. recently populated 12 headline indicators presented in the 2016 Lancet Commission on Adolescent Health and Wellbeing to examine trends in adolescent health over the past quarter century together with differences between countries and regions (Azzopardi et al. 2019). In that analysis, disability-adjusted life years (DALYs) estimates were used to assess variations in disease burden by country with disease burden grouped into three categories: group 1 conditions (infectious diseases, maternal health problems, undernutrition, and/or HIV) common in low-resource setting; injury and violence; and the NCDs of adolescence (chronic physical illness, mental disorders, and substance use disorders). Countries were classified as multiburden if they had high disease burden from all categories of disease: injury excess with high rates of injury and NCDs; and NCD predominant where the major disease burden around from chronic mental and physical health problems (Fig. 10.4.4).

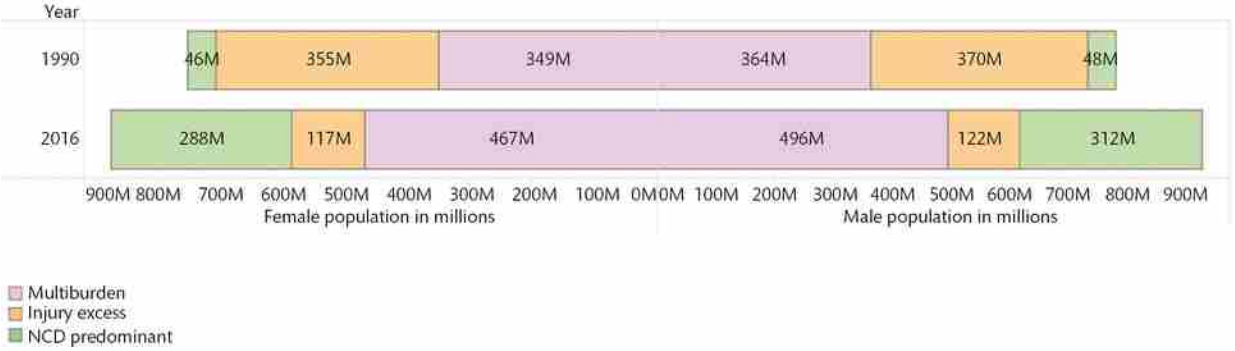
Panel A: Country groups 1990



Panel B: Country groups 2016



Panel C: Population share by group



Multiburden
Injury excess
NCD predominant

Fig. 10.4.4 Changing disease burden from 1990 to 2016 in 10- to 24-year-olds in countries classified as multiburden, injury excess, and NCD predominant.

Reprinted from Azzopardi, Peter S, et al. (2019), 'Progress in adolescent health and wellbeing: tracking 12 headline indicators for 195 countries and territories, 1990–2016', *The Lancet*, under a CC BY 4.0 License.

Figure 10.4.4 illustrates the extent of change in adolescent disease burden over the last 25 years. Particularly notable has been a reduction in adolescent injuries and violence in many places, with a majority of injury-excess countries in 1990 having become NCD predominant. There has been relatively less success in ensuring the transition of countries with a multiburden disease profile. Many of these countries are the countries that have seen the greatest increase in their adolescent population. For this reason the number of adolescents growing up in multiburden countries in 2016 was 35% greater than in 1990, suggesting that the challenges for health service systems in these places have become substantially greater.

Adolescent health risks

Although adolescent health risks contribute to disease burden during the adolescent and young adult years, their greatest contribution is often towards disease burden later in life or in relation to adverse consequences for the health growth and development of the next generation. Tobacco use is illustrative. It overwhelmingly has its onset in adolescence but causes little or no disease burden during these years. However, it is a major risk factor for NCDs in older adults and maternal smoking during pregnancy is a well-established risk factor for poor foetal growth and later life illness in offspring (Patton et al. 2018).

Overall tobacco use has declined in adolescents globally since 1990 in part due to the implementation of the Framework Convention on Tobacco Control. However, progress has been mixed. Rates of daily smoking remain above 15% in 10–24-year-olds across many European countries with one in four smoking daily in the Russian Federation. Across countries, tobacco use is almost always more common in adolescent males than in females. A number of countries in sub-Saharan Africa, Eastern Europe, and the Middle East have seen rises and in some low-resource countries (e.g. Indonesia and Papua New Guinea), rates of tobacco use are unacceptably high. Notably, non-signatories to the Framework Convention on Tobacco Control have generally seen lesser falls in adolescent and young adult tobacco use.

Alcohol use disorders typically begin during the young adult years, with a younger age of drinking onset being particularly hazardous (Bonomo et al. 2006; Viner and Taylor 2007). Alcohol use carries risks for injury, violence, suicide, and unwanted pregnancy during adolescence, is linked to a range of NCDs in adulthood (cancers, cardiovascular disease, liver disease, and pancreatitis) and greater use in pregnancy has prominent intergenerational harms (Riley and McGee 2005). Both alcohol use and binge drinking are generally more prevalent in males. Although still more common in resource rich countries, binge drinking is becoming prevalent in other places with the actual number of adolescent binge drinkers having increased slightly in recent decades (Azzopardi et al. 2019).

Overweight and obesity increase markedly across adolescence and young adulthood with very high persistence, particularly for obesity (Patton et al. 2011). The risks in later life include premature mortality, chronic disability, type 2 diabetes, ischaemic heart disease, hypertension, and cerebrovascular disease. Preconception maternal (adolescent) obesity increases risks for miscarriage, gestational diabetes, operative delivery, pre-eclampsia, infant perinatal mortality, and macrosomia (Yu et al. 2006). Adolescent overweight and obesity have increased markedly in prevalence across almost all countries since 1990, with dramatic annual increases of around 10% in China and Vietnam. The prevalence of overweight and obesity is around one in five young people in NCD Predominant and Injury Excess countries and is continuing to increase in all places. Since 1990, the increase in overweight and obesity in NCD-predominant and injury-excess countries has been 120% (Azzopardi et al. 2019).

Sexual and reproductive health risks also typically emerge during adolescence, with risks shifting in an environment of greater access to digital technology and social media (e.g. sexting, higher risk sex with multiple partners). Globally there have been strong trends towards reduced adolescent fertility, but these trends are less prominent in sub-Saharan Africa and Latin America. Rates of women giving birth before 18 years remain high across much of sub-Saharan Africa, and for a number of countries in South Asia. Overall around one in five women in multiburden countries report giving birth by the age of 18 years.

There is continued debate around the extent to which adolescent mothers are at greater risk of adverse maternal and infant outcomes. Some recent reports using verbal autopsy methods have failed to find differences in

mortality in young mothers (Blanc et al. 2013; Nove et al. 2014). However, a recent World Health Organization (WHO) multicountry survey of pregnancy complications found much greater pregnancy-related morbidity in adolescent mothers (Ganchimeg et al. 2014) and country level studies continue to report higher maternal mortality in young mothers, particularly from causes such as unsafe abortion that may be sensitive to reporting bias in household surveys. An estimated 22 million unsafe abortions occur annually, of which about 15% are among adolescent girls. Young women have a higher risk of abortion-related deaths than in those over 25, are more likely to terminate pregnancy after the first trimester, and are more likely to use unregulated providers (Woog et al. 2015). Even after controlling for various sociodemographic factors, knowledge of the legal status of abortion and partner support, adolescents had 77% lower odds of obtaining a safe abortion compared with women aged 30 and older.

Beyond adolescent birth rates, limited information is available on adolescent sexual and reproductive health in many countries. Across sub-Saharan African and Latin American countries with available data, between 10% and 20% of 15–24-year-olds report having first sexual intercourse before the age of 15 years. Condom use in the context of having had two or more partners in the past year shows great variability between countries, but remains under 50% for adolescents in multiburden countries where rates of adolescent pregnancy, HIV, and sexually transmitted infections (STIs) are high and access to contraception and facilities for menstrual management are low.

Data around adolescent mental disorder and mental health risks is perhaps some of the weakest in global health. One recent estimate of coverage of minimally sufficient data for estimating the prevalence of six major mental disorders of children and adolescents reported an overall global coverage of 6.5%; 124 countries had no usable data at all (Erskine et al. 2017). Yet mental health problems and mental health risks appear common in almost all places. Self-harm, for example, is a risk factor for premature death from suicide and accidental injury; in later life adolescents who self-harm also have higher mortality from cancer and cardiovascular disease (Bergen et al. 2012). At least 1 in 20 adolescents in the mid-teens in high-income countries report some self-harm, typically self-laceration, or self-poisoning, in the previous 12 months (Hawton and James 2005). Other risk factors for mental

disorders including bullying, sexual, and physical abuse and intimate partner violence are rarely monitored in adolescence.

Responding to adolescent health needs

Health services

Health services for adolescents play a number of roles ranging from acute healthcare such as the management of trauma; preventive healthcare such as the provision of preconception and antenatal care to young women; and the management of chronic health problems or health risks including HIV, mental and substance use disorders, and chronic physical illnesses.

Adolescents and young adults commonly fail to access healthcare at the same level as other age groups. Unmet versus met need for contraception are potentially useful indicators of the extent to which adolescents are able to access a sensitive health service. Just under two-thirds of young women aged 15–24 years have their demand for modern contraception satisfied with modern methods, with rates substantially higher in NCD-predominant countries than in multiburden countries where around half of young women have their needs met. Given the rate of demographic change in multiburden countries, the actual numbers of young women with unmet need have substantially increased, suggesting that health services have not kept pace with rapid population growth, particularly in sub-Saharan Africa. To a large extent, data on met need for contraception comes from girls who are married. Far fewer data are available on sexually active unmarried adolescents, particularly in the many countries where prescribing contraception to unmarried girls under the age of 18 years is illegal.

Health service interventions may differ in their effectiveness in adolescents in comparison to older adults for a range of reasons (Cipriani et al. 2016). Rates of contraceptive failure are, for example, higher in adolescents than in older women, with younger women more likely to abandon contraception despite ongoing need (Blanc et al. 2009). Reasons include poor understanding of pregnancy risks, concerns about the effect of contraceptives on health or fertility and opposition from partners. Lack of knowledge of services, cost, shyness, stigma around early sexual activity and judgemental providers are further barriers. One study in the United States found contraceptive failure was almost twice as high in under 21 year olds

using oral contraceptive pills, transdermal patches, or the vaginal ring. In contrast, the failure rate of long-acting reversible contraceptives (LARCs) is extremely low in all age groups (Winner et al. 2012). Yet despite increased use of LARCs in the United States since 2007, 15–24-year-olds remain less likely to use them than 25–39-year-old women (Finer et al. 2012).

Notwithstanding their increased prevalence across adolescence, most adolescents and young adults with mental health problems do not receive treatment from health professionals. One European study of mental healthcare in adults showed that about 6% of the sample was in need of mental healthcare but nearly half reported no formal healthcare use. Those aged 18–24 reported the least use of services (Alonso et al. 2007). Actions to reduce barriers include promoting mental health literacy and combatting stigma, particularly in LMIC where the scaling up of mental health services has not yet been a priority (Patel et al. 2007).

Barriers to healthcare

Two sets of barriers diminish access to quality healthcare for adolescents. One set are external to health services and include legal frameworks governing health actions, out of pocket costs, and cultural and community attitudes. Though not unique to adolescents, these barriers assume greater importance in the young.

Legal and financial barriers to health service access are greater for adolescents than other age groups and are increasingly out of step with the health risks that adolescents face. One example is lack of access to contraception in many countries for unmarried girls under the age of 18 who are sexually active. Even when national legislation allows a particular health response, provider attitudes and beliefs about the appropriateness of an action in the context of age, marital status, or partner/parental consent may affect their response to an adolescent (Calhoun et al. 2013). Financial barriers arise from adolescents' limited independent eligibility for tax and insurance-based funding schemes. Where health insurance exists, it is often through parents (or partners), presenting risks to confidentiality when accessing healthcare for sensitive problems, e.g. mental health, contraception, or substance use (Waddington and Sambo 2015). Socially marginalized adolescents often fall outside of financing systems for healthcare and are particularly vulnerable to catastrophic medical events.

Further barriers to access arise more from the developmental context of adolescence, including adolescents being less experienced users of healthcare. Adolescence brings an increasing desire for privacy and confidentiality, with embarrassment, shame, and fear of being judged presenting barriers to healthcare. In the United States, a leading concern of adolescents who had foregone healthcare was not wanting their parents to find out (Schoen et al. 1998). Inexperience in turn leads to many adolescents not appreciating that problems such as substance use, bullying or overweight can be discussed with health professionals (Sanci 2011).

Adolescents value respect, good coordination of care, appropriate provision of information, high-quality communication, involvement in decisions about care, and healthcare providers listening to their needs (Luxford et al. 2010). As such, the skills of clinicians differ from those needed for the healthcare of younger children which is typically mediated through parents. One important task for health professionals is to set expectations with adolescents and parents about an adolescent taking greater responsibility for healthcare. With increasing age, providers need to explicitly engage the young person themselves, which ideally includes consulting with them alone which is difficult in many cultural settings.

Improving access to quality healthcare requires actions beyond the health service sector. For making sexual and reproductive health services more adolescent-responsive, training of health workers alone does little to improve access or contraception use. However, high-quality training, combined with the provision of adolescent-responsive facilities, strategies to change community norms, and greater availability of LARCs can bring major shifts in access and quality (Chandra-Mouli et al. 2015). Further elements within a multifaceted approach might include promotion of adolescent health literacy, engagement of community leaders, having an appropriate package of services, monitoring provider competencies, promoting equity and non-discrimination, collecting data for quality improvement and engaging adolescents around practice policies (Nair et al. 2015).

Ultimately access to healthcare depends on the availability of platforms which are responsive to the life context of adolescents. School-based health services will have greater relevance as secondary education participation grows in many countries. Service models range from fully equipped and permanently staffed school centres with medical, nursing, and counselling staff to clinics sited near schools but with strong links. The predominant

focus of many school-based and school linked health services has been on provision of sexual and reproductive healthcare but there is clearly an opportunity for these services to extend to mental healthcare, treatment, and prevention of infectious diseases, management of anaemia, and other chronic physical health problems. Web-based technologies also have the potential to improve access to health services (Kauer et al. 2014).

Education settings as platform for health action

Education is a powerful determinant of health and human capital and socioeconomic progress, one in which the adolescent years are central (Viner et al. 2012). Those who are more educated live longer lives with less ill-health (Cutler and Lleras-Muney 2012). In high-income countries (HICs), the benefits are generally greater for women in terms of mortality, self-reported health, mental health, and obesity (Miyamoto and Chevalier 2010). Among adolescents in low- and middle-income countries (LMICs), higher education is associated with reduced teenage births and older age at marriage. The health and well-being benefits of expansion of secondary education accrue through multiple mechanisms, including healthier behaviours, greater cognitive capacity, and longer productive adult lives for the current generation, better health, and lower mortality among their children and overall greater productivity in the future workforce. Education brings profound intergenerational benefits; improved education for women may account for up to half of the global improvement in child mortality since 1970 (Patton et al. 2018).

With primary education now widespread in LMICs, expansion of participation in secondary education offers an achievable strategy for improving health across the life-course and into the next generation (Miyamoto and Chevalier 2010). In countries that already have high secondary education participation, facilitating schools to more explicitly promote health has the potential for leverage above and beyond the health benefits of educational participation alone.

However, many forces operate to divert adolescents from secondary education. Prominent among these are the costs of education and the opportunity costs to families of the loss of adolescent labour, especially in rural areas. For these reasons, poor adolescents in LMICs are less likely to attend secondary school. Early marriage accounts for higher drop-out rates

in girls in LMICs. Most interventions to increase access to and retention in education in these setting have been in younger adolescents, largely in primary schools. Scholarships, school fee reduction or removal, cash transfers conditional on remaining in school, providing school uniforms, decreases in grade repetition, school proximity, and mother tongue education are cost-effective strategies in different contexts for promoting school retention (Glewwe and Kremer 2006; UNICEF 2015). Addressing gender disparities in access, social norms around girls' participation in education, and targeting resources to the poorest regions as well as to disadvantaged students are important elements of increasing adolescent educational engagement in the poorest groups. There is also value in greater non-formal or flexible learning strategies for adolescents without access to mainstream education (e.g. child labourers and married adolescents who have left school) (Yasunaga 2014). Health interventions complement these strategies to promote educational engagement: these might include comprehensive sexuality education, providing access to modern contraception, treatment of micronutrient deficiencies and anaemia and management of mental disorders (Rosenberg et al. 2015).

The traditional way in which schools address health and health risks has been through health education delivered in classrooms. This approach can bring small to moderate benefits in knowledge and attitudes but implementation is often patchy and the effects are generally not sustained (Foxcroft and Tsertsvadze 2011). In contrast, the ways that schools operate more widely have great effects on health, behaviour, and well-being (Patton et al. 2003). For this reason actions that address the school ethos and environment have become much more a focus of recent evaluations and policy attention (Patton et al. 2006; Shinde et al. 2018). A school's environment extends to the physical and social environment, management, and organization, teaching, discipline, pastoral care, school health services, whole-school health promotion, and extracurricular activities. Positive effects from these interventions extend to student substance use, violence, and victimization, and mental health, particularly in groups facing discrimination and disadvantage (Shackleton et al. 2016).

The most effective actions for health in school settings will commonly be multicomponent with curriculum elements, a focus on a school's social and physical environment (e.g. around water, sanitation, and menstrual management facilities), together with engagement of families and/or the

community (Harden et al. 2009). These approaches consistently show positive outcomes for adolescent sexual health, violence, and tobacco smoking and may be beneficial for other health risks (Harden et al. 2009; Blank et al. 2010).

Peers, media, and marketing

The emergence of strong peer relationships is a central feature of early adolescent development (Viner et al. 2012). Such relationships may have either positive or negative effects on health. Peer connection, peer modelling, and awareness of peer norms can be protective against violence, mental health problems, substance use, and sexual health risks (Salazar et al. 2009). Equally, peers may also increase risks, with peer participation in risk behaviours likely to increase smoking initiation and persistence, alcohol initiation and use, self-harm, sexual risks, and violence (Fisher et al. 2007; Kretman et al. 2009). Other peer characteristics, such as sexual partner communication and negotiation skills, influence risks of unwanted pregnancy, STIs, and HIV (Taffa et al. 2002; Salazar et al. 2009).

Modern adolescent peer relationships differ markedly from a pre-industrial social context in both the number and diversity of peers. Prolongation of education and later marriage and parenthood have effectively expanded the role of peers within the lives of adolescents. Peer education, or the promotion of healthy behaviours *for* young people *by* young people, is one popular strategy adopted by the health sector. Such interventions have certainly shown positive outcomes for the young people conducting the programming (i.e. the peer educators themselves) (Maticka-Tyndale and Barnett 2010). Peer promoters themselves have greater self-confidence, improved communications, leadership and interpersonal skills, higher aspirations, and lower rates of health risk behaviours. However, there has been little testing of their effectiveness in changing health outcomes for the broader population of young people they are meant to reach, leading to suggestions that peer education is ineffective (Chandra-Mouli et al. 2015). This reflects a tension in resource-constrained settings between a wish to scale up promising approaches and the value of undertaking smaller scale interventions with rigorous methodologies (Maticka-Tyndale and Barnett 2010).

Mass media and more recently social media are further expanding the role of peers and youth cultures in the lives of adolescents in all places (Larson 2002). Online spaces have changed adolescent developmental tasks such as relationship and identity building which were previously mainly negotiated in face-to-face engagement with peers. For many adolescents, identity formation incorporates local influences with new elements derived from global cultures. There is continued debate over whether exposure to digital media, including a greatly expanded social network, may adversely affect adolescent social, emotional, and cognitive development (Bell et al. 2015). To date, the development of the new media has been so rapid that research efforts to understand their effects have failed to keep pace.

Few studies of the health effects of marketing to adolescents have extended either to LMICs or to new media. Much marketing through digital media extends beyond national borders and is more tailored to individuals. Given that marketing is likely to become even more powerful and increasingly cross-national borders, policy responses at both national and global levels will be needed. To date most research has been around mass media, where the effects on risks ranging from tobacco and alcohol use, to choice of diet and body image norms are clear. There are also good examples of interventions using mass media to promote adolescent health. The South African multimedia 'edutainment' programme Soul City, for example, helped change social norms around HIV/AIDS and domestic violence, contributed to increased individual knowledge about condom use and domestic violence and to the empowerment of local communities (Moyer-Gusé 2008).

While trends in using digital media are global, the means of accessing information varies widely (e.g. mobile phones, personal or public computers). In sub-Saharan Africa growth in access to mobile phones has meant that digital media have become available even in remote places (Lester and Karanja 2008). Young people are generally the earliest adopters of information and communication technologies (ICT) such as mobile phones, instant messaging, the internet, and social media, a phenomenon seen in LMICs as well as in HICs. These new media have the potential to provide a powerful voice for adolescents and young adults to actively engage with each other and their communities. Even in disadvantaged groups and remote locations mobile phones may offer a platform for communication and engagement around health (Alami et al. 2019). These

tools provide a platform for health education in their own right as well as the capacity to augment healthcare delivery and other preventive actions including those within schools. Social media offer the potential to reach diverse groups including those who are geographically isolated or facing discrimination and disadvantage.

Community interventions

The lives of young people are affected by the behaviour, norms, and values of adults and other adolescents in their communities. Community interventions commonly involve local government, families, youth-focused and religious organizations, and schools. Positive youth development (PYD) broadly refers to experiences that provide optimal preparation for the attainment of adult potential and well-being. These programmes are often community based and seek to promote life skills and positive attitudes including self-confidence and empowerment, social and emotional skills, and good problem solving. They aim to build the agency of young people so that they can positively influence the environments in which they live, learn, and play. The strategies employed range from sport and outdoor education to theatre, music, and art, survival skills, leadership training, and mentorship. Certain interventions, such as the promotion of sport in girls, have the potential to bring benefits in physical health and fitness, as well as empowering girls to deal with harmful gender norms (Saavedra 2009). PYD programmes aim to strengthen assets and skills, increasing youth agency and contribution, and foster an enabling environment. The most effective have generally built on available community structures, used good information on adolescent health and well-being, adopted a multicomponent strategy and monitored progress. One example has been the Communities that Care (CTC) framework that has been trialled in a number of United States sites with clear benefits and evidence for cost-effectiveness (Catalano et al. 2012).

One strength of PYD as a framework for intervention in LMICs is that it can be applied in different countries and contexts. One recent systematic review suggested that across LMIC, these programmes achieved gains across a range of youth outcomes, including education, sexual risk behaviour, early marriage, violence, and job skills and employment (Catalano et al. 2019). Several of the approaches also demonstrate promise that PYD increases gender equality.

Conditional cash transfer programmes (CCT) are another example of a community-based intervention that commonly targets adolescents. Examples include payment of school fees and supplemental cash to poor parents on the condition that they send their child to school. Trials in Malawi and Kenya found reduced self-reported sexual activity, pregnancy, and marriage in one or both countries, and reduced school drop-out among girls and young women who received the CCTs (Baird et al. 2010).

Structural interventions

Legislation, taxation, and implementation of policies bringing investment in families, communities, schools, and health services are essential elements of adolescent health action in all places. Laws commonly govern both access to resources for health and protection from hazards. Some specifically address health (e.g. access to healthcare including effective contraception); others address health risks (e.g. consumption of alcohol, access to tobacco); and others address social determinants of health (e.g. age of marriage, protection from hazardous work). Inadequate and inconsistent legal frameworks undermine the health, rights, and potential of adolescents and young adults. Particularly significant are laws that penalize groups facing discrimination and disadvantage where adolescents are more vulnerable that older age groups. One example are laws that diminish the rights of lesbian, gay, bisexual, transgender, and intersex (LGBTI) adolescents and young adults with particularly punitive laws in Nigeria, Somalia, and Sudan.

It was not until the twentieth century that more consistent international legal frameworks emerged for adolescents. The 1989 United Nations Convention on the Rights of the Child (UNCRC) recognized children as bearers of human rights rather than parental property, with liberties and responsibilities appropriate to their age (United Nations General Assembly 1989). The overwhelming majority (195) of countries have signed and ratified the UNCRC, with the United States the only exception. The UNCRC embodies adolescent development suggesting that where freedoms and liberties are restricted, this should be consistent with human rights principles and evolving capacities (Ruck et al. 2016). As adolescents acquire greater competency and maturity, there is a lesser need for adult guidance and an increased capacity for autonomy.

Despite these important developments in international legal frameworks, there remains great diversity in the laws that affect adolescents. Laws often fail both in their framing and implementation, reflecting neither the principles of the UNCRC or an understanding of adolescents' actual capacities. Rather, they mirror historical economic, social, cultural, and sometimes religious priorities (Lansdown 2015). In areas where legislation has a proven role in protecting health (e.g. minimum age to smoke tobacco, working age), laws commonly do not exist. In others, such as legal frameworks around sexual and reproductive health, there is little consistency between countries in the age of marriage, access to contraception, availability of safe abortion, and legality of same-sex relationships.

Child marriage jeopardizes the health, growth, and development of girls in many countries. Around 15 million girls around the world marry each year before 18 years and in lower-income countries, one-third of girls are married before 18, and one in nine before 15 years (UNICEF 2013). Too early marriage results in diminished opportunities for education, greater sexual exploitation, and violence that can sometimes extend to enslavement. Child brides are also exposed to health risks from early pregnancy, have greater maternal and infant mortality, and heightened vulnerability to HIV/AIDS and other STIs (Klugman et al. 2014).

The Convention on the Elimination of All Forms of Discrimination Against Women (CEDAW) requires states to ensure that men and women have equal rights to enter into marriage and to freely choose a spouse with their free and full consent. Yet 34 countries permit girls to marry below 18 years. Some still permit marriage below 15 years, and a few specify no minimum age (Forum on Marriage and the Rights of Women and Girls 2015).

More rational legal frameworks governing adolescents would take greater account of evolving adolescent cognitive and emotional capacities. With access to knowledge, adolescents from the age of 14 years demonstrate similar or even greater cognitive capacity than adults to make good judgements in calm and emotionally neutral contexts (Steinberg 2007). Yet in emotionally charged situations of stress or excitement, particularly with peers, emotions are more likely to drive decision-making (Steinberg 2007; Van Duijvenvoorde and Crone 2013). Legal frameworks would ideally take both factors into account in framing legislation around age of consent. Restricting an adolescent's right to smoke would therefore be justified from

their greater vulnerability to poor decision-making in emotionally charged contexts with peers, as well as from a still evolving capacity to understand the future health consequences. Such approaches would also support and promote adolescent autonomy. Thus voting from 16 years both empowers adolescents and promotes civic engagement. Permitting access to effective contraception including LARCs for sexually active adolescents is one example of how the law can take development into account in protecting an adolescent's health.

Investments in adolescent health should ideally extend from those directed towards conspicuous health problems to health risks that emerge during these years and the broader social determinants of health (Fig. 10.4.3). This ideal is rarely achieved as a result of major barriers to health action, whether through health services, community platforms, or structural initiatives. Adolescents and young adults are the age group with the poorest level of universal health coverage (Tylee et al. 2007). Social, structural, and environmental determinants of adolescent health lie largely outside the health service sector (Viner et al. 2012). There are shifts in these determinants as young people mature through adolescence, such that strategies suited to younger adolescents may be inappropriate or ineffective with young adults. So too strategies that are effective and appropriate for girls may be less effective for boys. The settings for health actions mirror those of determinants and extend from health services, schools, and education settings, to families and communities, places of employment, road transport, media, legal, and policy environments.

Youth engagement and empowerment

The United Nations defined youth participation as 'the active and meaningful involvement of young people in all aspects of their own, and their communities' development, including their empowerment to contribute to decisions about their personal, family, social, economic and political development' (United Nations (DESA) 2007). It embodies the ideas that adolescents and young adults make essential contributions to the design and implementation of programmes and policies that affect them and that meaningful engagement will lead to more equitable communities (Suleiman et al. 2006). In essence, meaningful engagement is not only a fundamental right but essential to capture the unique contributions of youth (Ruck et al.

2016). Developmental perspectives further support meaningful engagement as essential for positive social and emotional development during adolescence. Adolescents are emotionally 'primed' to engage so providing the structures, processes, and support for meaningful engagement itself shapes positive trajectories of development (Dahl 2004). Even though the rationale for youth engagement around health is strong, there has been little systematic research into models of youth engagement, leadership and participation, and their effectiveness.

The sociologist Karl Mannheim wrote early in the twentieth century about the capacity of youth to inspire social change and how the focus of efforts shifts across generations (Mannheim 1970). In recent years, adolescents and young adults have led movements for democracy in Asia and the Middle East, for lesbian, gay, bisexual, and transgender (LGBT) rights, for gender equality, and for action on climate change. Conversely, where opportunities for positive engagement are absent, we have seen adolescents and young adults drawn into violent extremism and gang violence with catastrophic health and societal consequences (Dahl 2004).

Successful engagement emphasizes the capacities of adolescents and young adults to work in partnership with adults. This generally requires both processes to promote active youth participation ('training young people to speak') and the responses of adults ('training adults to listen') (Patton et al. 2016). Adolescents and young adults depend on systems and structures for engagement that are put in place and supported by adults. To engage most effectively, youth need resources (including financial), training, ongoing mentorship, sensitization to political and management processes. They also need platforms and mechanisms for engagement. Existing models range from tokenism to those that are completely youth-led (Wong et al. 2010). More egalitarian relations between young people and adults are essential for meaningful engagement but decisions are still often taken through an adult lens that overlooks young people's perspectives (Gibbs et al. 2010). In general, the greater the level of youth control and responsibility, the greater is the effectiveness of engagement, whether around their own health or around policies and programmes that affect others (Wong et al. 2010).

Youth engagement in health-related advocacy has increased over the past two decades, fostered at a global level by the UN and other international organizations. Similar processes have occurred at country and community levels within organizations delivering services to young people.

There are reasons to be optimistic about the scope for extending meaningful youth engagement around their health and well-being. The growth of social networking and digital media make it possible to access information, communicate and mobilize with other advocates, as well as draw in resources in unprecedented ways. New media promote more active engagement as adolescents and young adults can independently generate their own content (Kaplan and Haenlein 2010). These possibilities allow broader collaboration which in turn can generate political leverage. Time, privacy, and the need to convince adolescents that their voices will be heard remain barriers. For these reasons capitalizing on these new possibilities is likely to require a shift in approaches to communication for both government and civil society organizations. Attitudes of adults towards young people and their capabilities, together with young people's lack of confidence in engagement processes, are a common barrier (Coates and Howe 2014). Attitudinal barriers commonly arise from cultural contexts where young people are perceived as subordinate to adults, that exclude young women and that further stigmatize socially marginalized groups (Harden et al. 2001; Iwasaki 2014). Inflexible bureaucratic arrangements, insufficient investment to support ongoing engagement, and restrictive policies, such as limiting the participation of adolescents in research, may be further limiting factors (Villa-Torres and Svanemyr 2015).

For these reasons, most youth engagement in health falls short of sustainability, and top-down, adult-driven approaches remain the most common.(Gibbs et al. 2010) In programmes that do engage young people, those selected are often already 'confident, articulate and naturally identified leaders'. In practice, few efforts meaningfully involve the most marginalized adolescents, who may include ethnic minorities, HIV-positive youth, married adolescents, youth with disabilities, LGBT youth, youth living in poverty, those with lower educational levels, and very young adolescents (Paterson and Panessa 2008). In many circumstances, adolescent girls and young women face inequitable systems and structures which limit their participation.

The investment case and accountability

There is also growing evidence for both cost-effectiveness and high benefit–cost ratios from investments in adolescent health. In the US, the benefits

from evidence-based prevention programmes were estimated to range between $2 and $42 for every dollar invested (Catalano et al. 2012).

A recent analysis of the investment case for adolescence in LMICs highlighted both high benefit–cost ratios from five investments (education, HPV vaccination, reducing rates of child marriage, prevention of road traffic injury and provision of youth-friendly healthcare) and their affordability even in low-resource setting (Sheehan et al. 2017). In providing health services for physical, mental and sexual health, an investment of US$4.65 per capita per annum had an unweighted mean ratio of benefit to costs (BCR) of over 10; for the prevention of road traffic injuries a BCR of 5.9 could be achieved through an annual per capita investment of US$0.60. Programmes to reduce child marriage (US$1.5 per capita per annum) showed a mean BCR of 5.7, with the benefits greatest in lower-income countries. Investment to improve secondary schooling were more expensive with a per capita investment of US$22.6 per annum generating a mean BCR of 11.8 (Sheehan et al. 2017).

Yet recent analyses of developmental assistance for health across the life-course suggests that there is scant investment in 10–24-year-olds (Skirbekk et al. 2017; Li et al. 2018). Only 1.6% of developmental assistance for health is targeted to 10 to 24-year-olds, despite this age band making up 30% of the population in low-income countries. Overwhelmingly the investments have been in HIV and reproductive health with little or no investment in adolescent nutrition, mental health, or injury prevention (Li et al. 2018).

The *2030 Agenda for Sustainable Development* and the *Global Strategy for Women's, Children's and Adolescents' Health* has outlined the need for better data to drive accountability for countries (United Nations 2014). Ideally, data should allow comparisons over time and tracking of inequalities within and between countries. In reality existing global data systems for adolescents remain poorly coordinated, inconsistent in coverage and timing, inadequately disaggregated, missing large groups of adolescents (e.g. out of school adolescents) and fail to deal with the spectrum of adolescent health problems and their determinants (Patton et al. 2012).

The implementation of adolescent health policies often requires engagement across government ministries. In many countries, this extends to international agencies and global funders and ideally includes civil society groups outside of government (Bhamra et al. 2015). The relevant constituents include professionals working with adolescents, academics and

private industry with young people and youth-focused organizations essential stakeholders in the implementation of effective policies and programmes. Their meaningful engagement can ensure a proper understanding of what data mean, inform the feasibility and appropriateness of potential actions, and support broader buy-in.

The most effective actions for adolescent health and well-being are often intersectoral and multicomponent, tailored to local needs and capacities. Ideally such interventions will have been adapted and trialled in similar contexts. Yet in some areas such as mental disorders, nutritional interventions, and sexual and intimate partner violence, there are not yet clearly effective and scalable actions for adolescents. In other areas, health actions such as comprehensive sexuality education have been evaluated in HICs, but neither adequately adapted nor evaluated in LMICs. Many innovative approaches have been piloted and sometimes rolled out on a larger scale but without evaluation. Examples include interventions based on digital media that offer great promise but in the absence of more systematic evaluation their broader utility remains uncertain (Patton et al. 2016).

Conclusion

Adolescent health now has a presence in global health and international development. The challenge now is to take responses to scale. We now have a good understanding of many approaches that are effective including the strategies for re-orienting health service responses to adolescents, intersectoral actions, and the value of partnerships with young people themselves. Next steps to more effective and accountable systems for adolescent health will need to include better information systems, development of a stronger evidence base for action and ensuring effective policy action around those facing the greatest levels of discrimination and disadvantage. The benefits will extend well beyond adolescence to health in later life and lay a foundation for the best possible start to life for the next generation.

REFERENCES

Alami, S.O., et al. (2019). Getting health information to internally displaced youth in Afghanistan: can mobile phone technology bridge the gap?. *Gates Open Research*, **3**, 1483.

Alonso, J., et al. (2007). Population level of unmet need for mental healthcare in Europe. *The British Journal of Psychiatry*, 190(4), 299–306.
Azzopardi, P.S., et al. (2019). Progress in adolescent health and wellbeing: tracking 12 headline indicators for 195 countries and territories, 1990–2016. *The Lancet*, 393(10176), 1101–18.
Baird, S., et al. (2010). The short-term impacts of a schooling conditional cash transfer program on the sexual behavior of young women. *Health Economics*, 19(S1), 55–68.
Bell, V., Bishop, D.V.M, and Przybylski, A.K. (2015). The debate over digital technology and young people. *British Medical Journal*, 351, h3064.
Bergen, H., et al. (2012). Premature death after self-harm: a multicentre cohort study. *The Lancet*, 380(9853), 1568–74.
Bhamra, A.S., Nagrath, K., and Niazi, Z. (2015). Role of non-state actors in monitoring and review for effective implementation of the post-2015 agenda. IRF2015. [Online] Available at: https://pubs.iied.org/pdfs/G04365.pdf
Blanc, A.K., Winfrey, W., and Ross, J. (2013). New findings for maternal mortality age patterns: aggregated results for 38 countries. *PloS One*, 8(4), e59864.
Blanc, A.K., et al. (2009). Patterns and trends in adolescents' contraceptive use and discontinuation in developing countries and comparisons with adult women. *International Perspectives on Sexual and Reproductive Health*, 35(2), 63–71.
Blank, L., et al. (2010). Promoting well-being by changing behaviour: a systematic review and narrative synthesis of the effectiveness of whole secondary school behavioural interventions. *Mental Health Review Journal*, 15(2), 43–53.
Bonomo, Y.A., Patton, G.C., and Bowes, G. (2006). What are the longer term outcomes of adolescent alcohol consumption in young adulthood? Results from a 10-year cohort study. *Alcoholism-Clinical and Experimental Research*, 30(9), 117A–17A.
Brooks-Gunn, J. and Warren, M.P. (1989). Biological and social contributions to negative affect in young adolescent girls. *Child Development*, **60**, 40–55.
Calhoun, L.M., et al. (2013). Provider imposed restrictions to clients' access to family planning in urban Uttar Pradesh, India: a mixed methods study. *BMC Health Services Research*, 13(1), 532.
Catalano, R.F., et al. (2019). Positive youth development programs in low-and middle-income countries: a conceptual framework and systematic review of efficacy. *Journal of Adolescent Health*, 65(1), 15–31.
Catalano, R.F., et al. (2012). Worldwide application of prevention science in adolescent health. *The Lancet*, 379(9826), 1653–64.
Chandra-Mouli, V., Lane, C., and Wong, S. (2015). What does not work in adolescent sexual and reproductive health: a review of evidence on interventions commonly accepted as best practices. *Global Health: Science and Practice*, 3(3), 333–40.
Cipriani, A., et al. (2016). Comparative efficacy and tolerability of antidepressants for major depressive disorder in children and adolescents: a network meta-analysis. *The Lancet*, 388(10047), 881–90.
Coates, D. and Howe, D. (2014). The importance and benefits of youth participation in mental health settings from the perspective of the headspace Gosford Youth Alliance in Australia. *Children and Youth Services Review*, **46**, 294–9.
Collishaw, S. (2015). Annual research review: secular trends in child and adolescent mental health. *Journal of Child Psychology and Psychiatry*, 56(3), 370–93.
Cutler, D.M. and Lleras-Muney, A. (2012). *Education and Health: Insights from International Comparisons*. Cambridge, MA: National Bureau of Economic Research (NBER).
Dahl, R.E. (2004). Adolescent brain development: a period of vulnerabilities and opportunities. Keynote address. [Review] [10 refs]. *Annals of the New York Academy of Sciences*, **1021**, 1–22.
Day, F.R, et al. (2015). Puberty timing associated with diabetes, cardiovascular disease and also diverse health outcomes in men and women: the UK Biobank study. *Scientific Reports*, **5**, 11208.

Erskine, H.E., et al. (2017). The global coverage of prevalence data for mental disorders in children and adolescents. *Epidemiology and Psychiatric Sciences*, 26(4), 395–402.
Every Woman Every Child (2015). *Global Strategy for Women's, Children's and Adolescents' Health.* New York, NY: Every Woman Every Child.
Finer, L.B., Jerman, J., and Kavanaugh, M.L. (2012). Changes in use of long-acting contraceptive methods in the United States, 2007–2009. *Fertility and Sterility*, 98(4), 893–97.
Fisher, L.B., et al. (2007). Predictors of initiation of alcohol use among US adolescents: findings from a prospective cohort study. *Archives of Pediatrics & Adolescent Medicine*, 161(10), 959–66.
Forum on Marriage and the Rights of Women and Girls (2015). Early marriage: whose right to choose?. [Online] Available at: https://www.crin.org/en/library/publications/early-marriage-whose-right-choose
Foxcroft, D.R. and Tsertsvadze, A. (2011). Universal school-based prevention programs for alcohol misuse in young people. *Cochrane Database of Systematic Reviews*, (5), CD009113.
Ganchimeg, T., et al. (2014). Pregnancy and childbirth outcomes among adolescent mothers: a World Health Organization multicountry study. *BJOG: An International Journal of Obstetrics & Gynaecology*, 121(s1), 40–8.
Gibbs, A., et al. (2010). Mismatches between youth aspirations and participatory HIV/AIDS programmes in South Africa. *African Journal of AIDS Research*, **9** (2), 153–63.
Glewwe, P. and Kremer, M. (2006). Schools, teachers, and education outcomes in developing countries. In: Hanushek, E. and Welch, F. (ed.) *Handbook of the Economics of Education* (1st ed.) (Volume 2). Amsterdam: Elsevier, pp. 945–1017.
Grazzi, L. (2004). MIDAS questionnaire modification for a new MIDAS junior questionnaire: a clinical experience at the Neurological Institute 'C. Besta'. *Neurological Sciences*, *25*(Suppl 3), S261–S62.
Harden, A., Oakley, A., and Oliver, S. (2001). Peer-delivered health promotion for young people: a systematic review of different study designs. *Health Education Journal*, 60(4), 339–53.
Harden, A., et al. (2009). Teenage pregnancy and social disadvantage: systematic review integrating controlled trials and qualitative studies. *British Medical Journal*, 339, b4254.
Hawton, K. and James, A. (2005). Suicide and deliberate self-harm in young people. *British Medical Journal*, **330**, 891–4.
Houghton, L.C., et al. (2014). A migrant study of pubertal timing and tempo in British-Bangladeshi girls at varying risk for breast cancer. *Breast Cancer Research*, 16(6), 469.
Iwasaki, Y. (2014). Reflection on Learnings from Engaging and Working with High-risk, Marginalized Youth. *Relational Child & Youth Care Practice*, 27(4), 24–35.
Jamison, D.T., et al. (2013). Global health 2035: a world converging within a generation. *The Lancet*, 382(9908), 1898–955.
Kaplan, A.M. and Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons*, 53(1), 59–68.
Kauer, S.D., Mangan, C., and Sanci, L. (2014). Do online mental health services improve help-seeking for young people? A systematic review. *Journal of Medical Internet Research*, 16(3), e66.
Klugman, J., et al. (2014). *Voice and Agency: Empowering Women and Girls for Shared Prosperity*. Washington, DC: World Bank Group.
Kretman, S.E., et al. (2009). Adolescent violence: risk, resilience, and prevention. In: DiClemente, R.J., Santelli, J.S., and Crosby, R.A.(eds.), *Adolescent Health: Understanding and Preventing Risk Behaviors*. San Francisco, CA: Jossey-Bass, pp. 213–32.
Lansdown, G. (2015). *Innocenti Insight: The Evolving Capacities of the Child*. Florence: Save the Children. [Online] Available at: http://www.unicef-irc.org/publications/pdf/evolving-eng.pdf
Larson, R. (2002). Globalization, societal change, and new technologies: what they mean for the future of adolescence. *Journal of Research on Adolescence*, **12**, 1–30.

Lester, R. and Karanja, S. (2008). Mobile phones: exceptional tools for HIV/AIDS, health, and crisis management. *The Lancet Infectious Diseases*, 8(12), 738–9.
Li, Z., et al. (2018). Global development assistance for adolescent health from 2003 to 2015. *JAMA Network Open*, 1(4), e181072–e72.
Luxford, K., et al. (2010). *Patient-centred Care: Improving Quality and Safety by Focusing Care on Patients and Consumers*. Canberra: Australian Commission on Safety and Quality in Healthcare. [Online] Available at: https://www.safetyandquality.gov.au/sites/default/files/migrated/PCCC-DiscussPaper.pdf
Mannheim, K. (1970). The problem of generations. *Psychoanalytic Review*, 57(3), 378–404.
Maticka-Tyndale, E. and Barnett, J.P. (2010). Peer-led interventions to reduce HIV risk of youth: a review. *Evaluation and Program Planning*, 33(2), 98–112.
Mayer-Davis, E.J, et al. (2017). Incidence trends of type 1 and type 2 diabetes among youths, 2002–2012. *New England Journal of Medicine*, **376** (15), 1419–29.
Miyamoto, K. and Chevalier, A. (2010). Education and health. In: OECD (ed.), *Improving Health and Social Cohesion through Education*. Paris: OECD Publishing.
Moyer-Gusé, E. (2008). Toward a theory of entertainment persuasion: explaining the persuasive effects of entertainment-education messages. *Communication Theory*, 18(3), 407–25.
Murray, C.J.L., et al. (2014). Global, regional, and national incidence and mortality for HIV, tuberculosis, and malaria during 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013. *The Lancet*, 384(9947), 1005–70.
Nair, M., et al. (2015). Improving the quality of health care services for adolescents, globally: a standards-driven approach. *Journal of Adolescent Health*, 57(3), 288–98.
Nove, A., et al. (2014). Maternal mortality in adolescents compared with women of other ages: evidence from 144 countries. *The Lancet Global Health*, 2(3), e155–e64.
Patel, V., et al. (2007). Mental health of young people: a global public-health challenge. *The Lancet*, **369**, 1302–13.
Paterson, B.L. and Panessa, C. (2008). Engagement as an ethical imperative in harm reduction involving at-risk youth. *International Journal of Drug Policy*, 19(1), 24–32.
Patton, G., et al. (2016). Our future: a Lancet commission on adolescent health and wellbeing. *The Lancet*, **387**, 2423–78.
Patton, G., et al. (2003). Changing schools, changing health? Design and implementation of the Gatehouse Project. *Journal of Adolescent Health*, 33(4), 231–9.
Patton, G.C., et al. (2006). Promoting social inclusion in schools: a group-randomized trial of effects on student health risk behavior and well-being. *American Journal of Public Health*, 96(9), 1582–7.
Patton, G.C. and Viner, R. (2007). Pubertal transitions in health. *The Lancet*, 369(9567), 1130–9.
Patton, G.C., et al. (2009). Global patterns of mortality in young people: a systematic analysis of population health data. *The Lancet*, 374(9693), 881–92.
Patton, G.C., et al. (2011). Overweight and obesity between adolescence and young adulthood: a 10-year prospective cohort study. *Journal of Adolescent Health*, 48(3), 275–80.
Patton, G.C., et al. (2012). Health of the world's adolescents: a synthesis of internationally comparable data. *The Lancet*, 379(9826), 1665–75.
Patton, G.C., et al. (2014). The prognosis of common mental disorders in adolescents: a 14-year prospective cohort study. *The Lancet*, *383*(9926), 1404–11.
Patton, G.C., et al. (2018). Adolescence and the next generation. *Nature*, 554(7693), 458–66.
Riley, E.P. and McGee, C.L. (2005). Fetal alcohol spectrum disorders: an overview with emphasis on changes in brain and behavior. *Experimental Biology and Medicine*, 230(6), 357–65.
Rosenberg, M., et al. (2015). Relationship between school dropout and teen pregnancy among rural South African young women. *International Journal of Epidemiology*, 44(3), 928–36.
Ruck, M.D., et al. (2016). The United Nations Convention on the Rights of the Child: its relevance for adolescents. *Journal of Research on Adolescence*, 26(1), 16–29.

Saavedra, M. (2009). Dilemmas and opportunities in gender and sport-in-development. In: Levermore, R. and Beacom, A. (eds.) *Sport and International Development*. Basingstoke: Palgrave, pp. 124–55.
Salazar, L.F., et al. (2009). Sexually transmitted disease transmission and pregnancy among adolescents. In: DiClemente, R.J., Santelli, J.S., and Crosby, R.A. (eds.), *Adolescent Health: Understanding and Preventing risk Behaviors*. San Francisco, CA: Jossey-Bass, pp. 274–302.
Sanci, L. (2011). Clinical preventive services for adolescents: facing the challenge of proving 'an ounce of prevention is worth a pound of cure'. *Journal of Adolescent Health*, 49(5), 450–2.
Sawyer, S.M., et al. (2018). The age of adolescence. *Lancet Child and Adolescent Health*, 2(3), 223–8.
Sawyer, S.M., et al. (2012). Adolescence: a foundation for future health. *The Lancet*, 379(9826), 1630–40.
Schlegel, A. and Barry, H. (1991). *Adolescence: An Anthropological Enquiry*. New York: Free Press.
Schoen, C., et al. (1998). *Health of Adolescent Boys: Commonwealth Fund Survey Findings* New York: Commonwealth Fund New York.
Shackleton, N., et al. (2016). School-based interventions going beyond health education to promote adolescent health: systematic review of reviews. *Journal of Adolescent Health*, 58(4), 382–96.
Sheehan, P., et al. (2017). Building the foundations for sustainable development: a case for global investment in the capabilities of adolescents. *The Lancet*, 390(10104), 1792–806.
Shinde, S., et al. (2018). Promoting school climate and health outcomes with the SEHER multi-component secondary school intervention in Bihar, India: a cluster-randomised controlled trial. *The Lancet*, 392(10163), 2465–77.
Skirbekk, V., et al. (2017). Vast majority of development assistance for health funds target those below age sixty. *Health Affairs*, 36(5), 926–30.
Steinberg, L. (2007). Risk taking in adolescence new perspectives from brain and behavioral science. *Current Directions in Psychological Science*, 16(2), 55–9.
Suleiman, A.B., Soleimanpour, S., and London, J. (2006). Youth action for health through youth-led research. *Journal of Community Practice*, 14(1–2), 125–45.
Taffa, N., et al. (2002). Psychosocial determinants of sexual activity and condom use intention among youth in Addis Ababa, Ethiopia. *International Journal of STDs & AIDS*, 13(10), 714–19.
Tylee, A., et al. (2007). Youth-friendly primary-care services: how are we doing and what more needs to be done?. *The Lancet*, **369**, 1565–73.
UNICEF (2015). *The National Strategy to end Child Marriage and Teenage Pregnancy.* Kampala: UNICEF.
UNICEF (2013). *State of the World's Children*. [Online] Available at: http://www.unicef.org/sowc2013/
United Nations (DESA) (2007). *World Youth Report 2007: Young People's Transition to Adulthood*. New York: United Nations.
United Nations (2014). *The Road to Dignity by 2030: Ending Poverty, Transforming All Lives and Protecting the Planet*. Synthesis report of the Secretary-General on the post-2015 sustainable development agenda. New York, NY: United Nations.
United Nations Children's Fund (2012). A report card on adolescence. In: UNICEF (ed.) *Progress for Children*. New York: UNICEF.
United Nations General Assembly (1989). Convention on the Rights of the Child. (1577: United Nations Treaty Series vol. 1577).
Van Duijvenvoorde, Anna CK and Crone, Eveline A (2013). The teenage brain a neuroeconomic approach to adolescent decision making. *Current Directions in Psychological Science*, 22(2), 108–13.
Villa-Torres, L. and Svanemyr, J. (2015). Ensuring youth's right to participation and promotion of youth leadership in the development of sexual and reproductive health policies and programs.

Journal of Adolescent Health, 56(1), S51–7.
Viner, R., et al. (2012). Adolescence and the social determinants of health. *The Lancet*, 379(9826), 1641–52.
Viner, R.M. and Taylor, B. (2007). Adult outcomes of binge drinking in adolescence: findings from a UK national birth cohort. *Journal of Epidemiology and Community Health*, **61** (10), 902–7.
Waddington, C. and Sambo, C. (2015). Financing health care for adolescents: a necessary part of universal health coverage. *Bulletin of the World Health Organization*, 93(1), 57–9.
Winner, B., et al. (2012). Effectiveness of long-acting reversible contraception. *New England Journal of Medicine*, 366(21), 1998–2007.
Wong, N.T., Zimmerman, M.A., and Parker, E.A. (2010). A typology of youth participation and empowerment for child and adolescent health promotion. *American Journal of Community Psychology*, 46(1–2), 100–14.
Woog, V., et al. (2015). *Adolescent Women's Need for and Use of Sexual and Reproductive Health Services in Developing Countries*. [Online] Available at: https://www.guttmacher.org/report/adolescent-womens-need-and-use-sexual-and-reproductive-health-services-developing-countries
World Bank (2007). *World Development Report 2007: Development and the Next Generation.* [Online] Available at: https://openknowledge.worldbank.org/handle/10986/5989
World Health Organization (WHO) (1989). *The Health of Youth*. Technical Discussions, document A42. Geneva: WHO.
Worthman, C.M. and Trang, K. (2018). Dynamics of body time, social time and life history at adolescence. *Nature*, **554** (7693), 451.
Yair, O. and Miodownik, D. (2014). Youth bulge and civil war: why a country's share of young adults explains only non-ethnic wars. *Conflict Management and Peace Science*, 33(1), 25–44.
Yasunaga, M. (2014). *Non-formal Education as a Means to Meet Learning Needs of Out-of-School Children and Adolescents*. Montreal: UNESCO Institute for Statistics (UIS).
Yu, C.K.H., Teoh, T.G., and Robinson, S. (2006). Review article: obesity in pregnancy. *BJOG: An International Journal of Obstetrics & Gynaecology*, 113(10), 1117–25.

10.5

Intersectional and social epidemiology approaches to understanding the Influence of race, ethnicity, and caste on global public health

Jennifer Beard, Nafisa Halim, Salma M. Abdalla, and Sandro Galea

Introduction

Health at the individual, interpersonal, and community level is a product of multiple, intersecting characteristics that comprise identity and contribute to privilege and vulnerability. Health, illness, well-being, and misery for all humans are shaped by a confluence of biological, social, and cultural factors. While traditional epidemiological approaches have measured associations by considering variables in isolation, there is growing recognition that multiple variables are important determinants of identity and, therefore, of health and well-being.

Social epidemiology, in particular, has aimed to more fully understand how multiple intersecting factors influence population health (Bowleg 2012; Bauer 2014; Larson et al. 2016). Social epidemiology the science of measuring the ways in which variables like race, ethnicity, sex, gender, sexual orientation, and physical ability are associated with illness and health at the population level (Berkman et al. 2014). It seeks to examine the role of material, behavioural, and psychosocial pathways which influence social inequality and health disparities over the life course, looking at early exposure to social toxins like racism and poverty (as well as biological toxins such as lead or air pollution) and their cumulative influence on health later in life (Cable 2014). Although the traditional methods of epidemiology remain limited to measuring single variables and their association with single outcomes, newer approaches have emerged that aim to understand the complexity of forces that shape health in populations (Galea, Riddle, et al. 2009; Diez Roux 2011). These innovations in social epidemiology are akin

to intersectional approaches to understanding health disparities within and across populations.

An intersectional approach challenges and complements traditional epidemiological methods by asserting the need to measure the impact of multiple, interlocking variables on the health of individuals and populations. As Amroussia et al. (2017) explain, 'The intersections of these categories create a complex web of social inequalities that are all experienced together by individuals. It is not possible to explore inequalities grounded in each category in isolation from each other, since that is not the way they are experienced by individuals' (Amroussia et al. 2017). Privilege, disadvantage, and bias exert pressure, influencing individual agency and power. These external forces are shaped by ambient societal factors like cultural and social norms, economic and political structures, institutions, and laws. For example, a Hispanic woman living in the US faces greater social inequities than a Hispanic man. That same man may face more social challenges than a White woman, particularly if he is poor, Black, or gay. Intersectionality takes all these ascribed, achieved, and external forces that shape identity and influence agency into consideration within a specific cultural, social, and historical context.

This chapter examines the ways in which the individual characteristics we are born with or acquire over the course of our lives intersect to offer us privileges or hamper us with disadvantages that dictate our positions in the social hierarchy and our health. We define common terminology and explore the intersections between social determinants and health outcomes in an illustrative case focused on India. Finally, we offer recommendations for making intersectionality central to social epidemiology and global health.

Terminology and definitions

Many of the concepts featured in this chapter (race, ethnicity, privilege, racism) are widely understood and used in common parlance, yet each is also abstract and nuanced. The following definitions aim to provide common understanding of the overlapping social positions and processes that contribute to individual and population health.

***Race and ethnicity*:** The Oxford English Dictionary (OED) defines 'race' as 'a group of people, animals, or plants, connected by common descent or origin. In its widest sense the term includes all descendants from an original

stock, but may also be limited to a single line of descent or to the group as it exists at a particular period' (Oxford English Dictionary Online 2018). The breadth of this basic definition, referring to vague categorization of all living entities based on context and other subjective criteria may startle first-time readers. Most of us have deeply ingrained ideas about race as an objective system grouping human beings according to intrinsic physical characteristics including skin colour, facial features, and hair texture. Yet, researchers unravelling the human genome have concluded that these external markers have very little grounding in genetic difference (Rosenberg et al. 2002; Yudell et al. 2016). This conclusion is consistent with decades of sociological theory and research (Berger and Luckmann 1966). Race is a social construct built on and enforced by differences that have no biologic grounding (Scott and Marshall 2009). James Scott and Gordon Marshall, editors of the Oxford Dictionary of Sociology, describe it as a term that should always be marked by inverted commas to indicate that 'this manner of categorizing individuals and population groups is not based on any biologically valid distinctions between the genetic make-up of differently identified "races'' (Scott and Marshall 2009).

The OED defines 'ethnicity' as 'status in respect of membership of a group regarded as ultimately of common descent, or having a common national or cultural tradition; ethnic character' (Oxford English Dictionary Online 2018). This definition closely mirrors that of 'race' but with emphasis on culture and other shared external identifiers such as country of origin, history, language, religion, and tribe. For many scholars and activists, 'ethnicity' is a more useful term than 'race' because it is free of biological connotations. But ethnicity is also a complex social construction dependent on external factors ranging from geopolitical boundaries to personal identification. For example, many Americans take great pride in their ancestors' pre-migration national heritage and identify as Irish or Italian even though they have never been to said country and have no personal connections there. Ethnicity, like race, also has a long history of being used to justify social exclusion and brutality.

Birth-ascribed status: The concept of birth-ascription is helpful for understanding aspects of identity into which one is born and upon which one is ranked within the social hierarchy. These demographic identifiers include race, ethnicity, sex, and caste. Gerald Berremann explains birth-ascription as a 'consequence of birth or ancestry and hence immutable', conferring a

‘societally defined and affirmed worth which is regarded as intrinsic to the individual that cannot be repudiated, relinquished, or altered’ (Berremann 1972). Birth-ascription, he suggests, is a prison without walls in which ‘everyone is sentenced for life to a social cell shared by others of like birth, separated from and ranked relative to all other social cells’ (Berremann 1972).

Raj Bhopal has described race and ethnicity as ‘complex, intertwining, and important concepts used by societies to identify and evaluate social groups and individuals’ (Bhopal 2015). His use of the word ‘evaluate’ is apt. Birth-ascribed status is a form of social evaluation imposed by those with privilege and power. In the US, being identified as Black, Hispanic, or more broadly as ‘not-White’, functions as a form of social evaluation used to justify slavery, brutality, codified social exclusion, and both explicit and implicit bias. In India, caste and sex are the primary drivers of social hierarchy.

Achieved status and agency: Achieved status, is a sociological term referring to ‘any social position held by an individual as a result of his or her personal accomplishments in open formal or market competition with others’ (Scott and Marshall 2009). Achievement refers to education, profession, income, marriage, parenthood, and other social positions that one earns, accomplishes, or chooses. The Oxford Dictionary of Sociology notes that the distinctions between achieved status and ascribed status are not absolute as one’s place within the social hierarchy based on race, ethnicity, gender, and caste further or limit opportunities for and scope of accomplishment. Agency or the ‘the capacity for willed (voluntary) action’, can be bolstered or diminished by education, social network, family relationships, professional path, and income (Scott and Marshall 2009). A Dalit woman living in a poor, rural region will face more barriers than her counterpart living in an urban area that will limit her agency or ability to transcend the social rank into which she was born.

Social position and identities: Social identity theory posits that individuals develop their sense of identity based on groups to which they belong. The development of social identity is a three-part process that starts with social categorization of the people around you as ‘us’ or ‘them’ (Tajfel and Turner 985). Step two, social identification, involves assuming the identities of the different groups to which we belong. Step three, social comparison, is the impulse to define oneself and one’s group as different

from and better from other groups. Brenda J. Allen outlines six primary identity groups: gender, race, social class, sexuality, ability, and age (Allen 2011). Some groups (male, White, wealthy, heterosexual, able-bodied, and adult) are typically more powerful than others and have historically built their identity around subjugating others.

Intersectionality: Intersectionality was first articulated by women of colour as a critique of the feminist, civil rights, and anticolonial movements of the last half of the twentieth century (Collins and Bilge 2016). Kimberlé Crenshaw, a Black feminist legal scholar, is generally credited with first using the term 'intersectionality' in 1989. Mainstream feminism in the US, she argued, was a movement created by and for White middle-class women. And the civil rights campaign to end racial discrimination likewise erased the compound sex- and race-based subordination of Black women with its focus on middle-class Black men (Crenshaw 1989). Although Crenshaw is often credited with first articulating the need for an intersectional approach to understanding 'multiply-burdened marginalization' (Crenshaw 1989), she was by no means alone. Her voice was one in a global chorus of women of colour frustrated with the binary thinking of social change movements that categorized marginalization as race- *or* gender-based, without considering what Patricia Hill Collins and Sirma Bilge describe as the 'many axes that work together and influence each other', enhancing the privilege and power of some and deepening the oppression of others (Collins and Bilge 2016).

Figure 10.5.1 illustrates intersectional identities as a cluster of interlocking ascribed, achieved, and geopolitical identities that lock together to form the position of each individual. Privilege, disadvantage, and bias exert pressure influencing each individual's agency and power. These external forces are shaped by ambient societal factors like culture, norms, economic and political structures, institutions, and laws.

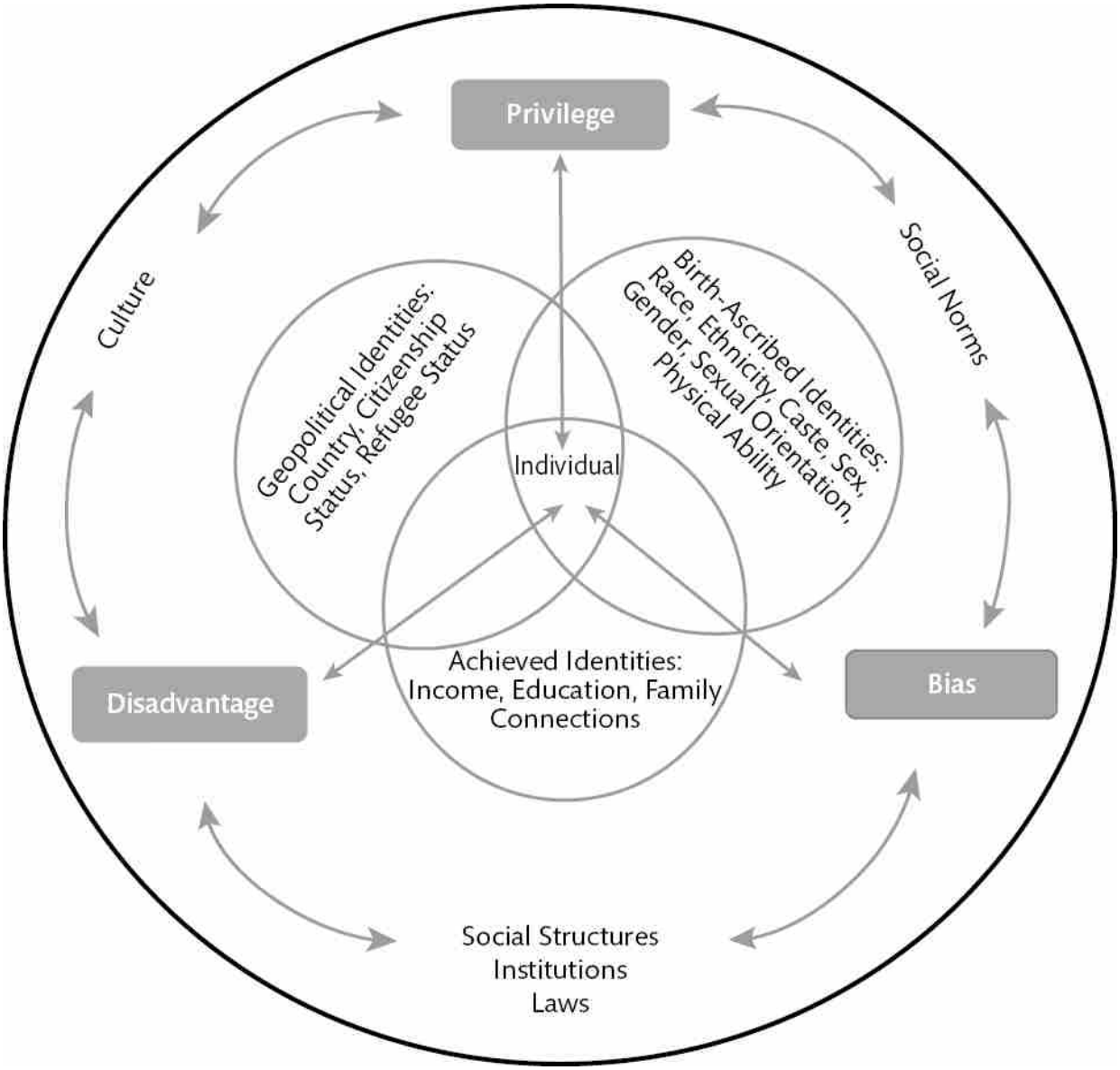


Fig. 10.5.1 Intersectional identities.

Racism and racialization: The Oxford English Dictionary definition of 'racism' provides a detailed overview of discriminatory social processes, beliefs, and actions based on race, ethnicity, an other defining characteristics:

> A belief that one's own *racial or ethnic group* is superior, or that other such groups represent a threat to one's cultural identity, racial integrity, or economic well-being; (also) a belief that the members of *different racial or ethnic groups* possess specific characteristics, abilities, or qualities, which can be compared and evaluated. Hence: *prejudice, discrimination, or antagonism directed against people of other racial or ethnic groups* (or, more widely, of other nationalities), esp. based on such beliefs. (Oxford University Press 2018, emphasis added)

From this broad perspective, skin colour and other physical features are inadequate as universal determinants of racialized discriminatory behaviour.

Racialization is the imposition of a 'racial interpretation' on groups of people seen as other, even when both parties share similar physical and social characteristics. In European and United States migration history, Caucasian Jews, Irish, and Italians have been depicted and shunned as separate and inferior races despite similar White skin and other physical attributes. In colonial Rwanda, Tutsis were favoured by German and Belgian colonizers over the more 'African' Hutus. Both ethnic groups are Black, but colonial rulers promoted the former as being closer to White because of their lighter skin, tall and slim stature, and facial features. Racialized discrimination has played out in its most extreme forms in, for example, the genocides in Europe, Rwanda, and Darfur, ethnic cleansing of the Rohingya from Myanmar, and tribal violence in Kenya. In India, the caste system has imposed discriminatory social stratification rendering millions 'untouchable' (Berremann 1972). In the US, structural racism is sometimes described as a caste system that has led to mass incarceration and police shootings of unarmed Black men (Alexander 2010; Sampath 2015; Demby 2018).

Privilege and disadvantage: Social privilege and disadvantage are the unearned consequences of birth-ascribed rank. Many variables are at play in the ratio of advantage and disadvantage each of us carry, including skin colour, religion, sex, gender, sexual orientation, body-type, and physical ability. The degree to which these qualities confer unearned advantage or disadvantage is dependent on the biases and norms of the context in which we live. In the 1989, Peggy McIntosh famously used the analogy of an 'invisible knapsack' filled with free passes, passports, and licences to describe White privilege and male privilege (McIntosh 1989). Intersectional approaches to understanding oppression shed light on many other determinants of privilege, power, disadvantage, and oppression. As McIntosh explained more recently: 'We are all put ahead and behind by the circumstances of our birth. We all have a combination of both. And it changes minute to minute, depending on where we are, who we're seeing, or what we're required to do' (Rothman 2014).

Social determinants of health: The US Centers for Disease Control and Prevention define social determinants of health as 'conditions in the places where people live, learn, work, and play that affect a wide range of health risks and outcomes' (Centers for Disease Control and Prevention 2018).

These daily conditions of life include the houses we live in, the type of food we eat, the schools we attend, the amount of money we have to meet our needs, the way we are treated by our parents as children, the degree to which we feel socially accepted, and more (Commission on Social Determinants of Health 2008). Likewise, race-based privilege, oppression, and other birth-ascribed attributes like ethnicity, sex, and caste are critical factors contributing to the health of individuals and populations. In the US, for example, infant mortality is two times higher for Black babies (11.7 deaths per 1,000 births) compared to White (4.8 per 1,000) (Riddell et al. 2017). And, in Brazil, low-birth weight is significantly higher among infants of African ancestry (12.4%) compared to Brazilians with European-only ancestry (8.10%) (Nyarko et al. 2013).

Intersectionality in global health research: conceptual framework and methodologies

Figure 10.5.2 illustrates the pathways connecting intersectional identities and health status across the life course. As articulated in Section 2, an individual's social position depends on his/her birth-ascribed and achieved attributes. Further, the individual's birth-ascribed attributes depend on his/her own characteristics (i.e. gender, birth-order, skin tone, sexual orientation) and the characteristics of his/her parents (wealth, income, religion, caste, race, and ethnicity). Intersectional theory posits that 'multiple social categories (e.g. race, ethnicity, gender, sexual orientation, socioeconomic status) intersect at the micro level of individual experience to reflect multiple interlocking systems of privilege and oppression at the macro, social-structural level (e.g. racism, sexism, heterosexism)' (Bowleg 2012). Birth-ascribed attributes largely shape an individual's achieved attributes later in life. Individuals are endowed with privileges and disadvantages along different dimensions, the net of which determine each individual's social position. History, culture, and social and economic context determine the identities associated with privileges and those with disadvantages.

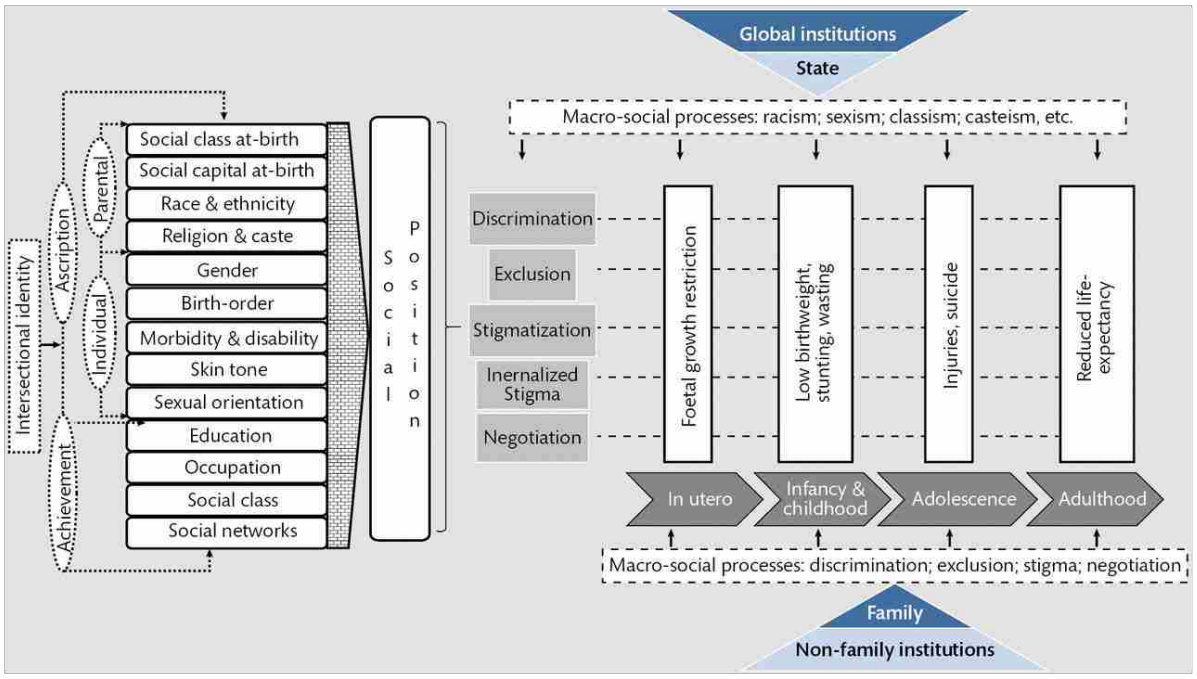


Fig. 10.5.2 Pathways linking intersectional identity and health outcomes throughout the life course.

Social positions affect health and well-being across the life course, as illustrated in the constructs on the far left in Figure 10.5.2. Depending on social position, each individual will or will not experience bias and discrimination in interactions with healthcare providers, educators, employers, and others. For example, clinicians influenced by unconscious bias may make assumptions and care decisions about poor patients, patients of another race or ethnicity, patients of a different caste, patients who are gay or lesbian, or patients with non-binary gender. Parents may rely on internalized social norms to make decisions on whether to invest in daughters as much as sons. And, educators or employers may perceive members of their own race or religion as having more positive attributes (e.g. cooperative, competent, honest, compliant, hard-working, ‘pure’) than those of another race or caste with whom they may associate negative qualities (e.g. competitive, incompetent, dishonest, deviant, lazy, ‘polluted’). These decisions or interactions can lead to discrimination and exclusion over the life course, affecting foetal growth, birth outcomes; childhood nutrition and development; sexual contraception choices and contraception use during adolescence, and pregnancy outcomes or life expectancy during adulthood.

The extent to which clinicians, educators, employers, and others engage in discriminatory practices will be shaped by the enactment and enforcement of national and international laws criminalizing discrimination. Whether or not a Black pregnant woman in the US and many other countries with a history of institutional racism or a low-caste pregnant woman in India and other South Asian countries will be able to leverage or negotiate respectful treatment will depend on her position. Leveraging or negotiating refers to using advantages along some social positions to compensate for disadvantages along others. For example, a wealthy Black pregnant woman in the US or a Brahmin pregnant woman in India will be better able to demand respectful treatment than a poor woman who is Black or from a Scheduled Caste. The more disadvantages individuals have, the less ability they will have to leverage power in social interactions. The ratio of disadvantages to privileges one carries, influences stress burden, perceived agency, and, ultimately, health and well-being.

Study design, measurement, and modelling

The advantages of incorporating intersectional approaches in health research have been widely acknowledged (Hankivsky et al. 2010; Bowleg 2012; Bauer 2014; Lopez and Gadsden 2016; Larson et al. 2016; Green et al. 2017). Yet, intersectionality frameworks have been underutilized due to limited progress in developing methodologically sound approaches for applying intersectionality in health research. Intersectional theory challenges and offers nuance to standard epidemiological methods for measuring associations between variables like race, ethnicity, caste, gender, economic status, and health outcomes by placing each individual at the centre of a busy thoroughfare with multiple roads leading in from all angles. Another helpful image is that of a knotted tangle of many strands of yarn in which each is spun from multiple types of fibre representing social identifiers and varying levels of access to privilege and power. The traditional methods of social epidemiology stop short of measuring and analysing what Lisa Bowleg has described as ‘multiple and interlocking social identities’ (Bowleg 2012), which is the goal of an intersectional approach.

Applying an intersectional approach to population health research requires using analytical methods that examine the complex characteristics of individuals that blur the fixed lines that define a population (Green et al.

2017). A number of social epidemiologists have incorporated intersectionality into their research, using it to inform study design, determine variables to measure, approach measuring multiple social positions, and analyse data (Weber and Parra-Medina 2003; Schulz and Mullings 2006; Warner 2008). An overview of some intersectional approaches to population health research follows:

Study approach/design: Eco-epidemiology is a methodological parallel to the intersectionality framework concerned with the determinants of disease in individual cases as well as the social contexts within which individuals live. To measure how race produces ill-health, for example, an eco-epidemiology approach takes into consideration that study subjects experience of racism at individual (i.e. internalized racism), interpersonal and community (discrimination and exclusion), and societal levels (institutionalized racism). Similarly, an eco-epidemiology approach to data analysis incorporates contextual factors pertinent to neighbourhoods and states:

Neighbourhood factors include violence rates, incarceration rates, income/wealth inequality, affordable housing, housing occupancy rates, and liquor outlet density;
State factors include public funding for healthcare and social programmes, human rights protections, discriminatory laws or policing practices, and employment protections.

Eco-epidemiology generally requires a large sample size, with samples organized across multiple levels (individuals; neighbourhoods; etc.) of the socioecological spectrum (Bauer 2014).

Operationalization of intersectional social positions: Intercategorical and anticategorical intersectionality are two methods researchers have used to measure multiple, interacting variables and their possible influence on health (McCall 2005; Wemrell et al. 2016; Green et al. 2017; Wemrell et al. 2017). The intercategorical approach uses traditional social categories such as race and gender as proxies of the social forces of racism or sexism. The anticategorical approach, emphasizes ‘the inherent fluidity and malleability of social categories, as. . . socially contingent constructions rather than mirrorings of fixed realities’ (Wemrell et al. 2017).

Wemrell et al. adopted the ‘anticategorical’ approach to measuring intersectional social positions and risk for ischaemic heart disease by

combining categories such as immigration status (yes, no), socioeconomic status (high, medium, low), marital status (yes, no), and mental illness (yes, no) separately for men and women. The research team constructed 48 variables, 24 for men and 24 for women (Wemrell et al. 2017). These variables included:

Males who had been registered inhabitants of Sweden for at least 10 years, had high income, cohabitated, and had been prescribed psychotropic medication.
Females who had been registered inhabitants of Sweden for at least 10 years, had high income, cohabitated, and had not been prescribed psychotropic medication. (Wemrell et al. 2017)

Modelling: Eco-epidemiologists primarily alternate between two approaches to model health effects of intersectionality. Green et al. (2017) advocate use of multilevel modelling approaches, which entail setting up the data in a way that nests individuals (level 1) within each intersectional identity (level 2) (Green et al. 2017). The authors suggest adding two sets of covariates to regression models: individuals' (a) gender, race, and other dimensions of social identity; (b) intersectional identity. This differs from traditional approaches of modelling intersectionality. In a traditional approach, a researcher uses an interaction variable for an intersectional identity (e.g. an interaction variable indicating being both Black and female) along with two variables, each for one identity (i.e. a variable indicating being female, and another indicating being Black).

Complementary use of complex systems approaches has strong potential utility for dealing with the complexities associated with intersectional health production processes (Galea et al. 2009). Individual and societal factors influence one another in the process of the production of health. For example, an individual's alcohol use and residence in a neighbourhood with a high density of alcohol shops can each adversely influence health. Yet, individuals who tend to engage in high levels of alcohol use may choose to live in neighbourhoods where alcohol outlets are easily accessed. Or, neighbourhoods where alcohol outlets are readily available may influence moderate or non-alcohol users to increase their use to the point of alcohol addiction. Each situation is an example of individual and neighbourhood-level factors influencing each other, and thereby confounding the associations estimated by regression models between single-level factors and

health outcomes. Such relationships between determinants are commonly referred to as dynamic relationships.

Individual, interpersonal, community, and societal indicators can be consequences as well as the causes of health outcomes. An individual's health can be a consequence as well as a 'complicating cause' of alcohol use disorder (Galea et al. 2009). An individual living with a chronic disease may turn to alcohol dependence among other harmful behaviours as a coping mechanism. Or many individuals' high demand for alcohol may influence availability of liquor stores in the neighbourhood. This phenomenon is commonly referred to as 'reciprocity' between health exposures and outcomes.

Complex systems computational models have the potential to simulate dynamic reciprocal interactions and interrelations between individuals, exposures, and outcomes that change over time and across the life course (Galea et al. 2009). Computer simulations that specify rules by which individual agent's interactions can be used to observe how intersecting variables at the individual level can translate into population health outcomes. In model specifications, researchers can let agents be adaptive (i.e. agents change their behaviours over time) or prescriptive (i.e. agents behave using pre-set rules). Agents may interact based on spatial proximity or memberships within a social network; can be clustered into groups at different levels (i.e. households, neighbourhoods, states); and, can be influenced by group-level factors (Galea, Riddle, et al. 2009; Galea, Hall, et al. 2009; El-Sayed et al. 2012).

Case study: intersectionality and health inequities in India

India is home to 1.2 billion people, ranked hierarchically by gender, religion, caste, and economic class, among other axes of power (World Bank Group 2017). Steady economic growth of 3–4% annually replaced the stagnation of the British colonial era, making famines a distant memory (Drèze and Sen 2013). India's economic growth has accelerated over the last three decades at an annual rate of 9%, making it one of the world's fastest growing economies (Paul and Mas 2016).

At independence in 1947, after two centuries of British colonial rule, India committed to using legislative policy within the framework of a representative democracy to improve the lives of women and historically

disadvantaged minority groups. To reduce caste-based inequalities, India has adopted a number of progressive laws designed to reduce structural inequities and equalize political representation. All 30 states have set aside 22.5% of seats in the legislature and government-sponsored educational institutions and public-sector jobs for the Scheduled Castes (16–17% of the population) and the Scheduled Tribes (7–8% of population) (Deshpande 2000; Pande 2003). Further, India enacted a number of laws and regulations aiming to eliminate gender-based discrimination and, thereby, improve women's status in society.

The 2006 Prohibition of Child Marriage Prevention Act criminalizes child marriage and imposes a jail sentence or fine. The 1994 National Pre-Conception and Pre-Natal Diagnostics Techniques (PNDT) Act bans sex-selective abortions. And, since 1993, Indian states have been constitutionally required to reserve for women one-third of all chief positions (Pradhan) in local village councils (Gram Panchayats). Notably, in a landmark ruling in 2014, India's Supreme Court recognized transgender people as members of a third gender, which shows the country's gradual progress toward gender inclusiveness.

Still, Amartya Sen has argued that health in India is in 'dire straits' due to severe underinvestment (India Express News Service 2017). In 2011, India spent 3.9% of its gross domestic product on health, a 28% decrease from 1995. India's health inequalities by gender, caste, and class are glaring, and particularly at the intersection where all three meet. A child from a Scheduled Caste, for example, is twice as likely to die in the first year of life compared to a Brahmin child (Thresia 2018). A female child is 1.7 times more likely to die in the first year than a male child (Chowdhury et al. 2017). Infants and children from the lowest standard of living quintile respectively have 2.73 and 4.61 higher odds of mortality than those from the highest quintile (Subramanian et al. 2006). Even Kerala, the southern-most Indian state with health indicators comparable to those in high-income countries, is beset by pervasive health inequalities (Thresia 2018).

Positions

Caste, gender, and class are the most salient axes of social position and power in present-day Indian society. Table 10.5.1 compares race in the US and caste in India as systems of social stratification.

Table 10.5.1 Race vs. caste comparisons

	Race in the US	Caste in India
Context of origin	Western colonialism and capitalism	Hindu society
Status assignment	At birth	At birth
Recognizable by:	Overt characteristics: skin colour; hair texture, facial features; stature, name	Overt and covert characteristics; skin tone; name; occupation; place of residence; etc.
Structural characteristics	Hierarchical	Hierarchical
	Endogamy* practised	Endogamy, a social norm
	Occupation, not race based	Occupation caste based
Social function	Maintains social order by reducing transaction costs	Maintains social order by regulating completion
Social meaning	Imposition of worth based on in or out-group membership	Imposition of worth based on in or out-group membership
Rationalized by:	Economic status and law	Religio-philosophical ideology
Affirmative action	Quotas	Quotas
Lived experiences	Differential access to goods; services; livelihood, respect, self-determination, peace of mind, pleasure, and other valued things including nourishment, shelter, health, independence, justice, security, and long life	

* Endogamy refers to marriage within one's own group based on kinship and/or shared customs, values, and beliefs (such as tribes or castes).

Ancient Hindu society stratified the population into four mutually exclusive, exhaustive, birth-ascribed, and endogamous castes (varna) (Desai and Kulkarni 2008; Vaid 2014), each associated with a type of occupation:

Brahmin (priests and teachers),
Kshatriya (rulers and warriors),
Vaishya (traders)
Shudra (artisans and manual labourers).

Each caste is further stratified into many smaller subcastes (*jatis*), and each subcaste into even smaller sub-sub castes (Vaid 2014). Scheduled Castes, the lowest in the hierarchy, tend to be employed in occupations considered profane (e.g. skinning animal carcasses, butchering animals, removal of human waste, attendance at cremation groups, washing clothes, fishing). The historical justification for 'untouchability' is 'they are so low that they do not deserve to be assigned a caste' (Desai and Kulkarni 2008).

The ritual of purity-pollution underlies the religious rationale for the caste system. Dharma (right conduct), karma (just deserts) and rebirth are the philosophies rationalizing caste assignments. At nearly 3,000 years, India's caste system is among the world's oldest forms of social stratification (Beteille 1969; Von Fürer-Haimendorf 1982; Galantar 1984; Omvedt 1993). Caste discrimination was legally abolished in 1950, but continues to vitally affect opportunities, rewards, and social roles. India's gender-based social stratification entitles men to more rights, resources, and responsibilities compared to women, regardless of ability. India's gender stratification and preference for sons has resulted in many families adopting sex-selective abortion skewing pro-male sex-ratio at birth and neglect of female infants and children (Miller 1989; Jeffery and Basu 1996; Murthi et al. 2018).

Patrilocality (i.e. a married couples living near or with the husband's parents), patrilineality (i.e. names and property pass to the next generation through male descendants), the role of sons in religious rituals, old-age support from sons, the dowry system, and desire to protect female safety and purity are the key cultural factors legitimizing gender stratification (Jayachandran 2015).

Intersectional processes

Returning to Figure 10.5.2, gender and caste interact in complex ways affecting illness and health in India. For the purpose of illustration, consider gender and caste as simple dichotomies. Compared to Priya, a Brahmin woman, Amitabh, a Brahmin man will likely be less vulnerable and susceptible to health risk factors. He will have better access to assets and

services that promote good health, and he will suffer fewer consequences of illness, violence, or disability. On the other hand, Priya will likely perform better than Prakash, a Scheduled Caste/Schedule Tribe (SC/ST) man, who will likely perform better than Anju, an SC/ST woman. While Amitabh and Anju hold extreme social positions on a continuum of intersecting gender and caste identities, Priya and Prakash belong to the middle group. Amitabh has only privileges accorded to him due to his male and Brahmin identities and Anju has only disadvantages. Priya can leverage her caste privileges to compensate for her gender disadvantages. And Prakash can leverage his gender privileges to partially compensate for his caste disadvantages. Although, both Priya and Prakash exist in the middle of the social continuum, Priya belongs to a dominant middle group and Prakash to subordinate middle group since gender identity is more fluid than caste identify in present-day India.

Intersectional health research in India

To date, only a handful of studies have adopted an intersectionality framework to explain health disparities in India. Sen and Iyer (2012) used household survey data on health seeking behaviour from 60 villages of Koppal district, Karnataka. They found that, of the six groups based on two categories of gender (male, female) and three categories of class (poorest, poor, and non-poor), middle-groups (non-poor women and poor-men) shared striking similarities in key health outcomes (Sen and Iyer 2012). Using the National Family Health Surveys of India 1992/93—2005/06, Joe found severe immunization shortfalls among SC/ST children from rural areas, particularly for girls (Joe 2015). Using the National Family Health Surveys of India 1992/93–2005/06, Mukhopadhyay (2015) found that ‘class inequality dominates caste inequality and caste inequality dominates gender inequality’ in childhood stunting in rural North India (Mukhopadhyay 2015). Finally, Scaria (2017) used data from a household survey from Wadakkanchery village of Kerala and found that morbidity is highest among women from Scheduled Castes and Tribes, followed by women from Other Backward Classes, and Christian and Muslim women (Scaria 2017).

Looking forward: embracing intersectional social epidemiology

Intersectionality offers social epidemiologists an opportunity to explore innovative research methods for exploring the systems driving health inequalities. It also offers a pathway for using our research to improve the underlying social and structural conditions contributing to disparate health outcomes (Galea 2013; Galea and Link 2013; Bauer 2014).

Embracing intersectionality will require a range of approaches, from ecological thinking to systems thinking. Complex systems approaches will require data tracing multiple positions and processes within and across levels (Galea 2017). At a minimum, we need to maximize our access to data on social positions (e.g. race, ethnicity, caste, etc.) and social processes (e.g. privilege/disadvantage, racial discrimination/racialization) across multiple levels of social-ecology. Big data from social media, consumer health-record platforms, and troves of administrative and corporate records presents unique opportunities to explore causes of disease, factors that protect health, and sources of resilience.

The conceptual nuance of measuring the contribution of multiple intersecting identities to both individual and population health offers social epidemiologists an opportunity to embrace theory and deepen our understanding of historical, structural, social, and cultural drivers of health and illness (Galea and Link 2013). Finally, and, perhaps most importantly, social epidemiologists need to continue to ask questions of consequence (Galea 2013). This means continuing to pursue knowledge of disease causality but not simply for the sake of knowing. Rather, we need to pursue this knowledge with the aim of intervening to mitigate or enhance individual and community-level factors driving health outcomes, both good and bad.

Conclusion

Global health activists and practitioners have been using intersectional approaches since at least 1978 and the Alma-Ata Health for All Declaration, which urged the world to eliminate ‘gross inequality in the health status of the people, particularly between developed and developing countries as well as within countries’ by promoting ‘maximum community and individual self-reliance and participation in the planning, organization, operation and control of primary healthcare’ (WHO and United Nations Children’s Fund 1978). As defined by Alma-Ata, primary care put equal emphasis on addressing social drivers of health and access to healthcare.

We have not yet achieved health for all. Health disparities continue to widen in countries at all income levels. The social margins team with individuals and populations excluded from economic productivity and good health based on their position in the social hierarchy and processes that assign privilege and disadvantage based on race, ethnicity, caste, sex, gender, sexual orientation, education, income, geography, and other characteristics. Social epidemiology and global public health have an opportunity to hasten social change by embracing intersectional complexities innovating our current research methods, and conducting research that facilitates social change. We must also recruit a new generation of public health researchers who represent multiple intersecting identities and encourage them to think critically and creatively about what our current research methods miss.

REFERENCES

Alexander, M. (2010). *The New Jim Crow: Mass Incarceration in the Age of Color Blindness*, New York: The New Press.

Allen, B.J. (2011). *Difference Matters: Communication of Social Identity*. Long Grove, IL: Waveland Press.

Amroussia, N., et al. (2017). 'Is the doctor God to punish me?!' an intersectional examination of disrespectful and abusive care during childbirth against single mothers in Tunisia. *Reproductive Health*, 14(1), 1–12.

Bauer, G.R. (2014). Incorporating intersectionality theory into population health research methodology: challenges and the potential to advance health equity. *Social Science and Medicine*, **110**, 10–17.

Berger, P. and Luckmann, T. (1966). *The Social Construction of Reality: A Treatise in the Sociology of Knowledge*. New York: Penguin.

Berkman, L.F., Kawachi, I., and Glymour, M.M. (eds.) (2014). *Social Epidemiology* (2nd ed.). Oxford: Oxford University Press.

Berremann, G.D. (1972). Race, caste, and other invidious distinctions in social stratification. *Race*, 13(4), 386–414.

Beteille, A. (1969). *Castes Old and New: Essays in Social Structure and Social Stratification*. London: Asia Publishing House.

Bhopal, R. (2015). Race, ethnicity, epidemiology, and public health. In: Detels, R., et al. (eds.) *Oxford Textbook of Global Public Health* (6th ed.) Oxford: Oxford University Press, pp. 1371–81.

Bowleg, L. (2012). The problem with the phrase women and minorities: intersectionality—an important theoretical framework for public health. *American Journal of Public Health*, 102(7), 1267–73.

Cable, N. (2014). Life course approach in social epidemiology: an overview, application and future implications. *Journal of Epidemiology*, 24(5), 347–52.

Centers for Disease Control and Prevention (2018). Social determinants of health: know what affects health. *Social Determinants of Health*. [Online] Available at: https://www.cdc.gov/socialdeterminants/

Chowdhury, R. et al. (2017). Gender differences in infant survival: a secondary data analysis in rural North India. *BMJ Open*, 7(8), 1–8.
Collins, P. H. and Bilge, S. (2016). *Intersectionality*. Cambridge: Polity Press.
Commission on Social Determinants of Health (2008). *Closing the Gap in a Generation: Health Equity Through Action on the Social Determinants of Health*. Geneva. [Online] Available at: http://apps.who.int/iris/bitstream/handle/10665/43943/9789241563703_eng.pdf;jsessionid=90882D00F0E5C98E262BB3F99D10733D?sequence=1
Crenshaw, K. (1989). Demarginalizing the intersection of race and sex: a black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *The University of Chicago Legal Forum*, 1989(1), 139–67.
Demby, G. (2018). How segregation shapes fatal police violence. *Code Switch: Race and Identity Remixed. National Public Radio*. [Online] Available at: https://www.npr.org/sections/codeswitch/2018/03/02/589483471/how-segregation-shapes-fatal-police-shootings
Desai, S. and Kulkarni, V. (2008). Changing educational inequalities in India in the context of affirmative action. *Demography*, 45(2), 245–70.
Deshpande, A. (2000). Does caste still define disparity ? A look at inequality in Kerala, India. *American Economic Review*, 90(2), 322–5.
Diez Roux, A.V. (2011). Complex systems thinking and current impasses in health disparities research. *American Journal of Public Health*, 101(9), 1627–34.
Drèze, J. and Sen, A. (2013). *An Uncertain Glory: India and Its Contradictions*, Princeton, NJ: Princeton University Press. [Online] Available at: https://press.princeton.edu/titles/10175.html
El-Sayed, A.M., et al. (2012). Social network analysis and agent-based modeling in social epidemiology. *Epidemiologic Perspectives & Innovations*, 9(1), 1–9.
Galantar, M. (1984). *Competing Inequalities: Law and the Backward Classes.*, Berkeley, CA: University of California Press.
Galea, S. (2013). An argument for a consequentialist epidemiology. *American Journal of Epidemiology*, 178(8), 1185–91.
Galea, S. (2017). Making epidemiology matter. *International Journal of Epidemiology*, 46(4), 1083–5.
Galea, S. and Link, B.G. (2013). Six paths for the future of social epidemiology. *American Journal of Epidemiology*, 178(6), 843–9.
Galea, S., Hall, C., and Kaplan, G.A. (2009). Social epidemiology and complex system dynamic modelling as applied to health behaviour and drug use research. *International Journal of Drug Policy*, 20(3), 209–16.
Galea, S., Riddle, M., and Kaplan, G.A. (2009). Causal thinking and complex system approaches in epidemiology. *International Journal of Epidemiology*, 39(1), 97–106.
Green, M.A., Evans, C.R., and Subramanian, S.V. (2017). Can intersectionality theory enrich population health research? *Social Science and Medicine*, **178**, 214–16.
Hankivsky, O., et al., 2010. Exploring the promise of intersectionality for advancing women's health research. *International Journal for Equity in Health*, 9(5), 1–15.
India Express News Service (2017). Healthcare in dire straits in India, says Amartya Sen. *The Indian Express*. [Online] Available at: http://indianexpress.com/article/cities/mumbai/healthcare-in-dire-straits-in-india-says-amartya-sen-4496610/
Jayachandran, S. (2015). The roots of gender inequality in developing countries. *Annual Review of Economics*, (7), 63–88.
Jeffery, R. and Basu, A.M. (1996). *Girls' Schooling, Women's Autonomy and Fertility Change in South Asia*. Thousand Oaks, CA: Sage Publications.
Joe, W. (2015). Intersectional inequalities in immunization in India, 1992–93 to 2005–06: a progress assessment. *Health Policy and Planning*, 30(4), 407–22.

Larson, E., et al. (2016). 10 Best resources on … intersectionality with an emphasis on low- and middle-income countries. *Health Policy and Planning*, 31(8), 964–9.
Lopez, N. and Gadsden, V.L. (2016). Health inequities, social determinants, and intersectionality. National Academy of Medicine. [Online] Available at: https://nam.edu/health-inequities-social-determinants-and-intersectionality/
McCall, L. (2005). The complexity of intersectionality. *Signs*, 30(3), 1771–800.
McIntosh, P. (1989). White privilege: unpacking the invisible knapsack. *Peace and Freedom Magazine*, 10–12. [Online] Available at: http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:White+Privilege:+Unpacking+the+Invisible+Knapsack#0
Miller, B.D. (1989). Changing patterns of juvenile sex ratios in rural India, 1961 to 1971. *Economic and Political Weekly*, 24(22), 1229–36.
Mukhopadhyay, S. (2015). The intersection of gender, caste and class inequalities in child nutrition in rural India. *Asian Population Studies*, 11(1), 17–31.
Murthi, M., Guio, A., and Drèze, J. (2018). Mortality, fertility, and gender bias in India: a district-level analysis. *Population and Development Review*, 21(4), 745–82.
Nyarko, K.A., et al. (2013). Explaining racial disparities in infant health in Brazil. *American Journal of Public Health*, 103(9), 1675–84.
Omvedt, G. (1993). *Reinventing Revolution: New Social Movements and the Socialist Tradition in India*. Armonk, ME: Sharpe.
Oxford English Dictionary Online (2018). Oxford English Dictionary. [Online] Available at: http://www.oed.com/
Pande, R.P. (2003). Can mandated political representation increase policy influence for disadvantaged minorities? Theory and evidence from India. *American Economic Review*, 93(4), 1132–51.
Paul, J. and Mas, E. (2016). The emergence of China and India in the global market. *Journal of East-West Business*, 22(1), 28–50.
Riddell, C.A., Harper, S., and Kaufman, J.S. (2017). Trends in differences in U.S. mortality rates between black and white infants. *JAMA Pediatrics*, 171(9), 911–13.
Rosenberg, N.A., et al. (2002). Genetic structure of human populations. *Science*, 298(5602), 2381–5.
Rothman, J. (2014). The origins of privilege. *New Yorker*. [Online] Available at: https://www.newyorker.com/books/page-turner/the-origins-of-privilege.
Sampath, R. (2015). Racial and caste oppression have many similarities. *The Conversation*. [Online] Available at: https://theconversation.com/racial-and-caste-oppression-have-many-similarities-37710
Scaria, S. (2017). Caste and gender in relation to health inequality: a village study from Kerala. *South Asia Research*, 37(1), 1–18.
Schulz, A.J. and Mullings, L. (2006). *Gender, Race, Class, and Health: Intersectional Approaches.* San Francisco, CA: Jossey-Bass.
Scott, J. and Marshall, G. (2009). *A Dictionary of Sociology*. Oxford: Oxford University Press.
Sen, G. and Iyer, A. (2012). Who gains, who loses and how: leveraging gender and class intersections to secure health entitlements. *Social Science & Medicine*, 74(11), 1802–11.
Subramanian, S.V., et al. (2006). The mortality divide in India: the differential contributions of gender, caste, and standard of living across the life course. *American Journal of Public Health*, 96(5), 818–25.
Tajfel, H. and Turner, J.C. (1985). The social identity theory of intergroup behavior. *Psychology of Intergroup Relations*, **5**, 7–24.
Thresia, C.U. (2018). Health inequalities in South Asia at the launch of sustainable development goals: exclusions in health in Kerala, India need political interventions. *International Journal of Health Services*, 48(1), 57–80.

Vaid, D. (2014). Caste in contemporary India: flexibility and persistence. *Annual Review of Sociology*, 40(1), 391–410.

Von Fürer-Haimendorf, C. (1982). *Tribes of India: The Struggle for Survival*. Berkeley, CA: University of California Press. [Online] Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.739.3059&rep=rep1&type=pdf

Warner, L.R. (2008). A best practices guide to intersectional approaches in psychological research. *Sex Roles*, 59(5–6), 454–63.

Weber, L. and Parra-Medina, D. (2003). Intersectionality and women's health: charting a path to eliminating health disparities. In: Texler Segal, M., Demos, V., and Kronenfeld, J.J. (eds.). *Gender Perspectives on Health and Medicine*. Bingley: Emerald Group Publishing Limited.

Wemrell, M., et al. (2016). Contemporary epidemiology: a review of critical discussions within the discipline and a call for further dialogue with social theory. *Sociology Compass*, 10(2), 153–71.

Wemrell, M., Mulinari, S., and Merlo, J. (2017). Intersectionality and risk for ischemic heart disease in Sweden: categorical and anti-categorical approaches. *Social Science and Medicine*, **177**, 213–22.

World Bank Group (2017). World Bank development indicators. *World Bank*. [Online] Available at: https://data.worldbank.org/products/wdi

World Health Organization and United Nations Children's Fund (1978). Alma-Ata Declaration: Primary Health Care. In: *International Conference on Primary Health Care*. Geneva, Switzerland: WHO. [Online] Available at: http://www.hauora.co.nz/assets/files/PHOInfo/Alma_Ata_conference_1978_report and declaration.pdf

Yudell, B.M., et al. (2016). Taking race out of human genetics. *Science*, 351(6273), 564–65.

10.6

The health of Indigenous peoples

Papaarangi Reid, Donna Cormack, Sarah-Jane Paine, Rhys Jones, Elana Curtis, and Matire Harwood

Defining Indigenous

The United Nations (UN) reports there are more than 370 million Indigenous peoples worldwide, residing in around 90 countries (UN Permanent Forum on Indigenous Issues (UNPFII) 2009, p. 1), with greater numbers in the Africa, Asia, and South America regions. These population statistics are fraught with significant challenges in how Indigenous peoples are defined, whether Indigenous peoples are recognized in all nation states, and data completeness and quality (Ohenjo et al. 2006; Gracey and King 2009; Davis 2016).

Defining the term 'Indigenous peoples' is contested and while it is often taken as an umbrella for other names such as Aboriginal, Native, First Nations, tribal peoples or original inhabitants (Gracey and King 2009), these terms are not universal and may fail to acknowledge the many and varied histories of coloniality across different continents in different eras (Ohenjo et al. 2006; Mohindra 2017). In discussing the challenges of defining Indigenous peoples, the UNPFII notes:

> Today, the prevailing view is that a single formal universal definition of the term is unnecessary for the recognition and protection of the rights of indigenous peoples, and will fail to capture the diversity, range and complexity of the situations of indigenous peoples throughout the world. (2015, p. 6)

The UN reiterates the importance of self-identification in recognition that some states may deny the existence of Indigenous peoples, and suggests that on an individual level:

> . . . an Indigenous person who self-identifies with an Indigenous population (group consciousness) and is recognised and accepted as one of its members (acceptance by group), which in turn reserves the sovereign right and power to decide who belongs without external interference. (UNPFII 2015, p. 4)

Despite the fundamental right for Indigenous peoples to be recognized, some states refuse to do so (Ohenjo et al. 2006), choosing instead to describe peoples as being marginalized by geographic, socioeconomic, linguistic or lifestyle factors (UNPFII 2015). Defining who is Indigenous is central to claiming Indigenous identity, history and rights and is also fundamental to 'being counted' within nation states. Reluctance to acknowledge Indigenous peoples amplifies challenges with data completeness and quality.

'Counting' Indigenous peoples in data

Indigenous peoples are often undercounted, miscounted, or not counted in state data systems (Anderson et al. 2016; Axelsson et al. 2016a; Davis 2016). In many jurisdictions, data for Indigenous peoples are not collected, not reported, or aggregated into broader groups, limiting visibility for Indigenous communities within the data. The ability to count, in a practical and technical sense, is deeply connected to the right to be counted as Indigenous peoples and nations (Robson and Reid 2001) and affirmed in human rights frameworks and conventions, including the United Nations Declaration on the Rights of Indigenous Peoples (UNDRIP) (UN 2007). Appropriate data collection and disaggregated data for Indigenous populations is an 'urgent priority' (UNPFII 2004).

Variable availability and quality of data across Indigenous populations also impacts on understanding the complex inequities in Indigenous health globally. Available data are not always relevant for Indigenous peoples' own priorities (Walter 2018), and limited by a focus on narrow biomedical measures rather than Indigenous conceptualizations of health and well-being (UNPFII 2009; Walker et al. 2017). This limits the ability to explore questions of importance to our own Indigenous communities, or to hold states to account for their failures to recognize Indigenous rights and to eliminate inequities.

The state of health of Indigenous peoples

UNDRIP reaffirms Indigenous rights, including the right to be actively involved in the planning, development, and implementation of health services and programmes, the right to traditional medicines and health practices and the right to equal health within the country of residence (UN 2007; UNPFII 2013; Axelsson et al. 2016b).

In contrast to this clear articulation of rights, and corresponding obligations for states to ensure the realization of these rights, many Indigenous peoples experience stark inequities in health, reflected in lower life expectancy, increased mortality and morbidity, and disproportionate numbers living with disability relative to other populations (Anderson et al. 2006; Ohenjo et al. 2006; UNPFII 2013; Axelsson et al. 2016b; Mohindra 2017). These outcomes are heavily influenced by the environments in which we live, including unequal access to conditions that are important for good health, such as safe water, secure and nutritious food, sanitation, and other social determinants of health (UNPFII 2013). Indigenous peoples experience multiple obstacles to healthcare, as well as discriminatory, inequitable, and culturally incompetent care (Allan and Smylie 2015; UNPFII 2015). Further threats to the health of Indigenous peoples include the ongoing impacts from the destruction of Indigenous systems and institutions, natural environments, and the effects of climate change (Gracey and King 2009; UNPFII 2013).

While there are publications describing important health issues across Indigenous communities, comprehensive health outcomes data, and peer-reviewed studies of Indigenous health at the state level predominantly focus on high-income countries, especially Australia, Canada, New Zealand, and the United States (Anderson et al. 2016; Axelsson et al. 2016a), with less research published for low- and middle-income countries (Mohindra 2017). Challenges also arise from international comparisons between Indigenous peoples that may not account for variations in territorial histories, social, political, and economic arrangements, and risk focusing attention on whose health status is worst among Indigenous communities. However, international comparisons can be useful in representing the international failure of many nation states to realise Indigenous peoples' rights to health. Mindful of these challenges, this section summarizes key health measures and outcomes for Indigenous peoples globally.

Stark, inequitable health outcomes for Indigenous peoples have been documented globally (e.g. Montenegro and Stephens 2006; Ohenjo et al. 2006; Hayman et al. 2015). An international collaboration recently published the most comprehensive review on Indigenous health to date, drawing on data from 28 populations in 23 countries, covering around 154 million Indigenous peoples globally (Anderson et al. 2016). The review confirmed previous work, documenting poorer outcomes for Indigenous peoples globally when compared with non-Indigenous populations, sustained across many health and social outcomes (Anderson et al. 2016). The review also identified variations in the magnitude of inequities between countries. Worse outcomes for Indigenous were typical but not universal, with some instances identified where there were no differences or, very rarely, better outcomes for Indigenous communities (Anderson et al. 2016). For example, data on Life-Expectancy-at-Birth (LEaB), covering 18 Indigenous populations from 14 countries, showed generally lower LEaB for Indigenous peoples. Inequities between Indigenous and non-Indigenous peoples varied from extreme inequities in some populations of 10 years or more, to eight Indigenous populations with gaps between 5 and 10 years, with the remaining countries having differences of two years or less, except for Sami in Sweden where the difference was not statistically significant (Anderson et al. 2016).

The review also reported that the health of Indigenous mothers and infants generally echoed socioeconomic conditions of their communities. In general, infant mortality rates were lower (less than 50/1,000 live births) in higher income countries than in lower income countries (more than 50/1,000 live births). In almost all populations where data were available, Indigenous infant mortality rates were higher than for non-Indigenous peoples with particularly high ratios in some countries (Anderson et al. 2016).

Maternal deaths should be infrequent and reflect the availability, accessibility, affordability, and safety of local health services. Data was unavailable or unstable for many nations in the review but consistent disparities in maternal deaths were noted in Australia, New Zealand, Panama, and Columbia (Anderson et al. 2016). The receipt of lower levels of maternity services by Indigenous mothers has been reported elsewhere (Rumbold et al. 2011; Paulino et al. 2019). In some nation states, Indigenous women in remote communities have to relocate to give birth

(UNPFII 2015), which may leave them without social support in health services that may be culturally unsafe.

Indigenous peoples continue to be disproportionately impacted by infectious diseases with mortality and morbidity generally reported to be higher, especially for HIV/AIDS, tuberculosis, malaria, syphilis, viral hepatitis, and rheumatic fever (Montenegro and Stephens 2006; Gracey and King 2009; UNPFII 2009; Russell et al. 2019). Infectious diseases have a significant impact, especially in lower income countries, among marginalized communities living in poverty, and where the disease or its risk factors are stigmatized. Indigenous health inequities for rheumatic fever exist in both developing (Steer et al. 2002) and wealthy countries (Carapetis et al. 2005; Abouzeid et al. 2019). Indigenous children in high-income countries also experience inequitably high rates of respiratory disease (Chang et al. 2009). The socioeconomic and health disadvantage common among Indigenous communities drives much of these inequities (UNPFII 2009, 2015), which are further exacerbated by inequitable health service access and care (Baker et al. 2012).

Non-communicable diseases (NCDs) are now deeply problematic for Indigenous peoples. NCDs account for significant mortality and morbidity globally (di Cesare et al. 2013), contributing to both health inequities within nations (di Cesare et al. 2013) and to the well-being and economic security of Indigenous communities (NCD Alliance 2012). The four most common NCDs are cardiovascular diseases, cancers, diabetes, and chronic respiratory diseases (WHO 2013), contributing significantly to inequities in life expectancy between Indigenous and non-Indigenous peoples. Cardiovascular disease (CVD) is the leading cause of mortality for most Indigenous peoples. Deaths from ischaemic heart disease, twice those of non-Indigenous peoples (Ministry of Health 2011; Espey et al. 2014), account for more than half of CVD deaths, followed by stroke, heart failure (Woods et al. 2012) and rheumatic heart disease (RHD) (Watkins et al. 2017). CVD inequities are evident between Indigenous and non-Indigenous peoples in age of onset and the underlying cause of the heart disease. For example, RHD causing heart failure is more common in young Indigenous Australians and Africans; hypertensive and ischaemic heart failure is more common in older non-Indigenous people (Bukhman and Ochieng 2017).

Cancer is underemphasized as a health issue for Indigenous peoples, particularly in low income countries, possibly due to overwhelming rates of

infectious diseases. Cancers have a broad range of risk factors including infections (e.g. Hepatitis B and C, HPV, HIV, and *Helicobacter pylori*), metabolic conditions, diet, and exposure to toxins and carcinogens (WHO 2019). The relative frequency of exposure to these factors has led to variation in the types of cancer impacting Indigenous peoples in different regions (Moore et al. 2014). However, the risk of dying from all cancers is significantly higher for Indigenous peoples in general (Cunningham et al. 2008; Parkin et al. 2008; Ministry of Health 2015). Timely access to cancer screening, diagnostic and treatment services (Moore et al. 2014) by Indigenous peoples show consistent inequities in regions where measured (UNPFII 2015).

Respiratory conditions, including chronic obstructive pulmonary disease (COPD) and asthma, are also common causes of mortality, morbidity, and hospitalizations for Indigenous peoples. Asthma is the second leading cause of hospitalization among Indigenous Australians and death rates are 3.2 times higher (NCD Alliance 2012). COPD hospitalization and mortality rates for Indigenous peoples are higher in most higher income nations (Australian Institute of Health and Welfare 2014; Gershon et al. 2014; Ministry of Health 2015) with differences in exposure to tobacco and environmental pollution, differences in disease management, and differential access to health services, explaining most of these inequities (NCD Alliance 2012; Chang et al. 2014). While allergic disorders are rare in traditional rural settlements, urbanization, and the concurrent move to overcrowded living conditions and increased exposure to environmental pollutants has seen increased rates for Indigenous communities (Mbugi and Chilongola 2010).

Disorders of nutrition, including undernutrition, malnutrition, and overnutrition, are also important. Anderson et al. (2016) note that child malnutrition is still a significant issue for Indigenous peoples, reporting higher levels of childhood malnutrition with stunting among Indigenous peoples residing in a number of nation states in South America. Rates of obesity among children and adults are increasing, including among those populations where childhood malnutrition is still disproportionately high (Anderson et al. 2016).

Type 2 diabetes is currently estimated to affect more than 50% of Indigenous adults aged over 35 years, with incidence rates climbing (NCD Alliance 2012). Indigenous peoples experience disproportionate rates of

diabetic complications including renal failure, lower extremity amputations, eye disease and macrovascular outcomes (Naqshbandi et al. 2008). While the high incidence of end-stage renal failure among Indigenous peoples is reported in major population centres across Oceania, Indigenous peoples receive fewer kidney transplants than non-Indigenous peoples and many experience geographic and economic difficulties accessing optimal treatments (Weil and Nelson 2006).

Diabetes inequities exist across the life course. For example, gestational diabetes mellitus (GDM) incidence rates among Native American (7.9%), Canadian First Nations (11.5%), Australian Aboriginal (8.4%), and Māori in New Zealand (6%) are higher than those for non-Indigenous women of those respective countries and the total international incidence (2–5%) (Lawrence et al. 2008; Porter et al. 2012; Jowitt 2016). GDM is a risk factor for the development of NCDs in the lifetime of both mothers (Nolan 2011) and their offspring (O'Reilly 2014). Despite suggestions that this might be a driver for the high prevalence of intergenerational diabetes (Osgood et al. 2011) and other NCDs in Indigenous peoples, screening for and management of GDM among Indigenous women is currently inadequate (McGrath et al. 2007; Daly et al. 2017). It is critical to acknowledge the lived context of Indigenous peoples who often have been excluded from traditional territories and food sources and reside in obesogenic environments and situations of food insecurity (Gracey and King 2009).

Finally, mental well-being is a global health priority for Indigenous peoples (Pan American Health Organization (PAHO) 2018). Mental health conditions are 'a complex challenge because of increased visibility, pervasive structural stigma, outdated practices and organizational fragmentation' (PAHO 2018, p. 2). Mortality data underestimate the importance of these conditions on the well-being and livelihoods of individuals and their families. In 2015, it was estimated that globally NCDs accounted for around 60% of Disability-Adjusted Life Years, of which mental, neurological., substance-use disorders, and self-harm (MNSS) accounted for 12% (PAHO 2018).

The legacies of colonization, dispossession, marginalization, systematic discrimination, and denial of Indigenous rights are linked to disproportionately high rates of stress and mental distress, highlighting the downstream impacts of these trauma on a number of parameters of Indigenous well-being. Anderson et al. (2016) note that the lack of

morbidity data for Indigenous peoples, including data on mental health conditions, was a key issue especially for countries outside of North America and Oceania, limiting the ability to fully explore the impacts of mental illness on Indigenous well-being.

Mental health conditions are disproportionately prevalent among young Indigenous peoples (WHO 2014). Suicide is a leading cause of death in young Indigenous peoples, particularly in Australia, Brazil, New Zealand, Taiwan, and circumpolar countries (Patel et al. 2007; WHO 2014). Indigenous youth are more likely to be exposed to risk factors for mental illness such as growing up in poverty, adverse conditions such as scarcity of food, poor nutrition, violence and aggressions, inadequate education, substance abuse and sexual and reproductive health challenges (Patel et al. 2007).

NCDs are clearly a challenge for Indigenous peoples and one that needs Indigenous voice, active participation, and leadership. Despite these inequities, international efforts to prevent and control NCDs have been slow to ensure Indigenous participation. Leaders in Indigenous health remain frustrated by this and note that working with Indigenous peoples is critical to ensure a reduction in NCD risk factors and NCD burden, and to avoid increasing inequity (NCD Alliance 2012; Finlay and Armstrong 2018).

Inequities have also been documented in both intentional and unintentional injury among Indigenous peoples, with higher rates of death, morbidity and disability reported by a number of countries illustrating differential exposure to violence and conflict (Montenegro and Stephens 2006), differential access to safe living and working conditions, and appropriate injury prevention strategies (Berger 2002).

Access to, and quality of, healthcare is an issue for Indigenous peoples globally (Montenegro and Stephens 2006; Ohenjo et al. 2006; UNPFII 2015). For some Indigenous peoples, healthcare is inaccessible for geographic and financial reasons, or because of discrimination (Montenegro and Stephens 2006; Ohenjo et al. 2006). In addition, many countries have healthcare systems that privilege non-Indigenous approaches to health and well-being, and marginalize Indigenous ways of healing, including denying access to traditional medicines. The decentring of Indigenous health knowledges and practices, coupled with discrimination, means that even where care is available, Indigenous peoples often experience culturally inappropriate or unsafe services.

The representation and framing of Indigenous health

Alongside the urgent need for improved data for Indigenous peoples, Indigenous scholars are concerned with the ways in which Indigenous peoples are represented in health research and statistics (Walter 2016, 2018). Representations of Indigenous health, often controlled by non-Indigenous peoples, invoke deficit, racialized narratives that contrast sharply with Indigenous narratives (Walter 2016). Moreover, data available to describe the health of Indigenous peoples often do not reflect historical and embodied contexts and circumstances, nor contemporary lived realities. In many states, non-Indigenous models of health are dominant in health systems (UNPFII 2015) and data, displacing Indigenous approaches to health and well-being.

Indigenous models of health recognize the interdependence of peoples, the connections between peoples and environments, and the importance of balance (Bird 2002; Smylie and Anderson 2006; Kaholokula et al. 2009; UNPFII 2009), and take into account how histories fundamentally shape social, cultural, political, legal, and economic institutions in societies, with flow on effects for social categorization and hierarchies such as socioeconomic status, and exposure to health-protective or health-damaging factors for individuals and communities (Walters et al. 2011). Current inequities in health are understood to be intimately linked to colonialism and racism, and the disruption to connections between people and environments from forced removal from lands, dispossession or destruction of territories and resources and environmental degradation and pollution (UNPFII 2009). The disruption of connection parallels a disruption to Indigenous systems of knowledge, ways of being, social institutions and processes of intergenerational transmission, and is understood as a root cause of current health inequities (Bird 2002; King et al. 2009; Wirihana and Smith 2014). Increasingly, Indigenous models of health are also articulating historical trauma and its intergenerational transmission, as a fundamental determinant of health for Indigenous peoples (Walter et al. 2011; Paradies 2016).

Interventions for equity in Indigenous health

Given the complex and self-replicating processes that contribute to the health inequity experienced by Indigenous peoples, effective interventions need to be comprehensive, connected, long-term, tailored, and responsive to change. The following are a range of important considerations for progress towards health equity for Indigenous peoples.

Self-determination

Self-determination is an essential component of Indigenous health development. It is affirmed as a prerequisite for the protection, promotion, exercise, and enjoyment of all other human rights and fundamental freedoms (UNPFII 2009). It is a pre-existing human right and is not dependent on the views of, or permission from, nation states (UNPFII 2009). Self-determination is intimately linked to a peoples' democratic entitlement and thus underpins the integrity of all relationships between Indigenous peoples and respective nation states. UNDRIP reaffirms Indigenous rights to self-determination, active participation in decision-making and the need of states to 'obtain free, prior and informed consent' before adopting and implementing legislative or administrative measures that may affect Indigenous peoples (UN 2007). Self-determination also applies to the right of Indigenous peoples to determine the structure and form of health systems and institutions in our territories.

Decolonisation of Indigenous knowledge and knowledge production

Human rights and fundamental freedoms detailed in international rights instruments do not always fully align with Indigenous worldviews. Indigenous values have greater focus on collective rather than individual rights, and these being balanced against responsibilities (UNPFII 2009). Further, Indigenous values including respect, cooperation, reciprocity, spirituality, deference to elders, and the importance of language and traditions are not captured. Most Indigenous peoples acknowledge the life force of all things, animate and inanimate and see as limited, international rights instruments that only acknowledge humans.

Royal (2009) promotes that the central tenet of indigeneity is an active and ongoing kin-based relationship between humankind and the natural

environment, noting that this is not about claiming resources but rather an affirmation that the earth is truly our mother. This alters relational dynamics between humans and the natural environment from one permitting extractive activities that degrade and endanger to one where humans are expected to respect and protect the environment. These worldviews are supported by Indigenous knowledge systems that are neither static nor finite, but have been undermined by processes of colonisation.

Indigenous knowledges form an essential part of the diverse knowledge systems on earth and are critical for important health challenges. UNDRIP specifically notes the right to protect Indigenous knowledges and manifestations of these knowledges including human and genetic resources, while also acknowledging rights to supporting customary processes, philosophies, and relationships.

Interventions on oppressive pathways

Acknowledge historical trauma, promote healing, and prevention

Historically, Indigenous peoples have survived horrific traumatic events, including genocides, ethnocides (destruction of cultures, languages, and ways of living), epistemicides (destruction of traditional knowledges and ways of knowing), and ecocides (destruction of environments including food sources) (Walters et al. 2011; Walters 2016). For many, these events have continued into recent history (forced adoptions or removal of children to boarding schools, long-term poverty, and discrimination) and the present (deforestation of tribal territories, pollution of waterways, continuous racialized microaggressions). Indigenous scientists are seeking to understand the role of historical trauma, and how contexts, processes and outcomes become embodied and contribute to patterns of illness among Indigenous peoples and subsequent health inequities (Walters et al. 2011).

Two growing bodies of evidence are important. First is the evidence that traumatic events are important in the life course by negatively and directly affecting the health of individuals and communities, especially if events are repeated (Walters et al. 2011). The second body of evidence demonstrates that significant trauma can have an intergenerational effect. These effects are mediated by both epigenetic pathways, affecting the way in which genes are regulated and expressed, as well as physiological mechanisms, usually mediated by stress hormones and metabolic or immune responses. It is

thought these mechanisms contribute to the disparate rates of diseases among Indigenous peoples including metabolic diseases, cardiovascular diseases as well as stress and coping disorders such as depression, anxiety, and substance abuse (Walters et al. 2011).

While the theory and evidence relating to historic trauma is still evolving, especially methodological issues (Walters et al. 2011), the complexities of the relationships become obvious. These complexities demand sophisticated research and we ought to rise to this challenge rather than dismiss the role of historic trauma in Indigenous health status.

While Indigenous peoples continue to exhibit courageous resilience in the face of these traumas, healing will require a dual focus. First, eliminating exposure to further stressors that could become embodied, such as poor nutrition, especially during pregnancy and childhood, poverty, lower quality services especially in education, health, justice, and welfare services, exposure to racisms, and higher levels of violence and loss. Second, Indigenous peoples must be permitted to heal and restore from these traumas. Some authors have noted the importance of traditional knowledge and practices to support this healing and how culture can be medicine (Walters et al. 2011, 2020).

Acknowledge racism, and intervene to mitigate and eliminate

Indigenous communities and scholars note the role of racism as a fundamental determinant of Indigenous health (Paradies et al. 2008; Allan and Smylie 2015; Paradies 2016; 2018). As a system of oppression, racism shapes health for Indigenous peoples through its structuring of histories as well as social, cultural, political, and economic institutions. Racism also operates through racially discriminatory practices at the level of individuals, intersecting with other oppressive systems to negatively influence health across the life course and intergenerationally (Paradies 2016).

Interventions are necessary to eliminate racism at interpersonal and structural levels, including action to address the role of racialized beliefs and assumptions among health providers in shaping interactions with patients and influencing clinical decision-making and healthcare outcomes (Paradies et al. 2013; Fitzgerald and Hurst 2017). There is an urgent need for the fundamental role of racism to Indigenous health to be acknowledged as a global public health issue, and for nation states to combat racism against Indigenous peoples within the health sector and more broadly in

society. Further, whether or not there are health impacts, Indigenous peoples have a right to freedom from all forms of discrimination (Paradies 2018).

Acknowledge and address health system bias, make (health) systems culturally safe

Although Indigenous peoples have long identified relationships between imposed colonial health systems and healthcare delivery, and health outcomes, these relationships have been underexamined in health research generally (Smedley et al. 2003), and for Indigenous peoples specifically (Paradies 2018). Acknowledgement of the contingent and culturally-embedded nature of health systems demands critical reflection on how organizational structures, policies, environments, practices, and provider behaviours reproduce and maintain colonial logics and power structures.

Cultural competency and cultural safety are proposed as possible interventions to eliminate ethnic health inequities. In some countries, they have been incorporated into regulatory and legislative frameworks governing healthcare workers (Bacal et al. 2006). However there are important differences in definitions of these terms, how they are interpreted, incorporated into practice, and what these interventions are hoping to achieve. While cultural competency interventions often focus on teaching healthcare workers awareness of Indigenous languages, cultures, and practices, these teachings do not address the fundamental causes of inequities and risk promoting stereotypes and therefore undermine the decolonizing and transformative work necessary to address biases (Truong et al. 2014).

Interventions are necessary that ensure health professionals develop a critical consciousness about historic processes and how they play out in contemporary practice today. Cultural safety was articulated by Dr Irihapeti Ramsden who insisted health workers reflect on their privilege and power within the health system and use it to eliminate injustice and inequity (Papps and Ramsden 1996). Cultural safety insists health workers reflect on themselves as power bearers and on their ‘own culture’ and ask what we need to change in our practice to achieve equitable outcomes for all patients. Ramsden’s work contributed to the current standards of critical self-reflection and reflective practice in the training of healthcare providers.

The importance of health systems where healthcare, interventions and policy are grounded in Indigenous culture and knowledge systems has been

acknowledged (Walters et al. 2020). There are many examples where Indigenous knowledge and healing practices have been incorporated into current health systems, and have contributed significantly to what is now called Westernized medicine (Oguamanum 2004; Maskarinec et al. 2014; Baddock et al. 2017). However, this incorporation can be problematic and damaging where it is extractive, performative, or tokenistic. While culturally grounded health services have been posited as a response to inequities in healthcare access and quality, they should be more rightly understood as a fundamental right of Indigenous peoples to have access to health services, interventions, and systems that are based on Indigenous knowledge and worldviews, and are designed and controlled by Indigenous peoples.

Build an Indigenous health workforce

A health workforce that reflects the community it will serve is a critical component of high quality and equitable healthcare (Betancourt et al. 2003; Lancet 2016). Indigenous health workforce development contributes both to Indigenous development and reducing ethnic inequities in health (Curtis 2016) and aligns with international workforce initiatives. There is evidence that if patients and healthcare workers are of the same ethnicity, described as ethnic or racial concordance (Burgess et al. 2004), interactions are marked by better communication, shared decision-making, improved respect, affiliation, and trust (Cooper et al. 2006), greater satisfaction with the care received and less likelihood of the patient perceiving racism in the interaction (LaVeist and Nutu-Jeter 2002).

Indigenous health workforce development also increases the potential for cultural competency of both health workers and health systems (Betancourt et al. 2005) as it requires healthcare providers to critically reflect on their own culture and biases and the impact these have on clinical interactions and health service delivery (Kumagai and Lypson 2009). Becoming an Indigenous health professional is also generally associated with improved financial independence (Curtis 2016), community respect and support, positive role modelling and intergenerational effects, access to health expertise and better system navigation (Ratima et al. 2007). Seeing Indigenous peoples as health professionals, as opposed to recipients of care, helps challenge colonial stereotypes and increases the potential for positive

collegial interactions which may in turn reduce implicit bias among non-Indigenous health workers (van Ryn et al. 2015).

Some countries are investing in Indigenous health workforce development and their experience has reinforced the need to adopt a 'whole pathway' approach that includes recognition of socioeconomic, cultural, and geographical barriers to accessing quality education at all levels, appropriate careers advice, and comprehensive academic, pastoral. and cultural support within tertiary contexts (Curtis et al. 2014).

Mitigate biases in research

Research is the process of knowledge production but knowledge, who is knowledgeable and knowledge creation are all situated—politically, economically, socially, culturally, and historically. Smith (2012) writes that many Indigenous peoples consider 'research' a dirty word. This references the alignment of research with colonization and its processes of appropriation and claiming. Colonization included the brutal suppression of Indigenous cultures, languages, customs, traditional knowledges and ways of knowing, and in this process, colonizers rebranded themselves as discoverers and authorities of Indigenous lives.

Colonial processes sought to replace what Indigenous peoples know with a belief in a single Euro-centric knowledge system. Imperialism now seeks to govern the way we think about knowledge, controlling what questions we ask, what we see and what we know (Smith 2012). Research is thus understood as a mechanism for establishing and maintaining power and control as part of the colonial project. Even supposedly 'objective' scientific methods can produce evidence that supports racist beliefs and further marginalizes Indigenous peoples (Te Rōpū Rangahau Hauora a Eru Pōmare 2002; Reid et al. 2017).

Indigenous researchers are increasingly rejecting approaches that seek to further entrench colonialism, deny the Indigenous experience, and reaffirm discriminatory views about Indigenous peoples. Further, Indigenous academics have been active in challenging Euro-centric research, revitalizing Indigenous research methodologies and supporting research institutions to do more for Indigenous peoples and communities (Smith 2012). However, it is necessary for all researchers, research organisations, and funding authorities to reject research projects and processes that fail to respect Indigenous knowledges, histories, ambitions, stories, voices, data,

bodies, genes, and rights (Reid et al. 2017; Walker et al. 2017). UNDRIP affirms the rights of Indigenous peoples to maintain, control, and protect traditional knowledges, develop, and incorporate traditional knowledges into our contemporary practices, advance new practices, and create new Indigenous knowledges.

Identify and mitigate future external risks to inequity

Climate change is the greatest threat to global health for the twenty-first century (Costello et al. 2009) working through multiple pathways both direct and indirect (Smith et al. 2014). Climate change acts as a threat multiplier, amplifying the health effects on structurally oppressed communities (Watts et al. 2018), with marginalized populations who already experience poorer health facing disproportionate risk (Levy and Patz 2015; Woodward and Porter 2016; Paavola 2017). Those who have contributed least to global greenhouse gas emissions (e.g. those in developing countries, low-lying island nations, and Indigenous peoples) will be hit first and worst.

Indigenous peoples are also more likely to depend on natural ecosystems for daily living and economic well-being. In addition, the capacity to respond and adapt to climate change is restricted by conditions of greater socioeconomic deprivation, ongoing social marginalization, and exclusion from political processes and structures (Ramos-Castillo et al. 2017).

Actions to tackle climate change could negatively impact Indigenous health and equity (Walpole et al. 2009; Robbins 2015) and further undermine Indigenous rights (Jones et al. 2014). However, climate action is also an opportunity for health development (Watts et al. 2015) and reducing health inequity. Despite the central importance of climate change to Indigenous well-being, international decision-making bodies and processes have not centralized Indigenous voices (Nilsson 2008).

Indigenous peoples have affirmed the pivotal importance of Indigenous ways of knowing and being to climate action. Whyte (2016) has shown that Indigenous cultural and political systems have in-built flexibility that allows societies to constantly adapt to environmental changes. The International Indigenous Peoples' Forum on Climate Change ((IIPFCC) 2012), stated, 'we reiterate the need for recognition of our traditional knowledge, which we have sustainably used and practised for generations; and the need to

integrate such knowledge in global, national, and subnational efforts. This knowledge is our vital contribution to climate change adaptation and mitigation' (p. 1).

To Indigenous peoples, climate change represents an extension and intensification of colonialism (Whyte 2017). Activities and industries that are major producers of greenhouse gas emissions have depended on the dispossession and appropriation of Indigenous lands and resources (Cameron 2012). Indeed, the values, systems, structures, and processes foundational in the colonization of Indigenous peoples also underpin the environmental changes that threaten local and global ecosystems (Wood 2015). That these same systems are being relied on provide solutions for the climate crisis is of concern to Indigenous peoples, as these systems have historically been tightly bound to philosophies that see the planet in extractive and exploitative ways (Jones 2019).

Indigenous worldviews, especially relationality with the natural environment, challenge systems where human rights dominate. The time has arrived to mitigate unprecedented damage to our planet and solutions based on Indigenous sciences provide balance to current debate, processes, and interventions.

Conclusion

Describing the health of Indigenous peoples has many challenges including definitions, enumeration, and analysis. Making sense of health data is also challenging as we seek to recognize the complex, layered, and multifactorial influences that play out through time and across geographies for Indigenous peoples. This chapter canvasses the inequitable health outcomes for Indigenous peoples globally, shaped by shared experiences of dispossession and oppression, and the significant and ongoing health impacts resulting from differential exposure to social, political, economic, and cultural determinants of health, particularly colonialism and racism. These data also speak to the ongoing resistance of Indigenous peoples, and aspirations for the recognition of our rights and restoration of relationships with our territories, environments, and relations. The state of Indigenous health challenges us to have a critical view of colonial histories and how these have become embedded in our lives. We also recognize the potential

of Indigenous knowledge systems, ontologies, and values to drive positive health and well-being into the future.

REFERENCES

Abouzeid, M., et al. (2019). Time to tackle rheumatic heart disease: data needed to drive global policy dialogues. *Global Public Health*, 14(3), 456–68.

Allan, B. and Smylie, J. (2015). *First Peoples, Second Class Treatment: The Role of Racism in the Health and Well-being of Indigenous Peoples in Canada*. Toronto: Wellesley Institute.

Anderson, I., Crengle, S., Kamaka, M.L., Chen, T.H., Palafox, N., and Jackson-Pulver, L. (2006). Indigenous health in Australia, New Zealand, and the Pacific. *The Lancet*, 367(9524), 1775–85.

Anderson, I., et al. (2016). Indigenous and tribal peoples' health (The Lancet–Lowitja Institute Global Collaboration): a population study. *The Lancet*, 388(10040), 131–57.

Australian Institute of Health and Welfare (AIHW) (2014). *Coronary Heart Disease and Chronic Obstructive Pulmonary Disease in Indigenous Australians.* Cat. no. IHW 126. Canberra: AIHW.

Axelsson, P., Kukutai, T., and Kippen, R. (2016a). Indigenous wellbeing and colonisation [editorial]. *Journal of Northern Studies*, 10(2), 7–18.

Axelsson, P., Kukutai, T., and Kippen, R. (2016b). The field of Indigenous health and the role of colonisation and history. *Journal of Population Research*, **33**, 1–7.

Bacal, K., Jansen, P., and Smith, K. (2006). Developing cultural competency in accordance with the Health Practitioners Competence Assurance Act. *New Zealand Family Physician*, 33(5), 305.

Baddock, S.A., et al. (2017). Wahakura versus bassinet for safe infant sleep: a randomized trial. *Pediatrics*, 139(2), e20160162.

Baker, M.G., et al. (2012). Increasing incidence of serious infectious diseases and inequalities in New Zealand: a national epidemiological study. *The Lancet*, 379(9821), 1112–19.

Berger, L.R. (2002). Injury prevention and indigenous peoples. *Injury Prevention*, **8**, 175–6.

Betancourt, J., Green, A., Carrillo, J., and Ananeh-Firempong, O. (2003). Defining cultural competence: a practical framework for addressing racial/ethnic disparities in health and health care. *Public Health Reports*, 118(4), 419–38.

Betancourt, J.R., Green, A.R., Carrillo, J.E., and Park, E.R. (2005). Cultural competence and health care disparities: key perspectives and trends. *Health Affairs*, 24(2), 499–505.

Bird, M. (2002). Health and Indigenous people: recommendations for the next generation [editorial]. *American Journal of Public Health*, 92(9), 1391–2.

Bukhman, G. and Ochieng, F. (2017). Universal coverage for heart failure diagnosis and treatment: looking from sub-Saharan Africa to central Australia. *Journal of Cardiac Failure*, 23(10), 743–5.

Burgess, D., Fu, S., and van Ryn, M. (2004). Why do providers contribute to disparities and what can be done about it? *Journal of General Internal Medicine*, 19(11), 1154–9.

Cameron, E. S. (2012). Securing Indigenous politics: a critique of the vulnerability and adaptation approach to the human dimensions of climate change in the Canadian Arctic. *Global Environmental Change*, 22(1), 103–14.

Carapetis, J., McDonald, M., and Wilson, N. 2005. Acute rheumatic fever. *The Lancet*, 366(9480), 155–68.

Chang, A.B., Brown, B., Toombs, M., Marsh, R.L., and Redding, G.J. (2014). Lung disease in Indigenous children. *Paediatric Respiratory Reviews*, 15(4), 325–32.

Chang, A.B., Chang, C.C., O'Grady, K., and Torzillo, P.J. (2009). Lower respiratory tract infections. *Pediatric Clinics of North America*, 56(6), 1303–21.

Cooper, L.A., Beach, M.C., Johnson, R.L., and Inui, T.S. (2006). Delving below the surface. Understanding how race and ethnicity influence relationships in health care. *Journal of General Internal Medicine*, 21(Suppl 1), S21–7.

Costello, A., et al. (2009). Managing the health effects of climate change: Lancet and University College London Institute for Global Health Commission. *The Lancet*, 373(9676), 1693–733.

Cunningham, J., Rombold, A.R., Zhang, X., and Condon, J. (2008). Incidence, aetiology, and outcomes of cancer in Indigenous peoples in Australia. *The Lancet Oncology*, 9(60), 585–95.

Curtis, E. (2016). *Kohi Maramara. The Effect of Tertiary Recruitment, Admission, Bridging/Foundation Education and Retention on Indigenous Health Workforce Development*. (Unpublished Doctor of Medicine thesis). Auckland: University of Auckland,

Curtis, E., et al. (2014). What helps and hinders indigenous student success in higher education health programmes: a qualitative study using the critical incident technique. *Higher Education Research & Development*, 34(3), 486–500.

Daly, B., Raiman, I., and Goodson, J. (2017). Screening for diabetes in pregnancy in a regional area with a high Māori population. *New Zealand Medical Journal*, 130(1450), 25–31.

Davis, M. (2016). Data and the United Nations Declaration on the Rights of Indigenous Peoples. In: Kukutai, T. and Taylor, J. (eds.) *Indigenous Data Sovereignty: Toward an Agenda*. Canberra: ANU Press, pp. 25–38.

Di Cesare, M., et al. (2013). Inequalities in non-communicable diseases and effective responses. *The Lancet*, 381(9866), 585–97.

Espey, D.K., et al. (2014). Leading causes of death and all-cause mortality in American Indians and Alaska Natives. *American Journal of Public Health*, 104(Suppl 3), S303–11.

Finlay, S.M., and Armstrong, K. (2018). Indigenous NCDs: a call for States to include Indigenous peoples in non-communicable disease Declaration. [Online] Available at: https://croakey.org/indigenousncds-a-call-for-states-to-include-indigenous-peoples-in-non-communicable-disease-declaration/

Fitzgerald, C., and Hurst, S. (2017). Implicit bias in healthcare professionals: a systematic review. *BMC Medical Ethics*, 18(1), 19.

Gershon, A.S., et al. (2014). Asthma and chronic obstructive pulmonary disease (COPD) prevalence and health services use in Ontario Metis: a population-based cohort study. *PloS One*, 9(4), e95899.

Gracey, M., and King, M. (2009). Indigenous health part 1: determinants and disease patterns. *The Lancet*, 374(9683), 65–75.

Hayman, N., Reid, P.M.J., and King, M. (2015). Improving health outcomes for Indigenous peoples: what are the challenges? *Cochrane Database of Systematic Reviews*, 8, ED000104.

Jones, R. (2019). Climate change and Indigenous health promotion. *Global Health Promotion*, 26(3 suppl), 73–81.

Jones, R., Bennett, H., Keating, G., and Blaiklock, A. (2014). Climate change and the right to health for Māori in Aotearoa/New Zealand. *Health & Human Rights*, 16(1), 54–68.

Jowitt, L. (2016.) Gestational diabetes in New Zealand ethnic groups. *Integrative Molecular Medicine*, 3(2), 583–9.

Kaholokula, J.K., Nacapoy, A., and Dang, K. (2009). Social justice as a public health imperative for Kānaka Maoli. *AlterNative*, 5(2), 116–37.

King, M., Smith, A., and Gracy, M. (2009). Indigenous health part 2: the underlying causes of the health gap. *The Lancet*, 374(9683), 76–85.

Kumagai, A., and Lypson, M. (2009). Beyond cultural competence: critical consciousness, social justice, and multicultural education. *Academic Medicine*, 84(6), 782–7.

Lawrence, J.M., Contreras, R., Chen, W., and Sacks, D.A. (2008). Trends in the prevalence of preexisting diabetes and gestational diabetes mellitus among a racially/ethnically diverse population of pregnant women, 1999–2005. *Diabetes Care*, 31(5), 899–904.

LaVeist, T., and Nutu-Jeter, A. (2002). Is doctor-patient race concordance associated with greater satisfaction with care? *Journal of Health and Social Behavior*, 43(3), 296–306.

Levy, B.S., and Patz, J.A. (2015). Climate change, human rights, and social justice. *Annals of Global Health*, 81(3), 310–22.

Maskarinec, G.G., et al. (2014). Patient perspectives on the Hula Empowering Lifestyle Adaptation Study: benefits of dancing hula for cardiac rehabilitation. *Health Promotion Practice*, 16(1), 109–14.

McGrath, N.M., Evans, C., and Holgate, A. (2007). Post-partum follow-up of women with gestational diabetes mellitus from Northland, New Zealand. *Diabetic Medicine*, **24**, 218–19.

Mbugi, C.V. and Chilongola, J.O. (2010). Allergic diseases in Africa and Africans: is it primarily a priority? *World Allergy Organization Journal*, **3**, 197.

Ministry of Health (2015). *Tatau Kahukura: Māori Health Chart Book* (3rd ed.). Wellington: Ministry of Health.

Ministry of Health (2011). *Tatau Kura Tangata: Health of Older Māori Chart Book* (2nd ed.). Wellington: Ministry of Health.

Mohindra, K.S. (2017). Research and the health of indigenous populations in low- and middle-income countries. *Health Promotion International*, **32**, 581–6.

Montenegro, R.A. and Stephens, C. (2006). Indigenous health 2: indigenous health in Latin America and the Caribbean. *The Lancet*, **367**, 1859–69.

Moore, S.P., Forman, D., Piñeros, M., Fernández, S.M., de Oliveira Santos, M., and Bray, F. (2014). Cancer in indigenous people in Latin America and the Caribbean: a review. *Cancer Medicine*, 3(1), 70–80.

Naqshbandi, M., Harris, S.B., Esler, J.G., and Antwi-Nsiah, F. (2008). Global complication rates of type 2 diabetes in Indigenous peoples: a comprehensive review. *Diabetes Research and Clinical Practice*, 82(1), 1–17.

NCD Alliance (2012). *Health Inequalities and Indigenous People*. [Online] Available at: https://ncdalliance.org/

Nilsson, C. (2008) Climate change from an Indigenous perspective: key issues and challenges. *Indigenous Affairs*, 1–2, 8–15.

Nolan, C.J. (2011). Controversies in gestational diabetes. *Best Practice & Research: Clinical Obstetrics & Gynaecology*, 25(1), 37–49.

Oguamanam, C. (2004). Localizing intellectual property in the globalization epoch: the integration of Indigenous knowledge. *Indiana Journal of Global Legal Studies*, 11(2), 135–69.

Ohenjo, N., Willis, R., Jackson, D., Nettleton, Cl., Good, K., and Mugarura, B. (2006). Indigenous health 3: health of indigenous people in Africa. *The Lancet*, **367**, 1937–46.

O'Reilly, S. (2014). Prevention of diabetes after gestational diabetes: better translation of nutrition and lifestyle messages needed. *Healthcare*, 2(1), 468–91.

Osgood, N.D., Dyck, R.F., and Grassmann, W.K. (2011). The inter- and intragenerational impact of gestational diabetes on the epidemic of type 2 diabetes. *American Journal of Public Health*, 101(1), 173–9.

Paavola, J. (2017). Health impacts of climate change and health and social inequalities in the UK. *Environmental Health*, 16(Suppl 1), 113.

Pan American Health Organization (PAHO). (2018). *The Burden of Mental Disorders in the Region of the Americas*. Washington DC: PAHO.

Papps, E., and Ramsden, I. (1996). Cultural safety in nursing: the New Zealand experience. *International Journal for Quality in Health Care*, 8(5), 491–7.

Paradies, Y., Harris, R., and Anderson, I. (2008). *The Impact of Racism on Indigenous Health in Australia and Aotearoa: Towards a Research Agenda*. Casurina, NT: Cooperative Research Centre for Aboriginal Health.

Paradies, Y., Truong, M., and Priest, N. (2013). A systematic review of the extent and measurement of healthcare provider racism. *Journal of General Internal Medicine*, 29(2), 364–87.

Paradies, Y. (2016). Colonisation, racism and indigenous health. *Journal of Population Research*, 33(1), 83–96.

Paradies, Y. (2018). Racism and Indigenous health. In: McQueen, D. (ed.) *Oxford Research Encyclopedia of Global Public Health*. Oxford: Oxford University Press.

Parkin, D.M., Sitas, F., Chirenje, M., Stein, L., Abratt, R., and Wabinga, H. (2008). Part I: Cancer in Indigenous Africans—burden, distribution, and trends. *The Lancet Oncology*, 9(7), 683–92.

Patel, V., Flisher, A.J., Hetrick, S., and McGorry, P. (2007). Mental health of young people: a global public-health challenge. *The Lancet*, 369(9569), 1302–13.

Paulino, N.A., Vázquez, M.S., and Bolúmar, F. (2019). Indigenous language and inequitable maternal health care, Guatemala, Mexico, Peru and the Plurinational State of Bolivia. *Bulletin of the World Health Organization*, 97(1), 59.

Porter, C., Skinner, T., and Ellis, I. (2012). The current state of Indigenous and Aboriginal women with diabetes in pregnancy: a systematic review. *Diabetes Research and Clinical Practice*, 98(2), 209–25.

Ramos-Castillo, A., Castellanos, E.J., and Galloway McLean, K. (2017). Indigenous peoples, local communities and climate change mitigation. *Climatic Change*, 140(1), 1–4.

Ratima, M., et al. (2007). *Rauringa Raupa: Recruitment and Retention of Māori in the Health and Disability Workforce*. Auckland: Taupua Waiora, AUT University.

Reid, P., Paine, S.J., Curtis, E., Jones, R., Anderson, A., Willing, E., and Harwood, M. (2017). Achieving health equity in Aotearoa: strengthening responsiveness to Māori in health research. *The New Zealand Medical Journal*, 130(1465), 96–103.

Robbins, A. (2015). Health consequences of climate change interventions. *The Lancet*, 386(10006), 1819.

Robson, B., and Reid, P. (2001). *Ethnicity Matters. Review of the Measurement of Ethnicity in Official Statistics: Māori Perspectives Paper for Consultation*. Wellington: Te Rōpū Rangahau Hauora a Eru Pōmare

Royal C. (2009). *Let the World Speak: Towards Indigenous Epistemology*. Wellington: Mauriora ki te Ao/Living Universe Ltd.

Rumbold, A.R., et al. (2011). Delivery of maternal health care in Indigenous primary care services: baseline data for an ongoing quality improvement initiative. *BMC Pregnancy and Childbirth*, 11(1), 16.

Russell, N.K., Nazar, K., del Pino, S., Alonso Gonzalez, M., Díaz Bermúdez, X.P., and Ravasi, G. (2019). HIV, syphilis, and viral hepatitis among Latin American indigenous peoples and Afro-descendants: a systematic review. *Revista Panamericana de Salud Pública*, 43, e17.

Smedley, B., Stith, A., and Nelson, A. (2003). *Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care*. Washington DC: National Academies Press.

Smith, K.R., et al. (2014). Human health: impacts, adaptation, and co-benefits. In: Field, C., et al. (eds.). *Climate Change 2014: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press, pp. 709–54.

Smith, L.T. (2012). *Decolonizing Methodologies* (2nd ed.). New York: Zed Books.

Smylie, J. and Anderson, M. (2006). Understanding the health of Indigenous peoples in Canada: key methodological and conceptual challenges. *Canadian Medical Association Journal*, 175(6),

602.
Steer, A.C., Carapetis, J.R., Nolan, T.M., and Shann, F. (2002). Systematic review of rheumatic heart disease prevalence in children in developing countries: the role of environmental factors. *Journal of Paediatrics and Child Health*, 38(3), 229–34.
Te Rōpū Rangahau Hauora a Eru Pōmare. (2002). *Mana Whakamārama—Equal Explanatory Power: Māori and non-Māori sample size in national health surveys*. Wellington: Ministry of Health.
The International Indigenous Peoples' Forum on Climate Change (IIPFCC) (2012). *IIPFCC Statement to the SPI. 36th Session of the UNFCCC—Subsidiary Body for Implementation (SBI)*. Bonn: IIPFCC.
The Lancet (2016). No health workforce, no global health security. *The Lancet*, 387(10033), 2063.
Truong, M., Paradies, Y., and Priest, N. (2014). Interventions to improve cultural competency in healthcare: a systematic review of reviews. *BMC Health Services Research*, 14(1), 99.
United Nations (2007). United Nations Declaration on the Rights of Indigenous Peoples: resolution/adopted by the General Assembly, 2 October 2007. [Online] Available at: http://www.refworld.org/docid/471355a82.html
UNFPII (2004). *Report of the Workshop on Data Collection and Disaggregation for Indigenous Peoples*. [Online] Available at: https://digitallibrary.un.org/record/517063?ln=en
UN Permanent Forum on Indigenous Issues (UNPFII). (2015). *State of the World's Indigenous Peoples (Volume 2): Indigenous Peoples' Access to Health Services*. New York: United Nations.
UN Permanent Forum on Indigenous Issues (UNPFII) (2013). Permanent forum on Indigenous issues: report on the 12th session (20–31 May 2013). [Online] Available at: https://www.refworld.org/docid/51f8e9274.html
UN Permanent Forum on Indigenous Issues (UNPFII). (2009). *State of the World's Indigenous Peoples*. New York: United Nations.
van Ryn, M., et al. (2015). Medical school experiences associated with change in implicit racial bias among 3547 students: a medical student CHANGES study report. *Journal of General Internal Medicine*, 30(12), 1748–56.
Walker, J., Lovett, R., Kukutai, T., Jones, C., and Henry, D. (2017). Indigenous health data and the path to healing. *The Lancet*, 390(10107), 2022–3.
Walpole, S.C., Rasanathan, K., and Campbell-Lendrum, D. (2009). Natural and unnatural synergies: climate change policy and health equity. *Bulletin of the World Health Organization*, 87(10), 799–801.
Walter, M. (2018). The voice of indigenous data: beyond the markers of disadvantage. *Griffith Review*, **60**, 256.
Walter, M. (2016). Data politics and Indigenous representation in Australian statistics. In: Kukutai, T., and Taylor, J. (eds.). *Indigenous Data Sovereignty: Toward an Agenda*. Canberra: ANU Press, pp. 79–98.
Walters, K.L., et al. (2018). Growing from our roots: strategies for developing culturally grounded health promotion interventions in American Indian, Alaska Native, and Native Hawaiian communities. *Prevention Science*, 21(Suppl 1), 54–64.
Walters, K. (2016). Keynote address at the Lowitja Institute International Indigenous Health and Wellbeing Conference, 8–10 November 2016, Melbourne. [Online] Available at: https://conference2016.lowitja.org.au/photos-and-videos/videos/keynote-address-professor-karina-walters/
Walters, K.L., Mohammed, S.A., Evans-Campbell, T., Beltrán, R.E., Chae, D.H., and Duran, B. (2011). Bodies don't just tell stories, they tell histories: embodiment of historical trauma among

American Indians and Alaska Natives. *Du Bois Review: Social Science Research on Race*, 8(1), 179–89.
Watkins, D.A., et al. (2017). Global, regional, and national burden of rheumatic heart disease, 1990–2015. *New England Journal of Medicine*, 377(8), 13722.
Watts, N., et al. (2015). Health and climate change: policy responses to protect public health. *The Lancet*, 386(10006), 1861–914.
Watts, N., et al. (2018). The Lancet Countdown on health and climate change: from 25 years of inaction to a global transformation for public health. *The Lancet*, 391(10120), 581–630.
Weil, E.J., and Nelson, R.G. (2006). Kidney disease among the indigenous peoples of Oceania. *Ethnicity & Disease*, 16(2 Suppl 2), S2–24.
Whyte, K.P. (2016). Is it colonial déjà vu? Indigenous peoples and climate injustice. In: Adamson, J. and Davis, M. (eds.) *Humanities for the Environment: Integrating Knowledge, Forging New Constellations of Practice*. London & New York: Taylor & Francis, pp. 88–105.
Whyte, K. (2017). Indigenous climate change studies: indigenizing futures, decolonizing the Anthropocene. *English Language Notes*, 55(1), 153–62.
Wirihana, R. and Smith, C. (2014). Historical trauma, healing and well-being in Māori communities. *MAI Journal*, 3(3), 197–210.
Woods, J.A., Katzenellenbogen, J.M., Davidson, P.M., and Thompson, S.C. (2012). Heart failure among Indigenous Australians: a systematic review. *BMC Cardiovascular Disorders*, **12**, 99.
Wood, L. (2015). The environmental impacts of colonialism. In: *BSU Honors Program Theses and Projects*. Item 119. Departmental Honors in Geography, Bridgewater State University. [Online] Available at: http://vc.bridgew.edu/honors_proj/119
Woodward, A. and Porter, J.R. (2016). Food, hunger, health, and climate change. *The Lancet*, 387(10031), 1886–7.
World Health Organization (WHO) (2013). *Global Action Plan for the Prevention and Control of Noncommunicable Diseases 2013–2020*. Geneva: WHO.
World Health Organization (WHO) (2019). *Cancer*. [Online] Available at: https://www.who.int/news-room/fact-sheets/detail/cancer
World Health Organization (WHO) (2014). *Preventing Suicide: A Global Imperative*. [Online] Available at: https://www.who.int/mental_health/suicide-prevention/exe_summary_english.pdf?ua=1

10.7

People with disabilities

Anne Kavanagh, Marissa Shields, and Alex Devine

> '*Disability doesn't make you exceptional, but questioning what you think you know about it does', Stella Young, 2014*
>
> (Palenzuela 2014)

Introduction

The health of people with disabilities has, until recently, been largely ignored by public health researchers, practitioners, and policymakers. We argue this is because public health largely concentrates on the prevention of disability and conceives of disability as a health deficit. This chapter will show why the health of people with disabilities is a significant public health issue that requires urgent attention.

People with disabilities are a substantial minority of the world's population. Approximately 15% of the world's population has a disability; that's over one billion people. Globally, people with disabilities are one of the most disadvantaged populations. They are more likely to have low levels of education, be unemployed, socially excluded, live in poverty, experience homelessness, and live in poor-quality housing (OECD 2010; WHO 2011). These disadvantaged circumstances impact on their physical and mental health. In many countries, in part due to their exclusion from public health and broader society, people with disabilities smoke at higher levels than their peers without disabilities (Courtney-Long et al. 2014; Australian Institute of Health and Welfare 2016b), are more likely to be overweight and physically inactive (Neter et al. 2011; Kim et al. 2015; Australian Institute of Health and Welfare 2016b).

People with disabilities are also more likely to experience discrimination, bullying, and physical and sexual violence (Hughes et al. 2012; Krnjacki et

al. 2016; Krnjacki et al. 2017; Pinquart 2017). This is further compounded by difficulty accessing quality health promotion and care. Remedying the disadvantage, discrimination, violence, risky health behaviours and poor access to healthcare and disability services people with disabilities experience will lead to improvements in their health. This means that much of the poorer health of people with disabilities is avoidable rather than a consequence of their underlying impairment.

In this chapter, we describe how disability is conceptualized and measured; what is known about the prevalence of disability internationally; the health outcomes for people with disabilities and the impact of the social determinants of health. We also discuss some of the extant literature on improving health outcomes for people with disabilities. The chapter concludes with a call for people with disabilities to be a priority population for public health and outlines an agenda for future public health research, policy, and practice.

Evolving conceptualizations of disability

Disability is a complex and contested concept. The United Nations Convention on the Rights of Persons with Disabilities (CRPD) describes disability as resulting from:

> *'the interaction between persons with impairments and attitudinal and environmental barriers that hinders their full and effective participation in society on an equal basis with others.'* (United Nations 2006, Article 1)

The importance of this definition is that it distinguishes between an impairment, such as paraplegia or vision loss, and barriers to participation such as discrimination, physical inaccessibility, and limited employment opportunities. The disability is not solely caused by an impairment *per se* but rather by the presence of the impairment and other characteristics of society that prevent, or make it difficult for, people with disabilities to participate in key life domains, attain their rights, and realize their aspirations.

The CRPD is the first human rights treaty to specifically define the rights of people with disabilities and outline state obligations to promote, protect and uphold these rights (United Nations 2006; Keogh and Down 2014). Among these rights is the recognition that people with disabilities have the

right ‘to the enjoyment of the highest attainable standard of health without discrimination on the basis of disability’ (United Nations 2006, Article 25).

Conceptualization of disability within public health

Public health researchers, practitioners, and policymakers have traditionally focused on the prevention of disability and ill-health through policy and programmes which reduce the risk of acquiring disability. Examples include seatbelts and speed limits to reduce road accident injuries, immunization, and smoking taxation. Prevention of injury and ill-health, alongside promoting health and well-being is central to public health and should remain so. Like the rest of the population, people with disabilities can benefit from such public health interventions. However, in most public health programmes disability is considered a ‘failure’ of public health. Further, interventions which work among the general population are often assumed to apply to and be effective for people with disabilities. This is despite limited evidence or knowledge on appropriate research designs which could measure the benefit of mainstream public health interventions for people with disabilities (Berghs et al. 2016).

Antiquated assumptions that conceptualize disability as a burden or deficit can reinforce this exclusion from public health. Burden of disease studies for example position disability as a health deficit. This means that years lived with disability are valued far less than years lived without disability. This positioning assumes that people with disabilities cannot be healthy, and, when used for resource allocation or cost-effectiveness analyses, emphasizes that the lives of people with disabilities are less effective at the outset (Arnesen and Nord 1999). Effectively, this means that an intervention targeted at women with disabilities such as mammographic screening would be considered less cost-effective than in the general population because their lives are assigned lesser value.

Addressing health deficits and disability aligns broadly with what is often referred to as the medical model of disability. The literature distinguishes between the medical (or individual) model of disability and the social model. Proponents of the social model portray the ‘medical model’ as conceiving of disability as a deficit in individual bodies that is addressed through prevention, medical treatments and, preferably, cured (Oliver 1981; 1990; 1996a; b). The social model, positioned as a radical departure from this

approach, identifies physical, social, cultural, and economic barriers that are disabling. The social model draws attention to the ways in which governments, organizations, workplaces, and communities can respond to overcome barriers to an individual's participation in society.

The social model has been important politically and widely embraced by advocacy groups as it calls government and societies to account and has helped pave the way towards the formal recognition of the rights of people with disabilities as articulated in the CRPD. Others have drawn attention to the shortcomings of both models, pointing out that while the social model is an important counterpoint to the medical model, it ignores the fact that individual bodies have limitations, experience pain, fatigue and that can indeed impact on participation (Shakespeare 2014).

The International Classification for Functioning, Disability, and Health

The International Classification for Functioning, Disability, and Health (ICF), developed by the World Health Organization (WHO 2002), is one of the most widely used frameworks for conceptualizing disability. The framework is illustrated in Figure 10.7.1. This framework is designed to guide measurement, research, service delivery, and policy development. It distinguishes between functional limitations related to an impairment affecting body functions and structures (e.g. loss of arm function), activity (e.g. execution of a task or action such as dressing) and participation in life (e.g. employment).

The ICF emphasizes that acting on the impairment through medical treatment alone is not sufficient. Society can reduce activity restrictions through technologies and other adaptions (e.g. screen readers, wheelchairs, hearing loops) and reducing participation restrictions (e.g. by reducing discrimination, referred to as environmental factors in the ICF). These domains are classified as bodily, individual, and societal perspectives in the ICF, respectively. In the ICF the 'functioning' refers to all body functions, activities, and participation, while disability is an umbrella term for impairments, activity limitations, and participation restrictions. The ICF also lists environmental (e.g. discrimination, policies) and personal (e.g. age) factors that interact with all these components.

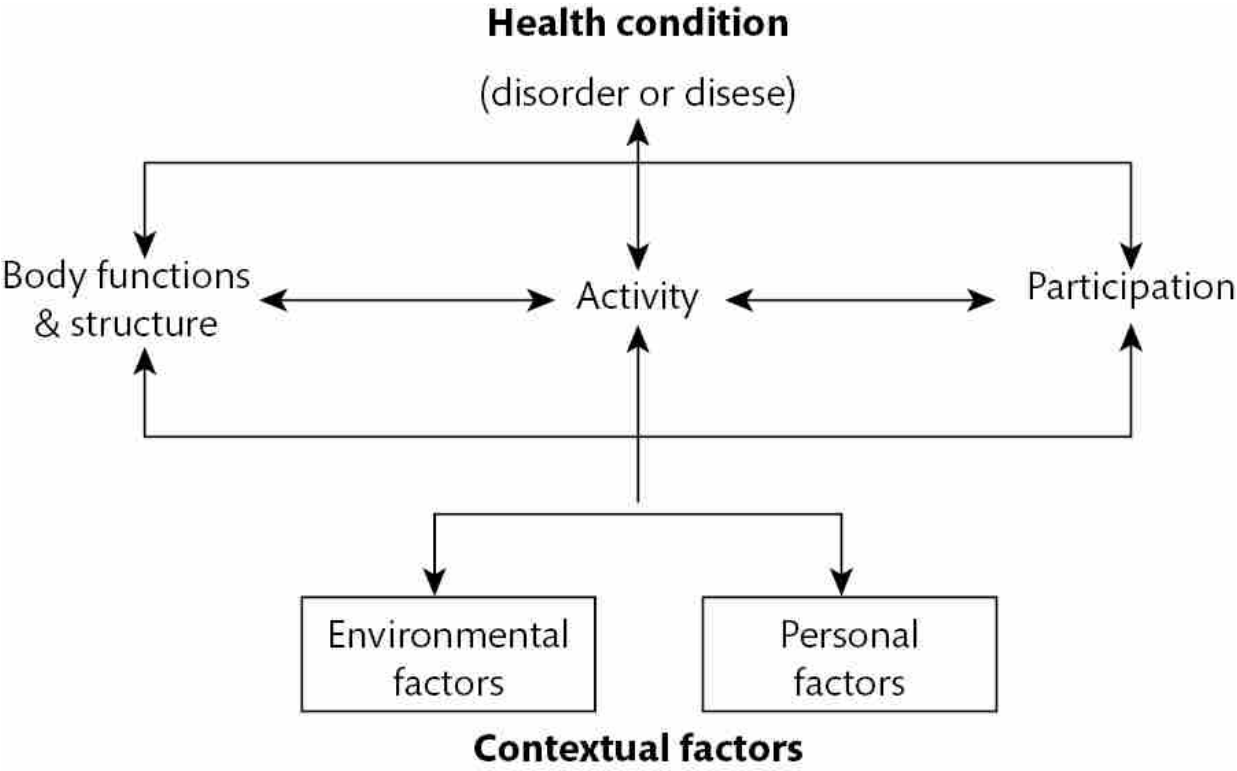


Fig. 10.7.1 Disability framework for the International Classification for Functioning.
WHO (2002) Towards a common language for functioning, disability, and health: The international classification of functioning, disability and health. Geneva: World Health Organization. p.9).

Measurement of disability

A recent scoping review of research using the ICF found that the ICF has been widely adopted in research and has led to better quality statistics on disability. Of particular note is the important role the ICF has played in monitoring community-based rehabilitation in low-resource settings and its use in policy and programme design and evaluation (e.g. education and social policy). Many countries have used the ICF in new surveys of disability, in particular conceptualizing the activity and participation domains of functioning. However, the authors of the review note that environmental factors have not been well conceptualized or measured (Madden and Bundy 2019).

Despite the widespread acceptance of ICF as a useful model for the collection of data on disability, it has been operationalized in many different

ways, resulting in challenges in the interpretation of disability statistics (Madans et al. 2015; Sabariego et al. 2015). For example, at an international level the Washington Group has developed a short and extended set of questions that can be used in censuses and surveys (see http://www.washingtongroup-disability.com/about/) (Madans et al. 2011; Madans and Loeb 2013). Over 30 countries have used the short set in their census, with the questions now globally endorsed for use in monitoring the Sustainable Development Goals. In parallel, the WHO developed the 'Model Disability Survey' (MDS). The MDS is considered a very comprehensive survey with approximately 320 questions and has been piloted in a small number of countries to allow for a more in-depth understanding of disability (WHO 2017a).

At a country level, a number of countries have developed instruments that broadly capture functioning in an ICF framework. For example, Ireland has developed a National Disability Survey which captures all elements of the ICF, including environmental factors which affect functioning, such as attitudes and flexible working arrangements (Kostanjsek et al. 2013). In Australia, the Australian Bureau of Statistics conducts a comprehensive survey of disability every three years—the Survey of Disability, Ageing, and Carers. This survey includes over 150 questions and provides a comprehensive picture of all aspects of the ICF. They have also produced a short module on disability which is used in a number of other surveys they administer (Harper 2010).

Prevalence of disability

Comparisons of different surveys designed to collect disability prevalence reveal markedly different prevalence estimates and have demonstrated that the concordance between surveys is poor (Mont 2007). For example, in Australia the prevalence of disability from the Survey of Disability, Ageing, and Carers was 18.3% compared to 6.7% when disability was measured using the Washington Group questions (Australian Bureau of Statistics 2016b). A Norwegian study investigated the prevalence of disability and the concordance based on eight different instruments applied to the same sample. They found that prevalence varied between 10 and 25% with instruments with more questions tending to yield higher estimates (Molden and Tossebro 2010).

The varying prevalence estimates, and the lack of comparability between the different instruments used to measure disability, presents major challenges for monitoring and evaluation. There is a need for methodological research that seeks to harmonize the different approaches so that disability is consistently captured.

Figure 10.7.2 shows the prevalence of disability across a selected range of countries that have measured disability using census or survey data. As discussed earlier, the Survey of Disability, and Carers in Australia yields a high prevalence of 18.5%. However, many other countries did not use such a comprehensive set of questions and prevalence was less than 2% in Togo, Iran, Philippines, and Nepal. These low prevalence estimates were ascertained through a single question on the country's census, which asked if the respondent or a person in the household had a disability, however the exact question varied between countries. For example, in Iran the respondent to the census was asked '*Is there anybody in the household with at least one of the following disabilities?*' with nine different disability types to select from. In Togo the respondent was asked for each household member '*Does [person] have a handicap?'* Until more consistent approaches to disability measurement are applied across countries it will be very difficult to make meaningful comparisons between countries.

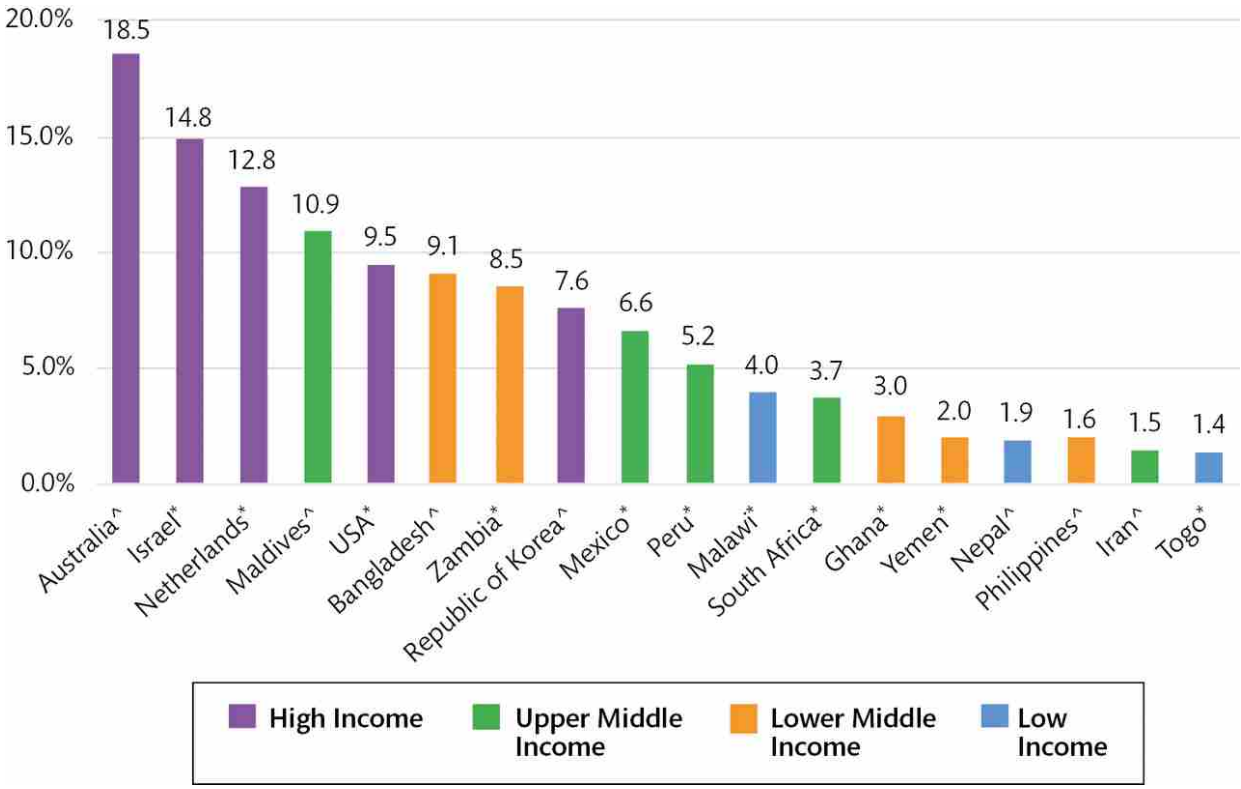


Fig. 10.7.2 International prevalence of disability across selected countries.
* Source data from Loeb (2016, Appendix 1). ^ Source: data from UNESCAP (2015, p.4).

The health of people with disabilities

People with disabilities generally experience poorer physical and mental health outcomes and have a higher prevalence of health conditions than the general population. People with disabilities are often at higher risk of experiencing secondary health conditions, compounding existing health conditions and disability, as well as increasing the likelihood of experiencing comorbidities such as diabetes and other non-communicable diseases (WHO 2011).

While it is clear that people with disabilities experience poorer physical and mental health outcomes compared to people without disabilities, there is uncertainty as to the true prevalence of health conditions among people with disabilities due to poor-quality research and data. For example, a systematic review on the prevalence of diabetes among people with intellectual disabilities found that the prevalence ranged from 0.4% to 25% across studies from Europe, North America, New Zealand, Australia, China, and

Hong Kong (MacRae et al. 2015). Similarly, a systematic review of mental health conditions among people with intellectual disabilities found the prevalence ranged from between 14% to 75%, again underscoring the challenges in ascertaining true and comparable prevalence estimates (Buckles et al. 2013).

Despite concerns regarding prevalence measurement of particular conditions, people with disabilities consistently report poorer self-rated health and increased psychological distress. Over half of adults with severe or profound disabilities in Australia reported fair or poor health, compared to only 6% of people without disabilities. Prevalence of psychological distress was also elevated in this group, with 34% experiencing high or very high distress. Only 5% of the population without disabilities experience such high levels of distress, clearly indicating a disparity in mental health (Australian Institute of Health and Welfare 2016b). These findings were replicated in the United States, as researchers found the number of conditions people experienced was directly associated with increased feelings of psychological distress (Okoro et al. 2014). While these perceptions of health and well-being are important in their own right, poor psychological well-being is associated with a higher prevalence of physical health conditions among people with disabilities, such as coronary heart disease, diabetes, and hypertension (Okoro et al. 2014).

Health outcomes are also related to health seeking behaviour, which in turn is influenced by disability. In a population-based case-control study in Cameroon and India, people with disabilities were significantly more likely to report significant health problems compared to people without disabilities, yet delayed seeking healthcare. The researchers found that awareness of and access to rehabilitation services were extremely low among people with disabilities (Mactaggart et al. 2016).

Health outcomes among specific groups of people with disabilities

Health outcomes within population of people with disabilities can vary however, with some population groups experiencing poorer health outcomes than others, depending on the type of disability, severity of disability, and geographic and cultural context. Adults with severe or profound disabilities in Australia for example are over three times as likely to have three or more

health conditions, such as back problems, arthritis, and cardiovascular disease, as compared to people without disabilities (Australian Institute of Health and Welfare 2016b). Individuals with particular disability types, such as autism spectrum disorder (ASD), or intellectual disability, have also been shown to have poorer physical health outcomes relative to the general population. A review of international literature found people with ASD had elevated levels of cardiovascular risk factors and more multimorbidities compared to people without ASD (Cashin et al. 2018).

People with intellectual disabilities in Scotland were also more likely to have poorer physical health outcomes, with only 39% of individuals reporting no physical health conditions. Overall, people with intellectual disabilities had a higher prevalence of 14 conditions, such as epilepsy, visual impairment, and hearing loss, in comparison with the general population (Cooper et al. 2015). These comorbidities are likely to become even more prevalent as individuals with disabilities age, as shown by research from the Netherlands wherein 80% of adults with intellectual disabilities over the age of 50 reported two or more physical health conditions, compared to 56% of community-dwelling older adults without disabilities (Hermans and Evenhuis 2014).

As highlighted by a national survey of psychosis in Australia, people with psychosocial disabilities experience higher risk of diabetes, metabolic disorders, heart disease, and obesity. They also have a significantly elevated mortality rate from all these conditions, compared with the general population. Their health outcomes are also impacted upon by the side effects of medication, as well as higher prevalence of drug and alcohol misuse which may aggravate mental health conditions (Morgan et al. 2012; National Mental Health Consumer and Carer Forum 2011; Simonsen et al. 2013). Compounding this is the insufficient resourcing of most mental health systems, and the fact that health practitioners do not always have the skills or resources to address the co-occurring health needs of people with disabilities.

Not only are people with psychosocial disabilities more likely to experience the ill-effects of co-occurring health conditions, but people with other types of disabilities are disproportionately impacted by certain co-occurring mental health conditions. Approximately 40% of mothers with disabilities in the United States for example, experienced depression prior to their pregnancy, and nearly one-third of women with disabilities reported

symptoms of post-partum depression following birth, compared to 16% of women without disabilities (Mitra et al. 2015). Adults with intellectual disabilities in Scotland have a sevenfold increase in the odds of schizophrenia and are twice as likely to have anxiety. Fifteen percent of people with intellectual disabilities were found to have depression, compared to 10% of the general population (Cooper et al. 2015). More generally, half of people with severe or profound disabilities in Australia had at least one mental health condition, compared to 8% of individuals without disability or restrictive long-term health conditions (Australian Institute of Health and Welfare 2016b).

Increased likelihood of engaging in risky health behaviours

People with disabilities are more likely to engage in some risky health behaviours, which in turn, contribute to poor health outcomes. Important health behaviours impacting people with disabilities include tobacco use, alcohol and substance misuse, and physical inactivity.

Smoking

Internationally, people with disabilities have higher rates of tobacco use compared to the general population. Across all disability types in the United States, 40% of men and 38% of women with disabilities aged 18–49 were regular smokers, while only 21% of adults without disabilities smoked regularly (Courtney-Long et al. 2014). Patterns are similar in Australia, where 30% of adults with severe or profound disabilities smoke; twice that of the general population (Australian Institute of Health and Welfare 2016b). People with disabilities are also less likely to report trying to quit smoking. In the Republic of Korea for example, 55% of adults without disabilities who smoked regularly made a past-year quit attempt, while only 45% of individuals with disabilities did (Lee et al. 2014b). Less than 20% of people with mobility impairments in the United States attempted to quit in the last year, compared to 27% in the general population (Borrelli et al. 2014).

Alcohol use

By contrast, people with disabilities appear to consume less alcohol than the general population (Glazier and Kling 2013; Pinquart and Pfeiffer 2015;

Australian Institute of Health and Welfare 2016b; Mitra et al. 2016). However, people with disabilities have an increased use of other substances. In the United States, the prevalence of any substance misuse in the past month, including prescription or illicit drugs, binge alcohol, or habitual cigarette smoking, is 40% for people with disabilities, and 34% for people without disabilities.

Substance misuse

Individual with disabilities in the United States have higher odds for using substances such as heroin, oxycodone, sedatives, and methamphetamine (Glazier and Kling 2013). Research into substance misuse has been less indicative for people with specific disability types, such as intellectual disabilities. While evidence from the United States suggests the overall prevalence of substance misuse is low among people with intellectual disabilities, the risk of developing a substance-related problem is comparatively high (Carroll Chapman 2013). As lower cognitive function is associated with increased drop-out from treatment (Brorson et al. 2013), issues relating to substance abuse may be particularly challenging to measure and treat among people with intellectual disabilities, as well as other disability groups.

When asked about substance misuse and treatment, people with physical disabilities in the United States were less likely to have received helpful treatment for substance misuse issues, with 50% reporting useful treatment compared to 57% of the general population (Anderson et al. 2014).

Physical inactivity

Participation in physical activity is an additional key concern for people with disabilities, as many individuals have very limited support or opportunity to engage in physical activity. Across all people 15–64 years with severe or profound disabilities in Australia for example, 46% reported no past-year physical activity. By comparison, 31% of people without disability reported no physical exercise (Australian Institute of Health and Welfare 2016b). Similarly, 30% of adults with disabilities 20 years and older in the Republic of Korea did not participate in physical activity within the last year (Kim et al. 2015). Rates of physical activity are especially poor for people with particular types of disability, such as people with intellectual disabilities.

Only 9% of adults with intellectual disabilities met minimum physical activity guidelines, according to a recent systematic review (Dairo et al. 2016). This pattern extends to children with disabilities, who also have lower rates of physical activity. For example, 70% of children with disabilities in the Netherlands did not participate in organized sports, and 80% played sports either not at all, or for less than 1 hour per week (Neter et al. 2011).

Disability and the social determinants of health

So why are the health behaviours and outcomes of people with disabilities so much poorer than people without disabilities? This can be answered in part by exploring the social determinants of health for people with disabilities. As discussed, disability is not simply the result of a particular impairment or condition. Rather, the experience of disability is largely determined by society and the environment. The social determinants of health reflect this by illuminating external factors which impact the circumstances, opportunities, and perception of a particular person (Emerson et al. 2011). Social determinants, such as violence, poverty, education and employment, and stigma and discrimination are crucial in understanding the experiences and health outcomes of people with disabilities.

It is likewise imperative to understand the processes by which people with disabilities are excluded from health systems, including public health, to the detriment of their health, well-being, and disability. International research continues to highlight that people with disabilities experience ongoing inequalities in relation to socioeconomic status, employment, educational attainments, lack of access to accessible and appropriate housing and surrounding environments, and restricted access to services (Groce et al. 2011; WHO 2011; Byrne 2018). These inequalities play a major role in influencing health behaviours and outcomes for people with disabilities.

Emerson et al. (2011) offer a useful explanation of how the relationship between disability and the social determinants of health are often established at an early age and differ across the life course. For young children, the most relevant social determinants often relate to ‘material and psychosocial’ hazards (e.g. poor housing, poor diet, economic insecurity, family violence, stressful life events, lack of supportive relationships with adults). As children age, these social determinants may be exacerbated by the costs of caring for a child with disabilities. Adults with disabilities face exclusion

from the labour force which may exacerbate existing social and economic challenges. While the mechanisms for these associations are not fully understood, it is clear that children and adults with disabilities are at increased risk of exposure to poorer living and working conditions and greater economic insecurity, which are associated with poorer health outcomes (Emerson et al. 2011).

Poverty and disability

Globally, people with disabilities are more likely to live in poverty compared to people without disabilities, and the situation appears to be worsening. In the United States, the overall poverty rate has increased for all individuals, but most dramatically for people with disabilities. Twenty-five per cent of people with disabilities were in poverty in 1981, rising to 32% in 2014 (Brucker and Houtenville 2015).

The link between poverty and disability is also evident in low and middle-income countries (LMICs), where a systematic review found over 80% of included studies identified a positive relationship between disability and poverty. Notably, the severity of disability was associated with the severity of economic poverty. The authors found the observed relationship remained despite adjustment for age, gender, area of residence, and education. This trend was consistent throughout regions across the globe, impairment types, study designs, and age groups (Banks et al. 2017).

Poverty among people with disabilities in LMICs has persisted despite many countries making gains in addressing poverty more broadly. In Cambodia for example, the country has made significant progress in addressing poverty through sustained economic growth and the implementation of social protection programmes. Poverty rates have declined over the past decade from estimates of 47% of the population in 2007, to current estimates of 13.5% (National Institute of Statistics et al. 2015; World Bank 2018a; b). However, reductions in poverty are not equal across households with and without family members with disabilities.

This situation was highlighted in a study analysing the Cambodian Socio-Economic Survey data from 2009 to 2014 (National Institute of Statistics et al. 2015). Analysis found that the poverty rate among households with members with disabilities was 18%, compared to 13.6% for households without any members with disabilities. When the cost of disability incurred

by households was included in the analysis, the poverty rate for households with a member with disabilities increased to 34%. An additional 17% of household income was required in order to achieve the same standard of living as similar households without disabilities (Palmer et al. 2016). The limited household resources of families who have a family member with a disability make it more difficult to meet the costs of accessing services, such as health and education (UNESCAP 2015).

Poverty, health, and disability are clearly inter-related. In the UK, it has been found that the greater risk of exposure to socioeconomic disadvantage may constitute 20–50% of the risk of decreased physical and mental health among children with intellectual disabilities (Emerson et al. 2011). Similarly, a cross-sectional survey of English men and women with mild and moderate intellectual disabilities found that self-rated health was strongly associated with material hardship, with individuals who had experienced material hardship in the past year more likely to report 'not good health' (Emerson and Hatton 2008). Most high income countries have income support for people with disability who are unable to work however many low-to-middle-income countries do not have such a safety net putting people with disability at even higher risk of poverty and ill-health (WHO 2011).

Access to nutrition

People with disabilities, especially children, experience poorer nutritional status than people without disabilities. Malnutrition and disability are intimately linked—with malnutrition contributing to the risk of impairment, and children with disabilities thought to be more at risk of experiencing malnutrition. The risk of malnutrition is often assumed to be linked to difficulty with feeding, or that children with disabilities are more likely to be unwell which impacts on their ability to eat. While these are important factors, a study conducted in Kenya in 2013 found malnutrition was also related to neglect, lack of awareness of the need for greater nutritional intake that some children with disabilities have, and intentional underfeeding (Groce et al. 2014; International Centre for Evidence in Disability et al. 2014). Identifying malnutrition in children with disabilities can also be challenging given certain impairments may impact on physical growth. As such it can be difficult to accurately use traditional malnutrition assessments such as measuring the mid-upper arm circumference. This may hinder the

identification of malnutrition among children with disabilities and reduce the likelihood of timely and adequate treatment for malnutrition.

Possible outcomes of poor diet and malnutrition include poorer health outcomes, missing or delay in developmental milestones, acquiring avoidable secondary conditions, stunting and wasting, and preventable death (Groce et al. 2014; International Centre for Evidence in Disability et al. 2014). Researchers in Iran found that more than 40% of girls and boys with disabilities assessed were underweight, while stunting was present in 42% of children with disabilities in the study. Notably, only 6% of children without disabilities were underweight. Further, 90% of the children with disabilities did not receive adequate vitamins and minerals, such as iron and calcium (Neyestani et al. 2010). In the Netherlands, children with disabilities were less likely to have breakfast, and were four and a half times less likely to eat fruit on all weekdays (Neter et al. 2011). Children with intellectual disabilities in France were also found to have poorer diets, with nearly half consuming soft drinks with meals at school, compared to only a quarter of students without disabilities (Mikulovic et al. 2011). Young people with intellectual disabilities in Taiwan also had a poorer diet, eating fewer fresh fruits and vegetables than their peers without disabilities (Yen and Lin 2010).

Poorer diet and lower rates of physical activity contribute to the higher prevalence of obesity and being overweight observed among adults and children with disabilities, particularly in high income countries. Forty-five per cent (45%) of adults with upper or lower limb disabilities in the Republic of Korea for example are reported to be obese (Oh et al. 2012), and nearly 80% of adults aged 18–65 with intellectual disabilities in the United States are reported to be overweight or obese (Barnes et al. 2013). Among school children in Northern Ireland, 33% of students with intellectual disabilities were overweight or obese, compared to 24% of their peers without disabilities (Slevin et al. 2014). In the Netherlands, 31% of children with disabilities were overweight, compared to 10% of children without disabilities. The difference in the prevalence of obesity was likewise striking, as 11% of children with disabilities were obese, compared to only 2% of children without disabilities (Neter et al. 2011).

Exclusion from education and work

People with disabilities in both LMICs and high-income countries (HICs) experience many barriers to education and to gaining and maintaining work (Saleh and Bruyere 2018). In relation to education, these inequalities arise due to negative attitudes or low expectations of the benefits of education for children with disabilities, the absence of necessary adjustments and accommodations, and limited access to inclusive education. In terms of employment, low expectations about the capacity of people with disabilities within the workplace, restricted access to important career development opportunities such as training and skills development, insufficient support to gain and maintain work, and precarious engagement with labour market (e.g. underemployment, insecure employment, exploitation) are major barriers (Saleh and Bruyere 2018).

People with disabilities in Guatemala for example, expressed their frustration at the unattainability of education, a result of poverty, social invisibility and stigma, and structural barriers to access. Lack of education was perceived as a primary barrier to alleviating poverty (Grech 2014). Alternatively, evidence from Nepal, Bangladesh, Cambodia, and the Philippines suggest that longer years of education significantly increase the opportunity for people with disabilities to get a job and engage in full time, white collar jobs, reducing the risk of poverty, improving participation and access to better health (Lamichhane 2015).

From the employment statistics of people with disabilities in the United States, it is clear that barriers to employment are still present. Full time, full year employment for people with disabilities has decreased markedly over time, from 12% in 1981 to less than 6% in 2014. For those employed, median earnings are $10,000 USD less for people with disabilities (Brucker and Houtenville 2015).

Similarly, of the 18.3% or 4.2 million Australians living with disabilities, only 53.4% are currently in the labour force (e.g. employed or looking for work). This is compared to 83.2% of Australians without disabilities. The unemployment rate for Australians with disabilities is 10%, nearly double the rate of that for Australians without disabilities (5.3%) (Australian Bureau of Statistics 2016a). Stigma and discrimination are recognized as one of the major barriers to improving employment outcomes for Australians with disabilities (Australian Human Rights Commission 2016). When people with disabilities are employed, they are often in insecure work, have poorer working conditions, less perceived fairness of pay and are more likely to be

underemployed (LaMontagne et al. 2016; Milner et al. 2015; Milner et al. 2017).

Likewise in the UK, adults with intellectual disabilities are more likely to be in non-standard employment and experience job insecurity. They are also more likely to transition into economic inactivity from standard employment (Emerson et al. 2018). Stigma and structural barriers also constituted major barriers to employment in Brazil. People with disabilities who did gain employment were typically paid less and given less skilled work, as their abilities were persistently doubted (Toldra and Santos 2013).

Additionally, people with disabilities struggle to enter the workplace, as shown by a study of young people with disabilities in Canada. Overwhelmingly, young people with disabilities had significantly poorer work and job search experiences. Eighty-three per cent of young people with learning or memory impairments were refused an interview, and 63% of young adults with mobility impairments were denied workplace accommodations. 77% of young people with learning or memory impairments reported discrimination at work due to their disability (Lindsay 2011).

Using longitudinal data from the US, the authors found that a third of young people with ASD had attended college (34.7%) and just over half had been in paid employment within the first 6 years after high school (55.1%). Over half of youth with ASD who had left high school in the past two years had no participation in employment or education. The authors found that young people with ASD had the lowest rate of employment participation and the highest rates of no participation as compared to young people with other disabilities (Shattuck et al. 2012).

Housing and the surrounding environment

Accessible, affordable, and secure housing are key determinants of health (Australian Institute of Health and Welfare 2016a). Conversely, reduced access to safe and accessible housing disproportionately impacts on the social and health outcomes of people with disabilities (Beer and Faulkner 2009). Researchers from Victoria, Australia found that people with disabilities were substantially less likely to own a home and were much more likely to experience housing stress and live in rental accommodation than the general public. People with disabilities may be additionally

hampered in their access to safe and accessible housing due to restrictions imposed by availability of transportation systems and access to services. Housing that is safe, accessible, and close to transportation and services may require people with disabilities to move away from family and friends, thereby leading to isolation from support networks (Beer and Faulkner 2009).

Community living is known to be essential for social inclusion and a key element of quality of life. Deinstitutionalisation has continued to occur in most countries, but this has not been sufficiently replaced with resources to support independent living within the community for people with disabilities. As such, adequate housing within the community remains out of reach for many people with disabilities, particularly for people with intellectual and psychosocial disabilities (Siska et al. 2018).

The surrounding environment to where people live and work is also important. A qualitative survey of adults with a variety of different disabilities and functional limitations in the US analysed the environmental factors influencing participation at the individual community, and societal levels. The eight major categories of factors were: built environment, natural environment, transportation, assistive technology, IT access, social support and societal attitudes, systems and policies, and economic environment. Particularly poignant was one participant's description of the immense process and organization that facilitates participation, wherein having physical access (such as a ramp) is not enough when one must consider the weather, if the personal assistant will come, if sidewalks and ramps are clear outdoors, and if transportation will arrive. This work highlights the interconnected nature of environmental barriers on health and participation. Individuals expressed a desire to work but were concerned that they would lose their disability pension, and if they lost their job, they would have no income at all. Individuals discussed challenges with negative, disabling societal attitudes, and of the importance of social facilitators such as a supportive family. The researchers emphasize that barriers can, and must, be addressed at the societal, community, and individual levels in order to positively influence participation (Hammel et al. 2015).

Violence

Globally, people with disabilities are disproportionately affected by violence. Despite this, violence prevention and response programmes all too often have neglected the needs or excluded people with disabilities, or do not understand how to accommodate or respond to the people with disabilities who have experienced violence (Devine et al. 2017). Twenty-eight per cent of people with disabilities in the United States have experienced intimate partner violence (Mitra and Mouradian 2014), and evidence from a systematic review of violence found people with disabilities were 50% more likely to experience violence compared to people without disabilities (Hughes et al. 2012). One-quarter of people with mental illness have experienced recent physical, sexual, or intimate partner violence (Hughes et al. 2012).

Women with disabilities are particularly at risk. In Australia, women with disabilities have a twofold increase in the odds of experiencing physical violence, intimate partner violence, emotional abuse, and stalking and harassment compared to women without disabilities (Krnjacki et al. 2016). In the United States, women with disabilities were fourfold as likely to experience rape and threefold as likely to experience other sexual violence as compared to women without disabilities (Breiding and Armour 2015).

Violence against people with disabilities is similarly prevalent in lower income countries. Eighty-four per cent of women with disabilities involved in a survey of disability and violence in Bangladesh had experienced at least one act of violence or abuse (Hasan et al. 2014). These experiences were echoed in Nepal and Cambodia, where nearly 60% of women with disabilities had ever experienced violence (Astbury and Walji 2014; Puri et al. 2015). Despite the alarmingly high prevalence of violence towards people with disabilities, especially women, a recent evaluation of interventions found that there is currently no clear evidence on the type of approaches that work to reduce violence (Mikton et al. 2014).

Stigma and discrimination

Stigma and discrimination play a key role in all the barriers encountered by people with disabilities. The negative attitudes of society at large exclude people with disabilities from education, employment, and healthcare, as shown throughout this chapter. Stigma deprives people with disabilities of their independence and rights, as discussed by individuals in a study from

Croatia, who typically have little say in their education, employment, or sexual and reproductive health (Buljevac et al. 2012). Negative attitudes have also been found to be the major barrier for people with disabilities in the Philippines to participate in work, community meetings and social activities (Marella et al. 2016). Stigma also contributes to neglect and exclusion, such as in the Solomon Islands, where children with disabilities are seen as a punishment for violating a taboo (Gartrell et al. 2018).

Stigma and exclusion trap people with disabilities in poverty (Grech 2014; Trani et al. 2015). For example, a study on people with severe mental illness in India highlights this 'negative feedback loop' wherein individuals with severe mental illness face stigma, especially for individuals from lower castes. This stigma, combined with the educational and employment limitations imposed on people from lower castes, feeds persistent poverty, decreases social involvement and self-esteem, and poses additional responsibilities on carers, further contributing to stigma and discrimination (Trani et al. 2015).

In summary, while public health action is needed across a multitude of health-related areas, eliminating the disparities in the social determinants and resultant health of people with disabilities can only be done through addressing stigma, prejudice, and discrimination and the subsequent exclusion experienced by people with disabilities.

Inclusion of people with disabilities within public health

The number of people living with disabilities is likely to increase in both LMIC and HIC over time due to communicable and non-communicable diseases; road traffic accidents and injuries; longer life expectancies/ageing; conflict; natural disasters; and growing socioeconomic inequities experienced in many countries (Marmot et al. 2010; WHO 2011; Mannan and MacLachlan 2013; Krahn et al. 2015; Durham et al. 2016). Exclusion from public health interventions, combined with the greater unmet health needs and barriers to health services of an increasing number of people with disabilities, will lead to poorer health outcomes at an individual and population level (Durham et al. 2016; Marella et al. 2016). This highlights the urgent need for public health and health systems more broadly to ensure people with disabilities are better included in and benefit from all aspects of public health.

This will require both disability-specific initiatives and disability-inclusive programmes. Disability-specific initiatives are designed to target people with disabilities directly, for example capacity development for people with disabilities to better understand the risks of certain health behaviours, promoting access to physical activity and recreation for people with disabilities, or, providing outreach screening programmes for people with disabilities who are not able to travel to health services. Disability-inclusive programmes are mainstream activities designed to ensure that they are accessible to people with disabilities and that they benefit people with and without disabilities equally (Kleinitz et al. 2011; Piatta and Guy 2010). Examples include health promotion messaging designed to include representation of people with disabilities which is accessible in a range of communication formats (e.g. Braille, easy-to-understand) (Piatta and Guy 2010; Kleinitz et al. 2011). Even with appropriate messaging, people with disabilities may experience additional barriers to making healthy choices due to issues such as mobility (e.g. unable to access parks, public transport, shops), difficulty to reading information (e.g. food labels) due to vision or cognitive impairments and community attitudes towards people with disabilities accessing resources in their community.

It also requires barriers to all aspects of public health and health systems experienced by people with disabilities to be identified and addressed. The section to follow focuses on key components of public health, outlining some of the barriers experienced by people with disabilities, as well as examples of how these barriers can be addressed.

Health promotion and prevention

Disease prevention, including screening for cancer and immunization, are crucial components of primary healthcare for all people. However, people with disabilities do not always receive disease prevention services to the same extent individuals without disabilities do. Access to disease prevention, especially screening, is often difficult, particularly for women with disabilities. Costs for screening may not be available through universal healthcare and costs may be prohibitive. Women with disabilities frequently report that medical professionals do not understand their disabilities, putting women in a position where they must explain themselves and advocate for their own care (Ramjan et al. 2016). Disability-specific information is

typically unavailable for screening processes, (Angus et al. 2012) and most healthcare providers report they need substantial additional training in order to provide adequate preventive care for people with disabilities (Merten et al. 2015).

Data from a national health survey in the United States revealed that women with disabilities are less likely to be up-to-date with cervical and breast cancer screening. Only two-thirds of women with mobility impairments were estimated to have received a Pap test within the previous three years, and 61% of women with cognitive limitations received a mammogram within the last two years. By comparison, 81% of women without disabilities had an up-to-date Pap test, and 73% had had a recent mammogram (Steele et al. 2017). Women in the United States with more complex, and more severe, disabilities are consistently less likely to be up-to-date with cancer screening (Horner-Johnson et al. 2014).

While breast and cervical cancer screening rates among women with disabilities are still lower than those of women without disabilities, there is some evidence for the effectiveness of educational programmes in increasing cancer screening rates among women with disabilities. *Women Be Healthy 2*, an educational programme about mammography and Pap tests, was appropriate for women with disabilities as materials were created to be hands-on, multimodal, and accessible to people with low literacy (Swaine et al. 2014). Similarly, a DVD about mammography preparedness for women with intellectual disabilities starred an actress with intellectual disability confidently navigating her first mammography appointment (Greenwood et al. 2014). Both programmes showed increases in cancer screening knowledge among participants. Despite small sample sizes, these programmes suggest that appropriate educational and support resources can help women with disabilities increase their knowledge of, and ability to access, cancer screening. For another example of disability-inclusive screening programmes, see the Case Study 10.7.1 next.

Encouraging research from the United States on colorectal cancer screening suggests similar rates of testing between people with and without disabilities, and that where disparities exist, people with disabilities are actually screened more (Iezzoni et al. 2016). While the exact explanation for good uptake of colorectal cancer screening among people with disabilities are unclear, increased contact with physicians, who may follow-up to ensure screening is completed, may be an important driver of this trend.

CASE STUDY 10.7.1 Just checking: increasing cancer screening for people with disabilities

While people with disabilities have generally lower cancer screening rates, the literature indicates mixed outcomes regarding immunization rates. Older people with disabilities in the United States were more likely to have received a flu vaccine in the past year and a pneumonia vaccine than their peers without disabilities (Pharr and Bungum 2012). Researchers in remote Ecuador found that children with disabilities were immunized at rates comparable with their siblings without disability, highlighting the effectiveness of the national immunization programme and importance of positive cultural attitudes towards disability (Groce et al. 2007). The relative ease of giving vaccinations, which typically do not require special equipment, may positively influence vaccination rates. However, some groups of people with disabilities, such as individuals with intellectual disabilities, may have vaccination rates lower than the general population. Data from Taiwan has shown that teenagers with intellectual disabilities were vaccinated against Hepatitis B at lower rates than the general population (Yen and Lin 2011). Young people with intellectual disabilities in Taiwan were also less likely to have received a flu vaccine than older people and very young children, although individuals with moderate and severe intellectual disabilities were more likely to have received the vaccine than individuals with mild intellectual disabilities (Yen et al. 2012; CDC 2010 cited in Yen et al. 2012). This suggests that although regular contact with the medical system may positively influence vaccination rates, people with disabilities do not consistently receive the same level of immunization as the general population.

Despite these factors, people with disabilities may be less likely to receive vaccinations associated with sensitive issues, such as sex and sexuality. Providers of the human papilloma virus (HPV) vaccination in the United States reported the main reason children with disabilities were not vaccinated was due to parental refusal (Cody and Lerand 2013). As such, immunization rates may differ between people with and without disabilities depending on the acceptability of particular vaccinations and (mis)perceptions about the relevance for people with disabilities. Identifying barriers to HPV vaccination among people with disabilities is critical and strategies to improve uptake of HPV need to be developed.

Health promotion is an area which requires significant new research and investment to improve diet, increase physical activity and reduce smoking rates. Despite the higher prevalence of smoking among people with disabilities, there are few programmes and little literature on successful tobacco cessation approaches and programmes, contributing to the decreased rate of quit attempts by people with disabilities. It is clear, however, that people with disabilities have a desire to quit: nearly 90% of people with mobility impairments in a study from the United States planned to quit within six months (Borrelli et al. 2013).

Some initial evidence from small tobacco cessation programmes in the United States emphasizes the importance of tailoring programmes to people with disabilities, through disability-specific examples and discussion of disability-relevant issues, such as the interaction between nicotine replacement therapy and medication (Pomeranz 2014; King et al. 2016). While these suggestions provide a foundation for creating effective cessation programmes, it is obvious people with disabilities are underserved in this area.

A recent review of intervention research focused on improving physical activity levels found that research in this area is increasing. Interventions were predominantly individually focused on behaviour change with mean incorporating technologies (e.g. information and interactive technologies). Most studies showed positive outcomes however 60% of studies were of groups were multiple sclerosis (41%), stroke (15%), and spinal cord injury (12%) limiting the generalisability of findings (Lai et al. 2017). Further research is needed with people with other types of disability.

Research should also concentrate on how population-based interventions in the general population can be adapted to be more appropriate and relevant to people with disability.

Access to health services

Access to healthcare services is a substantial issue for people with disabilities across the globe. The poorer physical and mental health outcomes of people with disabilities, as well as poorer health behaviours and discrepancies in disease prevention services, are related to barriers to health services. While the experiences and health of people with disabilities may vary widely across countries, issues of transport, lack and/or inaccessible health facilities and equipment (including toilets), discrimination by health workers, and cost of services and travelling to services are routinely reported as major barriers. Similarly, lack of knowledge on when, where and how to access health services, and inaccessible information and health messaging compound these barriers (MacLachlan et al. 2011; Kleinitz et al. 2012; Van Rooy et al. 2012; Vaughan et al. 2016).

Research from various contexts demonstrates these barriers. Nearly 15% of people with disabilities in the Republic of Korea for example reported difficulty accessing healthcare in the preceding year, with approximately 60% citing cost as their primary barrier to care (Lee et al. 2014a). Qualitative work in Namibia found cost was also a significant barrier to care for people with disabilities, although interviewees cited transport as the main barrier. Nearly 90% of respondents with disabilities reported walking, sometimes hours, to healthcare centres due to prohibitively expensive transport fees. For those who were able to reach the clinics, additional obstacles to care included language differences, a shortage of healthcare workers, and inaccessible facilities (Van Rooy et al. 2012).

Further qualitative work in Sudan, Namibia, Malawi, and South Africa reinforced these barriers, and respondents discussed problems with inaccessible equipment and poor availability of services (Eide et al. 2015). Children with disabilities and their parent carers in rural and remote Australia reported similar challenges in accessibility of health services due to long travel times and lack of services (Dew et al. 2013).

In addition to these barriers to general medical care, individuals often face further difficulties to effective sexual and reproductive care. Often policies and strategies do not recognize the unique needs of males and females with

disabilities and community perceptions may hold that people with disabilities are asexual or hypersexual. In Ghana, men and women with disabilities are typically grouped together as a homogenous group, despite differences in needs. The use of derogatory language, like 'idiot', when describing men and women with disabilities was found to further contribute to stigma and control over their sexual and reproductive health needs (Mprah et al. 2014).

Stigmatizing attitudes may also prohibit people with disabilities from having access to information about sexual health and reproduction, as was reported by people with physical disabilities in Vietnam (Nguyen et al. 2018). These negative attitudes, and the ubiquitous belief that sexuality and reproduction are taboo for people with disabilities can be further seen in a study of access to services in Uganda. Women with disabilities reported extreme stigma surrounding their pregnancy and sexuality due to their disabilities. Male partners with disabilities described total exclusion from the pregnancy and childbirth, as they were offered no assistance to reach antenatal and birthing wards on upper floors of the hospital (Ahumuza et al. 2014).

Pregnant women in higher income countries, such as the United States, still face humiliation and embarrassment in the doctor's office, such as when being transferred to inaccessible equipment, such as examination tables that cannot be adjusted for height (Iezzoni et al. 2015). For a case study on an effective sexual and reproductive health programme for women with disabilities, please refer to Case Study 10.7.2.

CASE STUDY 10.7.2 Promoting inclusive public health programmes

Women with Disability taking Action on REproductive and sexual health—W-DARE project

Compounding disability and gender based discrimination often mean that women with disabilities experience exclusion from the socioeconomic and political lives of their communities. Similarly, women with disabilities face barriers to accessing public health programmes such as health information, screening, prevention, and care services, as well as violence prevention and response programmes. As a result, women with disabilities experience poorer health outcomes across their life cycle when compared to women without disabilities, including in relation to sexual and reproductive health (SRH) and violence (MacLachlan and Swartz 2009; Lee et al. 2015; Vaughan et al. 2015; Marella et al. 2016; Vaughan et al. 2016; Devine et al. 2017).

Limited data, particularly from LMIC, on the prevalence of disability and the lived experience of women with disabilities in relation to their health and access to programmes, also makes it difficult

to plan and resource disability-inclusive public health programmes. The W-DARE project in the Philippines is one of the few examples of inclusive action research that has been undertaken to increase demand for and supply of quality SRH services and information for women with disabilities.

W-DARE was a 3-year project conducted by the University of Melbourne in Australia and De La Salle University in the Philippines, in collaboration with the Likhaan Center for Women's Health; Center for Women's Studies Foundation at the University of the Philippines; and two national disabled people's organizations (DPOs)—WOWLEAP (Women with disabilities Leap social and economic progress) and PARE (Persons with disability Advocating for Rights and Empowerment). Women with disabilities were resourced as co-researchers across the project, ensuring their lived experience informed the design, implementation, and evaluation of the project (Vaughan et al. 2015).

Key findings from phase one of the research highlighted women with disabilities reported poorer well-being when compared to women without disabilities, and were more likely to report reduced access to health services, work, education, and social activities. Barriers to inclusion reported included cost of services, lack of information about services, and difficulty getting to health facilities. Levels of SRH knowledge were generally low, especially among women who are Deaf or hard of hearing and women with intellectual disabilities. Specific barriers to SRH services reported by women with disabilities included discrimination; limited capacity of service providers; inaccessible transport; economic barriers; and difficulty accessing sign language interpreters. Reports of violence and abuse against women with disabilities were common. What was also evident was the resilience of many of the women with disabilities interviewed.

In response, the W-DARE team developed and implemented a number of interventions aimed at promoting access to quality SRH programmes. Demand side interventions included peer-facilitated Participatory Action Groups for women with disabilities to enhance knowledge and demand for SRH services and develop networks of peer support. Supply side interventions included workshops for service providers to increase awareness of the needs and rights of women with disabilities; collaboration with the Department of Health to increase disability inclusion in primary healthcare settings; project partner exchange visits to share experiences of disability inclusiveness within local health and development programmes; and, working with violence prevention and response services to strengthen disability inclusion including through referral practices. Recognition that sustained change requires a major shift in society's perception of disability, led to the production of a number of social media videos of women with disabilities talking about their lives and their rights.

The W-DARE interventions led to positive outcomes for women with disabilities and highlighted that increased access to information on rights and disability inclusion for both women with disabilities and service providers; enhanced social networks and opportunities for social inclusion for women with disabilities; and consumer participation in the design of more inclusive health services, are essential for upholding the right to SRH for women with disabilities.

For the W-DARE project to have a sustained impact however, it is vital address systemic barriers to participation in all other life domains experienced by women with disabilities such as education, employment and general health programmes are addressed This will require governments, service providers, DPOs, public health and development programmers to work collaboratively with women with disabilities to identify context-specific barriers and solutions to promote inclusion of people with disabilities within their communities.

Health-related rehabilitation

While not relevant for all people with disabilities, health-related rehabilitation is not only an important right articulated in the CRPD, but recognized as a key component of Primary Health Care and a critical part of

the broader health system (Pryor et al. 2017; Bezmez et al. 2018). However, many people with disabilities, particularly those living in LMICs, have limited access to health-related rehabilitation services, resulting in unmet health and rehabilitation needs, reduced social and economic participation, and reduced quality of life (Gargett et al. 2016; Pryor et al. 2017; WHO 2017b).

Similar to broader health services, individual level barriers to rehabilitation services in LMICs include difficulty travelling long distances to services when roads and transport are often inaccessible or unaffordable and services are mainly located in urban areas, such as was identified by Gargett et al. in their review of rehabilitation workforce needs in the Pacific (Gargett et al. 2016); the prohibitive cost of rehabilitation services and assistive devices as identified by Palmer et al. in their study of the economic challenges experienced by people with disabilities in Vietnam (Palmer et al. 2015); and lack of knowledge on when, where, and how to access rehabilitation services, such as highlighted by Mactaggart in their study of access to health and rehabilitation services in Cameroon and India (Mactaggart et al. 2016).

At a systems level, access to rehabilitation is hampered by the neglect of rehabilitation within broader health systems strengthening efforts (Pryor et al. 2017); lack of rehabilitation specialists within an already limited pool of health workers who can respond to the rehabilitation needs of people with disabilities (Gargett et al. 2016); limited inclusion of measures of disability and rehabilitation within Health information Systems in many low- and middle-income countries (McPherson et al. 2017); and extensive gaps within referral processes across and within health systems (WHO 2017b).

As like all public health interventions that aim to improve the capacity to meet the needs of people with disabilities, health-related rehabilitation services need to actively and respectfully involve people with disabilities in all aspects of decision making and service delivery; and, ensure a holistic understanding of rehabilitation that responds to the complex needs of individuals (Bezmez et al. 2018).

Health information systems and data transparency for monitoring and evaluation

There is still a need for greater context-specific information on the lived experience of people with disabilities in relation to their health and inclusion within the public health sphere (Vaughan et al. 2016). Very few health systems in LMIC and HIC collect information on people with disabilities and their access to health. For example, a study of health information systems in Lao found indicators related to disability or functioning were not included, and overall health data was often incomplete, inaccurate or poorly used (Durham et al. 2016). This makes it difficult to monitor progress in access or health outcomes. Enhancing the focus on disability within public health could provide an opportunity to strengthen health information systems (Durham et al. 2016).

Similar to the issue of measurement of disability more broadly, this requires an accepted definition of disability and standard measure of disability that health information systems can readily and affordably adopt. As highlighted earlier, better disability measures within population and health surveys are also required to better monitor health behaviours, health outcomes, access to health and impact of disability on other key domains such as education and employment (Durham et al. 2016). In Australia the Australian Institute of Health and Welfare have developed a set of questions that can be routinely implemented in administrative data (such as primary care and hospitals). It is hoped in the future that the implementation of these questions will enable enhanced monitoring and evaluation of health service use for people with disabilities (Australian Institute of Health and Welfare 2016c).

Disability services and supports

Disability services and supports (both within and external to health systems) for people with disabilities are critical for health and well-being. Services and supports include personal assistance with self-care (e.g. dressing, eating, showering) as well as support to participate in the community and paid and unpaid work. Access to therapeutic supports such as speech therapy is also critical to enabling people with disabilities to attain their potential. Basic supports such as hearing aids and wheelchairs are also critical for full participation in society. New technologies such as robots (van den Heuvel et al. 2017), IT assisted technologies including smartphone apps (Mavrou et al. 2017), and technological innovation in assisted and augmented

communication (Miksztai-Rethey and Farago 2018) hold considerable promise for improving participation and access to health. In all settings, including education, workplaces, and health services, reasonable adjustments and accommodations are critical to enable people with disabilities to access and participate. This can range from ramps and lifts to improve the physical accessibility, to flexible working conditions, such as working from home. Further, ensuring the development, implementation and monitoring of workplace disability strategies (including mental health) will not only promote healthier workplaces, but will promote better understanding of the importance of inclusion (International Labour Office et al. 2014).

Internationally, disability services are often underfunded and poorly resourced with considerable inequities in access to services. Approaches to the funding and delivery of services vary widely. Given the barriers that many people with disabilities experience in accessing mainstream services, one of the most important approaches to improving health and social outcomes in LMIC is community-based rehabilitation (CBR). CBR is a multisectoral approach to improve the equalization of opportunities for, and social inclusion of, persons with disabilities. CBR is implemented through the combined efforts of people with disabilities, their families and communities, and relevant government and non-government health, education, vocational, social, and other sectors (WHO 2010).

In Cambodia for example, a number of non-government organizations have been working with communities in rural and remote areas to improve understanding of disability, promote early identification, and, develop locally based approaches to support people with disabilities and their families to address impairment related conditions (i.e. through training families on appropriate rehabilitation techniques and exercises that they can do with their family members with disabilities), and, working with local authorities and other stakeholders to promote community inclusion (i.e. promoting access to local schools). The government of Cambodia has been working with these organizations to develop National CBR guidelines. As with many countries of all income levels, sustainable financial and human resources is an ongoing challenge (Bailey and Nguon 2014).

A review of disability services in Australia by the Productivity Commission found that services were ‘inequitable, underfunded, fragmented, and inefficient and give people with disabilities little choice’. The Commission recommended a doubling of funding and significant

changes to how services were provided (Productivity Commission 2011). This led to the introduction of the National Disability Insurance Scheme which provides individuals with disability-individual budgets and they can purchase the services and support they need. See Case Study 10.7.3.

CASE STUDY 10.7.3 Australia's National Disability Insurance Schemes

Prior to the introduction of the National Disability Insurance Scheme (NDIS) in Australia, the disability services were delivered by state and territory government and access, quality and coverage varied significantly between states. Funding was provided by the federal government to the states and territories to fund disability services but the delivery of services was a state or territory responsibility. This meant that service access varied according to what state people lived in and whether they lived in cities or rural and remote areas and there was no guarantee of long-term security of funding. People also had little choice of the services they were provided.

The Productivity Commission of Australia was tasked with reviewing the current disability service system and to consider alternative models of delivery. In particular they considered the implementation of personalized budgets, such as introduced in the United Kingdom, where people with disabilities and their families and carers are provided with individual budgets, depending on their particular support needs, to buy the services and supports they need (Needham and Dickinson 2017). The Commission also recommended the adoption of an insurance approach whereby people with disability are assured of funding to cover the lifetime costs of their disability. The Productivity Commission recommended the implementation of the National Disability Insurance Scheme and a doubling of funding to the disability service sector (Productivity Commission 2011).

The Productivity Commission recommended that the NDIS include three tiers of support.

Tier 1 provides insurance (i.e. guaranteed support) for all Australians who acquire a significant disability. Tier 1 was also designed to promote opportunities for people with disability to participate in the community.

Tier 2 included anyone with a disability and their primary carers (estimated by the Productivity Commission at around 4.8 million people). Tier 2 involves working with mainstream services, provided information and referral services.

Tier 3 involves the provision of long-term care and support, through individualized funding packages, to people with a significant and ongoing disability and who were less than 65 years of age and Australian citizens or residents. At the time of Commission's report this was estimated be around 410,000 people, about ten percent of the disabled population. To be eligible for Tier 3 support the disability needs to be permanent as well as associated with 'significantly reduced functioning in self-care, communication, mobility or self-management and require significant ongoing support' and/or be assessed as belonging to a group 'for whom there was good evidence that [early] intervention would be safe, significantly improve outcomes and would be cost-effective'. To be eligible for packages individuals are assessed to identify their goals and needs, are provided with certainty of funding based on needed and have choice over the services and supports they purchase (Productivity Commission 2011). The services and support were to be provided by a range of not-for-profit, governmental, and private organizations.

The NDIS reform received strong support from both major political parties as well as State and Territory governments. It was implemented in four trial sites across Australia from 1 July 2013, with

another three sites commencing on 1 July 2014. It began full roll-out across Australian in July 2016 with full roll-out due to be completed in 2019/2020 (Buckmaster and Clark 2018).

The NDIS is the largest social policy reform implemented in Australia since the introduction of Medicare—the universal health scheme. It has faced a number of significant implementation issues including clarity around eligibility for individualized services and supports particularly for people with psychosocial disabilities, ICT issues, lack of services for people with disabilities to choose from ('market failure'), problems with the processes of planning with participants and difficulties faced by individual's navigating the complexities of the new system. There are also concerns related to the different capacities of individuals to negotiate and navigate such a complex system resulting in significant inequities in funding packages (Carey et al. 2017). A further tension is in the role of governments from 'light touch' 'market regulation', where government identifies fraudulent practices and sets some minimum standards, to 'market stewardship', where government plays an active role in supporting innovation, stimulating markets particularly where there are shortages, and ensuring equity and choice and control. While it is clear that there are gaps in the NDIS in terms of access to services for people with different types of disability or people regional and remote Australia, partly driven by lack of financial sustainability, the role of government in supporting markets to develop in these areas is underdeveloped with the potential for considerable inequities to arise (Carey et al. 2017; Carey et al. 2018).

The initial evaluation of the scheme showed that participants were generally satisfied (Mavromaras et al. 2018) however it will be many years before we know just how well Australia's bold reforms in the funding and delivery of disability services delivers on its aspirations.

Public health research and disability

To date, research on the health of people with disabilities has been sparse, however it is clear that public health research could do much to advance the evidence-base about the health of people with disabilities as well as identify evidence-based interventions and strategies to improve their health. For example, public health researchers could advance the measurement and reporting of disability-related information, the development of interventions and programmes to address health disadvantage (e.g. targeted tobacco cessation programmes) and policy analysis. Importantly this research needs to be conducted from multiple disciplinary perspectives and employ both qualitative and quantitative methods. It is also critical that people with disabilities are central to all research conducted and at the core of solutions to problems identified. Best practice in this area involves co-design of research projects where people with disabilities are appropriately resourced to conduct the research.

The W-DARE project, presented earlier, is an example of co-designed research. As such, women with disabilities were resourced as co-researchers to work alongside researchers without disabilities at all stages of the project. This included initial two-way capacity development, design and

implementation of all stages of the research, ongoing training of health workers and other women with disabilities on attaining the rights to sexual and reproductive health, analysis of data, and presentation of findings. Not only did this ensure the research benefit from insider knowledge of the lived experience of women with disabilities in the Philippines, but it also demonstrated the capabilities of women with disabilities as central to research that is about them (Vaughan et al. 2015).

The Centre for Research Excellence in Disability and Health is example of a multidisciplinary programme of research on the health of people with disabilities that draws on the disability sector to drive its research priorities and dissemination activities. See Case Study 10.7.4.

CASE STUDY 10.7.4 National Health and Medical Research Council, Centre for Research Excellence in Disability and Health

The Centre of Research Excellence (CRE) in Disability and Health ((http://credh.org.au/) is funded by Australia's National Health and Medical Research Council. It is a world first; it aims to gather the evidence needed to guide social and health policy reform in Australia.

The CRE brings together an interdisciplinary group of researchers with stakeholders from the health and disability sectors, state and federal governments, and disability consumer representative groups—who are embedded in the research design and development. In collaboration with people with disabilities, service providers, non-government organizations and other stakeholders, the CRE will identify *the key and emerging issues affecting Australians with disabilities*. The research is designed to respond to the priorities identified by stakeholders in the disability sector.

The research involves four work programmes:

1. Mapping the spatial distribution of social, economic, and health inequities between people with and without disabilities
2. The application of cutting-edge epidemiological methods to establish the main social determinants of health of people with disability
3. The construction of the first ever cost-effectiveness model to estimate the health impacts and value for money of policy interventions for people with disabilities
4. An innovative policy lab where stakeholders are embedded in the research process where they are interviewed throughout the life of the CRE, given the opportunity to reflect on others' input and respond to the findings of the other three work programmes.

The CRE-DH is also engaged in developing a skilled research workforce in disability and health by training new early-career researchers and PhD students as well as building the research capacity of academics and policymakers across sectors to respond to the pressing problem of people with disabilities' poor health.

Conclusion and ways forward

In this chapter we have shown that while the evidence-base on the health of people with disability is incomplete, there is enough evidence to conclude that people with disabilities have poorer mental and physical health, on average, than non-disabled people. These poorer health outcomes are in part due to the underlying health conditions but are also a result of socioeconomic disadvantage, higher levels of health damaging behaviours, poorly targeted health promotion, and preventive healthcare, lack of access to quality health, disability and rehabilitation services, and higher exposure to discrimination, stigma, and violence.

Turning around the health disadvantage for people with disabilities requires population-based and targeted approaches that address the specific circumstances of people with disabilities. As one of the leading researchers into the social determinants of health highlights: '*To reduce the steepness of the social gradient in health, actions must be universal, but with a scale and intensity that is proportionate to the level of disadvantage.*' (Marmot et al. 2010). It can therefore be argued that people with disabilities have a greater need for all public health functions (e.g. immunizations, support for healthy behaviours) and it is time for public health researchers, policymakers and practitioners to address the exclusion of people with disabilities within public health and the broader health system (United Nations 2006; WHO 2011; Mannan and MacLachlan 2013; Krahn et al. 2015).

Public health approaches to improving the health of people with disabilities requires a social determinants approach to primary prevention, health promotion and support for healthy lifestyles, through to care and rehabilitation for those who experience health conditions and/or disability. Public health approaches must also consider social protection and welfare programmes which aim to reduce poverty and more effective interventions that promote the social and economic participation of people with disabilities (Skempes and Bickenbach 2015; Carey 2016).

Health and rehabilitation system strengthening is required to improve health outcomes of people with disabilities. Systems strengthening must have good governance; coherent and explicit policies which articulate and resource for targeted programmes for people with disabilities; appropriate resourcing to implement inclusive policies and programmes; better coordination of care and referral process throughout health and rehabilitation

systems; reduce the current gap in public health human resources for health that can support coordination of care as well as provide services that meet the needs of people with disabilities (Krahn et al. 2015; Mannan and MacLachlan 2013; McVeigh et al. 2016).

Essential to all of these processes is greater recognition that people with disabilities have the same (and often additional) health needs and rights as people without disabilities. The right to health is enshrined in the CRPD (United Nations 2006). A critical illustration of why this right and all other intangible rights of people with disabilities required its' own UN convention, is the fact that people with disabilities were excluded from international action, including public health action, in response to the Millennium Development Goals. With significant advocacy by and with people with disabilities and their representative organizations, the inclusion or people with disabilities is clearly articulated in the Global Goals for Sustainable Development 2015–2030 (SDGs) (Marella et al. 2016).

Global commitment to the CRPD, the SDGs, as well as platforms such as the WHO Alma-Ata Primary Health Care Principles (WHO 1978), mandates public health practitioners, researchers, and policymakers in all countries to ensure people with disabilities are both meaningfully included and benefit in all aspects of public health. This will require more inclusive approaches to address existing barriers to inclusion within mainstream public health policy and programming (i.e. health promotion messaging, disease outbreak and disaster responses), as well as disability-specific actions (i.e. promotion and resourcing of CBR in low- and middle-income countries (United Nations 2006; WHO 2011; MacLachlan et al. 2012; MacLachlan and Mannan 2014; Krahn et al. 2015).

We have described how the measurement of disability and the quality of information collected on disability is poor. There is an urgent need for better context-specific information and data on the lived experience of people with disabilities in relation to their health and inclusion within the public health sphere (Vaughan et al. 2016). Improved context-specific disability data collection and disaggregation including within Health Information Systems is critical to monitor progress towards inclusion, as well as identify barriers to improving health outcomes for people with disabilities (Durham et al. 2016). This includes identifying gaps in knowledge, practices, and attitudes of public health practitioners to improve their capacity for inclusion, alongside strategies to enhance the knowledge and capacity of people with

disabilities to improve their access to public health programmes (Mannan and MacLachlan 2013; Krahn et al. 2015; Vaughan et al. 2016).

Meaningful leadership and participation of people with disabilities in policy development is needed so that policy is responsive to the needs of people with disability (Love et al. 2018). Furthermore, we need to ensure that people with disabilities are supported to understand their rights to health, as well as provided with information, resources, and support to access public health programmes (Devine et al. 2017). In terms of research, we need to build the capacity of people with disabilities to take leaderships roles in public health research and ensure that all research projects are responsive to the priorities of people with disabilities.

Public health has much to contribute to alleviating the health disadvantage that people with disabilities face. To do this, public health needs to move beyond understanding disability as a health deficit and take a social determinants approach to the health of people with disability. Reducing disability-related health inequalities requires addressing the root causes of poor health including social determinants such as poverty and discrimination. Doing so would enable public health to more truly achieve its goals of promoting the health of *all* individuals and their communities.

REFERENCES

Ahumuza, S.E., et al. (2014). Challenges in accessing sexual and reproductive health services by people with physical disabilities in Kampala, Uganda. *Reproductive Health*, **11**, 59.

Anderson, M.L., et al. (2014). Posttraumatic stress disorder and substance use disorder comorbidity among individuals with physical disabilities: findings from the National Comorbidity Survey Replication. *Journal of Traumatic Stress*, 27(2), 182–91.

Angus, J., et al. (2012). Access to cancer screening for women with mobility disabilities. *Journal of Cancer Education*, 27(1), 75–82.

Arnesen, T. and Nord, E. (1999). The value of DALY life: Problems with ethics and validity of disability adjusted life years. *British Medical Journal*, 319(7222), 1423–5.

Astbury, J. and Walji, F. (2014). The prevalence and psychological costs of household violence by family members against women with disabilities in Cambodia. *Journal of Interpersonal Violence*, 29(17), 3127–49.

Australian Bureau of Statistics (2016a). *Disability, Ageing and Carers, Australia: Summary of Findings, 2015*. Canberra: Australian Bureau of Statistics.

Australian Bureau of Statistics (2016b). *Supplementary Disability Survey, 2016: Key Findings, 2016*. [Online] Available at: http://www.abs.gov.au/AUSSTATS/abs@.nsf/Lookup/4450.0Main+Features12016?OpenDocument

Australian Human Rights Commission (2016). *Willing to Work: National Inquiry into Employment Discrimination Against Older Australians and Australians with Disability*. Australian Human

Rights Commission. [Online] Available at: https://humanrights.gov.au/our-work/age-discrimination/projects/willing-work-national-inquiry-employment-discrimination
Australian Institute of Health and Welfare (2016a). *Australia's Health 2016*. Canberra: AIHW.
Australian Institute of Health and Welfare (2016b). *Health Status and Risk Factors of Australians with Disability 2007–2008 and 2011–2012*. Canberra: AIHW.
Australian Institute of Health and Welfare (2016c). *Standardised Disability Flag: data Collection Guide. Cat. no. 6. DAT 6*. Canberra: AIHW.
Bailey, S.R. and Nguon, S. (2014). *Situation Analysis for Disability-inclusive Governance and Community Development in Cambodia.* UNICEF Cambodia & Local Governance and Child Rights. [Online] Available at: https://www.dfat.gov.au/sites/default/files/cambodia-disability-inclusive-governance-community-development-sit-analysis.pdf
Banks, L.M., et al. (2017). Poverty and disability in low- and middle-income countries: A systematic review. *PLoS One*, 12(12), e0189996.
Barnes, T.L., et al. (2013). Physical activity in a large sample of adults with intellectual disabilities. *Journal of Physical Activity & Health*, 10(7), 1048–56.
Beer, A. and Faulkner, D. (2009). *The Housing Careers of People with a Disability and Carers of People with a Disability*. AHURI Research Paper. [Online] Available at: https://www.ahuri.edu.au/__data/assets/pdf_file/0007/3103/AHURI_Research_Paper_The-housing-careers-of-people-with-a-disability-and-carers-of-people-with-a-disability.pdf
Berghs, M., et al. (2016), Implications for public health research of models and theories of disability: a scoping study and evidence synthesis. *Public Health Research*. Southampton (UK): NIHR Journals Library; 2016 Jul. PMID: 27512753.
Bezmez, D., et al. (2018). Rehabilitation as a disability equality issue: A conceptual shift for disability studies? *Social Inclusion*, 6(1), 61–72.
Borrelli, B., et al. (2014). Cigarette smoking among adults with mobility impairments: a US population-based survey. *American Journal of Public Health*, 104(10), 1943–9.
Borrelli, B., et al. (2013). Methods used to quit smoking by people with physical disabilities. *Rehabilitation Psychology*, 58(2), 117–23.
Breiding, M.J. and Armour, B.S. (2015). The association between disability and intimate partner violence in the United States. *Annals of Epidemiology*, 25(6), 455–7.
Brorson, H.H., et al. (2013). Drop-out from addiction treatment: a systematic review of risk factors. *Clinical Psychology Review*, 33(8), 1010–24.
Brucker, D. L. and Houtenville, A. J. (2015). People with disabilities in the United States. *Archives of Physical Medicine & Rehabilitation*, 96(5), 771–4.
Buckles, J., Luckasson, R., and Keefe, E. (2013). A systematic review of the prevalence of psychiatric disorders in adults with intellectual disability, 2003–2010. *Journal of Mental Health Research in Intellectual Disabilities*, 6(3), 181–207.
Buckmaster, L. and Clark, S. (2018). *The National Disability Insurance Scheme: A Chronology*. Canberra: Department of Parliamentary Services.
Buljevac, M., et al. (2012). The stigma of disability: Croatian experiences. *Disability & Rehabilitation*, 34(9), 725–32.
Byrne, B. (2018). Dis-equality: exploring the juxtaposition of disability and equality. *Social Inclusion*, 6(1), 9–17.
Carey, G. (2016). *Grassroots to Government: Creating Joined-up Work in Australia*. Melbourne: Melbourne University Publishing.
Carey, G., et al. (2018). The vexed question of market stewardship in the public sector: Examining equity and the social contract through the Australian National Disability Insurance Scheme. *Social Policy & Administration*, 52(1), 387–407.
Carey, G., et al. (2017). Redressing or entrenching social and health inequities through policy implementation? Examining personalised budgets through the Australian National Disability Insurance Scheme. *International Journal for Equity in Health*, 16(1), 192.

Carroll Chapman, S.L., and Wu, L. (2013). Substance use among individuals with intellectual disabilities. *Research in Developmental Disabilities*, 33(4), 1147–56.
Cashin, A., et al. (2018). A scoping review of what is known of the physical health of adults with autism spectrum disorder. *Journal of Intellectual Disability*, 22(1), 96–108.
Cody, P.J. and Lerand, S.J. (2013). HPV vaccination in female children with special health care needs. *Journal of Pediatric & Adolescent Gynecology*, 26(4), 219–23.
Cooper, S.A., et al. (2015). Multiple physical and mental health comorbidity in adults with intellectual disabilities: Population-based cross-sectional analysis. *BMC Family Practice*, **16**, 110.
Courtney-Long, E., et al. (2014). Disparities in current cigarette smoking prevalence by type of disability, 2009–2011. *Public Health Reports*, 129(3), 252–60.
Dairo, Y.M., et al. (2016). Physical activity levels in adults with intellectual disabilities: A systematic review. *Preventive Medicine Reports*, **4**, 209–19.
Devine, A., et al. (2017). 'Freedom to go where I want to go': Qualitative valuation of the W-DARE peer facilitated Participatory Action Groups to improve access to sexual and reproductive health for women with disabilities in the Philippines. *Reproductive Health Matters Disability Special Issue*, **25**, 55–65.
Dew, A., et al. (2013). Addressing the barriers to accessing therapy services in rural and remote areas. *Disability & Rehabilitation*, 35(18), 1564–70.
Durham, J., et al. (2016). Health information systems and disability in the Lao PDR: A qualitative study. *The International Journal of Health Planning and Management*, 31(4), 446–58.
Eide, A.H., et al. (2015). Perceived barriers for accessing health services among individuals with disability in four African countries. *PLoS One*, 10(5), e0125915.
Emerson, E. and Hatton, C. (2008). Socioeconomic disadvantage, social participation and networks and the self-rated health of English men and women with mild and moderate intellectual disabilities: cross-sectional survey. *European Journal of Public Health*, 18(1), 31–7.
Emerson, E., et al. (2018). The association between non-standard employment, job insecurity and health among British adults with and without intellectual impairments: Cohort study. *SSM Population Health*, **4**, 197–205.
Emerson, E., et al. (2011). The health of disabled people and the social determinants of health. *Public Health*, 125(3), 145–7.
Gargett, A.L., et al. (2016). Identifying rehabilitation workforce strengths, concerns, and needs: a case study from the Pacific Islands. *Disability, CBR & Inclusive Development*, 27(2), 19–36.
Gartrell, A., et al. (2018). Social determinants of disability-based disadvantage in Solomon islands. *Health Promotion International*, 33(2), 250–60.
Glazier, R.E. and Kling, R.N. (2013). Recent trends in substance abuse among persons with disabilities compared to that of persons without disabilities. *Disability & Health Journal*, 6(2), 107–15.
Grech, S. (2014). Disability, poverty and education: Perceived barriers and (dis)connections in rural Guatemala. *Disability and the Global South*, 1(1), 128–52.
Greenwood, N.W., et al. (2014). Testing the feasibility of a DVD-based intervention to promote preparedness for mammography in women with intellectual disabilities. *Journal of Cancer Education*, 29(1), 99–105.
Groce, N., et al. (2014). Malnutrition and disability: Unexplored opportunities for collaboration. *Paediatrics & International Child Health*, 34(4), 308–14.
Groce, N., et al. (2011). Disability and poverty: the need for a more nuanced understanding of implications for development policy and practice. *Third World Quarterly*, 32(8), 1493–513.
Groce, N.E., et al. (2007). Immunization rates among disabled children in Ecuador: Unanticipated findings. *Journal of Pediatrics*, 151(2), 218–20.
Hammel, J., et al. (2015). Environmental barriers and supports to everyday participation: A qualitative insider perspective from people with disabilities. *Archives of Physical Medicine &*

Rehabilitation, 96(4), 578–88.
Harper, P. (2010). Information paper: ABS sources of disability information, Australia, 2003–2008, Cat. 4431.0.55.002. Australian Bureau of Statistics. [Online] Available at: https://www.ausstats.abs.gov.au/Ausstats/subscriber.nsf/0/69F4AB340D15511ACA25778900119EC6/$File/attqvre7.pdf
Hasan, M.T., Muhaddes, T., Camellia, S., Selim, N., and Rashid, S.F. (2014). Prevalence and experiences of intimate partner violence against women with disabilities in Bangladesh: results of an explanatory sequential mixed-method study. *Journal of Interpersonal Violence*, 29(17), 3105–26.
Hermans, H. and Evenhuis, H.M. (2014). Multimorbidity in older adults with intellectual disabilities. *Research in Developmental Disabilities*, 35(4), 776–83.
Horner-Johnson, W., et al. (2014). Breast and cervical cancer screening disparities associated with disability severity. *Women's Health Issues*, 24(1), e147–53.
Hughes, K., et al. (2012). Prevalence and risk of violence against adults with disabilities: A systematic review and meta-analysis of observational studies. *Lancet*, 379(9826), 1621–9.
Iezzoni, L.I., et al. (2016). Trends in colorectal cancer screening over time for persons with and without chronic disability. *Disability Health Journal*, 9(3), 498–509.
Iezzoni, L.I., et al. (2015). Physical accessibility of routine prenatal care for women with mobility disability. *Journal of Women's Health (Larchmt)*, 24(12), 1006–12.
International Centre for Evidence in Disability, et al. (2014). *Childhood Disability and Malnutrition in Turkana Kenya; A Summary Report for Stakeholders and Policy*. CRIN. [Online] Available at: https://www.lshtm.ac.uk/sites/default/files/2019-06/Malnutrition-and-Disability-Report-Full-Report.pdf
International Labour Office, et al. (2014). *Business as Unusual: Making Workplaces Inclusive of People with Disabilities*. Geneva: ILO.
Keogh, M. and Down, E. (2014). What are the enablers and barriers to implementation of the United Nation Convention on the Rights of Persons with Disabilities (CRPD) in the Indo-Pacific? CBM Australia.
Kim, Y., et al. (2015). Correlates of physical activity among people with disabilities in South Korea: A multilevel modeling approach. *Journal of Physical Activity & Health*, 12(7), 1031–8.
King, J.L., et al. (2016). Evaluation of a newly developed tobacco cessation program for people with disabilities. *Disability & Health Journal*, 9(1), 145–9.
Kleinitz, P., et al. (2011). *Disability Inclusive Practices in Eye Health Development.* Bensheim: CBM.
Kleinitz, P., et al. (2012). *2012: Barriers to and Facilitators of Health Services for People with Disabilities in Cambodia*. Canberra: AusAID Knowledge Hubs for Health.
Kostanjsek, N., et al. (2013). Counting disability: and global national estimation. *Disability & Rehabilitation*, 35(13), 1065–9.
Krahn, G.L., et al. (2015). Persons with disabilities as an unrecognized health disparity population. *American Journal of Public Health*, 105(Suppl 2), S198–206.
Krnjacki, L., et al. (2016). Prevalence and risk of violence against people with and without disabilities: findings from an Australian population-based study. *Australian & New Zealand Journal of Public Health*, 40(1), 16–21.
Krnjacki, L., et al. (2017). Disability-based discrimination and health: findings from an Australian-based population study. *Australian & New Zealand Journal of Public Health*.
Lai, B., et al. (2017). Current trends in exercise intervention research, technology, and behavioral change strategies for people with disabilities: a scoping review. *American Journal of Physical Medicine & Rehabilitation*, 96(10), 748–61.
Lamichhane, K. (2015). *Disability, Education and Employment in Developing Countries.* Cambridge: Cambridge University Press.

LaMontagne, A.D., et al. (2016). Psychosocial job quality in a national sample of working Australians: a comparison of persons working with versus without disability. *SSM Population Health*, **2**, 175–81.
Lee, J.E., et al. (2014a). Accessibility of medical services for persons with disabilities: Comparison with the general population in Korea. *Disability & Rehabilitation*, 36(20), 1728–34.
Lee, J.E., et al. (2014b). Smoking behaviors among people with disabilities in Korea. *Disability & Health Journal*, 7(2), 236–41.
Lee, K., et al. (2015). Sexual and reproductive health services for women with disability: a qualitative study with service providers in the Philippines. *BMC Women's Health*, 15(87).
Lindsay, S. (2011). Discrimination and other barriers to employment for teens and young adults with disabilities. *Disability & Rehabilitation*, 33(15–16), 1340–50.
Love, L., et al. (2018). Achieving disability equality: empowering disabled people to take the lead. *Social Inclusion*, 6(1), 1–8.
MacLachlan, M., et al. (2012). Inclusive health. *Tropical Medicine & International Health*, 17(1), 139–41.
MacLachlan, M. and Mannan, H. (2014). The World Report on Disability and its implications for rehabilitation psychology. *Rehabilitation Psychology*, 59(2), 117–24.
MacLachlan, M., et al. (2011). Access to health care of persons with disabilities as an indicator of equity in health systems. *Open Medicine*, 5(1), e10–e12.
MacLachlan, M. and Swartz, L. (2009). *Disability & International Development*. New York: Springer-Verlag.
MacRae, S., et al. (2015). Diabetes in people with intellectual disabilities: a systematic review of the literature. *Research in Developmental Disabilities*, **47**, 352–74.
Mactaggart, I., et al. (2016). Assessing health and rehabilitation needs of people with disabilities in Cameroon and India. *Disability and Rehabilitation*, 38(18), 1757–1764.
Madans, J.H. and Loeb, M. (2013). Methods to improve international comparability of census and survey measures of disability. *Disability and Rehabilitation*, 35(13), 1070–3.
Madans, J.H., et al. (2011). Measuring disability and monitoring the UN Convention on the Rights of Persons with Disabilities: the work of the Washington Group on disability statistics. *BMC Public Health*, 11 Suppl 4, S4.
Madans, J.H., et al. (2015). Comments on Sabariego et al. Measuring disability: comparing the impact of two data collection approaches on disability rates. *International Journal of Environmental Research & Public Health*, 13(1), ijerph13010065.
Madden, R.H. and Bundy, A. (2019). The ICF has made a difference to functioning and disability measurement and statistics. *Disability & Rehabilitation*, 41(12), 1450–62.
Mannan, H. and MacLachlan, M. (2013). Disability and health: a research agenda. *Social Inclusion*, 1(1), 37–45.
Marella, M., et al. (2016). Rapid assessment of disability in the Philippines: understanding prevalence, wellbeing and access to the community for people with disabilities to inform the W-DARE. *Population Health Metrics*, 14(26), 1–11.
Marmot, M., et al. (2010). *Fair Society, Healthy Lives: The Marmot Review*. London: Strategic Review of Health Inequalities in England PostT2010.
Mavromaras, K., et al. (2018). *Evaluation of the NDIS*. Adelaide: Flinders University.
Mavrou, K., et al. (2017). Opportunities and challenges related to ICT and ICT-AT use by people with disabilities: an explorative study into factors that impact on the digital divide. *Technology and Disability*, 29(1), 63–75.
McPherson, A., et al. (2017). Strengthening health information systems for disability-related rehabilitation in LMICs. *Healthy Policy and Planning*, 32(3), 384–94.
McVeigh, J., et al. (2016). Promoting good policy for leadership and governance of health related rehabilitation: a realist synthesis. *Globalization and Health*, 12(1), 1–18.

Merten, J.W., et al. (2015). Barriers to cancer screening for people with disabilities: a literature review. *Disability & Health Journal*, 8(1), 9–16.
Miksztai-Rethey, B. and Farago, K. (2018). An exploratory case study on letter-based, head-movement-driven communication. *Technology and Disability*, 29(4), 153–61.
Mikton, C., Maguire, H., and Shakespeare, T. (2014). A systematic review of the effectiveness of interventions to prevent and respond to violence against persons with disabilities. *Journal of Interpersonal Violence*, 29(17), 3207–26.
Mikulovic, J., et al. (2011). Prevalence of overweight in adolescents with intellectual deficiency. Differences in socio-educative context, physical activity and dietary habits. *Appetite*, 56(2), 403–7.
Milner, A., et al. (2015). Perceived fairness of pay among people with and without disabilities: a propensity score matched analysis of working Australians. *Scandinavian Journal of Work, Environment & Health*, 41(5), 451–9.
Milner, A., et al. (2017). Underemployment and its impacts on mental health among those with disabilities: evidence from the HILDA cohort. *Journal of Epidemiology & Community Health*, 71(12), 1198–202.
Mitra, M., et al. (2016). Disparities in adverse preconception risk factors between women with and without disabilities. *Maternal & Child Health Journal*, 20(3), 507–15.
Mitra, M., et al. (2015). Prevalence and risk factors for postpartum depression symptoms among women with disabilities. *Maternal & Child Health Journal*, 19(2), 362–72.
Mitra, M. and Mouradian, V.E. (2014). Intimate partner violence in the relationships of men with disabilities in the United States: Relative prevalence and health correlates. *Journal of Interpersonal Violence*, 29(17), 3150–66.
Molden, T.H. and Tossebro, J. (2010). Measuring disability in survey research: Comparing current measurements within one data set. *ALTER-European Journal of Disability Research/Revue Européenne de Recherche sur le Handicap*, 4(3), 174–89.
Mont, D. (2007). *Measuring Disability Prevalence*. Washington, DC: The World Bank.
Morgan, V.A., et al. (2012). People living with psychotic illness in 2010: The second Australian national survey of psychosis. *Australian & New Zealand Journal of Psychiatry*, 46(8), 735–52.
Mprah, W.K., et al. (2014). Does disability matter? Disability in sexual and reproductive health policies and research in Ghana. *International Quarterly of Community Health Education*, 35(1), 21–35.
National Institute of Statistics, et al. (2015). *Cambodia Demographic and Health Survey 2014*. [Online] Available at: https://dhsprogram.com/pubs/pdf/fr312/fr312.pdf
National Mental Health Consumer and Carer Forum (2011). *Unravelling Psychosocial Disability: A Position Statement of the National Mental Health Consumer and Carer Forum on Psychosocial Disability Associated with Mental Health Conditions*. Canberra: NMHCCF.
Needham, C. and Dickinson, H. (2017). 'Any one of us could be among that number': Comparing the policy narratives for individualized disability funding in Australia and England. *Social Policy & Administration*, 52(3), 731–49.
Neter, J.E., et al. (2011). The prevalence of overweight and obesity and its determinants in children with and without disabilities. *Journal of Pediatrics*, 158(5), 735–9.
Neyestani, T.R., et al. (2010). Nutritional status of the Iranian children with physical disability: a cross-sectional study. *Asia Pacific Journal of Clinical Nutrition*, 19(2), 223–30.
Nguyen, T.T.A., Liamputtong, P., Horey, D., Monfries, M. (2018). Knowledge of sexuality and reproductive health of people with physical disabilities in Vietnam. *Sexuality & Disability*, **36**, 3–18.
OECD (2010). *Sickness, Disability and Work: Breaking the Barriers*. [Online] Available at: https://www.oecd.org/publications/sickness-disability-and-work-breaking-the-barriers-9789264088856-en.htm

Oh, M.K., et al. (2012). Differences in obesity rates between people with and without disabilities and the association of disability and obesity: a nationwide population study in South Korea. *Journal of Preventive Medicine & Public Health*, 45(4), 211–18.
Okoro, C.A., et al. (2014). A triple play: Psychological distress, physical comorbidities, and access and use of health services among U.S. adults with disabilities. *Journal of Health Care for the Poor & Underserved*, 25(2), 814–36.
Oliver, M. (1981). A new model of the social work role in relation to disability. In: Campling, J. (ed.), *The Handicapped Person: A New Perspective for Social Workers*? London: RADAR.
Oliver, M. (1990). *The Politics of Disablement*. Basingstoke: Macmillan.
Oliver, M. (1996a). A sociology of disability or a disablist sociology?. In: Barton, L. (ed.), *Disability and Society: Emerging Issues and Insights*. New York: Longman.
Oliver, M. (1996b). *Understanding Disability*. Basingstoke: Macmillan.
Palenzuela, K. (2014). Questioning assumptions on disability, 2014. [Online] Available at: http://www.abc.net.au/rampup/articles/2014/05/01/3996019.htm
Palmer, M., et al. (2015). The economic lives of people with disabilities in Vietnam. *PLoS One*, 10(7), e0133623.
Palmer, M., et al. (2016). The cost of disability in a low income country, 2016. [Online] Available at: http://dx.doi.org/10.2139/ssrn.2856285
Pharr, J.R. and Bungum, T. (2012). Health disparities experienced by people with disabilities in the United States: a Behavioral Risk Factor Surveillance System study. *Global Journal Health Sciences*, 4(6), 99–108.
Piatta, F. and Guy, M. (2010). *Rights in Action—Good Practice for Inclusive Development*. Lyon: Handicap International.
Pinquart, M. (2017). Systematic review: Bullying involvement of children with and without chronic physical illness and/or physical/sensory disability: A meta-analytic comparison with healthy/nondisabled peers. *Journal of Pediatric Psychology*, 42(3), 245–59.
Pinquart, M. and Pfeiffer, J.P. (2015). Alcohol use among students with and without hearing loss. *Journal of Deaf Studies & Deaf Education*, 20(1), 82–90.
Pomeranz, J.L., et al. (2014). Creating a tobacco cessation program for people with disabilities: a community based participatory research approach. *Journal of Addiction Research and Therapy*, 5(204). doi:10.4172/2155-6105.1000204.
Productivity Commission (2011). *Disability Care and Support, Report no. 54*. Canberra: The Productivity Commission.
Pryor, W., et al. (2017). Compliance with standards of practice for health-related rehabilitation in low and middle-income settings: development and implementation of a novel scoring method. *Disability & Rehabilitation*, 41(19), 2264–71.
Puri, M., Misra, G., and Hawkes, S. (2015). Hidden voices: Prevalence and risk factors for violence against women with disabilities in Nepal. *BMC Public Health*, 15(261).
Ramjan, L., et al. (2016). Barriers to breast and cervical cancer screening for women with physical disability: a review. *Women Health*, 56(2), 141–56.
Sabariego, C., et al. (2015). Measuring disability: Comparing the impact of two data collection approaches on disability rates. *International Journal of Environmental Research & Public Health*, 12(9), 10329–51.
Saleh, M.C. and Bruyere, S.M. (2018). Leveraging employer practices in global regulatory frameworks to improve employment outcomes for people with disabilities. *Social Inclusion*, 6(1), 18–28.
Shakespeare, T. (2014). *Disability Rights and Wrongs Revisited* (2nd ed.). London: Routledge.
Shattuck, P.T., et al. (2012). Postsecondary education and employment among youth with an autism spectrum disorder. *Pediatrics*, 129(6), 1042–9.
Simonsen, M., et al. (2013). Youth and adults with psychiatric disabilities: Improving career and employment. In: Strauser, D.R. (ed.) *Career Development, Employment, and Disability in*

Rehabilitation: From Theory to Practice. New York: Springer Publishing Company, p. 401.
Siska, J., et al. (2018). Social inclusion through community living: current situation, advances, and gaps in policy, practice and research. *Social Inclusion*, 6(1), 94–109.
Skempes, D. and Bickenbach, J. (2015). Strengthening rehabilitation for people with disabilities: a human rights approach as the essential next step to accelerating global progress. *American Journal of Physical Medicine & Rehabilitation*, 94(10), 823–8.
Slevin, E., et al. (2014). Obesity and overweight in intellectual and non-intellectually disabled children. *Journal of Intellectual Disability Research*, 58(3), 211–20.
Steele, C.B., et al. (2017). Prevalence of cancer screening among adults with disabilities, United States, 2013. *Preventing Chronic Disease*, 14, E09.
Swaine, J.G., et al. (2014). Test of an intervention to improve knowledge of women with intellectual disabilities about cervical and breast cancer screening. *Journal of Intellectual Disability Research*, 58(7), 651–63.
Toldra, R.C. and Santos, M.C. (2013). People with disabilities in the labor market: Facilitators and barriers. *Work*, 45(4), 553–63.
Trani, J.F., et al. (2015). Mental illness, poverty and stigma in India: a case-control study. *BMJ Open*, 5(2), e006355.
UNESCAP (2015). *Disability at a Glance.* Review of the economic costs of exclusion of persons with disabilities in Asia and the Pacific. [Online] Available at: https://www.unescap.org/sites/default/files/SDD%20Disability%20Glance%202015_Final.pdf
United Nations (2006). *Convention on the Rights of Persons with Disabilities*. Geneva: United Nations.
van den Heuvel, R., et al. (2017). Robots supporting play for children with physical disabilities: exploring the potential of IROMEC. *Technology and Disability*, 29(3), 109–20.
Van Rooy, G., et al. (2012). Perceived barriers to accessing health services among people with disabilities in rural northern Namibia. *Disability & Society*, 27(6), 761–75.
Vaughan, C., et al. (2016). Building capacity for a disability inclusive response to violence against women and girls: Experiences from the W-DARE project in the Philippines. *Gender and Development*, 24(2), 245–60.
Vaughan, C., et al. (2015). W-DARE: a three-year program of participatory action research to improve the sexual and reproductive health of women with disabilities in the Philippines. *BMC Public Health*, 15(984).
World Health Organization (WHO) (1978). Declaration of Alma-Ata: International Conference on Primary Health Care. *WHO Chronicles*, 32(11), 428–30.
World Health Organization (WHO) (2002). *Towards a Common Language for Functioning, Disability, and Health: The International Classification of Functioning, Disability and Health*. Geneva: WHO.
World Health Organization (WHO) (2010). *Community-based Rehabilitation: CBR Guidelines*. Geneva: WHO. [Online] Available at: http://www.who.int/disabilities/cbr/en/
World Health Organization (WHO) (2011). *World Report on Disability*. Geneva: WHO.
World Health Organization (WHO) (2017a). *Model Disability Survey (MDS): Survey Manual*. Geneva: WHO.
World Health Organization (WHO) (2017b). *The Need to Scale Up Rehabilitation*. Geneva: WHO.
World Bank (2018a). Poverty & Equity Data Portal: Cambodia, 2018a. [Online] Available at: http://povertydata.worldbank.org/poverty/country/KHM
World Bank (2018b). *The World Bank in Cambodia.* [Online] Available at: http://www.worldbank.org/en/country/cambodia/overview
Yen, C.F., et al. (2012). Analysis of seasonal influenza vaccine uptake among children and adolescents with an intellectual disability. *Research in Developmental Disabilities*, 33(2), 704–10.

Yen, C.F. and Lin, J.D. (2010). Factors for healthy food or less-healthy food intake among Taiwanese adolescents with intellectual disabilities. *Research in Developmental Disabilities*, 31(1), 203–11.

Yen, C.F. and Lin, J.D. (2011). Factors influencing administration of hepatitis B vaccine to community-dwelling teenagers aged 12–18 with an intellectual disability. *Research in Developmental Disabilities*, 32(6), 2943–9.

10.8

Health of older people

Samir K. Sinha and Brittany Ellis

Introduction to health of older people

This chapter considers the phenomenon of population ageing, the unique health needs of older people, and specific approaches to providing health and social care for them. Ageing reflects biological, psychological, and social processes that occur across the life course, and at individual rates. Indeed, these processes are associated with ageing, but are not synonymous with chronological age. They vary between individuals according to many factors such as gender, ethnic background, lifestyle, socioeconomic, geographic location, and historical factors. This helps to explain why certain countries or populations are ageing faster than others.

How ‘old age’ has been defined continues to vary across settings, over time, and from person to person. In the past, the United Nations (UN) and other international agencies used 60 years or older (60+) to define ‘the elderly’, but more recently there has been a growing consensus that 65+, or even older ages, should be used in general to describe this segment of the population, while recognizing that disadvantaged populations may experience processes associated with ageing earlier.

While the process of ageing is universal, it is considered by many to be anything but uniform. Indeed, regardless of any current chronologically based descriptions, it is clear that people over 65 years of age are not one homogenous group and that there exists a great variance of experience, capacity, and health status between individuals at any age. As people age, some will age well and with little disability for significant periods of their later years while others may experience most of their later years as living with increasing frailty and a higher degree of dependence. It is essential to recognize this heterogeneity in defining the health and social needs of our ageing populations, designing and evaluating interventions, and planning for the future for this rapidly growing and increasingly diverse segment of our populations.

Demographic projections related to population ageing

The age of the world's population is rapidly increasing both structurally and numerically, and by 2050 the number of people aged 60 years or older worldwide is expected to exceed the number of people aged 15 years or younger. In 2019, there were an estimated 694 million people who were aged 65 years or older worldwide, representing 9% of the total population. This population is projected to number an estimated 1.55 billion persons, representing 15.9% of the total population in 2050 (UN DESA 2019).

Population ageing, which refers to an increasing median age, is occurring. In 2015, the median age of the world's population was estimated to be 30 years. The country with the youngest population by median age was Niger (14 years), and the countries with the oldest populations by median age were Germany, Italy, and Japan (46 years). By 2050, it is anticipated that the world median age will increase to 36 years, with Taiwan and the Republic of Korea projected to both reach a median age of 54 years (UN DESA 2017).

While much attention is focused around population ageing in high-income countries, this phenomenon is currently occurring far more rapidly in lower-income countries. Indeed, between 2017 and 2050, the number of older adults in high-income countries is expected to increase by 38%, which is much less than the expected 260% increase for lower-income countries (UN DESA 2017). In general, the more developed a country is, the more likely it is that population ageing will occur within an affluent environment, with higher levels of overall health and social care resources and infrastructure being available to support this phenomenon. However, in developing countries, ageing can still occur successfully even in the context of the emerging burden of chronic diseases, industrialization, and changing living patterns.

Figures 10.8.1 and 10.8.2 illustrate the global phenomenon of population ageing at the country level in the year 2020 and again in 2050. As can be appreciated across these two figures, a significant number of more developed and developing countries will be experiencing major demographic shifts over the coming decades with older people representing greater proportions of their ageing populations.

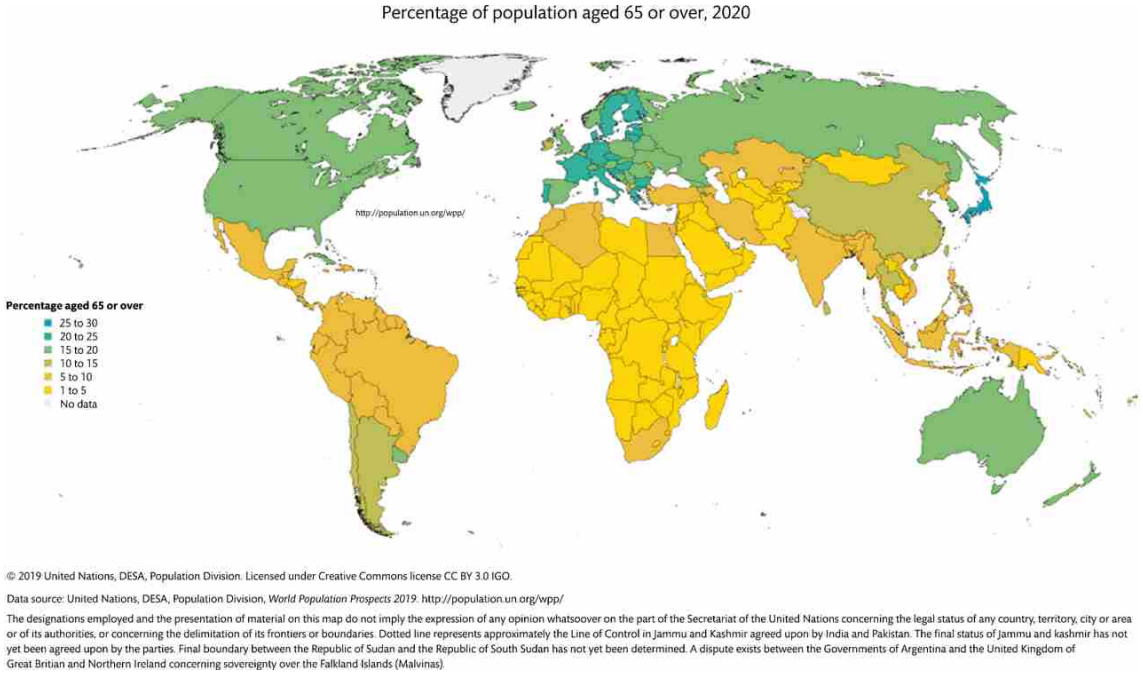


Fig. 10.8.1 Percentage of population aged 65 or over in 2020.

Source: data from United Nations (2019) Department of Economic and Social Affairs: Population Dynamics, available from https://population.un.org/wpp/Maps/.

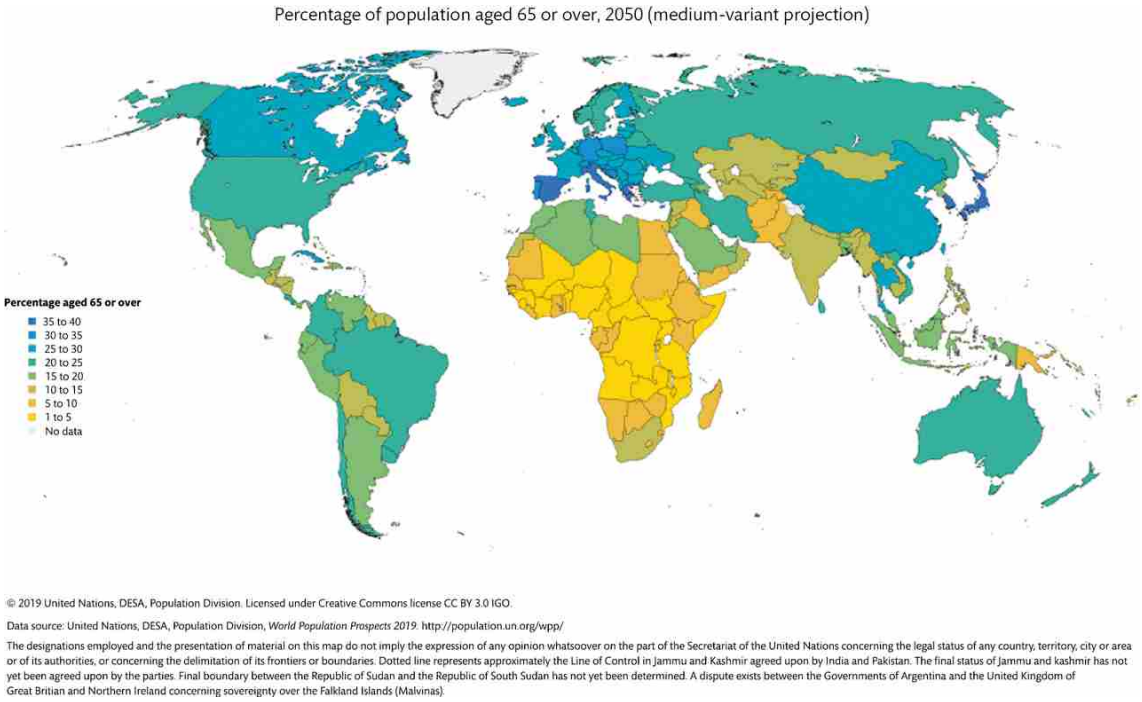


Fig. 10.8.2 Percentage of population aged 65 or over in 2050.
Source: data from United Nations (2019) Department of Economic and Social Affairs: Population Dynamics, available from https://population.un.org/wpp/Maps/.

Determinants of population ageing

The achievement of population ageing can be considered a major success story, having come largely as a result of gains in infant and child survival, and, more recently, to gains in life expectancy at older ages accompanied by reductions in fertility.

The earliest observed influence on population ageing occurred in Europe as a result of declines in fertility and childhood mortality which began in the nineteenth century and accelerated in the twentieth century initially due to improved living conditions and sanitation, and then to better nutrition, immunizations, and antibiotics. From the middle of the twentieth century, reductions in deaths from heart disease, stroke and cancer have further significantly extended adult life expectancy, primarily in developed countries.

In developing countries, fertility and mortality rates have also fallen, but to a lesser degree than among their more developed counterparts. The adoption of modern health practices, including successful infant and child health programmes, are probably most responsible for the rapid rates of mortality decline achieved at younger ages. As a result, globally, over two-thirds of adults over the age of 60

years are now living in developing or low- and middle-income regions (UN DESA 2017).

The dramatic declines in fertility being witnessed throughout the world have mainly been attributed to urbanization, industrialization, and social mobility which have all had an influence on family structures. In addition, the 'emancipation' of women resulting in their greater chances for education, later age of marriage, and access to contraception have all contributed as well (Kinsella and He 2009). China's one-child policy, introduced in 1979, also had a dramatic impact on fertility rates in that country which fell from 4.77 between 1970 and 1975 to 1.51 between 1995 and 2000 only increasing slightly to 1.60 in between 2010 and 2015—further contributing to its more accelerated rate of population ageing (UN DESA 2017a).

Currently, close to half of the world's people live in a region in which fertility rates are below what is considered the population replacement level of there being at least 2.1 children per woman (UN DESA 2019). Currently, the country with the highest fertility rate is Niger (7.2) and the lowest is Taiwan (1.2) while fertility rates are considered to be below replacement levels in North America (1.9), Europe (1.6), and many countries in Latin America, Asia and the Caribbean (UN DESA 2017a).

As fertility rates decline, life expectancy begins to have a greater impact on the rate of population ageing. While human life expectancy has increased dramatically over the twentieth century and the beginning of this century, new data is starting to show a deceleration and even more recently a demonstrable decline in life expectancy in some developed countries like Canada, the United States, and the United Kingdom. For example, in 2016 the average life expectancy in the United States of America reportedly decreased for a second year in a row, driven largely by a 50% gradual rise since 2005 in what are being coined 'deaths of despair'—the combined deaths from suicide, alcohol and drug use (Radley et al. 2018). Nevertheless, current conventional thinking predicts that our actual potential life-span as humans remains at around 120 years, with centenarians—those living to ages greater than 100 years—still representing the fastest growing age group in many developed countries.

In some populations, including Canada and the United States, international migration has also had a role in changing age (and gender) distributions, with large influxes of relatively young immigrants helping to influence population ageing rates. In the case of many of these countries, progressive immigration policies have been developed to help partially address the issue of declining fertility and population ageing rates. Furthermore, dramatic shifts have been seen in some nations as social and political conflicts have led to significant emigration. Nowhere was this more recently evident than in Syria where death and emigration have led to a population decrease by over three million people or roughly 20% of the country's population between 2010 and 2019 (UN DESA 2019). However, in most countries,

migration has had less of an overall impact than changing fertility and mortality rates.

Specific factors that further influence population ageing

Ageing and gender

Women have always tended to live longer than men, and globally, with the exception of Mali and Swaziland, this continues to be the case with the current average life expectancy of women 75.0 years and men 70.2 years (UN DESA 2019). The reasons behind this difference in life expectancy between men and women is complex and likely results from an interplay between differences in attitudes, social roles, biological risks, and particular behaviours including the higher incidence of alcohol and drug use and smoking in men, leading to their higher rates of accidents, violence, and cardiovascular disease. Many theories behind biological risk differences exist, however, all lack firm evidence. They include more active female immune functioning, the protective effect of oestrogen, compensatory effects of the second X chromosome, and the influence of oxidative stress (Austad 2006). Regardless, as a result of their tendency to live longer, women are more likely to experience a greater burden of the disease and disability associated with advanced age.

While conditions that affect people at older ages are similar among men and women (except for breast cancer and reproductive cancers), there are specific gender differences in the incidence of these conditions and their peak age of onset. For instance, men experience their peak incidence of coronary artery disease earlier than women. Lung cancer is more common among men who have long had higher rates of smoking, although lung cancer has now become the second most common cancer among women, reflecting their increasing smoking rates. Also, while there are few differences noted on self-rated health, older women have higher rates of reported functional disability than men, likely due to the higher prevalence of arthritis, asthma, depression, anxiety, and cognitive impairment that occur among women as they age.

Women continue to also be at increased risk of violence and abuse, both as children and adults, and these acts, which come in a variety of forms from domestic and sexual violence, to human trafficking, honour crimes and female genital mutilation, are associated with an increased incidence of anxiety, depression and post-traumatic stress disorder (Oram et al. 2017) and risk of premature death. It has been further argued that research in this area continues to be flawed as it often reports issues in a gender-neutral fashion, not recognizing the important differences

between men and women, while regularly excluding important variables such as intimate partner violence (Howard et al. 2017).

Gender differences in terms of evolving social roles and human rights will continue to have significant effects on current and future experiences of ageing for women as well. For example, during the life course women are still more likely to be excluded from schooling, receive lower quantities of nutritious food, have no control over household decisions, and receive no health insurance from an employer. Their greater longevity also means that they are more likely to experience the death of their spouse and to spend a greater proportion of their later years widowed, very often with an accompanying loss of social status and reduced economic circumstances. The highest levels of widowhood (above 40% among women aged 60–64) are found in some sub-Saharan countries that have experienced political unrest or high prevalence rates of HIV (UN DESA 2015). Furthermore, in parts of Africa, widowed women have no rights to own their late husband's land, which often leaves them in poverty. Older women are also especially vulnerable in the event of crises such as war, natural disaster, and famine, partly because they generally control fewer economic resources than older men, and also because much of the care of children and other vulnerable people tends to fall on older women.

Finally, historically, clinical research, including disease modelling, management, and pharmacologic treatment, has been conducted in predominantly younger male populations and had its findings generalized to women and older people. For example, despite the facts that women are twice as likely to live with two or more chronic diseases when over 80 years old, and that heart disease kills more women than men, protocols for diagnosing and treating heart disease have been largely based upon research with middle-aged, white, male patients with only 27% of participants being female (Kim et al. 2008). As a result, women tend to be diagnosed later with more advanced disease and have poorer treatment outcomes.

Ageing among ethnic minorities and indigenous peoples

With increasing numbers of immigrants reaching old age in their host countries, racial and ethnic minorities are now beginning to constitute the fastest growing segment of the older populations within many Western countries (Zubair and Norris 2015). These demographic changes will continue to have important implications, especially since morbidity and mortality rates among older populations have also been shown to be unevenly distributed across racial and ethnic groups. In the United Kingdom, for example, data from the Black and Minority Ethnic Elders Study demonstrated that older adults from ethnic minority groups, particularly those of African/Caribbean or South Asian ethnicity, were more likely to report worse health outcomes than the white ethnic majority population (PRIAE Policy

Research Institute on Ageing and Ethnicity 2005). Many older people in these ethnic groups are first-generation immigrants and may have encountered particular issues as they age that relate to migration stress, their reasons for having had to leave their country of birth, and having to cope with language limitations and the separation from family members and networks in their country of origin.

Noted differences for older ethnic minority populations in general have also been extended to physiological ageing, access to social and healthcare services, income and educational levels, health habits, disease prevalence, progression, and response to treatment. Similar disparities among health and disease in older people in ethnic minority groups are evident in both the Netherlands (Denktaş et al. 2009) and the US (Federal Interagency Forum on Aging-Related Studies 2016). Despite these recurrent findings, there remains a lack of understanding as to the exact causes of these differences as, historically, 'minority' ethnic groups have been excluded from studies or their heterogeneity diminished through being classified occasionally as part of broader 'non-majority' ethnic groups.

Other research has demonstrated that minority ethnic communities sometimes have little confidence that mainstream services will meet their diverse cultural, linguistic, or religious needs and uncertainty around what services exist and how to access them. The challenge of having public sector providers assuming, often incorrectly, that minority ethnic communities have strong community/family support networks that they can rely upon is an additional barrier towards accessing the care and support needed (PRIAE 2005).

Significant health disparities also exist among indigenous minority groups throughout the world especially where their continued lower socioeconomic status, as compared to the majority non-indigenous population, and persisting effects of colonialism, remain two of the key drivers of their disparities. Their achievement of a poorer overall health status has been linked to being disproportionately plagued by factors like poverty, malnutrition, inadequate housing, and overcrowding. This in turn has been exacerbated by the transition from a traditional to more modern lifestyle, systemic racism, and discrimination, and has led to high rates of chronic diseases including diabetes, cardiovascular disease, obesity, and mental health disorders and functional limitations at an earlier age than the general population.

Access to care issues for indigenous populations has been associated with a significant health divide especially characterized by a decreased life expectancy of between 5 and 10 years compared to the general population in Canada (Statistics Canada 2005) and of similar ranges in both New Zealand and Australia (Australian Bureau of Statistics (ABS) 2011; Phillips et al. 2017). Nevertheless, while the life expectancy gap is expected to continue to shrink over time in Canada, the numbers of older indigenous people greater than 65 years of age is expected to triple over the next two decades, compared to only a doubling of this age cohort for the overall population (Government of Canada 2009). Indeed, shifting demographics, coupled

with the more vulnerable health status of older indigenous peoples, will demand that their needs are responded to with increasingly culturally safe and appropriate health interventions to further address and reduce health disparities.

Ageing among LGBTQ peoples

While between an estimated five to ten per cent of the population identifies as being lesbian, gay, bisexual, transgender or queer (LGBTQ), many may have lived a portion or their entire lives, to varying degrees, in hostile environments, prior to experiencing more inclusive modern-day standards. For some, this could result in feelings of great stigma and shame that continue to shape their lives, often requiring them to keep their sexual orientation hidden as a strategy for survival. In part due to their increased experiences of victimization and discrimination, individuals who identify as LGBTQ often face greater health disparities than those who do not. Older LGBTQ persons are more likely to live in poverty and may also have fewer future caregiver options as they are more likely to live alone, are less likely to have children or grandchildren (Barker et al. 2006), or they may remain isolated from their families of origin. Furthermore, current or past negative experiences with health, social, and community care professionals may also lead to hesitation on the part of older LGBTQ adults to seek appropriate supportive services, leaving some with specific healthcare needs that also go largely unrecognized by healthcare professionals (Brotman and Ryan 2008).

There are certain medical issues that require particular attention in LGBTQ older adults such as regular screening for HIV and other sexually transmitted infections among gay and bisexual men; greater risks of cardiovascular disease due to higher rates of obesity and metabolic syndrome, smoking, and alcohol use for lesbian and bisexual women (Valanis et al. 2000); and a higher prevalence of psychological disorders and distress overall (Mays and Cochran 2001). This latter issue is particularly important as experiencing social stressors such as prejudice, stigmatization, violence, and internalized homophobia over long periods of time can result in higher risks of depression, suicide, risky behaviour, and substance abuse for LGBTQ older adults. These disparities will likely continue as they age, but it is unclear if these disparities will increase, remain stable, or decrease with time. Further recognition of these challenges and work in this area is needed.

Ageing in rural areas

In 2010 the World Health Organization (WHO) recognized that, for the first time in history, the majority of the world's population were now living in urban settings. As early as 1960 it was estimated that 66% of the global population lived rurally, but this has been steadily declining to 45% by 2018 (UN DESA 2018). Rapid

urbanization has led to difficulties in providing appropriate urban infrastructure as many urban centres struggle to manage issues related to pollution, sanitation, transportation, and the provision of health and education. Health challenges with rapid urbanization have included increases in rates of non-communicable diseases (cardiovascular disease, diabetes, cancers, and chronic respiratory disease), violence, physical inactivity, substance abuse, and communicable diseases.

The shift of younger generations to urban environments principally to find work, coupled with the greater movement of some older people into rural areas in retirement, has led to unusually high levels of population ageing in rural regions. Further, living in rural areas has been associated with greater economic hardship, less access to basic health and social services and quality housing, and higher commodity prices (Phillipson and Scharf 2005). Although older people in rural settings might report higher levels of social cohesion, they may live more undersupported lives as their younger family members, who previously may have acted naturally as unpaid caregivers for them, migrate to urban areas in search of greater economic opportunities.

Demographic and epidemiological transitions

Population ageing and changes in mortality and life expectancy have also been accompanied by changes in patterns of prevalent and incident diseases. Omran (1971) described the epidemiological transitions we have been experiencing over time as (i) the age of pestilence and famine; (ii) the age of receding pandemics; and (iii) the age of degenerative and man-made diseases. These transitions have been witnessed at an accelerated rate in rapidly developing countries like Brazil, China, and India, as they have experienced substantial changes in economic development, increasing urbanization, and changes in the health status of their citizens resulting in rapid population ageing.

Globally, non-communicable diseases, particularly cardiovascular diseases and cancers, now represent the current leading causes of morbidity and mortality with prevailing risk factors for developing these conditions including an unhealthy diet, physical inactivity, obesity, hypertension, and tobacco use (Li et al. 2012). Even though the proportion of later-life fatalities caused by heart disease, strokes, and cancer significantly fell over the last 25 years, this decrease was partially offset by a concurrent increase in the number of deaths attributable to other chronic non-communicable diseases, such as diabetes, chronic lung disease, renal disease, and dementia (Table 10.8.1).

Table 10.8.1 Ten leading causes of death of those aged 60 years or over globally, by sex, 2016

	Males			Females		
	Cause of death	**Deaths (thousands)**	**%**	**Cause of death**	**Deaths (thousands)**	**%**
1	Ischaemic heart disease	3,545,283	20.1	Ischaemic heart disease	3,708,354	21.1
2	Stroke	2,440,336	13.9	Stroke	2,828,193	16.2
3	COPD[i]	1,575,380	9.0	COPDi	1,311,550	7.5
4	Lung cancer[ii]	919,887	5.2	Alzheimer disease[iii]	972,608	5.6
5	Lower respiratory infections	823,234	4.7	Lower respiratory infections	911,094	5.2
6	Diabetes mellitus	558,075	3.2	Diabetes mellitus	700,998	4.0
7	Alzheimer disease[iii]	547,096	3.1	Hypertensive heart disease	453,076	2.6
8	Tuberculosis	431,954	2.5	Lung cancer[ii]	412,480	2.4
9	Kidney disease	402,816	2.3	Kidney disease	412,463	2.4
10	Stomach cancer	370,263	2.1	breast cancer	304,989	1.7

Data source: World Health Organization (2016). Global Health Estimates 2015: Deaths by Cause, Age, Sex, by Country and by Region, 2000–2015 Geneva

[i] Chronic obstructive pulmonary disease

[ii] Includes trachea and bronchus cancers

[iii] And other dementias

Reprinted with permission from United Nations, Department of Economic and Social Affairs, Population Division (2017). World Population Ageing 2017 (ST/ESA/SER.A/408).

Nevertheless, stark differences in mortality causes in older people between developed and developing regions do remain. For example, in higher-income countries over 90% of deaths among people aged 60 years or over are due to non-communicable diseases, whereas in lower-income countries almost 20% of deaths in this age group are still due to communicable diseases. Many rapidly ageing lower-income countries are also still experiencing high overall levels of communicable diseases including childhood infections and HIV/AIDS (WHO 2011).

Variations in life expectancy

While mortality rates tend to increase at older ages, in recent decades they have slowed at all ages. Life expectancy is defined as the average number of additional years of life that a person at a given age could expect to live, providing age-specific mortality levels remain constant. Table 10.8.2 shows life expectancy at age 60 for males and females in different regions of the world. In 2015, a woman in Africa who had already lived to 60 years could expect to live to 77.5 years, while one in Oceania could expect to live to 85.4 years. Similar differences are reflected in male life expectancy at age 60.

Table 10.8.2 Life expectancy at age 60, by sex, for the world and regions, 1950–1955, 2010–2015, and 2045–2050

	Females					Males				
	Life expectancy at age years (years)					Life expectancy at age years (years)				
	1950–1955	2010–2015	2045–2050	Change between 1950–1955 and 2010–2015 years	Change between 2010–2015 and 2045–2050 years	1950–1955	2010–2015	2045–2050	Change between 1950–1955 and 2010–2015 years	Change between 2010–2015 and 2045–2050 years
World	15.1	21.6	24.3	6.5	2.6	13.1	18.8	21.5	5.7	2.7
Africa	13.2	17.5	20.5	4.4	3.0	12.2	16.0	18.3	4.0	2.3
Asia	13.0	20.7	23.6	7.7	2.9	11.4	18.2	20.9	6.8	2.7
Europe	17.8	23.9	27.5	6.1	3.6	15.5	19.9	23.8	5.7	3.9
Latin America and the Caribbean	15.8	23.4	26.9	7.5	3.5	14.4	20.1	23.7	7.3	3.6
Northern America	19.1	24.9	28.0	5.8	3.1	16.0	21.9	26.3	4.3	4.4
Oceania	18.1	25.4	27.8	7.3	2.4	14.9	22.3	25.1	6.0	2.7

Data Source: United Nations (2017). World Population Prospects: The 2017 Revision

Reprinted with permission from United Nations, Department of Economic and Social Affairs, Population Division (2017). World Population Ageing 2017 (ST/ESA/SER.A/408).

Inequalities in life expectancy exist between countries and regions, and between different population groups within countries. Between-country reductions were largely achieved between 1950 and 1990 due to improved access to safe water and sanitation, childhood immunization, improved childhood nutrition, and declining fertility (Hosseinpoor et al. 2012). However, large inequalities in life expectancy still remain, with life expectancy remaining low at 63 years across lower-income countries (UN DESA Population Division 2017a).

Within-country differences in life expectancy have been further observed according to gender, socioeconomic status, and other factors. A major study of 1.7 million people across seven countries compared the impact of low socioeconomic status with other major risk factors on health, such as physical inactivity, smoking, diabetes, high blood pressure, obesity, and high alcohol intake. It found that when compared with their wealthier counterparts, people with low socioeconomic status were almost 1.5 times (46%) more likely to die before they were 85 years old—estimating that low socioeconomic status was associated with a reduced life expectancy of 2.1 years, similar to being inactive (2.4 years) and more impactful than respectively having high blood pressure, obesity and high alcohol consumption (1.6, 0.7, and 0.5 years) (Stringhini et al. 2017).

Healthy life expectancy

As well as adding years to life, it is also important to add life to years. Healthy life expectancy (or health-adjusted life expectancy (HALE)) estimates life expectancy, but with an adjustment for time spent living with a disability and/or other impairment to good health. Within the context of population ageing, age-associated limitations in performing everyday activities can have substantive effects on healthcare utilization. Healthy life expectancy estimations can therefore help answer questions raised regarding whether extra years of life achieved are being spent in poor or good health and their resulting associated impact around the demand for health and social care services.

The level of inequalities in both life and healthy life expectancies that can exist across a region is well studied. One study looking at these two data points across 25 countries in the European Union observed a greater variation in healthy life expectancy than for life expectancy. It found that, on average, a 50-year-old man could expect to live until 67.3 years free of activity limitations, and a woman to 68.1 years with observed variations in healthy life expectancy ranging from a low of 9.1 to a high of 23.6 years, respectively, for men in Estonia and Denmark. Higher healthy life expectancies were associated with higher gross domestic product, lower unemployment, and with life-long learning (Jagger et al. 2008).

When healthy life expectancy increases relative to life expectancy, then people will spend a greater proportion of their life in good health and a ‘compression of

morbidity' will be experienced (Fries et al. 1989). If people live longer due to a reduction in fatal illness, but with an increase in disability due to chronic or degenerative diseases, then there will be 'expansion of morbidity' experienced. A third scenario, that of 'dynamic equilibrium', could, for example, occur if the population prevalence of some chronic diseases increases with ageing, but the progression of some degenerative diseases reduces.

While the achievement of a compression of morbidity may be prevailing in some populations, data to test this hypothesis are difficult to find and hard to interpret. Very few countries are capable of making reliable assessments over time, between geographic regions or socioeconomic groups. In some settings, there is evidence that the prevalence of disability at older ages may be decreasing, but other studies are finding an increase in disability. For example, a study which assessed disability trends in 12 Organization for Economic Cooperation and Development (OECD) countries among individuals aged 65 years and over, reported mixed results (Lafortune et al. 2007). Disability declined among elderly people in five out of the 12 countries (Denmark, Finland, the Netherlands, the United States, and Italy). At the same time, three countries demonstrated increased rates of severe disability (Belgium, Sweden, and Japan), indicating that increased longevity can still lead to increased numbers of older people in poor health and in need of long-term care.

If a widespread compression of morbidity is to occur at a population level, greater health promotion and risk reduction activities will be needed for groups that have high levels of health risks. For example, preventing smoking is likely to result not only in more years of life, but also more years in good health. However, even if preventive activities were highly successful in reducing or postponing disability, it is expected that increasing numbers of people living to very old ages would mean that there will also be greater numbers of people eventually living with disabling conditions. At a population level, even if a compression of morbidity dominates, there will still be an increasing need for health and other forms of care as the population ages.

Differential, usual, and successful ageing

Ageing is believed to be a slow, continuous process. Therefore, some of its effects can be observed only when they have progressed sufficiently to induce identifiable alterations that can be validated by available testing methods. An illustrative example is atherosclerosis, the consequences of which manifest during middle and older ages even though an atherosclerotic lesion may start early in life. Whatever organ or tissue is considered, timetables of ageing represent an approximation, as the onset of ageing cannot be pin-pointed precisely by a specific physiological sign.

While a number of functions undoubtedly decline as we age, the extreme heterogeneity of an overall achieved functional status, even in the oldest cohorts, supports the view that ageing must be evaluated on an individual basis. In addition to genetic makeup, an older individual's health and social status, economic, and environmental conditions all contribute to how they age. Ageing processes have thus been distinguished across three distinct categories: categories (1) 'disease/disabled', characterized by the presence of pathology, disability, or both; (2) 'usual' (normal) ageing, characterized by the absence of overt pathology but the presence of functional declines; and (3) 'successful' ageing, characterized by few or none of the physiological losses seen in the usual ageing group. This concept of differential ageing is based on (1) the substantial heterogeneity that exists among older individuals; (2) the common observation that ageing is not a uniform or inevitable process of disease or disability; (3) the persistence of 'plasticity' well into advanced age, that is, a continuing capacity for adaptive and compensatory rehabilitation; and (4) the identification of predictors of various patterns of ageing, especially factors that contribute to more successful trajectories of ageing (Seeman 1994).

Extrinsic factors appear to also play an important role in age-associated functional decline. In one study, factors promoting the maintenance or improvement of functioning included younger age, higher income, higher education, high cognitive performance, being Caucasian, being of lower body weight, maintaining good lung function, living without diabetes or hypertension, participating in exercise, and receiving emotional support (Seeman 1994). Other correlates of improved survival at later ages even among those with chronic conditions were the absence of smoking, regular participation in physical activities, and the maintenance of a strong social network (Rizzuto et al. 2012). Taking all such factors into account, the prospects for avoidance, or eventual reversal, of functional loss with age are vastly improved, while the risks of adverse consequences are reduced (Rowe and Kahn 1987). Furthermore, it is increasingly appreciated that interventions may be beneficial at all ages provided they are customized to the individual. Indeed, physical exercise, dietary, surgical, and pharmacological interventions have proved successful in enhancing health even at older ages and thus should never be restricted on the basis of age alone. Our growing ability to identify those factors that can maximize health and physiological competence can support an overall enhancement of an older person's quality of life, a reduction in their burden of disease and disability, while decreasing the potential need for healthcare resources.

Societal perceptions of ageing

Images and stereotypes related to ageing

‘Ageism’ was defined in 1969 by US gerontologist Robert Butler as a combination of prejudicial attitudes towards ageing and older people, discriminatory practices against older people, and institutional practices and policies that perpetuate older age stereotypes (Butler 1969). A report by the American Association of Retired Persons (AARP) concluded that many Americans had misconceptions about ageing and older people, including that they are senile, asexual, and miserable (Abramson and Silverstein 2006). Although some changes have occurred for the better in terms of depictions of ageing, older adults are still generally viewed less favourably than the young.

Older people’s own perceptions of ageing

The Growing Old in America report (Taylor et al. 2009) explored ‘markers of old age’. The study reported that ‘turning 85’ was the most commonly endorsed marker of what it meant to be old, with the second and third markers being ‘can’t live independently’ and ‘can’t drive a car’, respectively. Other markers included ‘has trouble walking up stairs’, ‘retires from work’, and ‘has grey hair’. The English Longitudinal Study of Ageing found that self-assessed health was strongly related to how old individuals felt, with study participants often reporting being in better health and feeling younger than their actual age (Demakakos et al. 2007).

Interviews with older adults in Australia demonstrated the wide acceptance of ageist stereotypes. Participants aged 65 and 89 years described oldness as ‘not trying, withdrawn, isolated, irritating, self-oriented, living outside the mainstream, unattractive, uninteresting, frail, senile, silly, over the hill, narrow-minded, a burden, lonely, vulnerable, dowdy, and unproductive’ (Minichiello et al. 2000). Such ‘internalized ageism’ can, however, be dangerous in that older adults may, as a result, experience worse functional health and even a reduced will to live. Levy et al. (2002) found that older individuals who held more positive self-perceptions of ageing (measured over 20 years earlier) tended to live 7.5 years longer than individuals with less positive self-perceptions of ageing. This advantage remained after accounting for age, gender, socioeconomic status, feeling lonely, and level of functional health. Within the context of Western societies, where older age is portrayed more negatively, while the post-war ‘baby boomers’ are said to be challenging old age stereotypes, there still remains a high level of diversity in the health and social circumstances of these birth cohorts.

Evolving concepts of later-life productivity and social contribution

With a societal emphasis pointed towards the anticipated demands that the ageing population will make, the significant value of the contributions older people make through continued paid employment, but also as unpaid caregivers, volunteers, and mentors are easily overlooked. Indeed, the way in which population ageing is evolving and being experienced is helping to challenge well-established concepts. The '*old-age dependency ratio*', for example, is a common demographic metric which compares the ratio of the economically 'dependent' part of the population, or those 65 years or older, with those who are 20–64 years old, who are considered to be working age and therefore economically 'active'. The metric, while providing some utility, is increasingly leading to erroneous conclusions as it specifically underestimates the contributions of individuals over the age of 65 years where increasing numbers of older people are remaining in paid employment, or still acting as significant economic contributors through their spending, while also neglecting to consider that not all persons of traditional working ages are active in the labour force, with some being economically dependent. The concept also negates the value of the contributions older people who perform unpaid work, including organized volunteer positions, child-minding, and unpaid caregiving activities, all of which have significant overall economic and societal benefits and implications. Several alternative measures are now being proposed to better account for the growing diversity of capacities and dependencies across age (UN DESA 2017a).

Ageing-related concepts in health

Multimorbidity and polypharmacy

The prevalence of diseases and morbidity among older people has increased over time, partly due to the improved and earlier diagnosis of diseases such as hypertension, diabetes, and dementia, but also due to the rise of disabling non-fatal chronic conditions such as arthritis.

Multimorbidity is commonly defined as the co-occurrence of two or more chronic conditions whose prevalence increases with age and negatively associated with many important health outcomes such as increased health service utilization and decreased life expectancy. It also profoundly affects an individual's overall well-being, quality of life, and ability to function normally. Patients with multimorbidity further struggle to manage the associated polypharmacy and medication burden of having to take the multiple medications prescribed to them, leading to difficulties with treatment adherence and a worsened quality of life.

Despite the increasing numbers of individuals living with multimorbidity, clinical practice guidelines and care delivery models are still primarily built around

the management of single diseases rather than the interplay between them. Thus approaches to the care and treatment need to shift to become more relevant to the actual needs and preferences of patients with multimorbidity, which will require complex treatment decisions to be made sometimes in the absence of high-quality or direct evidence.

While medications can be beneficial in the management of specific health problems, they also have the potential to cause serious adverse effects. Older people have a reduced volume of distribution for drugs, and slower rates of drug metabolism and elimination, and so are more vulnerable to overdosing or the side effects of certain medications that can make them less appropriate to prescribe in older adults. Furthermore, polypharmacy increases the risk of adverse drug reactions or related events including hospitalization and death.

In high-income countries like Sweden a recent study reported a prevalence of polypharmacy of 44.0% (5+ medications), and excessive polypharmacy (10+ medications) of 11.7% among older adults (Morin et al. 2017); while another study from Canada noted that nearly 40% of older adults were currently taking one inappropriate medication while 12% were taking multiple inappropriate medications (CIHI 2014). Adverse drug reactions have been associated with around 30% of the hospital admissions among older people and it is estimated that up to half these adverse reactions could have been prevented (Simonson and Feinberg 2005). With the support of medication reviews that promote more appropriate prescribing and even de-prescribing that is aligned with an older person's actual treatment objectives and preferences, increased compliance with an overall treatment regimen could likely be achieved, accompanied by a concomitant decrease in the risk and occurrence of adverse events and their associated morbidity and mortality.

Disability and function

As people age they are more likely to experience challenges with their ability to function, which can impact their capacity to mobilize, stay socially engaged, and live independently. Activities of daily living (ADLs) and instrumental activities of daily living (IADLs) are two classes of functional measures that have been developed, based on an individual's ability to perform specific tasks, and are widely utilized to describe one's capacity for independent living. ADL tasks refer to those that are required to carry out basic self-care activities, such as transferring out of a bed or a chair, walking, toileting, bathing, dressing, and eating, while IADL tasks refer to those that are required to maintain an independent household, such as doing light housework, preparing meals, using the telephone, shopping for personal items, managing money, and getting around the community.

ADL dependency is less common than IADL dependency, and in one survey of community-dwelling population, 85 years of age and older, only 16% reported requiring help with one ADL, 10% with two or three ADLs, and 9% with four or more ADLs, while just under half reported needing help in performing one or more IADLs (Kane et al. 2003). The monitoring of an individual's IADL and ADL performance, which generally demands a combination of both cognitive and physical abilities, has also become a useful way to measure and monitor disease severity, progression of certain neurodegenerative conditions such as dementia, and in certain cases recovery as well such as after a stroke.

Finally, the WHO estimates that approximately 33% of adults over 65 years of age have a disabling hearing impairment, while 81% of the world's visually impaired people are aged 50 years and above (WHO 2019). Both hearing and visual impairments can impact a person's ability to function independently, and can also negatively affect their economic opportunities, quality of life, mental health, physical health, and mortality risk. Furthermore, the burden associated with dual sensory impairment, which is prevalent in between 5 and 20% of people over the age of 70 years, is likely more profound (Brennan and Bally 2007).

Frailty

Frailty is a term that is increasingly being used to describe older people who are at risk of poor outcomes due to their diminished adaptive reserve and resiliency. Fried et al. (2001) developed an operational definition of frailty—a 'frailty phenotype' that includes decreased resiliency and reserve, cycles of decline across multiple systems, negative energy balance, sarcopenia, and decreased strength and exercise tolerance. Another common way to define frailty, uses the concept of a 'frailty index' to include the accumulation of a broader range of limitations and deficits that accumulate exponentially with age and corresponds with an increasing level of vulnerability to stressors (Rockwood and Mitnisky 2011). While frailty is not age specific, it more commonly occurs with increasing age. Furthermore, at any given age, women on average compared to men have been found to be living with more deficits than do men. However, women appear to tolerate their deficits better than men, and appear to have a lower risk of adverse outcomes in association with higher frailty index scores.

Increasingly, frailty is becoming recognized as an important risk factor for increased morbidity and mortality across a range of clinical settings. From emergency departments, to inpatient medical and surgical stays, increasing frailty has been associated with an increased risk of experiencing falls, disability, hospitalization, the need for long-term care and mortality (Clegg et al. 2013). However, more work needs to be done to develop better ways to more easily

measure and detect frailty and determine the levels at which certain therapies or interventions may or may not be useful in an individual's care.

Healthcare for older people

As ageing populations carry the greatest overall burden of illness and disability, healthcare planners are often challenged by concerns about increasing demands and rising costs of health and social care in the face of population ageing. Older adults visit doctors more often, use more medications, have more admissions to hospital, tend to have longer stays in hospital and pose the most complex diagnostic and therapeutic problems, but also have the potential to benefit the most from the provision of appropriate and timely health and social care. As greater consideration is given about how and for whom treatment interventions should be extended to mitigate morbidity and mortality within older populations, overcoming ageism around treating illness at later ages will be just as important.

For many, health concerns and symptoms in the elderly are overlooked or dismissed as part of the normal ageing process and thus may go underdiagnosed and undertreated. While healthcare professionals do not necessarily intend to behave in a discriminatory fashion, it is clear that a lack of skills and confidence in working with older people can lead to behaviour which may be perceived as discriminatory. Furthermore, denying access to services on the basis of age alone is becoming increasingly recognized as not being acceptable. Decisions about treatment and healthcare should always be made on the basis of a patient's health needs, the overall health status of the individual, their own wishes and aspirations and, where appropriate, those of their caregivers, and the patient's ability to benefit.

Increasingly, strategies that can both promote healthy ageing, prevent, and treat functional decline and enable people to more successfully manage and live with their chronic health conditions and functional impairments are becoming recognized and understood. These can both improve an older person's longevity and their overall quality of life and help them compensate for functional impairments, maintain their independence, and prevent future complications. Furthermore, as the health needs of older people living with multimorbidity can be particularly complex to manage, their care should be approached when possible from a multidisciplinary perspective and include them and their families and caregivers as partners in their own health management to enable the delivery of more efficient and effective care.

Health promotion activities to support healthy ageing

A useful framework for promoting healthy ageing has been described in three stages by Kalache and Kickbusch (1997). The *first* is increasing capacity for health in early life and includes good maternal nutrition, and optimizing the development of vital organs such as brain, muscle, bone, and blood vessels during childhood and early adulthood to build resources that may affect later capacity (such as increasing peak bone density to protect against osteoporosis). The *second* stage occurs in adult life and involves strategies to reduce damage (such as avoiding smoking), to protect against damage (such as eating a balanced diet) or to prevent loss through lack of use (such as with regular physical activity). The *third* stage is in late life and involves the continued primary prevention and minimizing the secondary progression of disease and disability, protecting against increased environmental demands or stresses, and compensating for lost capacity. Examples of these approaches in health promotion include routine immunization, exercise and strength training and rehabilitation when necessary, maintaining and strengthening social supports, correction of deficits in vision and hearing, making modifications to domestic and outdoor environments to enable more independent living and trying to address as possible important social determinants of health such as poverty.

Indeed, several key health behaviours are associated with diseases that are common in older age and thus suggest a primary preventive approach is likely to be the most effective way to reduce their incidence. A common example of this is the prevention of smoking which is linked to cardiovascular and respiratory diseases, many cancers, and ultimately a decreased life expectancy. While older people may find it difficult to maintain high levels of exercise, regular physical activity is strongly associated with lower rates of obesity, cardiovascular disease, diabetes, depression, and osteoporosis, and an increased life expectancy in many observational studies. Observational studies also suggest an association between good nutrition and better health at older ages (Tucker et al. 2005) as well as the importance of staying socially connected with others and an increased life expectancy (Rizzuto et al. 2012)

Much of the evidence for promoting vaccinations among older adults comes from observational studies and may reflect other differences between vaccinated and non-vaccinated groups. As the ageing process is associated with immunosenescence and subsequently an increased risk of infection, as well as diminished response to vaccination, the value of still promoting vaccinations in later life becomes apparent. Accordingly, it is currently recommended that in addition to updating routine vaccinations that may have been missed in earlier life, routine vaccinations for all older people over 65 years should include: influenza, pneumococcal, shingles (zoster) and tetanus, diphtheria, and pertussis (TdaP of Td) (CDC 2019).

It is a common assumption that prevention in general—and screening in particular—is always in the patient's best interest. In young healthy populations,

benefits of early cancer detection and treatment are considered to outweigh its harms. In older populations, mortality benefits may be less pronounced, as the screening tests themselves may pose a greater threat to one's health, while the presence of comorbidities may make treatment options untenable (Eckstrom et al. 2013). Furthermore, older adults are often underrepresented in both cancer screening and treatment trials, making recommendations around either, especially older adults with multimorbidity, functional impairments, and shorter life expectancy more difficult.

Currently in the US routine screening for cervical and breast cancers in women is still recommended for those up to the age of 65 and 74 years, respectively; screening for colorectal cancer is recommended for all adults over the age of 50 and up to the age of 75 years depending on the person's health and comorbidities; meanwhile, in men 70 years and older there remains insufficient evidence to support prostate cancer screening using the prostate-specific antigen (PSA) test (US Preventive Services Task Force 2019). In any case appropriate testing should occur for appropriate symptomology in line with an individual's overall goals of care.

Specific conditions predominantly affecting older people

A number of conditions affect older people with greater frequency and impact than those who are younger. This includes dementias, falls and osteoporosis, incontinence, frailty, and loneliness, as well as elder abuse. Further, older adults are also at risk of diseases that affect all populations, including mental health conditions such as depression and anxiety. The unique manifestations and impact of these conditions in older people has long been neglected in research, and it is increasingly evident that such issues impact the lives of older people in significant ways.

Dementia

The growing global prevalence of dementia is partially the result of population ageing, combined with improving awareness and diagnosis of the disease. With the greatest risk factor for developing dementia, with Alzheimer's disease being the most common form, being age, women remain more likely to develop dementia than men, due to their greater longevity. It is currently estimated that there are more than 50 million people currently living with dementia worldwide, and that this figure will reach 131.5 million in 2050 (Alzheimer's Disease International 2019).

Advances in the prevention and management of cardiovascular diseases over the past few decades, and improvements in educational attainment, has likely contributed to a lower current prevalence of vascular and some others forms of dementia (Livingston et al. 2017). While there currently exists no cure for

dementias, key reasons to support early diagnosis include the ability to provide pharmacological and social interventions that can help manage the manifestations of dementia, and better support for involved caregivers and individuals with dementia to live with as much independence as possible.

Falls and osteoporosis

Falls are a common and major cause of disability in older adults, with both the risk of falls and the risk of injury from falls increasing with both age and frailty. Approximately 30% of community-dwelling older adults report sustaining a fall on an annual basis, while between 5 and 10% of falls result in serious injuries such as osteoporosis-related fractures (Gillespie et al. 2012). Falls are often attributed to a complex interaction of risk factors that include chronic and acute illnesses, mobility or balance issues, medications, and cognitive impairment. Falls are also a growing concern because injuries from falls can threaten an older person's mobility, independence, and confidence. A number of interventions have been shown to reduce the risk of falls, including health/environment risk factor screening/intervention programmes, muscle strengthening, and balance retraining, and home hazard assessment and modification (Gillespie et al. 2012). Furthermore, the increased recognition of osteoporosis as a preventable and treatable disease common in older men but especially older women would aid in the dissemination of proven medical therapies that could significantly reduce a person's future risk of sustain a fragility fracture.

Incontinence

While urinary and faecal incontinence are not diseases of older age *per se*, the likelihood of having these conditions significantly increases with age, especially in older adults. The prevalence of any degree of urinary incontinence in older adults is reported to be as high as 5.4% for men and 29% for women (Herschorn et al. 2008), while the overall prevalence of faecal incontinence is 15.3% (Whitehead et al. 2009). Incontinence impacts all parts of a person's, and by extension, their caregiver's life, including staying active and socially engaged, and can lead to increased social isolation, depression, physical deconditioning, falls, fractures, and skin breakdown. Despite the high prevalence of these issues, older adults often do not seek treatment due to embarrassment and social stigma, or because they assume this is a normal aspect of ageing, and left untreated, incontinence can worsen over time. Educating both the public and professionals on how to engage more around this issue will allow more persons with these issues to be supported.

Elder abuse and neglect

Elder abuse is defined as ‘a single, or repeated act, or lack of appropriate action, occurring within any relationship where there is an expectation of trust which causes harm or distress to an older person’ (WHO/INPEA 2002). It can take various forms: physical, verbal, psychological/emotional, sexual, and financial, and also reflect intentional or unintentional neglect, as well as self-neglect. The self-reported global annual prevalence rate of elder abuse in 2015 was found to be 15.7% affecting an estimated 140 million older people (Yon et al. 2017). Psychological and financial abuse were the most commonly reported forms (11.6% and 6.8%, respectively), with neglect, physical, and sexual abuse representing 4.2, 2.6, and 0.9% of cases (Yon et al. 2017).

The most likely persons to commit elder abuse are family members and friends which creates especially complex situations when an older person is also dependent on their abuser for having their living or care needs met. With population ageing, the likelihood exists that elder abuse will increase unless it is more comprehensively recognized and addressed, especially when it can result in higher rates of depression, anxiety, hospitalization, nursing home placement, and mortality (Lachs and Pillemer 2015).

Depression and anxiety

The WHO estimates 15% of people over the age of 60 live with mental health conditions such as depression and anxiety (WHO 2017), which in older people can present atypically and is likely under-recognized. There are a number of reasons to expect why people may experience poorer mental health in later life, including the presence of more potential psychosocial stressors such as financial stress, social isolation, a lack of social support and losses, as well as the increasing likelihood with age of experiencing chronic illnesses, multimorbidity, and living with frailty. It is worthwhile, to screen for depression and anxiety, as there are many effective non-pharmacological and pharmacological treatments.

Models of care to support older people

Primary care for older people

It is essential that the evolving needs of older people are reflected in the routine provision of primary care by ensuring its practitioners are appropriately trained to have a solid understanding of normal ageing as well as the common problems and issues including those that can present more atypically in older adults. In addition to becoming more aware of effective disease prevention interventions,

understanding the importance of screening for and how to manage common geriatric issues like dementia, falls, osteoporosis, incontinence, constipation, frailty, multimorbidity, and polypharmacy, will be essential towards the provision of effective primary care of older people. Indeed, the reality is that when older persons with complex needs are paired with the inadequate provision of primary care, this tends to lead to more physician visits, more prescribed medications, more visits to emergency departments, more hospital admissions, and worse overall outcomes (Nie et al. 2008.).

Traditional clinic-based primary care models will also need to adapt to better support older adults with more complex health and social needs to receive ongoing care by the same providers and where possible with an interprofessional team-based approach to care. The advent of primary care models like the 'patient-centred medical home' as well as the provision of 'house calls' are becoming increasingly recognized as a cost-effective alternative way of providing primary care for patients for whom this form of care is a necessity rather than a convenience that can not only improve healthcare experiences and outcomes but also their overall healthcare costs (Stall et al. 2014). Lastly, the advent of new technologies that support remote patient monitoring and the provision of 'virtual' care using various technologies is also gaining prominence as a way to better deliver care to frail older populations who are unable to access conventional models of care.

Specialty care for older adults

In more developed nations the medical specialities of geriatric medicine and psychiatry have emerged to meet the changing needs of particularly frail older adults across the continuum of care. These fields' increasing expertise lay in their ability to understand the differences between normal and abnormal ageing and how typical ways of approaching the care of older persons may need to be further adapted to achieve the most optimal outcome. The value of a Comprehensive Geriatrics Assessment (CGA), the main skilled activity geriatricians provide has also become well-established as a highly effective way to minimize overall functional decline, help older adult remain living in the community rather than in an institution and improve an older person's chance of survival and overall quality of life (Ellis et al. 2017).

Despite their growing perceived value as specialist providers, even in most developed countries an extreme shortages of geriatricians still exist. In Canada for example, while older Canadians began to outnumber younger Canadians in 2016, there are currently nine times the number of as paediatricians as there are geriatricians. As the specialties of geriatric medicine and psychiatry gain more traction and support for their growth, other medical subspecialty areas of care are quickly emerging like geriatric cardiology, surgery, oncology, and emergency

medicine. Furthermore, the advent of specialist gerontological nurses, social workers, therapists, and pharmacists is also occurring as well in order to fill the rapidly growing need for more specialist expertise in the care of older adults.

Long-term institutional and community-based care

Many older persons as they age will require some form of long-term care during the course of their life. A common approach until recently to address the growing need for long-term care associated with population ageing in middle and high incomes countries was to prioritize the provision of long-term care in more institutional-based residential care settings like nursing homes as opposed to through the provision of more home and community care. While there will always remain a role for the provision for some long-term care in institutional settings, technological advances to better support home diagnosis, treatment, and monitoring is enabling a much broader range of care to be provided in one's home. In high-income countries, as a result of evolving societal preferences and the realization that community-based care can be far more cost-effective that institutional-based care, there is a trend towards care being increasingly provided in community settings. In many countries, a growing shortage of available paid long-term care providers and available family caregivers is placing additional demands on other health services such as acute care hospitals when inadequate long-term care capacity exists. As a result, more funding is being prioritized to better support the development and provision of more cost-effective and sustainable home and community-based long-term care.

Acute care for older people

Older adults compared to the rest of the population, not only use hospital services more often, but also on a more urgent and critical basis and are at least twice as likely as younger populations to be admitted to the hospital with acute presentations (Keim and Sanders 2004). While older adults represent the greatest users of hospital services overall, a number of longitudinal studies have consistently demonstrated that only a small proportion of older adults are actually high users of hospital services. Furthermore, three factors appear to characterize these high users: multimorbidity, having pre-existing functional impairments that make one more dependent, and being socially frail in having inadequate social and community supports at home (Wolinsky et al. 1995).

In recognizing that typical older patients that present to hospitals are characterized with more complex and often inter-related health and social care issues, compared to younger patients, we can appreciate that the management of their acute care needs often necessitate a different overall approach. Indeed, frail

older adults have been found to be at greater risk of experiencing functional decline, delirium, longer lengths of stay, and a greater risk of becoming institutionalized or dying during or after a hospitalization. In response, the Acute Care for Elders (ACE) model of care has emerged as a response designed to improve functional outcomes and to improve the processes for the care of older patients. This model emphasizes: an environment of care designed to promote improved function for older patients; an interdisciplinary team that works together to identify/address the vulnerabilities of older patients; nursing care plans to for the prevention of disability; early planning to help prepare the patient to return home; and a review of medical care to prevent iatrogenic illness (Malone et al. 2015). Implementing specific models and point-of-care interventions and practices that encompass ACE principles in locations such as the emergency department, inpatient, transitional, outpatient, prehospital, or community settings can improve overall outcomes and reduce lengths of stay, admissions, readmissions, and inappropriate resource use (Sinha et al. 2018).

A growing interest in helping frail older patients to minimize their need to receive acute care in hospitals has led to the emergence of more proactive prehospital models of care such as ‘community paramedicine’ that have been shown to meaningfully reduce emergency calls and in turn emergency department visits and hospitalizations (Verma et al. 2017) or alternative delivery models of acute care such as ‘hospital at home’ can result in less clinical complications such as hospital-acquired infections, delirium and functional decline, lower mortality rates, and for less overall costs (Shepperd and Iliffe 2005). These two examples represent just some of the emerging acute care models that could allow both more and less developed countries with their growing and evolving acute care systems to re-think their traditional acute care approaches in order to promote the achievement of better outcomes through the better use of limited healthcare resources.

Rehabilitative care and reablement services

The loss of one’s functional abilities can represent the largest threat to an older individual’s independence which can often necessitate increased support to maintain one’s overall capacity for independent living. Although care that prioritizes ‘early mobilization’, can mitigate some degree of the functional loss that is commonly associated with episodes of acute or chronic illness, some frail older patients will inevitably suffer from a loss of function to such a degree that, without restorative interventions, they may not be able to return to their previous level of functional independence. Therefore, having timely access to targeted and appropriate preventive, and restorative care services is critical to minimizing the frequency and extent of functional loss experienced by at-risk older adults.

Traditionally, individuals needing more intensive rehabilitation could only receive it in hospitals or other institutional settings, but more recently community-based rehabilitation options like geriatric day hospitals, in-home and community-based therapy options including exercise and falls prevention programmes are being developed and demonstrating their ability to better prevent and manage function decline. Furthermore, the provision of traditional home care services for those with functional limitations is increasingly being delivered using a 'reablement' approach that recognizes the 'resiliency' of many and enables them re-learn or develop new skills to once again complete their functional tasks independently. Not only does this approach further enable ageing-in-place, but it has appropriately reduced the need for ongoing long-term care services and their associated costs (Tessier et al. 2016).

The growing importance of unpaid caregivers and social support

Within the context of ageing populations, the role of unpaid caregivers is paramount, but often still remains socially and politically invisible. In Canada, for example, half of all Canadians aged 15 and over have provided care to a family member or friend, at some point in their lives (Sinha 2013). The contributions of unpaid caregivers represent an important component of many national economies. In the United States, the provision of unpaid care has an estimated annual economic value of $470 billion, which in the same year was more than six times the total amount spent on paid home care and health services (Reinhard et al. 2015).

With current societal, political, and economic imperatives encouraging a shift away from institutional care and towards 'care at or closer to home' there exists an assumption that in addition to having enough accessible and affordable housing and community-based healthcare providers, families and friends will be willing and able to continue to provide a substantial level of unpaid care. However in many countries, as their overall populations age, their population of potential unpaid caregivers is predicted to shrink. What is further complicating matters is that constantly shifting economic opportunities and the growing ease of mobility is contributing to the growing distance between family members and lessening their potential to be family caregivers.

While both caregivers and care recipients generally regard home-based care as preferable to institutionalization, caregiving can be physically, emotionally, and financially taxing. Legg et al. (2013) reported an association between provision of care and self-reported poor health of caregivers, that also increased with the amount of care being provided. Women, who tend to spend more of their time than others providing unpaid care, are also more likely to feel the financial effects of caregiving, contributing to higher rates of later-life poverty.

The tradition of being able to rely on family support to care for older people is changing in many countries as a result of changing economic and political conditions, the increased participation of women in paid work, and the impact of certain diseases. Economic development, urbanization, and internal and international migration, leading to demographic transitions and changing social norms are all significantly impacting on the traditional provision of intergenerational support. In what some view as an extreme response to this challenge, a new Chinese 'Elderly Rights Law' was instituted in 2013 requiring grown children to provide physical and emotional care for their elderly parents, including a requirement to visit their parents regularly and look after them financially, or face fines or jail sentences. HIV/AIDS has so significantly affected middle-aged populations in many regions that it has left both older people and children without support and further led to the care of orphaned children being undertaken by their grandparents.

There is a growing recognition that caregivers must ideally be able to access training, services, and financial supports to enable and encourage them to take on and maintain caregiving roles that enables the delivery of more personalized care and enable lower health system expenditures. The UK, New Zealand, and Australia have all implemented national strategies to better support caregivers, which include information services, financial supports, community supports, legislation to allow caregivers to request flexibility in the workplace, and working with employers to show the financial and organizational benefits of better supporting caregivers in the workplace. Indeed, recognizing the important role of unpaid caregivers, and ensuring that their evolving needs are taken into account in planning and allocating resources will be needed to enable the care of ageing populations.

Age-friendly cities and communities

The WHO over the past decade has been actively promoting the creation of age-friendly cities and communities to promote a more thoughtful approach that could promote the health and well-being of people of all ages across the life course. An age-friendly community is defined as one that recognizes the great diversity among older persons, promotes their inclusion and contributions in all areas of community life, respects their decisions and lifestyle choices, and anticipates and responds flexibly to ageing-related needs and preferences (WHO 2007). Through its work, the WHO has identified eight domains of community life that influence the quality of life and health of older persons and people of all ages: respect and social inclusion; social participation, communication, and information; civic participation and employment; outdoor spaces and buildings; transportation, housing, and community support and health services. The consideration of these eight domains

into the design and development of communities, has supported the needed transformation of communities into age-friendly ones that encourage active and healthy ageing by optimizing opportunities for health, participation, and security.

The development of the WHO *Global Network for Age-friendly Cities and Communities* was recently identified as one of its ten priorities to enable healthy ageing. It currently has over 500 member cities and communities across 37 countries. Among communities that have taken part in age-friendly community development activities at various levels, many have learned to assess their level of 'age-friendliness', how to integrate ageing perspectives into urban planning efforts, and how to create age-friendly environments that consider not only design and physical features, but also infrastructure, policies, and service provision. Further increased concern towards ensuring older people have appropriate and affordable housing and transport options and that pedestrian safety and safety from violence and crime are better addressed will ensure that the social involvement and the health and well-being of older people can be better supported.

Public health implications of population ageing

As seen in this chapter, population ageing is a major global phenomenon that will continue to have a major impact on disease and disability patterns and on the use of health and social services. A coordinated and appropriate public health response, therefore, is essential if this phenomenon is to translate into an opportunity for longevity and healthy ageing, rather than a threat to global health and resources.

Public health is important in monitoring trends and transitions in risk factors, disease, and disability, encouraging effective health promotion practices that promote healthy ageing, and in setting policy for equitable and efficient use of healthcare resources for this growing population that requires different and more customized approaches to their care. The fact that vaccinations rates among the elderly lag far behind those of children speaks to just some of the significant public health gaps that currently exist. For example, no country other than South Korea has achieved the WHO-recommended target influenza vaccination rate of 75% for older adults (National Institute on Ageing 2018).

Health promotion to support healthy ageing begins in early life and continues throughout the life course to the oldest ages. As noted earlier, exercise, not smoking and maintaining strong social networks can allow older adults to live significantly longer that others who do not engage in these practices. Increasingly, it is also becoming recognized that dementia is not an inevitable consequence of ageing and that nine potentially modifiable health and lifestyle factors from different phases of life that, if eliminated, might prevent the development of dementias (Livingston et

al. 2017). Much more knowledge is needed as to how we can optimize the conditions and experiences of ageing.

Finally, better acknowledgement and recognition of the continued contributions made by older people themselves, as well as unpaid caregivers, to their families and communities and national economies is also of fundamental importance in addressing ageism and better meeting the challenges of an ageing population and their associated health needs.

Summary

Ageing is universal, unavoidable, and not uniform.

Population ageing is occurring at far greater rates in low-and middle-income countries than in high-income countries, and is placing considerable strains on health and social care sectors in varying ways.

Increased longevity has not been experienced uniformly across socioeconomic and ethno-cultural groups or in different countries.

Severe disability is becoming less common in successive cohorts of older people indicating that a compression of morbidity and disability is occurring, although not with certainty in a growing number of cases.

Advances in healthcare for older people have more recently had dramatic effects in improving survival and reducing disability associated with the inevitable increases in chronic diseases at older ages, but has also allowed the emergence of new chronic diseases like Alzheimer's and other related and incurable dementias that now appears to be emerging as one of the greatest global challenges for health and social care in the twenty-first century.

Families and increasingly friends in all countries are increasingly providing a major part of the care needed by older people—as unpaid caregivers—but do require more knowledge and support of formal health and social services.

Public health has major roles in surveillance, disease prevention and the promotion of healthy ageing, and in the development of health and social services for older people.

REFERENCES

Abramson, A. and Silverstein, M. (2006). *AARP Images of Aging in America, 2004*. Washington, DC: AARP and the University of Southern California.

Alzheimer's Disease International (2019). *Dementia Statistics*. [Online] Available at: http://www.alz.co.uk/research/statistics

Austad, S. (2006). Why women live longer than men: sex differences in longevity. *Gender Medicine*, 3(2), 79–92.

Australian Bureau of Statistics (2011). *Deaths, Australia, 2010*. Canberra: ABS.

Barker, J.C., Herdt, G., and de Vries, B. (2006). Social support in the lives of lesbians and gay men at midlife and later. *Sexuality Research & Social Policy: A Journal of the NSRC* 3(2), 1–23.
Brennan, M. and Bally, S.J. (2007). Psychosocial adaptations to dual sensory loss in middle and late adulthood. *Trends in Amplification*, 11(4), 281–300.
Brotman, S. and Ryan, B. (2008). Healthy aging for gay and lesbian seniors in Canada: an environmental scan. [Online] Available at: http://lgbtqpn.ca/wp-content/uploads/woocommerce_uploads/2014/08/Healthy%20Aging%20Environmental%20Scan.pdf
Butler, R. (1969). Age-ism: another form of bigotry. *The Gerontologist*, 9(4), 243–6.
Canadian Institute for Health Information (CIHI) (2014). *Drug Use Among Seniors on Public Drug Programs in Canada, 2012*. Ottawa: CIHI.
Centers for Disease Control and Prevention (CDC) (2019). *Recommended Adult Immunization Schedule for Ages 19 Years or Older*. Atlanta: CDC. [Online] Available at: https://www.cdc.gov/vaccines/schedules/downloads/adult/adult-combined-schedule.pdf
Clegg, A., Young, J., Iliffe, S., Rikkert, M.O., and Rockwood, K. (2013). Frailty in elderly people. *The Lancet*, 381(9868), 752–62.
Demakakos, P., Gjonca, E., and Nazroo, J. (2007). Age identity, age perceptions and health. *Annals of the New York Academy*, 1114(1), 279–87.
Denktaş, S., Koopmans, G., Birnie, E., Foets, M., and Bonsel, G. (2009). Ethnic background and differences in health care use: a national cross-sectional study of native Dutch and immigrant elderly in the Netherlands. *International Journal for Equity in Health*, **8**, 35.
Eckstrom, E., Feeny, D.H., Walter, L.C., Perdue, LA., and Whitlock, E.P. (2013). Individualizing cancer screening in older adults: a narrative review and framework for future research. *Journal of General Internal Medicine*, 28(2), 292–8.
Ellis, G., et al. (2017). Comprehensive geriatric assessment for older adults admitted to hospital. *Cochrane Database of Systematic Reviews* 9, 1–82.
Federal Interagency Forum on Aging-Related Statistics (2016). *Older Americans 2016: Key Indicators of Well-Being*. Washington, DC: US Government Printing Office.
Fried, L.P., et al. (2001). Frailty in older adults: evidence for a phenotype. *The Journals of Gerontology*, 56A(3), M146–56.
Fries, J., Green, L., and Levine, S. (1989). Health promotion and the compression of morbidity. *The Lancet* 333 (8636), 481–3.
Gillespie, LD., et al. (2012). Interventions for preventing falls in older people living in the community. *Cochrane Database of Systematic Reviews*, (2), CD007146.
Government of Canada, Special Senate Committee on Aging (2009). *Canada's Aging Population: Seizing the Opportunity*. Otttawa: Government of Canada.
Herschorn, S., Gajewski, J., Schulz, J., and Corcos, J. (2008). A population-based study of urinary symptoms and incontinence: the Canadian Urinary Bladder Survey. *BJU International*, 101(1), 52–8.
Hosseinpoor, A., Harper, S., Lee, J., Lynch, J., Mathers, C., and Abou-Zahr, C. (2012). International shortfall inequality in life expectancy in women and men, 1950–2010. *Bulletin of the World Health Organization*, 90(8), 588–94.
Howard, L., Ehrlich, A., Gamlen, F., and Oram, S. (2017). Gender-neutral mental health research is sex and gender biased. *Lancet Psychiatry*, 4(1), 9–11.
Jagger, C., et al. (2008). Inequalities in healthy life years in the 25 countries of the European Union in 2005: a cross-national meta-regression analysis. *The Lancet*, 372(9656), 2124–31.
Kalache, A. and Kickbusch, I. (1997). A global strategy for healthy ageing. *World Health*, 50(4), 4–5.
Kane, R.L., Ouslander, J.C., and Abrass, I.B. (2003). *Essentials of Clinical Geriatrics* (5th ed.). New York: McGraw Hill.
Keim, S. and Sanders, A. (2004). Geriatric emergency department use and care. In: Meldon, S., Ma, O.J., and Woolard, R. (eds.) *Geriatric Emergency Medicine*. New York: McGraw Hill, pp. 1–4.
Kim, E.S., Carrigan, T.P., and Menon, V. (2008). Enrollment of women in national heart, lung, and blood institute-funded cardiovascular randomized controlled trials fails to meet current federal mandates for inclusion. *Journal of the American College of Cardiology*, **52**, 672–3.
Kinsella, K. and He, W. (2009). *US Census International Population Reports. An Ageing World: 2008.* Washington, DC: US Census Bureau.
Lachs, M.S. and Pillemer, K.A. (2015). Elder abuse. *New England Journal of Medicine*, **373**, 1947–56.

Lafortune G., Balestat G., and the Disability Study Expert Group Members (2007). *Trends in Severe Disability Among Elderly People: Assessing Evidence in 12 OECD Countries and Future Implications*. Paris: OECD.

Legg, L., Weir, C.J., Langhorne, P., Smith, L.N., and Stot, D.J. (2013). Is informal caregiving independently associated with poor health? A population-based study. *Journal of Epidemiology and Community Health*, 67(1), 95–7.

Levy B., Slade M., Kunkel S., and Kasl S. (2002). Longevity increased by positive self-perceptions of aging. *Journal of Personality and Social Psychology*, 83(2), 261–70.

Li, L., Guo, Y., Chen, Z., Chen, J., and Peto, R. (2012). Epidemiology and the control of disease in China, with emphasis on the Chinese Biobank Study. *Public Health*, 126(3), 210–13.

Livingston G., et al. (2017). Dementia prevention intervention, and care. *The Lancet*, 390(10113), 2673–734.

Malone, M.L., Capezuti, E.A., and Palmer, RM. (2015). *Acute Care for Elders: A Model for Interdisciplinary Care*. New York: Springer.

Mays, V.M., and Cochran, S.D. (2001). Mental health correlates of perceived discrimination among lesbian, gay, and bisexual adults in the United States. *American Journal of Public Health*, 91(11), 1869–76.

Minichiello, V., Browne, J., and Kendig, H. (2000). Perceptions and consequences of ageism: views of older people. *Ageing and Society*, 20(3), 253–78.

Morin, L., Johnell, K., Laroche, M.L., Fastboom, J., and Wastesson, J.W. (2017). The epidemiology of polypharmacy in older adults: register-based prospective cohort study. *Clinical Epidemiology*, **10**, 289–98.

National Institute on Ageing (2018). *The Underappreciated Burden of Influenza Amongst Canada's Older Population. And What We Need to Do About It*. Toronto: National Institute on Ageing.

Nie, J.X., Wang, L., Tracy, C.S., Moineddin, R., and Upshur, R.E. (2008). Health care service utilization among the elderly: findings from the Study to Understand the Chronic Condition Experience of the Elderly and the Disabled (SUCCEED project). *Journal of Evaluation in Clinical Practice*, 14(6), 1044–9.

Omran, A.R. (1971). The epidemiologic transition: a theory of the epidemiology of population change. *The Milbank Memorial Fund Quarterly*, 49(4), 509–38.

Oram, S., Khalifeh, H., and Howard, L.M. (2017). Violence against women and mental health. *Lancet Psychiatry*, 4(2), 159–70.

Phillips, B., Daniels, J., Woodward, A., Blakely, T., Taylor, R., and Morrel, S. (2017). Mortality trends in Australian Aboriginal peoples and New Zealand Māori. *Population Health Metrics*, 15(25).

Phillipson, C. and Scharf, T. (2005). Rural and urban perspectives on growing old: developing a new research agenda. *European Journal of Ageing*, 2(2), 67–75.

PRIAE Policy Research Institute on Ageing and Ethnicity (2005). *Black and Minority Ethnic Elders' in the UK: Health and Social Care Research Findings*. Leeds: PRIAE.

Radley, D.C., McCarthy, D., and Hayes, D.L. (2018). *2018 Scorecard on State Health System Performance*. New York: The Commonwealth Fund.

Reinhard, S.C., Feinberg, L.F., Choula, R., and Houser, A. (2015). *Valuing the Invaluable: 2015 Update*. Washington, DC: AARP Public Policy Institute.

Rizzuto, D., Orsini, N., Qiu, C., Wang H.X., and Fratiglioni, L. (2012). Lifestyle, social factors, and survival after age 75: population based study. *British Medical Journal*, **345**, 1–10.

Rockwood, K. and Mitnisky, A. (2011). Frailty defined by deficit accumulation and geriatric medicine defined by frailty. *Clinics in Geriatric Medicine*, 27(1), 17–26.

Rowe, J.W. and Kahn, R.L. (1987). Human aging: usual and successful. *Science*, 237(4811), 143–9.

Seeman, T.E. (1994). Successful aging: reconceptualizing the aging process from a more positive perspective. In: Vallas, B.J., Albarede, J.L., and Garry, P.J. *Epidemiology and Aging: Facts and Research in Gerontology*. New York: Springer, pp. 61–73.

Shepperd, S. and Iliffe, S. (2005). Hospital at home versus inpatient hospital care. *Cochrane Database of Systematic Reviews*, 20(3), CD000356.

Simonson, W. and Feinberg, J. (2005). Medication-related problems in the elderly: defining the issues and identifying solutions. *Drugs & Aging*, 22(7), 559–69.

Sinha, M. (2013). *Portrait of Caregivers, 2012*. Ottawa: Statistics Canada, Social and Aboriginal Statistic Division.

Sinha, S.K., Bennett, J., Ramsden, R., Bon, J., and Chalk, T. (2018). Delivering improved patient and system outcomes for hospitalized older adults through an acute care for elders strategy. *Health Management Forum*, 31(4), 126–32.

Stall, N., Nowaczynski, M., and Sinha, S.K. (2014). A systematic review of outcomes from home-based primary care programs for homebound older adults. *Journal of the American Geriatrics Society*, 62(12), 2243–51.

Statistics Canada, Demography Division (2005). *Projections of the Aboriginal Populations, Canada, Provinces and Territories, 2001 to 2017*. Ottawa: Statistics Canada.

Stringhini, S., et al. (2017). Socioeconomic status and the 25x25 risk factors as determinants of premature mortality: a multicohort study and meta-analysis of 1.7 million men and women. *The Lancet*, 389(10075), 1229–37.

Taylor, P., Parker, K., Wang, W., Morin, R., and Cohn, D.V. (2009). *Growing Old in America: Expectations vs. Reality*. Washington, DC: PEW Research Center.

Tessier, A., Beaulieu, M-D., Mcginn, C.A., and Latulippe, R. (2016). Effectiveness of Reablement: Systematic Review. *Health Policy*, 11(4), 49–59.

Tucker, K.L., Hallfrisch, J., Qiao, N., Muller, D., Andres, R., and Fleg, J.L. (2005). The combination of high fruit and vegetable and low saturated fat intakes is more protective against mortality in aging men than is either alone: the Baltimore Longitudinal Study of Aging. *The Journal of Nutrition*, 135(3), 556–61.

United Nations Department of Economic and Social Affairs, Statistics Division (UN DESA) (2015). *The World's Women 2015: Trends and Statistics*. New York: UN DESA.

United Nations Department of Economic and Social Affairs, Population Division (UN DESA) (2017). *World Population Ageing*. New York: UN DESA.

United Nations Department of Economic and Social Affairs, Population Division (UN DESA) (2017a). *World Population Prospects: The 2017 Revision*. New York: UN DESA.

United Nations Department of Economic and Social Affairs, Population Division (UN DESA) (2018). *World Urbanization Prospects: The 2018 Revision*. New York: UN DESA.

United Nations Department of Economic and Social Affairs, Population Division (UN DESA) (2019). *World Population Prospects: The 2019 Revision*. New York: UN DESA.

US Preventive Services Task Force (2019). Published final recommendations. [Online] Available at: https://www.uspreventiveservicestaskforce.org/BrowseRec/Index/browse-recommendations

Valanis, B.G., Bowen, D.J., Bassford, T., Whitlock, E., Charney, P., Carter, R.A. (2000). Sexual orientation and health: comparisons in the women's health initiative sample. *Archives of Family Medicine*, 9(9), 843–53.

Verma, A., et al. (2017). Paramedic-initiated home care referrals and use of home care and emergency medical services. *Prehospital Emergency Care*, **15**, 1–6.

Whitehead, W., et al. (2009). Fecal incontinence in U.S. adults: epidemiology and risk factors. *Gastroenterology*, 137(2), 512–17.

Wolinsky, F.D, Stump, T., and Johnson, R. (1995). Hospital utilization profiles among older adults over time: consistency and volume among survivors and decedents. *Journals of Gerontology Series B, Psychological Sciences and Social Sciences*, 50(2), S88–100.

World Health Organization (WHO) (2007). *Global Age-friendly Cities: A Guide*. Geneva: WHO.

World Health Organization (WHO) (2011). *Global Status Report on Noncommunicable Diseases, 2010*. Geneva: WHO.

World Health Organization (WHO) (2017). *Fact Sheet on Mental Health of Older Adults*. Geneva: WHO.

World Health Organization (WHO) (2019). *Fact Sheets on Blindness and Vision Impairment & Deafness and Hearing Loss*. Geneva: WHO.

WHO/INPEA (2002). *Missing Voices: Views of Older Persons on Elder Abuse*. Geneva: WHO.

Yon Y., Mikton CR., Gassoumis ZD., and Wilber KH. (2017). Elder abuse prevalence in community settings: a systematic review and meta-analysis. *The Lancet* 5 (2), e147–56.

Zubair, M. and Norris, M. (2015). Perspectives on ageing, later life and ethnicity: ageing research in ethnic minority contexts. *Ageing and Society*, 35(5), 897–916.

10.9

Forced migrants and other displaced populations

Catherine R. Bateman Steel and Anthony B. Zwi

Introduction to forced migrants and other displaced populations

Migration and displacement have always been features of human life but the intensification of globalization, and increased ease of travel and interconnectedness, has seen the effects of displacement reaching far beyond the immediate contexts in which they occur. Forced migration is displacement that involves coercion and the implication that those involved have moved against their will. The previous few decades with the end of the cold war, rapid globalization, increasing inequality between and within nations, ongoing wars, and conflict (United Nations High Commissioner for Refugees (UNHCR) 2006), have created a landscape where forced migration affects all populations to some extent, whether as displaced or as hosts to the displaced. Health professionals will undoubtedly be confronted with the effects of forced migration and called upon to engage in responses for those affected regardless of where they practice. This chapter aims to provide a global overview of forced migration and the public health challenges facing populations forced to move from their usual homes, in order to assist the public health professional in framing their own responses. It begins by presenting a brief history of forced migration and the legal definition of ‘refugee’. We describe the emergence of categories of forced migrants that do not meet the legal definition of ‘refugee’, describe how these groups are treated by states, and the implications for people’s lives and health. The chapter then explores the social and political factors that have influenced the treatment of refugees by states. We lay out some of the major health issues that refugees face in both developing and developed countries. The wide variety of actors involved in both protecting forced migrants and providing

health services is presented, along with a discussion of the contextual conditions that have an impact, often detrimental, on the health of forced migrants. As well as describing the structural factors that create and limit health outcomes, this analysis seeks to highlight the agency and voice of forced migrants, exploring the importance of understanding and projecting these, and of appreciating the rationales for how forced migrants may act in the often chaotic situations in which they find themselves. We consider how interactions with other actors, within the complex social, political, and legal context of forced migration, may lead to seemingly unhealthy choices, and at times, barriers to engaging with services. Yet, forced migrants make choices, when possible, seeking to protect their well-being and fight for better, and healthier, outcomes.

Causes of displacement

The global context

Forced migration reached record levels in 2011 with more people becoming refugees in that year than in any year since the start of the new millennium. Humanitarian crises in countries such as Sudan, Libya, Somalia, and Côte d'Ivoire, as well as ongoing instability in Afghanistan and Iraq among others, contributed to a staggering 4.3 million people being newly displaced from their homes (UNHCR 2011). While war, violence, and natural disaster are major contributors, and the causes most commonly associated with forced migration, it is important to consider the broader global dynamics that shape and create the forced migration situation.

Globalization, inequality, and perceptions of crisis

Forced migration cannot be fully understood without considering the global context and the shifting global forces that shape the form and patterns of migration. Globalization and its role in shaping the dynamics of forced migration need to be considered. Although globalization is a contested concept with disputed definitions, most people accept it as a global process of integration which affects sociocultural, economic, and political domains (see Chapter 1.5). The increased ease of travel, and the cultural interconnectedness that projects visions of opportunities in distant lands,

provides a context which may facilitate migration among those with powerful forces driving them from their homes.

The economic implications of the globalized landscape are also worthy of mention in any analysis of forced migration. There is a polarized debate over the effects of globalization on human development and health. Some commentators argue, like Feachem (2001), that the forces of economic globalization may be beneficial to the poor, increasing wealth and therefore health; while others (Baum 2001; MacDonald 2007) highlight the poor health situation in many developing countries and the vast inequities present, and suggest that globalization plays a part in generating these conditions by creating an uneven playing field and weighting the economic benefits towards those that already have power and wealth. They argue that globalization cannot be separated from economic policies that serve the interests of the wealthy, reinforcing or exacerbating inequalities, fuelling conflicts, and contributing to public health crises in developing countries. This analysis suggests that the only way to improve health is to seriously restructure the dominant international financial institutions. Taking the middle ground, Lee (2004) suggests that rigorous, ongoing analysis is required to assess the positive and negative impacts on health within the overall conceptual domain of globalization. Zwi and Alvarez-Castillo (2003) argue that globalization has often led to poor health outcomes for vast populations and that this in turn is responsible for fuelling forced migration. Regardless of our standpoint on globalization, what it is, how it impacts on health, and its influence, what is clear is that global inequities exist and are dramatic, the impact of such inequities cannot be ignored as major contributors to forced migration, whether or not associated with violence.

Castles (2004) provides an analysis that calls for careful examination of the global context when describing forced migration dynamics. He argues that the current perception of a global ‘asylum crisis’ should be seen more as a crisis of North–South relationships, within which ‘asylum’ is but one aspect. The current picture of forced migration, he argues, is a result of disparities between ‘North’ and ‘South’. Although the North–South conceptualization does not perhaps accurately represent the divisions we see today, we can see the dynamics he describes as ‘North–South’ as being between the wealthy and less-wealthy nations. While disparities exist and people search for a better life, or simply for survival, forced migration will continue. The disparity is also reflected in how the effects of forced

migration are absorbed globally, further deepening the crisis in North–South relations. He shows that while there has repeatedly been a perception of impending asylum crises ever since the 1800s, which largely do not fully emerge as feared, there are some new features. He cites the increase in endemic violence and human rights violations as a feature that has increased the volume of forced migration in recent times. Castles (2004) shows, however, that for the North, there is no significant economic or social crisis resulting from forced migration and contrary to popular belief, the North has not had to absorb large influxes of people. The situation is nonetheless construed by those in the North as a crisis because it highlights the 'erosion of nation-state sovereignty'. The South, however, has absorbed the vast majority of forced migrants from neighbouring countries (Allotey and Reidpath 2003). This combined with increasingly restrictive policies of wealthy countries that seek to limit refugee absorption from the South, leads to a tension in North–South relations, and opens space for people smuggling and trafficking, further threatening the sense of control over borders that the North is seeking to maintain, and therefore deepening the sense of crisis perceived by the North. Hence the burden has been left largely with the South, while in the North, incomplete pictures of fear that exclude vulnerable people have been painted and promoted (Grove and Zwi 2006). In their review of globalization, forced migration, and health, Zwi and Alvarez-Castillo (2003) suggest that public health professionals have a responsibility to examine the context and whole picture, and in so doing, to break down these 'incomplete pictures of fear' and advocate for a more humane globalization that has a chance of decreasing the poor health outcomes associated with forced migration.

Immediate drivers of displacement

There is general agreement on the range of factors that contribute to displacement and forced migration. However, the importance of particular factors, and the linkages between them, varies in different contexts. Identification of specific causes has implications for both the states involved and the migrants themselves, given that the perceived cause of displacement has an impact on the international response, the level of funding support provided, and the attitudes of host populations to those seeking protection. The health consequences of forced migration will be shaped by not only the

causes of displacement, but also by the perceptions of all actors regarding these causes and their relationship to them. In this respect, the terms and context within which forced migrants are described and 'presented' greatly affects the societal response and opportunities, or constraints, operating.

Forced migration has traditionally been associated with war and violence. The conception of a refugee as defined by the United Nations (UN) Convention, fleeing from persecution to a place of safety in another country, was clear-cut and well understood. As Moore and Shellman (2004) describe, the central model was one of:

Violence in = Forced migration out.

However, while the main causes of involuntary displacement remain armed conflict, repression, and war, other factors such as natural disasters and development projects are also important, as are the more complex set of economic, social, and political circumstances that may lead to trafficking and people smuggling.

Martin (2002) focuses attention on natural and man-made disasters and border changes leading to statelessness, as factors which lead to forced migration, alongside the well-recognized factors related to persecution, human rights violations, repression, and conflict. Environmental factors, such as rising sea levels and extreme weather conditions, are also increasingly recognized as important. Some argue that these causes may, over time, supersede conflict as a primary cause of forced migration (Myers and Kent 1995). Resource constraints, including depleted access to land and water, and conflict over precious minerals, are likely to become increasingly important, especially as the global population increases from around 6 billion people at present to nearly 9 billion by 2050.

While all these factors are important, quantifying their effects is difficult. Data are available regarding numbers of 'Convention' refugees (see 'Categories of forced migrants' section) and also, more recently, regarding internally displaced persons (IDPs). These data may give some indication of the level of displacement produced by conflicts and violence, alongside immediate natural disasters such as tsunamis and earthquakes. However, more insidious environmental and economic effects are far harder to quantify.

The increased complexity of recent times in describing and understanding factors that cause displacement raise new concerns that go beyond the simple 'violence migration' model therefore calls for new approaches. Major

challenges include an increased focus on economic migrants and those who are forced to move as a result of environmental and climate changes; there is also a rise in ‘mixed migration’, a term which recognizes that many migrants fall somewhere on the spectrum between forced and voluntary migration and are hard to categorize as one or the other.

Particularly important is a call for a shift in policies of protection for host countries to developing policies which address the root causes within countries from which people are forced to migrate; and a growing recognition that understanding causes may allow predictions of impending displacement as well as the identification of the means to avert large-scale displacement. Some analyses continue to support the primacy of violent conflict, or threat of conflict as the major predictor of populations leaving their homes. Moore and Shellman (2004) analysed a global sample of 40 countries and using a model that incorporated a wide range of indicators (government threat, dissident threat, government-dissent interaction, ethnic or non-ethnic civil wars, war on territory, income opportunity, political freedom and the rule of law, and cultural and family ties), they found that the first three were the primary determinants of forced migration flows. These were more important than other factors such as size of economy or type of political institution. Apodaca (1998) analysed a variety of models developed to predict forced migration and concluded that human rights violations are one of the most important predictors.

Economic migration has often been portrayed in the media as freely chosen migration that should not confer a responsibility to protect on accepting states. However, analysts such as Castles (2003) argue that forced migration is linked to economic migration and that the two are ‘closely related (and indeed often indistinguishable) forms of expression of global inequalities and societal crises’. These inequities have intensified since the end of the cold war. He argues that studies of the causes of forced migration have neglected the overarching economic and social structures which should be seen in relation to ‘US political and military domination, economic globalization, North–South inequality and transnationalism’ (Castles 2003).

Debates regarding causes of displacement have particular salience in relation to the protection offered to forced migrants. As subsequent sections of this chapter describe, the nature and extent of protection is often determined by perceived cause. Therefore, understanding these debates is

crucial as they may have effects on what and how protection, if any, is offered to forced migrant populations.

Categories of forced migrants

Definitions of forced migration have been debated for more than 50 years, dating back to the first UN Convention on Refugees in 1951. The legal framework that emerged under the Convention centred on ‘refugees’ as a specially defined group and is based upon the 1951 Convention and a subsequent 1967 protocol. According to these, a refugee is a person who:

> owing to a well-founded fear of being persecuted for reasons of race, religion, nationality membership of a particular social group, or political opinion, is outside the country of his nationality, and is unable to or, owing to such fear, is unwilling to avail himself of the protection of that country. (UNHCR 2009)

This definition has been modified by regional instruments in Africa and Latin America, as described in a later section. The International Association for the Study of Forced Migration (IASFM), however, defines forced migration much more widely as:

> a general term that refers to the movements of refugees and internally displaced people (those displaced by conflicts) as well as people displaced by natural or environmental disasters, chemical or nuclear disasters, famine, or development projects. (Forced Migration Online 2008)

This definition encompasses a wide range of forced migrants, and these can be distinguished into several distinct groups (Table 10.9.1).

The definitions of other forced migrant groups, and their sources, are outlined in the following subsections. They highlight the increased attention devoted to describing, as well as defining, the many forms of forced migration that exist. The discourse and narratives presented have a marked impact on how people see, and respond to, those forced to leave their usual homes and livelihoods.

Internally displaced persons

Unlike refugees, there is no legal definition of an IDP. It is generally accepted that if a person meets the definition of a refugee, but their

displacement does not take them across an international border, then they are internally displaced. The Secretary General of the UN has appointed a representative to deal with IDPs and the Office of the United Nations High Commissioner for Human Rights (OHCHR) has produced a set of Guiding Principles to address the specific needs of IDPs who are described as:

> persons or groups of persons who have been forced or obliged to flee or to leave their homes or places of habitual residence, in particular as a result of, or in order to avoid the effects of, armed conflict, situations of generalized violence, violations of human rights or natural or human-made disasters, and who have not crossed an internationally recognized State border. (Deng 1998)

Asylum seekers

Asylum seekers are those who are fleeing persecution and are seeking refugee status in another country. According to the UNHCR, an asylum seeker is:

> Someone who has made a claim that he or she is a refugee, and is waiting for that claim to be accepted or rejected. The term contains no presumption either way—it simply describes the fact that someone has lodged the claim. Some asylum seekers will be judged to be refugees and others will not. (UNHCR 2009)

The application referred to relates to being recognized under the 1951 Convention as a refugee and therefore being entitled to the protection and support stipulated by the Convention. They may or may not arrive in a country with authorization, nevertheless, they have a legal right to claim asylum and are therefore not ‘illegal’.

Trafficked persons

The 2003 Protocol to Prevent, Suppress and Punish Trafficking in Persons, especially women and children, describes trafficking as:

> recruitment, transportation, transfer, harbouring, or receipt of persons, by means of the threat or use of force or other forms of coercion, of abduction, of fraud, of deception, of the abuse of power or of a position of vulnerability or of the giving or receiving of payments or benefits to achieve the consent of a person having control over another person, for the purpose of exploitation. Exploitation shall include, at a minimum, the exploitation of the prostitution of others or other forms of sexual exploitation, forced labour or services, slavery or practices similar to slavery, servitude or the removal of organs. (UN 2003, Resolution 25)

Table 10.9.1 Categories of forced migrants

Refugee	A person who 'owing to a well-founded fear of being persecuted for reasons of race, religion, nationality, membership of a particular social group, or political opinion, is outside the country of his nationality, and is unable to or, owing to such fear, is unwilling to avail himself of the protection of that country' (UNHCR 2009)
Internally displaced persons	'Persons or groups of persons who have been forced or obliged to flee or to leave their homes' (Deng 1998). Or places of habitual residence, in particular as a result of, or in order to, avoid the effects of armed conflict, situations of generalized violence, violations of human rights or natural or human-made disasters, and who have not crossed an internationally recognized state border
Asylum seeker	'Someone who has made a claim that he or she is a refugee, and is waiting for that claim to be accepted or rejected. The term contains no presumption either way—it simply describes the fact that someone has lodged the claim. Some asylum seekers will be judged to be refugees and others will not' (UNHCR 2009)
Trafficking in persons	'Recruitment, transportation, transfer, harbouring or receipt of persons, by means of the threat or use of force or other forms of coercion, of abduction, of fraud, of deception, of the abuse of power or of a position of vulnerability or of the giving or receiving of payments or benefits to achieve the consent of a person having control over another person, for the purpose of exploitation' (United Nations 2003)
Smuggling in persons	'The procurement, in order to obtain, directly or indirectly, a financial or other material benefit, of illegal entry of a person into a state, party of which the person is not a national or permanent resident' (United Nations 2000) [Article 3 (a)]
Environmentally displaced person	'A person displaced owing to environmental causes, notably land loss and degradation, and natural disaster' (Deng 1998)
Development displacees	'People who are compelled to move as a result of policies and projects implemented to supposedly enhance "development" ' (Forced Migration Online 2008)

Text extracts from United Nations High Commissioner for Refugees (UNHCR), *Protecting Refugees and the Role of UNHCR*, Copyright © UNHCR 2009, available from http://unhcr.org.au/unhcr/images/protecting%20refugees.pdf, reproduced with permission from the United Nations.

Text extracts from Deng, F., *The Guiding Principles on Internal Displacement*, United Nations, New York, Copyright © United Nations 1998, reproduced with permission from the United Nations.

Text extract from United Nations, *Protocol to Prevent, Suppress and Punish Trafficking in Persons, Especially Women and Children, Supplementing the United Nations Convention Against Transnational Organized Crime*, G.A. Res. 25, Annex II, U.N. GAOR, 55th Session, Copyright © United Nations 2003, reproduced with permission from the United Nations. Text extract from United Nations,

Protocol Against the Smuggling of Migrants by Land, Sea and Air, Copyright © United Nations 2000, reproduced with permission from the United Nations.

Text extract from United Nations, *Glossary of Environment Statistics, Studies in Methods, Series F*, United Nations, New York, USA, Copyright © United Nations 1998, reproduced with permission from the United Nations.

Text extracts from Forced Migration Online, *What is Forced Migration?*, Copyright © 2010 Forced Migration Online and Contributors, available from http://www.forcedmigration.org/whatisfm.htm, reproduced with permission from Forced Migration Online.

Smuggled persons

Smuggling of aliens or 'illegal migrant smuggling' is defined by the UN 2000 Protocol Against Smuggling of Migrants by Land, Sea and Air, which supplemented the UN Convention Against Transnational Organized Crime, to mean:

> the procurement, in order to obtain, directly or indirectly, a financial or other material benefit, of illegal entry of a person into a state, party of which the person is not a national or permanent resident. (UN 2000, Article 3(a))

Environmentally displaced persons

The Organisation for Economic Cooperation and Development's online *Glossary of Statistical Terms* describes an environmental refugee as 'a person displaced owing to environmental causes, notably land loss and degradation, and natural disaster' (United Nations 1997). However, the term was first described by El-Hinnawi (1985), a researcher from the United Nations Environment Programme (UNEP) in 1985, who defined environmental refugees as:

> those people who have been forced to leave their traditional habitat, temporarily or permanently, because of a marked environmental disruption (natural and/or triggered by people) that jeopardized their existence and/or seriously affected the quality of their life.

'Environmental disruption' in this definition is taken to mean any physical, chemical, and/or biological changes in the ecosystem (or resource base) that render it, temporarily or permanently, unsuitable to support human life. The validity of the term 'environmental refugees' has been debated greatly since that time. Actors emerge from a realm of perspectives including

environmentalists, economists, and refugee advocates, each with a different interpretation of the term. Some prefer to use the term ‘environmental migrant’ or ‘environmental displacee’; different terms establish different implications for both the individuals and states involved.

Development displacees

There is no legal definition for development displacees; however, Forced Migration Online suggests they are:

> people who are compelled to move as a result of policies and projects implemented to supposedly enhance ‘development’. These include large-scale infrastructure projects such as dams, roads, ports, airports; urban clearance initiatives; mining and deforestation; and the introduction of conservation parks/reserves and biosphere projects. Affected people usually remain within the borders of their country. People displaced in this way are sometimes also referred to as ‘oustees’, ‘involuntarily displaced’ or ‘involuntarily resettled’. (Forced Migration Online 2008)

Interestingly, large-scale agricultural projects do not seem to be specifically mentioned.

Forced migration statistics

Assessing the volume of global flows of forced migrants is valuable in identifying trends and the patterns of distribution with respect to where the burden for providing assistance and care lies. In specific countries and contexts, where people are forced to migrate from their place of origin, the first step in addressing public health and other more general concerns is often to define the size of the population at risk. Donors, governments, aid agencies, and community organizations require basic demographic and related data to inform planning and service delivery. The collection of data is complicated and surrounded by controversy, having the potential to immediately reframe what looks like a technical issue into a political one.

The two main sources of statistics on forced migrants are the United Nations High Commissioner for Refugees (UNHCR) and the US Committee for Refugees and Immigrants (USCRI). These organizations produce differing estimates for different types of forced migrants in different geographical contexts. The logistic difficulties of collecting reliable data from shifting forced migrant populations leaves space for political interests

to enter, and seek to shape, the field. This is compounded by disagreement over definitions of forced migrants and which groups to include in which categories, as mentioned earlier, with significant implications for the rights and entitlements of such groups, as well as for who bears the responsibilities for addressing their needs. Despite contested data, available sources provide valuable information on trends and distribution.

UNHCR forced migration statistics

The UNHCR provides data based on 'persons of concern', an increasingly wide-ranging term. 'Persons of concern' includes refugees (as defined by the 1951 Convention or the 1969 Organization of African Unity (OAU) Convention); persons granted complementary protection (alternative forms of protection for those that fall outside of the terms of the 1951 Convention) and those with forms of temporary protection; asylum seekers who in the year of data collection are still pending a final decision on an application for protection; stateless persons; refugees who returned to their country of origin in the year of data collection; and IDPs (see Table 10.9.2). For the purposes of UNHCR the latter are limited to conflict-generated IDPs to whom UNHCR protection or assistance is extended, but does not include all IDPs.

UNHCR reported that in 2011 the number of forcibly displaced persons exceeded 42 million for the fifth consecutive year (UNHCR 2011). This figure included refugees, IDPs, and asylum seekers. It was estimated that 4.3 million people were newly displaced in 2011 due to conflict or persecution, with 800,000 of these displaced across international borders. This was a 20% increase since 2010 of new displacement. There was also an increase of 700,000 in refugees and IDPs protected by the UNHCR and 7.1 million, three-quarters of the UNHCR refugee population, were reported to be in a protracted refugee situation. Developing countries hosted four-fifths of the world's refugees (UNHCR 2011). The countries of origin for refugees are shown in Figure 10.9.1. Afghanistan produced the greatest number of refugees followed by Iraq and Somalia. UNHCR statistics demonstrate that the majority of refugees remain within their region of origin. Pakistan was the country that hosted the most refugees in 2011, followed by Iran and Syria. Almost half of the world's refugees falling under the UNHCR mandate were hosted in countries with a GDP per capita of less than $3,000.

Table 10.9.2 Persons of concern to United Nations High Commissioner for Refugees (UNHCR 2011)

	2011
Newly displaced	4.3 million
Refugees	10.4 million
IDPs	15.5 million
Asylum seekers	876,100
Stateless	12 million
Resettlement	79,800[a]
Unaccompanied children	17,700
Protected by UNHCR	25.9 million[b]
Forcibly displaced people worldwide	42.5 million

[a] Resettlement—with or without UNHCR assistance—by 22 countries.

[b] Protected by UNHCR—refugees and IDPs.

Source: data from United Nations High Commissioner for Refugees (UNHCR), A Year of Crises: *UNHCR Global Trends 2011*, UNHCR, Geneva, Switzerland, Copyright © UNHCR 2012, available from http://www.unhcr.org/4fd6f87f9.html.

It is important to note that Palestinian refugees falling under the mandate of the United Nations Relief and Works Agency for Palestinian Refugees in the Near East (UNRWA) are not included in the UNHCR refugee figures, but are included in the 42 million forced migrants. The Palestinian community has been one of the largest groups of refugees that has remained without being fully resettled or absorbed into host nations; a reflection of how community members on the ground suffer while local, national, regional, and global powers resolve, or fail to resolve, higher-level strategic and political issues.

Given the importance of these numbers, they are invariably contested and debated; with key actors seeking to shape what data are collected and presented, who are counted, and what the implications are for future activities. Indeed each ‘player’ in the system, whether they be a UN or non-

governmental agency, a recipient or state of origin, a local community or professional group, and many others, has their own rationale for inflating or deflating the numbers of people whether they be refugees or IDPs, in or outside of camps, dependant or free of other major influences.

Internally displaced persons

The global number of IDPs at the end of 2011 was estimated by the UNHCR to be 26.4 million, with 15.5 million benefiting from UNHCR assistance. The figure for 2011 was the second highest on record, according to the UNHCR this was partly due to displacement from Afghanistan, Côte d'Ivoire, Libya, South Sudan, Sudan, and Yemen (UNHCR 2011). Figure 10.9.2 shows the distribution of IDPs globally.

Asylum seekers

In 2011, 876,100 applications for asylum or refugee status were submitted to UNHCR offices or governments. One-tenth of all these applications were registered in South Africa, making it the top asylum destination by a clear margin, more than half these claims were from Zimbabwean nationals. The United States was the next highest receiver of asylum applications with almost half coming from China, Mexico, and El Salvador. The UNHCR noted significant differences between countries in recognition rates suggesting differences in how asylum seekers are treated. For example, between eight European countries the recognition rate for Afghan asylum seekers ranged from 3 to 33% (UNHCR 2011).

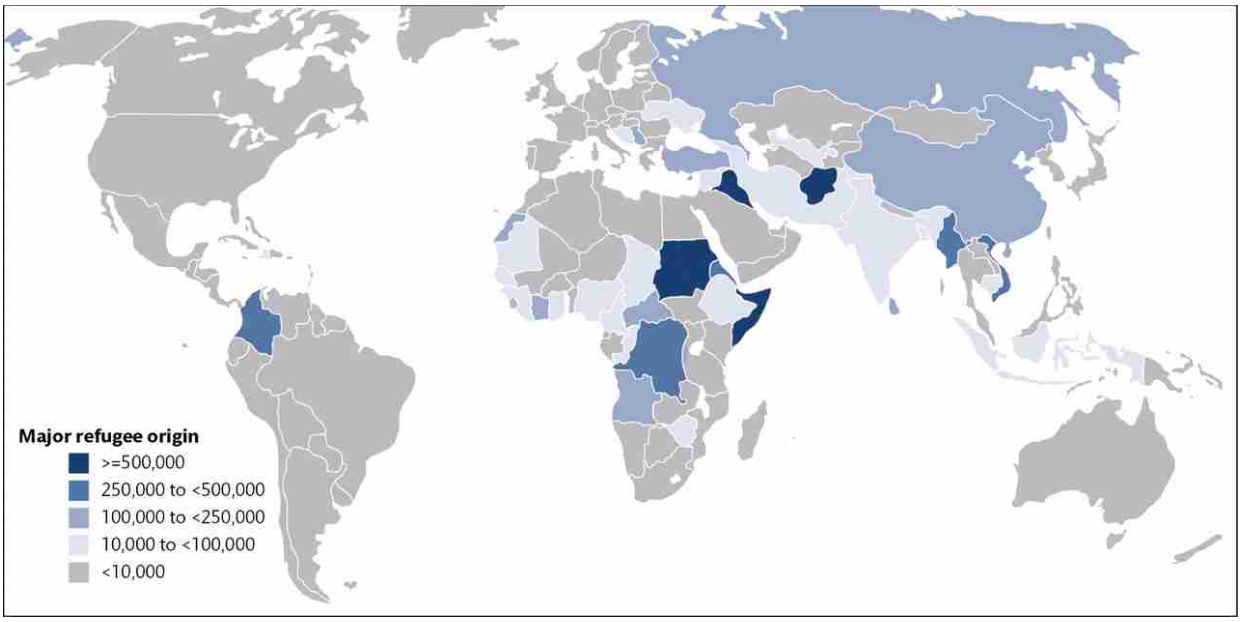


Fig. 10.9.1 Country of origin for refugees under UNHCR mandate 2011.
Reproduced from United Nations High Commissioner for Refugees (UNHCR), A Year of Crises: UNHCR Global Trends 2011, UNHCR, Geneva, Switzerland, Copyright © United Nations High Commissioner for Refugees 2012, available from http://www.unhcr.org/4fd6f87f9.html.

Legal framework

Refugees

The 1951 Convention, which forms the basis of all international refugee law, was drafted in the aftermath of the Second World War in response to the large numbers of displaced people caused by the war. The Convention included restrictions on who would be classified as a ‘refugee’ based on geographical and temporal factors that were specific to the European context. As a result, the 1967 Protocol relating to the Status of Refugees was adopted which removed these restrictions and allowed for more general application of the Convention. The Convention and Protocol laid out not only the definition of a refugee but also the scope of protection to which refugees were entitled. The most fundamental of these is *non-refoulement*, the right to remain rather than be returned to a place where their lives or freedoms would be at risk. Debate has raged over the continuing relevance of the Convention definition, given that it excludes large numbers of people in need of international protection, notably IDPs and those that have

migrated as a result of generalized (as opposed to individualized) violence, civil war, and widespread human rights abuses.

Africa and Latin America responded to the limitations of the Convention and Protocol by adopting their own regional laws and standards. The Organization of African Unity (OAU) (now known as the African Union) adopted the 1967 Convention Governing the Specific Aspects of Refugee Problems in Africa (Barnett 2002). This is the only legally binding regional instrument and has a refugee definition similar to that of the Convention but includes those who flee due to 'external aggression, occupation, foreign domination or events seriously disturbing public order in either part or the whole of his country of origin or nationality' (OAU 1969).

In Latin America, the Cartagena Declaration, adopted in 1984, included a definition of refugee similar to that of the OAU Convention but which also included reference to 'massive violation of human rights' (Regional Refugee Instruments & Related 1984). The latter is not legally binding but has been adopted by many states and incorporated into some national legislation in the region.

Persons who have participated in war crimes and violations of humanitarian and human rights law, including the crime of terrorism, are specifically excluded from the Convention and cannot be recognized as refugees.

The UNHCR has been given a mandate to provide international protection to refugees and seek permanent solutions to their problems through its Statute, adopted by the UN General Assembly (UNGA) in December 1950. Over the years, the UNGA has allowed UNHCR to extend its responsibilities and is increasingly focused not only on refugee protection but also on managing refugees in the field.

Internally displaced persons

While IDPs do not have the same legal framework of protection, they have become included in the UNHCR's 'people of concern' and are therefore often covered by its mandate. The United Nations Office for Coordination of Humanitarian Affairs (OCHA) has produced a set of Guiding Principles to address the specific needs of IDPs (Deng 1998). The guidelines identify rights and guarantees that should be afforded to IDPs although they are not legally binding.

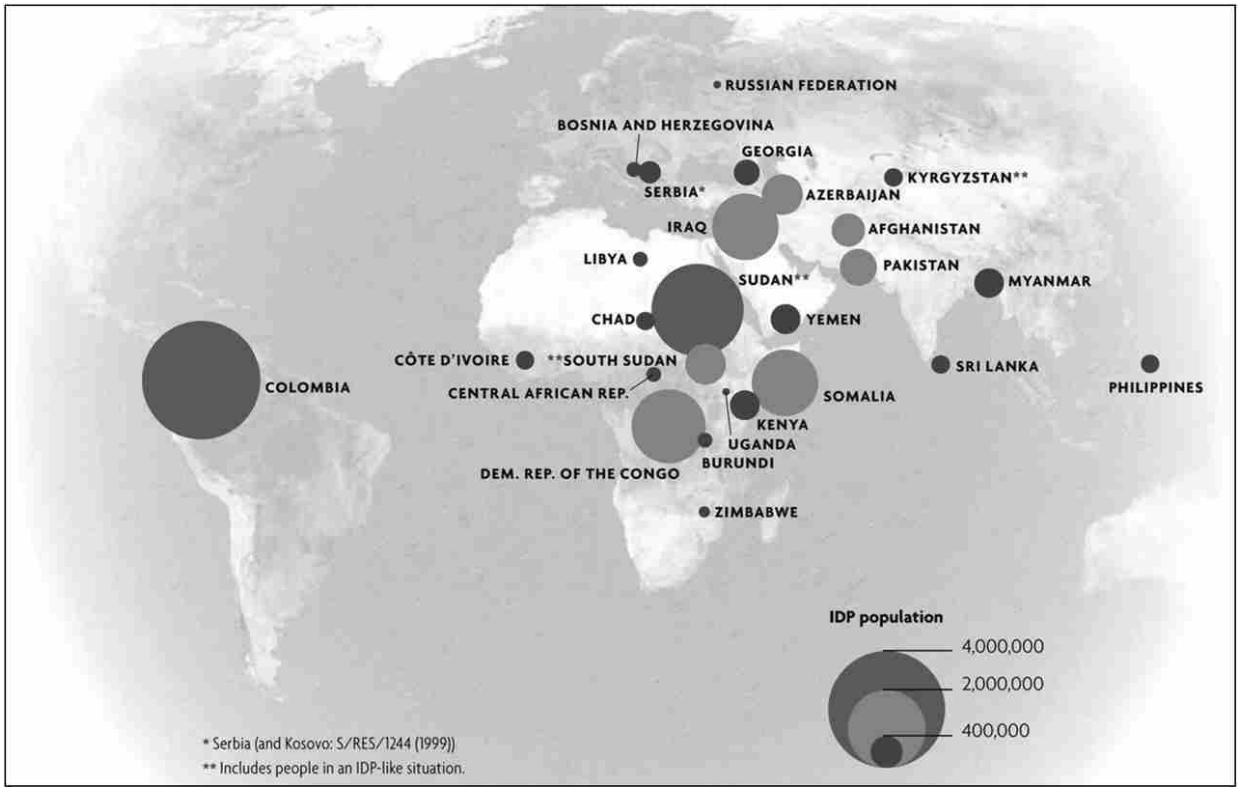


Fig. 10.9.2 Distribution of IDP populations globally 2011.

Reproduced from United Nations High Commissioner for Refugees (UNHCR), A Year of Crises: UNHCR Global Trends 2011, UNHCR, Geneva, Switzerland, Copyright © United Nations High Commissioner for Refugees 2012, available from http://www.unhcr.org/4fd6f87f9.html.

Asylum seekers

The Declaration of Territorial Asylum (DTA) adopted by the UNGA in 1967 is based on article 14 of the Universal Declaration of Human Rights, which declares that ‘everyone has the right to seek and to enjoy in other countries asylum from persecution’. The principles of the DTA are that all states recognize asylum granted by one State, that states themselves evaluate the grounds for asylum, and that no one seeking asylum should be rejected at borders or if already within a country should not be compulsorily returned to any state in which they may be subject to persecution. Exceptions can only be made in situations of national security or for asylum seekers that have committed a crime against humanity, against peace, or a war crime.

Asylum may be granted by states on the basis of the 1951 Refugee Convention, in which an individual has been deemed a 'refugee' as defined by the Convention. However, given the restrictions of this definition and the increasing numbers of people who are seeking protection but do not fit the Convention definition, alternative forms of protection have developed. Often termed 'complementary' or 'subsidiary' protection, McAdam (2006) describes the plethora of alternative forms of protection for 'non-Convention refugees' that previously had been treated much the same as Convention refugees by states.

'Non-convention' refugees, such as those fleeing widespread violence rather than targeted individual persecution, have to look to instruments of human rights law, rather than refugee law, for protection. However, human rights instruments do not attach the same clear rights and standards as the Refugee Convention and it has also been argued that human rights law can be weak on implementation (McAdam 2006). This has led many to argue that complementary protection is a lesser form of protection, and there is a danger of states using these forms of protection to decrease the numbers of Convention refugees and therefore their level of commitment. Debate continues as to whether expanding the current Convention definition, or expanding systems of complementary protection, will weaken or broaden protection to those seeking asylum.

Trafficking and smuggled persons

Although the reasons for a person to be trafficked or smuggled are very similar, international agreements and national laws treat them quite differently. Unlike trafficking, which may occur internally in a country, smuggling always involves the illegal crossing of an international boundary. Smuggling is seen as a crime against a *state* whereas trafficking is a crime against an *individual*. Consequently persons that are trafficked or smuggled are likely to be treated quite differently. The UN Convention on Trafficking mandates states to consider granting victims permanent or temporary resident status, whereas the Smuggling of Migrants Protocol calls on states to facilitate the return of smuggled persons to the country from which they came.

Environmental and developmental displacees

The concept of environmental refugees or displacees is particularly contentious and the subject of much current debate. There is little agreement on whether environmental factors can be seen as primary causes of displacement and equally no consensus on a definition for those displaced from their homes by these factors. As a result the legal protection available is marginal. The current legal regime has no specific mandate to protect environmental displacees, if they are displaced across international borders; however, 'natural or human-made disasters' are covered by the Guiding Principles on Internal Displacement (Deng 1998). These place the primary duty for protection with the national authority; however, no duty is placed on other states to recognize the need for protection. A growing body of opinion has begun calling for a system of global governance that recognizes environmental factors as a cause of displacement and for the displacees to be protected under international law (Courtland Robinson 2003; Biermann and Boas 2007).

Similarly, developmental displacees do not have frameworks of protection of their own but are covered by the Guiding Principles on Internal Displacement. There is a need to recognize the needs of this group and for research on displacement not to fall into a dichotomy that separates conflict-related displacement from development-related displacement (Cernea 2007).

Evolving rights of protection

Protection based on a definition of refugees fleeing from the effects of violent conflict has been the prevailing paradigm, and while the centrality of conflict-related displacement has been debated, conflict and violence clearly remain significant causes of forced migration and suffering. Analysis reflecting the increasingly complex drivers of forced migration, including economic and environmental concerns, while vital, should be careful not to compromise existing rights of protection. These rights of protection, based on conflict as the central tenet, should not be further eroded by the recognition that there are other causes of forced migration that need to be seriously considered.

Public health situation of forced migrants

Forced migrants live in a range of different social, political, and geographical conditions. They may form distinct populations or be dispersed among host populations; they may be displaced within their own country, to a nearby or distant country of asylum; and may dwell anywhere from a developing country experiencing conflict, to a more peaceful but poverty-stricken developing country, to a wealthy developed nation. The health of each forced migrant population, and each individual within the group, will be shaped by these factors, as well as their prior health and social condition, the journeys they have had to take, the stability and security present in their new environments, the social structures within which they live, and the access they have to services and social support.

Gushulak and MacPherson (2000) describe migration health in terms of 'three distinct but interdependent undertakings: The pre-departure phase, the migratory journey itself, and the arrival at destination', each of which influences health. Factors such as social networks, assets, language, knowledge, information or access to services, food security issues, poor shelter, sanitation, and availability of safe water will influence health outcomes.

In this section we describe the major health challenges facing forced migrant populations. This is followed by an overview of the public health responses and an indication of where more detailed technical advice may be found.

Refugees in developing countries

Developing countries continue to absorb the vast majority of forced migrants, in particular as developed nations have become more restrictive. Within developing countries, people who are forcibly displaced either within or across borders may find themselves in a variety of possible contexts, settling short or long term either in a camp situation or dispersed within a host population. Each presents threats and opportunities.

The type of settlement formed has implications for health and well-being. The typical refugee camp setting has presented a range of problems, although improvements in public health knowledge and experience, and informing policy and practice with evidence, have led to better outcomes.

The impact of forced migration falls heavily on developing nations, and within these countries, on host populations. This is particularly so where

forced migrants reside in non-camp situations.

Camps and non-camp settlements

Camps, while constructed with the intention of providing protection, safety, and access to services, may also pose risks. Poorly planned refugee settlements may have inadequate infrastructure, may pose risks to safety, or set up new inequalities. The combination of overcrowding, poor hygiene and sanitation, lack of water, poor nutrition, and inadequate shelter provides conditions for spread of many diseases, particularly communicable disease. Refugees may arrive in a poor condition, suffering from malnutrition, and with poor immunity, reflecting limited access to immunization services. Disrupted social structures decrease the capacity to deal with sickness. Communicable disease spreads quickly and may be particularly severe. Psychological distress is also a major issue due to the insecure, restricted environment and lack of resources.

Debates over the utility of camps are highlighted by Harrell-Bond (2000) who argues that as a concept they have failed. Originally based on modernization theory, the premise was that camp settlements would become new agricultural centres and eventually be integrated with host populations to the benefit of all. However, this was abandoned as the theory became discredited and was replaced by a view that refugees should be seen as temporary. Harrell-Bond suggests that those in camps are in ‘prison-like’ settlements with poor conditions, increased vulnerability to disease, and often poor nutrition. She also argues that camps do not help host populations, as large sums of money are diverted to camps which may be demolished if refugees do move on, as well as diverting scarce human and material resources from the host population. She suggests that it would be far more appropriate and cheaper to help integrate refugees into host populations and invest in the country and strengthening existing social structures, as a whole.

In many situations of forced migration, for example, when the Kosovar Albanians fled from Kosovo over the border into Albania, integration with host populations there was the norm. Following political instability and violence in Timor-Leste in 2006, 4 years after independence, makeshift camps were established within Dili, the capital city, for IDPs, while many also fled and were integrated with host communities in the districts. The

latter necessarily absorbed a portion of the social and economic costs associated with accommodating those who fled.

However, Salama et al. (2004) show that, despite the problems associated with camps, there are also benefits. The increased knowledge base and coordination of humanitarian interventions have led to improvements in health, with better nutrition, sanitation, and health services. The burden of ill health may shift from displaced populations who are relatively well serviced, especially if they cross international borders and become refugees, relative to host populations and IDPs whose needs may be greater and who have less access to assistance.

Mortality

Crude mortality rates (CMRs) are often used as an indicator of the immediate impact of complex emergencies. Toole and Waldman (1990) proposed a definition of the acute phase of a complex emergency as greater than one death per 10,000 people per day, based on a doubling of the baseline mortality rate in sub-Saharan African countries. Surveillance systems to monitor the impact of emergencies have been established in the past 15 years based on this definition. In recent decades, it has not been unusual to see CMRs in refugee populations of 30 times the baseline in the country of origin (Toole and Waldman 1997); some of the highest recorded were among Rwandan refugees who flooded into North Kivu in 1994 where average CMRs were 20–35 deaths per 10,000 (Salama et al. 2004). In refugee camps, CMRs of 2 per 10,000 are now rarely seen; however, the situation in non-camp situations has deteriorated (Salama et al. 2004). Refugees are usually at greatest risk in early days of displacement, with mortality rates subsequently diminishing. The highest risk group is young children with most deaths occurring in the under-5 age group, demonstrated dramatically in one Somali camp in which 70% of children under 5 appear to have died over an 8-month period (Moore et al. 1993; Toole and Waldman 1997).

Causes of death

The main causes of death in developing country refugee situations are similar to those of developing country populations in general. The most

common causes of death in Africa are diarrhoeal disease, measles, acute respiratory infections, and malaria (where endemic), exacerbated by a high prevalence of acute malnutrition (Toole and Waldman 1997). These diseases typically account for the vast majority of reported deaths in refugee populations, at times up to 95% thereof. This is in contrast to European countries affected by war, where the most common causes of death are injuries, as well as communicable diseases, neonatal problems, and nutritional deficiencies. Recent analyses suggest that chronic diseases require attention, given the ageing population, and violence, instability, and forced migration in populations in which communicable diseases have been relatively well controlled and chronic diseases have emerged as problems (Chan and Sondorp 2007).

The conditions affecting refugees may have been brought with them from the country of origin, may be acquired in the host country (with refugees susceptible due to lack of sufficient immunity), or may result from the camp conditions themselves.

Diarrhoeal disease

Possibly the greatest threat to life among refugees is diarrhoeal disease. The main pathogens of concern are usually cholera and shigella. In some refugee camps, more than 40% of deaths in the acute phase of an emergency can be attributed to diarrhoeal disease, and over 80% of these deaths are in children under the age of 2 years (Salama et al. 2004). Outbreaks are exacerbated by factors such as the sharing and contamination of water sources and food, the scarcity of soap, and inadequate sanitation.

Among Rwandan refugees in camps in North Kivu, 85% of the 50,000 deaths among the 800,000 people that flooded into the area in early July 1994 were caused by diarrhoea, with 40% from shigella dysentery and 60% from cholera (Salama et al. 2004). The primary drinking source, which was lake water, became rapidly contaminated, and as water was also scarce, infection spread swiftly through the refugees. Deaths in Goma were associated with a number of preventable causes (Siddique et al. 1995). Among Kurdish refugees, 70% of deaths were a result of diarrhoea (Toole and Waldman 1997). In the elderly and young children, dysentery case-fatality rates can be as high as 10%. Important public health interventions include the need to supply sufficient water, control excreta and hygiene, and

increase public awareness of basic rules of hygiene. Evidence-informed case management is necessary to avoid unnecessary deaths.

Measles

Another of the most feared diseases in camps is measles—a particularly large problem in refugee situations in the 1970s and 1980s until it was recognized that vaccination should be accorded priority. The risk of death from measles is exacerbated in the presence of malnutrition, a common occurrence among refugee populations, as are secondary complications. Globally, measles kills hundreds of thousands of people each year. In 2000 there were 562,000 measles deaths but with improved surveillance and vaccination rates this was reduced to 122,000 deaths in 2012 (WHO 2014). While it is less of a problem in previously well-vaccinated populations, in developing countries, measles death rates can range from 1 to 5%, and in refugee settings among malnourished children, the death rate may reach 10–30%.

In 1985, 53 and 42% of refugee deaths in eastern Sudan and Somalia, respectively, were caused by measles (Salama et al. 2004). The situation has improved in refugee camps due to mass vaccination, but IDPs still suffer as they have less access to vaccinations and, since 1990, it is more common to find high rates of measles in IDPs than in refugees. Measles is included in the Expanded Programme on Immunization but there is often low coverage in young children especially where human resources for health are limited or conflict affects normal service delivery. Increasing attention is being devoted to ensuring that services can continue to be supplied in fragile states.

Malaria

In areas where malaria is endemic, it can cause high rates of mortality and morbidity among refugees. This is particularly the case when people are displaced from an area of low endemicity to one of high endemicity. Drug resistance also plays a large part with chloroquine resistance and more recent resistance to fansidar reported in Rwandan refugees in Zaire. Local communities in low endemic areas may be placed at risk by refugees arriving from high endemic areas, especially if local conditions favour the mosquito. In Africa, 30% of annual malarial deaths occur in complex

emergency settings. This is because in complex emergencies (CEs) overcrowding increases the frequency of bites and also delays intervention thereby increasing the time that the parasite is in the blood (Salama et al. 2004). An example is Burundi which had a major epidemic in 2000–2001 when the population of 7 million reported 2.8 million cases of malaria, the war in 1993 caused control efforts to break down, populations were moving, and there was widespread resistance to chloroquine.

Acute respiratory infections

Refugee camp conditions facilitate the spread of acute respiratory infections (ARIs) and rank among the leading causes of death in displaced populations. The lack of shelter, overcrowding with poor ventilation, and exposure that characterize the refugee existence contribute to the susceptibility to ARIs with poor outcomes. Young children are particularly affected. In Kabul, in 1993, 30% of under-5 deaths and 23% of deaths in displaced persons were a result of ARIs (Salama et al. 2004). Vaccinations against malaria, diphtheria, and pertussis have led to decreased mortality from ARIs. It is not just that these pathogens cause ARIs but they also render hosts less able to defend themselves from bacterial infection.

Tuberculosis

Tuberculosis (TB) is becoming increasingly important in CEs and situations of forced migration—in part as a result of disrupted therapy (Connolly et al. 2004). TB can become a major problem, leading to a large numbers of deaths. Poor nutrition and overcrowding exacerbate the risk of TB, as may underlying HIV in some populations. A recent study of instability and internal displacement in Timor-Leste revealed a number of problems in maintaining continuity of treatment for those with TB. As a result, many may have incomplete or inadequate treatment.

Meningococcal disease

Meningococcal disease is a potentially major problem for refugees when they are in overcrowded settlements in an endemic area, such as the 'meningitis belt' of sub-Saharan Africa that stretches from Senegal to Ethiopia. In Africa, a threshold incidence of 15 cases per 100,000 population per week is used to define the presence of an 'epidemic'. In CEs, large epidemics have also been reported outside the traditional malaria belt. The Democratic Republic of the Congo had six epidemics in 6 months in 2002 and epidemics have also occurred in Uganda, Rwanda, and Tanzania (Salama et al. 2004). Immunization has proved to be an effective control measure but is only effective in epidemics, routine immunization is not effective as the current vaccines only produce immunity for 3–5 years and may not be used in children under 2 years. Therefore, the World Health Organization (WHO) only recommends mass immunization in areas where epidemic proportions have been reached, and they estimate that this can reduce deaths by up to 70%.

Hepatitis

Outbreaks of hepatitis E have occurred in African refugee camps. It is transmitted by the entero-faecal route, and therefore water supplies are often involved. Case-fatality rates vary from 1–4% but there have been particularly high attack rates and case-fatality rates (up to 20%) reported among pregnant women. The high level of previous exposure to hepatitis A and B in developing countries makes it likely that any hepatitis epidemic with high attack rates in adults is likely to be caused by hepatitis E.

HIV/AIDS and other sexually transmitted infections

The pattern of HIV/AIDS and sexually transmitted infections (STIs) among refugee populations in developing countries will depend on many factors, including the level of infection prior to displacement (Dualeh and Shears 2002). Speigel (2004), working at the UNHCR, has highlighted the key preventive and other responses to HIV/AIDS. Conflict situations and camp conditions have often been thought to exacerbate the spread of STIs, due to breakdown of social structures, behavioural change, sexual violence, and lack of access to services and testing (Zwi and Cabral 1991). However,

Speigel (2004) describes other factors inherent in refugee situations that may retard the spread of these diseases, such as decreased mobility, and the possibility of a host country having increased services in comparison to the country from which people have fled. He argues that each situation should be examined independently and that the spread of STIs is most likely to be determined by the complex interactions between a number of groups: refugees and their hosts, the original conflict-affected population, IDP populations in the country of origin, and host communities in country of origin. He argues for public health responses to be broad and focus not only on camps but the entire situation, including host populations.

Other communicable diseases

Outbreaks of other less common communicable diseases have been reported in some refugee camps, as highlighted by Dualeh and Shears (2002). Typhoid fever is endemic in many tropical areas and may occur sporadically in refugee situations. Leishmaniasis and trypanosomiasis have also been reported in epidemic proportions among displaced populations in Uganda and Sudan. Viral haemorrhagic fevers occur in selected areas and may place refugee communities at risk. Rift Valley fever has been documented among Somali refugees in Kenya, yellow fever occurred among displaced populations from Ethiopia to Sudan, and dengue fever affects refugees in endemic areas of South East Asia.

Connolly et al. (2004) provide an excellent overview of the literature on communicable diseases and CEs.

Malnutrition—protein energy deficiencies

Moderate to severe acute malnutrition is a common occurrence in refugee settings. Children are particularly affected, with some deteriorating in terms of nutritional status after arrival in camps, due to inadequate nutrition or diarrhoeal disease. Surveys undertaken between 1988 and 1995 documented a range of 11–81% of children under 5 years suffering acute malnutrition in conflict-affected populations (Toole and Waldman 1990). Acute malnutrition can be both the cause and effect of other illnesses, and contributes to mortality both directly and indirectly. Malnutrition and measles are

particularly linked with malnutrition exacerbating measles, and measles itself contributing to high malnutrition rates in those that survive it (Toole and Waldman 1990). This ‘synergistic relationship’ that malnutrition has with communicable disease requires an integrated response aimed at improving nutrition combined with efforts to improve public health services.

Micronutrient deficiencies

As a result of inadequate food rations micronutrient deficiencies occur frequently in refugee populations. The most common diseases seen are iron and vitamin A deficiencies but other diseases include scurvy (vitamin C deficiency), pellagra (niacin and/or tryptophan deficiency), and beriberi (thiamine deficiency). As Prinzo and Benoist (2002) discuss, the solutions are not simple and each setting may require a different approach, including providing fortified blended cereal and finding ways to exchange rations for locally produced fresh fruit and vegetables.

Food and nutritional interventions need to be responsive to cultural as well as economic issues.

Reproductive health

The reproductive health needs of refugees have gained recognition since the mid-1990s. The burden is undeniably greater for women as a result of insecurity in their social situations and gender-based violence. Women lack basic necessities such as sanitary napkins and condoms, and pregnancy itself can be a serious health threat in settings in which services are lacking or overwhelmed (Krause et al. 2000). Referral services are often non-existent and birthing or abortion may be undertaken in unsafe situations. Women may have no, or limited, access or means to carry out family planning, rape and sexual coercion are typically increased, and there may be pressures on women to rebuild the population. Female genital mutilation may occur and STIs, including HIV, may be relatively common (Krause et al. 2000). A particularly important set of risks relate to sexual and other forms of gender-based violence, common in all societies and increased where societal norms are less able to be enforced or reinforced, and in which men with guns can demand whatever they like in return for precious resources such as safety,

food, and onward passage. Intense efforts have focused on improving all aspects of reproductive health for women affected by CEs and forced migration—and a Minimum Initial Services Package has been produced to address these needs. A study by Hynes et al. (2002) showed that due to the strength of the response there have actually now been better reproductive health outcomes in refugee camps than in host population or origin countries. However, the lessons learned require expansion across settings.

Mental health and psychosocial needs

Mental health problems are recognized as an important source of morbidity globally (Prince et al. 2007) and are likely to be even greater among displaced populations. Epidemiological assessments have documented high rates of mental illness in these populations. Studies have tended to focus on trauma-related conditions, although low prevalence conditions such as psychosis have also proven significant in post-conflict and refugee settings. There is much debate about the validity of Western-derived measurement and about the trauma focus, with further development of relevant culturally valid ways to measure psychiatric morbidity still required. Despite the measurement debate, there is clearly a substantial burden of mental distress in these settings, and support for vulnerable people is vital in situations of social breakdown, especially where usual systems of social support and healing are disrupted (Silove et al. 2000a). It may be of value to identify and respond to the needs of three different groups of people who experience psychosocial and mental distress in these situations: those who are mentally ill, who need treatment; those who have had particularly horrific experiences and need care and support in order to recover; and what tends to be the majority who need a more stable social and political environment in which they can resume their normal day-to-day activities necessary to promote mental health.

Internally displaced persons

While IDPs in developing countries have very similar health problems to refugees, they also have particular vulnerabilities. Some of these result from falling through gaps in a system geared to protect a particular group of forced migrants (i.e. ‘Convention refugees’). Particularly in developing

countries, but also in more developed states, the limited data available suggest higher morbidity and mortality in IDPs compared with refugees (Salama et al. 2001). With limited protection granted by the international community, the responsibility for protection and assistance often lies with their own governments. Yet, it is these governments themselves that have often contributed to the displacement experienced. Governments may not even officially recognize the existence of IDPs for fear of international scrutiny or interference. A growing number of countries have developed laws and policies to govern responses to IDP populations but not all make concerted attempts to implement these policies, and even when there is political will the states involved are often the poorest countries of the world, limiting their ability to respond. There is no mandated international body although UNHCR may include them among their persons of concern and they are increasingly supported by agencies which follow the UN's Guiding Principles on Internal Displacement (Deng 1998).

IDP populations rarely experience access to health services at the same standard as offered to other citizens and are likely to be markedly poorer than those available to refugees. Recent efforts to improve quality and responsiveness of interventions in humanitarian crises and CEs have contributed to better health outcomes. IDPs, however, have poor access to these improved systems of protection and assistance.

Collection of data is more difficult among IDP populations and relatively few studies focus specifically on IDPs. However, the IDP-specific health data that are available suggest that access to healthcare for IDPs is often extremely limited. Salama et al. (2001) show how some of highest mortality rates in humanitarian emergencies have been found in IDP populations; Sudan and Somali, for example, reached 8 per 10,000/day and 17 per 10,000/day respectively (Salama et al. 2001). In Ethiopia and the Democratic Republic of the Congo, delays in vaccinations led to measles outbreaks. In Ajiep (southern Sudan), severe acute malnutrition among IDPS reached 36% during a famine in 1998.

Although international humanitarian law applies to all civilians affected by armed conflict and therefore IDPs should be afforded the same protection and assistance as the un-displaced host population, there may be multiple reasons why IDPs face discrimination and therefore do not receive the help they need.

Even in developed countries IDPs may have poor health outcomes. They are often subject to the discrimination and exclusion experienced by their developing country counterparts. The Internal Displacement Monitoring Centre (IDMC) describes the situation for a number of IDP groups in developed nations:

> While IDPs in the Balkans generally have satisfactory access to water, sanitation and healthcare, they are more likely than the local population to suffer from trauma-related problems. Roma IDPs usually live in informal settlements with very poor sanitary conditions. IDPs in Azerbaijan and the Russian Federation access healthcare less easily than the local population in some areas, due to administrative inconsistencies, lack of healthcare facilities and the demand for informal payments for medical treatment. (IDMC 2008)

IDPs, whether in developing or developed countries, are clearly living with a higher burden of illness and mortality than many other forced migrant groups. Initiatives are emerging that aim to support this extremely vulnerable population, but they are yet to be supported by any legal framework.

Refugees and asylum seekers in developed countries

Refugees and asylum seekers often reach destinations in developed countries after long, precarious journeys. Prior to the journey, they will have faced many of the routine health problems that present themselves in disrupted developing country settings. The journey itself will have produced health challenges; people may have been detained at points along the way or experienced significant anxiety and distress through difficult and dangerous journeys and experiences. On arrival, the range of situations they find themselves in will further shape their health condition. Those who have applied for refugee status prior to arrival may have undergone health checks already in the country of origin, as is the practice of many developed countries; therefore, certain health conditions may have been picked up. Depending on the country they arrive in, they will then have a different range of options available to address their health problems. However, those applying for refugee status on arrival have an even more complicated journey; they may be detained for short or long periods, they may have extensive periods on temporary visas or they may find themselves living illegally without authorization from the host country. Typically asylum seekers are often isolated with difficulty in communication, in a situation of

poverty, discrimination, and a lack of social structure, causing many to suffer from trauma-related problems and many general health problems, exacerbated by poor provision of health information and barriers to accessing to services.

Biggs and Skull (2003) discuss the specific medical conditions that may be present among refugees in developed countries. These include communicable disease such as TB, HIV, STIs, hepatitis B and C, and parasitic diseases, all compounded by incomplete immunization. In addition, nutritional deficiencies (vitamin D in particular), mental health problems, and other non-communicable diseases such as diabetes and reproductive health issues are common.

Screening for health problems among newly 'arrived' refugees, or prior to arrival for those applying for refugee status overseas, is common practice. Biggs and Skull (2003) point out that there is little evidence demonstrating that screening has proved effective as many people have limited contact with services and many are lost to follow up. Refugees seeking asylum in the United States are screened before arrival to identify 'inadmissible conditions' such as HIV, TB, and leprosy. They may also be screened for diseases that are known to be prevalent and targets of public health intervention. For example, Somali refugees were mass treated for malaria and intestinal parasites. The types of health problems sought and treated in newly arrived immigrants also reflect particular areas of concern for asylum seekers, including infectious disease, gynaecological health (rape, genital surgery, STIs), and mental health (torture, trauma) (Adams et al. 2007).

Burnett and Peel (2001) evaluated the health needs of asylum seekers and refugees in Britain. They point out that these groups are subject to 'poverty, dependence, and lack of cohesive social support' as well as racial discrimination. They consider a range of studies examining the health status of refugees and found, for example, that 21% of migrants to Spain from sub-Saharan Africa were chronic carriers of hepatitis B, and 3.4% of refugees arriving in the United States in 1988 had TB. They highlight the lack of screening programmes available and how the focus has tended to be on protecting hosts rather than a responsibility to provide healthcare to all.

One specific refugee and asylum health condition that has been investigated somewhat more extensively than other health issues is mental health (see Chapter 8.7). Gerritsen et al. (2004) describe how early research focused on those who used services, or victims of trauma, followed by early

population-based studies in Western settings which have tended to focus on psychiatric disorders such as post-traumatic stress disorder (PTSD), anxiety, and depression. In a review of the literature, Fazel et al. (2005) show documented rates of PTSD of 4–70%, depression 3–88%, and anxiety 2–80%. Their meta-analysis suggests that one in ten refugees in Western countries have PTSD, one in 20 has major depression, and one in 25 has a generalized anxiety disorder. Although lower than some estimates, this is still ten times higher than rates of PTSD in the general population in the United States. Data for asylum seekers alone is sparse but there are reasons to believe they have even higher rates due to their vulnerability because of disruption, separation from family members, increased levels of torture and trauma, lack of language skills, and the often hostile reception from host population fuelled by media perceptions. The current trend towards detention, and temporary rather than permanent protection has added extra stress, as evidenced by studies of asylum seekers in detention centres and on temporary protection visas showing very high rates of mental illness (Silove et al. 2000b, 2007).

Problems that asylum seekers in particular face, rather than refugees, relate to issues of having a temporary and often contested legal status, and in some cases being considered illegal, with implications (real or perceived) for access to health services. Many countries such as the United Kingdom and Australia have fluctuating policies with regards to what asylum seekers, of differing categories, may be entitled to in terms of healthcare. This can act as a barrier to health for many asylum seekers. The increased use of detention for asylum seekers, while claims are processed, has been challenged heavily by health professionals as a result of evidence that it is detrimental to health. The other popular policy of extending only temporary protection to some asylum seekers, therefore also causing confusion with regards to health entitlements, and long-term insecurity with associated health risks, has also caused concern to health professionals (Silove et al. 2000b). Australia has passed legislation to allow off-shore processing for asylum seekers. A policy of this kind raises more complex questions in terms of entitlement and availability of healthcare. Health professionals have called for screening to ensure those with the most severe mental and physical health problems are identified and if necessary processed on the mainland.

Healthcare for asylum seekers may be difficult to access on many levels: policies of active deterrent of asylum seekers may mean they are not entitled

to care, or there may be lack of prioritizing refugees; health professionals often misunderstand what the policies are and may refuse to register asylum seekers, they also fear the burden of vulnerable groups; those that do offer treatment may lack understanding of the refugee health situation; temporary registration with GPs may be offered rather than permanent (limiting access to immunizations and screening); and language problems are also a great barrier, with many practitioners having inadequate access to interpreting services.

Harris and Telfer (2001) estimated that in 1999–2000 there were over 10,000 asylum seekers living in the community in Australia. Many had to survive for many months or years with no right to work and no medical cover through the national Medicare scheme, resulting in untreated medical conditions. A study in Sydney suggested that most presenting to an Asylum Seeker Centre providing health services had fairly serious psychological or physical symptoms. Some related to trauma and torture, others were infectious diseases, musculoskeletal complaints, and gynaecological problems. Up to 40% of the asylum seekers were denied Medicare. A few are picked up by the Asylum Seeker Assistance Scheme and the Red Cross assists others. Hospital fees are particularly high and case studies have shown that some people are denied hospital treatment until assurance of payment is provided.

Trafficked and smuggled people

People that are trafficked or smuggled are not an entirely distinct group from other types of forced migrants already described. Trafficked people have been coerced or tricked into allowing another person to organize their passage to another country; also smuggled across borders are those who ‘voluntarily’ migrate for a vast range of reasons but feel unable to do so through legal channels. Although these groups may appear distinct from refugee populations they may include Convention refugees and others whose circumstances are very close to those of refugees. Resorting to illegal and clandestine methods of migrating to another country, either through being coerced and misled or through having such limited choices as to choose this route, is often a result of many of the factors behind the forced migration of refugees and internally displaced people.

Gushulak and MacPherson (2000) describe some of the reasons why people may fall into a trafficking situation. Poverty, exposure to violence, restrictive barriers—for example, screening for certain diseases or security barriers, slow processing of routine immigration applications which may cause delay to family reunification—all may encourage people to take alternative routes—especially when traffickers may offer what appear to be attractive possibilities. They state that the reasons why someone is more likely to migrate through these illegal methods are also reasons that place a person at higher risk of having poor health prior to migration. The journeys faced by trafficked and smuggled people are often extremely precarious, with high levels of morbidity and mortality attached as a result: unseaworthy and overcrowded boats on long journeys, stowed in containers, attached to carriers, attempts to cross inhospitable land with extremes of temperature, drowning from swimming attempts, violence from traffickers, and illegal activities such as smuggling drugs with people. On arrival people are then subject to many of the difficulties already outlined that face asylum seekers. Some trafficked and smuggled people may seek asylum and depending on the policies of the countries may incur some health benefits; however, many will be forced into illegal employment and be either too scared to make or prevented from making an approach to health services. Disease prevention and control in this population will be extremely poor—leading to increased prevalence of disease in the trafficked community as a whole.

Busza et al. (2004) examine the need for careful understanding of trafficking situations and to unravel the local dynamics. They state that even when intermediaries are involved, and even when they are taking people into exploitative arenas such as sex work, these situations remain a complex mixture of voluntary and forced migration. While economic disadvantage may fuel the choice, there is still a choice being made in some contexts and organizations seeking to stem the flow or break up these systems may compromise the health or the individuals involved more than they help. Attempts to ‘save’ people may make them even more vulnerable—for example, the raiding of brothels and other such moves can force vulnerable people to go underground making engagement with health services difficult.

The health of trafficked and smuggled people is affected by the fact that they cannot easily be separated out from each other or from other types of ‘voluntary’ migration that are not truly voluntary but have strong forces pushing people on. The policies developed by states towards any of these

groups are likely to impact on the capacity of other groups to access healthcare. Displaced people are more vulnerable to trafficking and may in desperation turn to smugglers. This should be kept in mind by health professionals to avoid vulnerable people missing out on essential services and support.

Public health solutions and interventions

Forced migration clearly has major impacts on public health. Specific solutions and interventions need to be sensitive to culture and context, and include a critical assessment of the political dynamics that may affect how assistance is actually received and impacts on the lives of forced migrants. Harrell-Bond et al. (1992), the doyens of analysts examining forced migration, along with other key commentators such as Crisp (1999) and Bakewell (n.d.), demonstrate how resources to deal with forced migration are structured by broader ideological, political, and economic decisions. Harrell-Bond et al. (1992) argue that refugee camps and many of the structures and organizations set up are part of a system of control of those in need. She argues that the allocation of resources (whether aid, political or military resources) is a 'central component in an ideology of control'. The granting of aid often legitimizes a group and their struggles; politics influences the status and visibility accorded to different groups. Even where humanitarian and development assistance are well targeted at forced migrants in need, the impact of services, systems, and funds, may be both positive and negative. While the forced migrants in great need may be able to access resources, these are not distributed equitably, and local populations, which often bear the brunt of providing for such community members, frequently are marginalized and unable to access new resources or services.

A valuable compilation of insights (Essed et al. 2004) highlights how forced migrants interact with agencies and the implications for ethics, policies, and politics. A case study by Turner (2004) in the same volume, for example, demonstrates how UNHCR effectively displaces older men as providers in their families and in so doing creates new dependencies. UNHCR becomes 'a better husband'—at the same time undermining and transforming traditional household and family relationships. These political

complexities are vital to keep in mind when planning and implementing a public health response with vulnerable communities.

Key actors and organizations in refugee health

There is an extensive range of organizations involved in addressing forced migration. The UN has designated the UNHCR as the key multilateral agency for identifying and dealing with refugee issues. More recently, the mandate of UNHCR, as described earlier, has been extended to include other 'people of concern' including IDPs and asylum seekers. Other agencies with a specific mandate for dealing with forced migrants include the International Organization on Migration, the International Centre for Migration and Health, the American Refugee Committee International, the US Committee on Refugees, and the Global IDP Project.

A wide range of multilateral agencies including the WHO, UN Population Fund (UNFPA), UNICEF, International Committee of the Red Cross (ICRC), International Federation of Red Cross and Red Crescent Societies (IFRC), and UNAIDS has some concern with forced migrants. Key non-governmental organizations (NGOs) include Merlin, Médecins Sans Frontières (MSF), International Medical Corps (IMC), International Rescue Committee (IRC), OXFAM, Save the Children Fund, CARE, Caritas, CONCERN, and the Norwegian Refugee Council. Academic bodies include the Oxford University Refugee Studies Programme, the Centers for Disease Control and Prevention (CDC), and the Refugee Studies Programme in Cairo. A number of important journals, notably the *Journal of Refugee Studies*, *Disasters* and *Conflict and Health*, and the website of Forced Migration Online, provide avenues for academic debate.

Developing countries and complex emergency settings

The principles of public health and epidemiology only began to be systematically applied to complex emergency situations from the early 1970s (Salama et al. 2004). Since then, new academic and practitioner fields have developed, including emergency public health and public health nutrition. With the introduction of an indicator to measure the impact of a complex emergency through CMRs, alongside the development of simplified measurements of malnutrition, and threshold rates for epidemics,

it has become increasingly possible to draw meaningful inferences from data collected.

An understanding of the dynamics of disease and related mortality in CEs has grown, and practice improved on the basis of experience and research evidence. Guidelines and technical manuals have been produced and efforts enhanced to promote organizational learning and enhanced service delivery, building on a growing evidence base. The call for evidence with which to inform policy and practice (Banatvala and Zwi 2000) reinforced what was already happening in the field where practitioners and academics were drawing lessons from experiences and enhancing the quality of their responses.

Mortality rates are now much lower than previously seen, malnutrition is declining in many settings, and measles, a major killer in the past, is less catastrophic following efforts to ensure early immunization of children. Enhanced understanding of risks of communicable diseases has decreased risks, although failures to apply evidence, such as in Goma following the Rwandan genocide, still lead to tens of thousands of preventable deaths. Epidemics of some communicable diseases are better controlled and may be less serious and last for shorter periods given improved surveillance and earlier detection of outbreaks. The pattern seen is often similar to the baseline in developing countries, in terms of the major killers. However, the shift has been to seeing many of the highest mortality rates outside of camps, where the public health response has not been as well developed.

Toole and Waldman (1997) describe the need for primary, secondary, and tertiary prevention. In primary prevention, they argue that political and diplomatic mechanisms should be developed to prevent conflicts evolving to the point of mass displacement, and that epidemiology may be able to add value in studying some of the dynamics that indicate conflict developing on a large scale. Secondary prevention comprises early detection of population movements, contingency planning so that relevant actors are prepared for the scale and nature of likely public health crises, having well-trained personnel with knowledge and experience in dealing with the health problems associated with refugee situations and able to rapidly assess and intervene in partnership with local organizations, along with the capacity to monitor and evaluate programmes swiftly and effectively. Tertiary prevention involves channelling resources into combating the sources of morbidity and mortality already outlined. This involves providing shelter and protection, adequate

food rations, adequate clean water and sanitation facilities, programmes for preventing specific communicable diseases, preparing for epidemics, specific programmes such as maternal and child health or mental health, managing diarrhoeal disease, and setting up health information systems.

The social, cultural, and political challenges of instituting these levels of prevention, alongside the practical implications of working in such settings, are substantial. Salama et al. (2004) describe how the field has been growing over 30 years, with increasing attention by humanitarian actors since the 1990s to codify technical guidelines. In the early 1990s, there was a growing disappointment and recognition among the actors attempting to engage with these complexities, and provide emergency assistance to refugees, that the field was uncoordinated and lacking in technical direction. Research, adverse publicity, and heightened calls for accountability around standards of service delivery and to the supposed beneficiaries of humanitarian interventions, all pointed to the need for significant improvements.

Initiatives were beginning to emerge that were investigating ways to address these issues, but the field was ultimately galvanized by events in 1994 in Goma, Zaire. The unprecedented rapid influx of refugees and the resulting massive cholera epidemic, producing one of the highest mortality rates recorded, led to a serious reappraisal of humanitarian work in these settings. The Joint Evaluation of Emergency Assistance to Rwanda was an impressive, multiagency, multidonor-funded initiative, paving the way to a deep review of the humanitarian and relief field and the standards of care and conduct.

In terms of future public health directions, Salama et al. (2004) describe the need to move beyond the current refugee and IDP paradigms, as non-camp populations affected by conflict are now often greater in number and burden, and mechanisms to deliver services to such populations are less clear. There is a need for enhanced coverage in all settings of known effective public health interventions, and for the more upstream causes of CEs to be addressed. The technical needs identified include the development of more context-specific methods for assessing mortality rates as an indicator based on surrounding baseline CMRs, and considering which other indicators may point to a deteriorating situation such as fluctuations in disease incidence, levels of displacement, and decreasing food security. They suggest that implementation of control measures should be early, not when mortality and malnutrition are already high, and that while communicable

disease control should continue to be prioritized along with nutrition and food security, that enhanced efforts at addressing mental health, reproductive health, and neonatal health should also be incorporated. Important policy issues include better means of determining direction and coordination of public health interventions, early identification of a lead agency, and greater effort to support fledgling governments to take the lead with the support of a coherent international system. They suggest that the skills of relief workers should be strengthened and broadened to include knowledge of assessment and prevention of disease, monitoring and evaluation, and international systems, policies, and regulatory bodies.

Ensuring equity in the response and equity in access to services and resources made available remains a considerable and underemphasized challenge.

Developed countries

In developed countries, the complex political arena and the increasingly restrictive views of host populations towards asylum seekers and refugees complicates the task of addressing the public health of forced migrants. While an array of motivated and effective actors plays a number of different roles in this field, there are few central bodies charged with overseeing the public health of forced migrants as a whole. The policies of the host nations themselves are likely to shape the public health landscape and in situations where these policies seem to be detrimental to the health of this population, public health professionals may find themselves having to take increasingly political roles to fulfil their public health duties.

Advocacy groups and health professional organizations have set up refugee and asylum seeker health networks and centres through which to provide services and support and advocate for more inclusive social and health policies.

The Health of London Project established a health centre in 1999 to provide access to good-quality primary care for forced migrants. The service aims to provide a thorough initial health assessment, along with effective responses to issues such as mental health problems and communicable diseases. In Europe, Médecins du Monde, an NGO, has stepped in to provide care for groups with restricted access to healthcare, particularly failed asylum seekers.

The best way in which to organize healthcare for resettled refugees, and for asylum seekers, remains contentious. A key issue point of debate is whether they should be integrated within mainstream services, or offered separately through distinct services for these marginalized groups (Finney Lamb and Cunningham 2003). Such centres are often voluntarily run and unable to cover the full extent of need. Hull and Boomla (2006) discuss the implications of such NGO provision, including some of the unanticipated negative aspects.

Another potential theoretical divide that can emerge with health approaches to asylum seekers and refugees is that between public health and human rights frameworks. Governments are more responsive to public health arguments whereby health problems of forced migrant populations may be addressed for the benefit of the wider population, often highlighting the risk brought with incoming refugees and forced migrants of communicable diseases. However, these approaches focus more on protecting the host than the forced migrants' right to receive healthcare (Burnett and Peel 2001) and may reinforce discrimination and marginalization of the migrants who are seen as posing a risk to the general community and demanding excessive use of public services (Grove and Zwi 2006). Tarantola et al. (2008) argue for much closer links between public health, human rights, and development—all equally relevant to forced migration.

Hull and Boomla (2006) describe recent UK government proposals to deny general practitioners the discretion of registering overseas visitors including failed asylum seekers, and to provide care to them. They highlight the possibility that such a policy will inadvertently increase costs through more emergency admissions, and they stress that such a policy is a breach of human rights, ethically unsupportable, and may place clinicians in conflict with their professional duties.

Ashcroft (2005) describes the language of the UK and Australian governments as 'Orwellian' in an overview of their approaches to asylum seekers and health. In discussing the UK government's move to make failed asylum seekers pay for all non-emergency care, despite not having the right to work, he states that the treatment of asylum seekers is an ethical issue and that denial of medical treatment should not be used as a lever to exclude people or get them to leave the country. In some cases, the presence of HIV seropositivity has been used to exclude refugees from the opportunity to

resettle in a third country; this is of particular concern given the heightened risk of violence and infection that may characterize the lives of many refugees. Using this against such vulnerable community members is a double victimization.

As Silove et al. (2000b) stated:

> the medical profession has a legitimate role in commenting on the general and mental health risks of imposing restrictive and discriminatory measures on asylum seekers, especially when some of these administrative procedures threaten one of the fundamental principles underpinning the practice of medicine: primum non nocere [do no harm].

In response to current debates in Australia, the Australian Medical Association issued a position statement on the Health Care of Asylum Seekers and Refugees (Australian Medical Association 2011). This lays out not only the health vulnerabilities of these populations but also that medical practitioners should insist on the rights of their patients being respected and lays out certain competencies necessary for treating asylum seekers. These include identifying victims of torture, assessing and managing trauma, identifying suicide risk, and recognizing health conditions specific to these groups such as TB.

Further study is required to document the patterns of disease, the links to the asylum experience, and impact of policies, not only on forced migrant health but also the implications for host populations are required. Given the more restrictive policy environment, health professionals require a strong evidence base from which to lobby for improved access to services as a first step in improving the health of refugees and asylum seekers in developed countries.

Emerging issues for public health professionals

Many of the issues raised in earlier sections have a technical focus, even if their links with broader sociopolitical issues have been acknowledged. In this concluding section of the chapter, we highlight some of the many issues which challenge professionals and societies, not only on a technical level, but in terms of the values underpinning how we, individually and collectively, whether as citizens or professionals, see and respond to others, especially those who have been forced to flee.

This section seeks to highlight, briefly, a number of key themes: globalization and poverty, quality, and accountability, ethics, and researching with forced migrants, voice, visibility, and agency. They are central to understanding the context within which forced migration arises and health needs should be understood and addressed.

Globalization and poverty

There is increasing evidence of widening gaps in income and basic needs between the well-off and the impoverished in many settings. Labonte and Schrecker (2005) identify the key links between current patterns of globalization and the influence these have on the social determinants of health. The challenges are evident in the failure to address the Millennium Development Goal (MDG) targets, while extreme wealth continues to be generated and monopolized by small fractions of the world's population. Resource constraints are likely to increase; countries engaging actively with global economic and political opportunities may do well but not all their citizens will necessarily benefit.

Evidence of gaps is widespread; and it is apparent that the MDGs, a major driving force at present in international development, do not adequately address issues of equity (Attaran 2005). Such concerns are likely to be amplified in relation to especially vulnerable and socially excluded communities such as forced migrants, data on whom may be unavailable and systems of response to which will be fragmented, uncoordinated, and inadequately monitored. Colson (2007) argues cogently that much greater emphasis should be placed on understanding the 'factors that provide the impetus to leave' and not focusing only on the aftermath of such forced migrations.

Quality and accountability

A key challenge in engaging with situations of violence, social exclusion, and disempowerment is in ensuring that work undertaken to assist those affected does no harm, or at least strives to limit the potential harm and maximize the potential value. The international community, through the UN, other multilateral and bilateral agencies, and international NGOs, all ostensibly seek to assist those forced to flee or migrate.

One way in which evidence improvements are underway is in relation to enhancing and seeking to assure the quality of humanitarian and relief work. Major quality and accountability initiatives such as SPHERE, the Humanitarian Code of Conduct, the Humanitarian Accountability Partnership, and many others contribute to enhanced practice. There remains a need to facilitate self-reflection and open review, transparency of difficulties limitations and failures, and promotion and funding for the search to do better.

The vast majority of incentives currently in place, however, encourage the hiding, rather than the declaration, of weaknesses, so as to elicit the next tranche of funding. This is counterproductive when seeking to reflect and improve services and their responsiveness to need. New thinking is required: how do we develop systems of trust and collegiality, mutual support between affected communities, services providers, and agencies, to ensure that all incentives operate to maximize the gain for those at greatest risk?

New initiatives in the humanitarian and relief areas of activity seek to enhance practice and open out debate and public scrutiny. In so doing, weaknesses are identified, guidelines and lessons for better practice formulated, and processes which seek to engage and creatively respond to problems, developed. Taking forward such action requires promoting an ability to report and record limitations and inadequacies, and to carefully and systematically document strengths and limitations, weaknesses, and potentials, to address real problems. Services are never ideal and always have constraints and limits; it is crucial to establish ways of examining them such that opportunities to change, to experiment, to creatively engage with solving problems are established.

New and enhanced knowledge management systems would be most valuable if they assisted in finding innovative means of conceptualizing problems, sharing information and insights of what works in different contexts, and developing systems to ensure accountability to the intended beneficiaries, the forced migrants themselves.

Ethics and researching with forced migrants

Ethical issues require considerable attention especially given the power imbalances present and in relation to research. Forced migrants may be especially vulnerable and risk being abused, either through their stories and

survival strategies being exposed and used by their opponents to undermine them; the very practice of talking with an outsider or sharing information may heighten risk, vulnerability, and the attention of those seeking further to undermine and disempower. Real risks may result from security lapses, confidentiality lapses, exposure of coping strategies and adaptations; as well as stigma and reinforcement of difference and of 'otherness'.

A major challenge in research with refugees and other forced migrants is to assist them in documenting their experiences of forced migration (Mackenzie et al. 2007). In so doing, the locus of power in the research relationship has the potential to shift from the researcher him or herself, to the communities being researched. The project may shift from research *on* refugees to research *with* refugees and research *by* refugees for action and change.

While this may be conceptually and morally attractive, the practicalities of achieving this are impeded at multiple levels, not least of which is establishing a trusting relationship between the researcher and the communities with whom they research. Engaging with forced migrants is itself a major challenge: Such communities are often highly dependent on others, and may be reliant on outside agencies with expertise and resources and/or on local power-brokers in control of the basic necessities of life and survival—food, water, shelter, and security.

Zwi et al. (2006) have highlighted the importance of reconceptualizing such research to consider the potential to ensure benefit and reciprocity for the subjects of research and to shift the locus of decision-making in their direction. They offer a model to describe this process, while Mackenzie et al. (2007) present a case study of applying many of these considerations to work with forced migrants on the Thai–Burmese border. Of note is the transfer of power to community organizations, with women's groups from the Thai–Burmese border having control over the stories and narratives collected from them and academics being required to obtain permission from them to publish and distribute and write up such material.

Innovative approaches to research are emerging and warrant support. Youth-focused research highlights the importance of young people in exploring and documenting their own lives and experiences. Young people play a central role in refining and redefining the issues, identifying the challenges and the potential approaches to solving them through new models

of participatory action research. Such collaboration needs to build skills, resilience, and opportunities to shape the future.

Voice, visibility, and agency

There is increasing recognition of the importance of hearing from those most affected by any public health condition and ensuring that their voices are heard; empowering them to help shape the proposed solutions. Black (2003) emphasizes the importance of advocacy as a central component of work with marginalized and disempowered groups and reinforces the centrality of ensuring that the voices of those affected, and of the range of stakeholders engaged, is heard.

The voices of forced migrants may be silenced while those of powerful agencies and governments are privileged in shaping our understanding both of the problems and of the solutions which surround forced migration (Harrell-Bond and Voutira 2007). Power imbalances, vulnerability, and cultural and linguistic differences are among the many barriers to creating an environment in which refugees and IDPs are able to take an active part in shaping solutions. As with the youth, truly participatory models of action and research require sensitivity to the challenges of the context and understanding of the conditions that usually deny refugees a voice.

It is crucial that we hear the voices of those most affected by forced migration if we are to recognize the range of experiences, and the positive and negative change and transformation which results. As Eastmond (2007) states eloquently: ‘while transformation and change are part of the refugee experience, not all change is perceived as loss or defined as problematic or unwelcome by all individuals involved. Nor are refugees necessarily helpless victims, but rather likely to be people with agency and voice’ (p. 253). Learning and recognizing how agency, transformation, resilience, adaptation, and change are shaped, notably, but not only, in relation to gender and power, would not be possible without hearing from and learning with, those most affected.

Conclusion

Forced migrant populations frequently suffer vast health challenges. Their experiences need to be understood within the context of accelerating

globalization, with its positive and negative features transforming the lives of millions of people.

The role of public health professionals in working alongside forced migrants to bring about improvements in their health is demanding and complex. A comprehensive understanding of the dynamics of forced migration and sensitivity to the constraints placed on individuals and populations is vital to gain trust and provide opportunities for transforming health outcomes. Particularly, given the increasingly restrictive environments that forced migrants find themselves in, public health professionals can have a crucial role in advocating for forced migrant health, enabling forced migrants to speak and be heard, and to bring about improvements in their own health.

Improving knowledge and understanding is central to affirming the rights and the agency of those affected. Public health can play an important role, and indeed has done so, in improving technical quality of service delivery. Key gaps remain, however, in understanding, appreciating, and responding to the varied ways in which those forced to move seek to shape, transform, and better their own lives and experiences.

ACKNOWLEDGEMENTS

Text extracts from United Nations High Commissioner for Refugees (UNHCR), *Protecting Refugees and the Role of UNHCR*, Copyright © UNHCR 2009, available from http://unhcr.org.au/unhcr/images/protecting%20refugees.pdf, reproduced with permission from the United Nations.

Text extracts from Deng, F., *The Guiding Principles on Internal Displacement*, United Nations, New York, Copyright © United Nations 1998, reproduced with permission from the United Nations.

Text extract from United Nations, Protocol to Prevent, Suppress and Punish Trafficking in Persons, Especially Women and Children, Supplementing the United Nations Convention Against Transnational Organized Crime, G.A. Res. 25, Annex II, U.N. GAOR, 55th Session, Copyright © United Nations 2003, reproduced with permission from the United Nations.

Text extract from United Nations, *Protocol Against the Smuggling of Migrants by Land, Sea and Air*, Copyright © United Nations 2000,

reproduced with permission from the United Nations.

Text extract from United Nations, *Glossary of Environment Statistics, Studies in Methods, Series F*, United Nations, New York, USA, Copyright © United Nations 1998, reproduced with permission from the United Nations.

Text extracts from Forced Migration Online, *What is Forced Migration?*, Copyright © 2010 Forced Migration Online and Contributors, available from http://www.forcedmigration.org/whatisfm.htm, reproduced with permission from Forced Migration Online.

FURTHER READING

The following manuals give technical advice in managing public health aspects of forced migration:

Médecins Sans Frontières (MSF) (1997). *Refugee Health: An Approach to Emergency Situations*. London: Macmillan. [Online] Available at: http://www.refbooks.msf.org/msf_docs/en/Refugee_Health/RH1.pdf

World Health Organization (WHO) (2006). *Communicable Disease Control in Emergencies. A Field Manual*. [Online] Available at: https://www.who.int/diseasecontrol_emergencies/publications/9241546166/en/

RELEVANT WEBSITES

ALNAP—'learning, accountability, performance in humanitarian action' http://www.alnap.org
Forced Migration Online—http://www.forcedmigration.org
SPHERE—http://www.sphereproject.org
The Humanitarian Accountability Partnership—http://www.hapinternational.org
United Nations High Commissioner for Refugees—http://www.unhcr.org

REFERENCES

Adams, K., Gardiner, L., and Assefi, N. (2007). Healthcare challenges from the developing world: post-immigration refugee medicine. *British Medical Journal*, **328**, 1548–52.

Allotey, P.A. and Reidpath, D.D. (2003). Refugee intake: reflections on inequality. *Australian and New Zealand Journal of Public Health*, **27**, 12–16.

Apodaca, C. (1998). Human rights abuses: precursor to refugee flight? *Journal of Refugee Studies*, **11**, 80–93.

Ashcroft, R. (2005). Standing up for the medical rights of asylum seekers. *Journal of Medical Ethics*, **31**, 125–6.

Attaran, A. (2005). An immeasurable crisis? A criticism of the millennium development goals and why they cannot be measured. *PLoS Medicine*, 2, e318.

Australian Medical Association (2011). Health care of asylum seekers and refugees—2011 [Online]. Available at: https://ama.com.au/position-statement/health-care-asylum-seekers-and-refugees-2011

Bakewell, O. (n.d.). Can we ever rely on refugee statistics? [Online]. Available at: http://www.radstats.org.uk/no072/article1.htm

Banatvala, N. and Zwi, A.B. (2000). Public health and humanitarian interventions: developing the evidence base. *British Medical Journal*, **321**, 101–5.

Barnett, L. (2002). Global governance and the evolution of the international refugee regime. *International Journal of Refugee Law*, **14**, 238–62.

Baum, F. (2001). Health equity justice and globalisation: some lessons from the People's Health Assembly. *Journal of Epidemiology and Community Health*, **55**, 613–16.

Biermann, F. and Boas, I. (2007). *Preparing for a Warmer World—Towards a Global Governance System to Protect Climate Refugees.* Global Governance Working Paper. Amsterdam: Global Governance Project.

Biggs, B. and Skull, S. (2003). Refugee health: clinical issues. In: Allotey, P. (ed.) *The Health of Refugees: Public Health Perspectives from Crisis to Settlement*. Melbourne: Oxford University Press, pp. 54–67.

Black, R. (2003). Ethical codes in humanitarian emergencies: from practice to research? *Disasters*, **27**, 95–108.

Burnett, A. and Peel, M. (2001). Asylum seekers and refugees in Britain: health needs of asylum seekers and refugees. *British Medical Journal*, **322**, 544–7.

Busza, J., Castle, S., and Diarra, A. (2004). Trafficking and health. *British Medical Journal*, **328**, 1369–71.

Castles, S. (2003). Toward a sociology of forced migration and social transformation. *Sociology*, **37**, 13–34.

Castles, S. (2004). Confronting the realities of forced migration. [Online] Available at: https://www.migrationpolicy.org/article/confronting-realities-forced-migration

Cernea, M. (2007). Development-induced and conflict-induced IDPs: bridging the research divide. *Forced Migration Review*, Special Issue, 25–7.

Chan, E.Y. and Sondorp, E. (2007). Medical interventions following natural disasters: missing out on chronic medical needs. *Asia Pacific Journal of Public Health*, **19**, 45–51.

Colson, E. (2007). Linkage methodology: no man is an island. *Journal of Refugee Studies*, **20**, 321–33.

Connolly, M.A., Gayer, M., Ryan, M.J., Speigel, P., Salama, P., and Heymann, D.L. (2004). Communicable diseases in complex emergencies: impact and challenges. *The Lancet*, **364**, 1974–83.

Courtland Robinson, W. (2003). *Risks and Rights: The Causes, Consequences, and Challenges of Development-Induced Displacement*. Washington, DC: The Brookings Institution-SAIS Project on Internal Displacement.

Crisp, J. (1999). *'Who has Counted the Refugees?' UNHCR and the Politics of Numbers. New Issues in Refugee Research*. Geneva: UNHCR.

Deng, F. (1998). *Guiding Principles on Internal Displacement*. New York: OCHA.

Dualeh, M. and Shears, P. (2002). Refugees and other displaced populations. In: Detels, R., McEwen, J., Beaglehole, R., and Tanaka, H. (eds.) *Oxford Textbook of Public Health* (4th ed.). Oxford: Oxford University Press, pp. 1737–56.

Eastmond, M. (2007). Stories as lived experience: narratives in forced migration research. *Journal of Refugee Studies*, **20**, 248–64.

El-Hinnawi, E. (1985). *Environmental Refugees*. Nairobi: United Nations Environment Programme.

Essed, P., Frerks, G., and Schrijvers, J.E. (eds.) (2004). *Refugees and the Transformation of Societies. Agency, Policies, Ethics and Politics*. New York: Berghahn Books.

Fazel, M., Wheeler, J., and Danesh, J. (2005). Prevalence of serious mental disorder in 7000 refugees resettled in western countries: a systematic review. *The Lancet*, **365**, 1309–14.
Feachem, R. (2001). Globalisation is good for your health, mostly. *British Medical Journal*, **323**, 504–6.
Finney Lamb, C. and Cunningham, M. (2003). Dichotomy or decision-making: specialization and mainstreaming in health service design for refugees. In: Allotey, P. (ed.) *The Health of Refugees. Public Health Perspectives from Crisis to Settlement*. Melbourne: Oxford University Press, pp. 156–68.
Forced Migration Online (2008). *What is Forced Migration?* [Online]. Available at: http://www.forcedmigration.org/whatisfm.htm
Gerritsen, A., Bramsen, I., Deville, W., Van Willigen, L., Hovens, J., and Van Derploeg, H. (2004). Health and health care utilisation among asylum seekers and refugees in the Netherlands: design of a study. *BMC Public Health*, **4**, 1–10.
Grove, N. and Zwi, A.B. (2006). Our health and theirs: forced migration, othering and public health. *Social Science & Medicine*, **62**, 1931–42.
Gushulak, B. and MacPherson, D. (2000). Health issues associated with the smuggling and trafficking of migrants. *Journal of Immigrant Health*, **2**, 67–78.
Harrell-Bond, B. (2000). *'Are Refugee Camps Good for Children?' New Issues in Refugee Research*. New York: UNHCR.
Harrell-Bond, B. and Voutira, E. (2007). In search of 'invisible actors': barriers to access in refugee research. *Journal of Refugee Studies*, 20(2), 282–98.
Harrell-Bond, B., Voutira, E., and Leopold, M. (1992). Counting the refugees: gifts, givers, patrons and clients. *Journal of Refugee Studies*, 5(3/4), 205–25.
Harris, M.F. and Telfer, B.L. (2001). The health needs of asylum seekers living in the community. *Medical Journal of Australia*, **175**, 589–92.
Hull, S. and Boomla, K. (2006). Primary care for refugees and asylum seekers. *British Medical Journal*, **332**, 62–3.
Hynes, M., Sheik, M., Wilson, H.G., and Spiegel, P. (2002). Reproductive health indicators and outcomes among refugee and internally displaced persons in postemergency phase camps. *Journal of the American Medical Association*, **288**, 595–603.
Internal Displacement Monitoring Centre (2008). *Internal Displacement: Global Overview of Trends and Developments in 2007*. Geneva: Internal Displacement Monitoring Centre.
Krause, S.K., Jones, R.K., and Purdin, S.J. (2000). Programmatic responses to refugees' reproductive health needs. *International Family Planning Perspectives*, **26**, 181–7.
Labonte, R. and Schrecker, T. (2005). Globalization and social determinants of health: promoting health equity in global governance. *Globalization and Health*, 3(7).
Lee, K. (2004). Globalisation: what is it and how does it affect health? *Medical Journal of Australia*, **180**, 156–8.
MacDonald, T.H. (2007). *The Global Human Right to Health: Dream or Possibility?* Oxford: Radcliffe Publishing.
MacKenzie, C., McDowell, C., and Pittaway, E. (2007). Beyond 'do no harm': the challenge of constructing ethical relationships in refugee research. *Journal of Refugee Studies*, **20**, 299–319.
Martin, S. (2002). Averting forced migration in countries in transition. *International Migration*, **40**, 25–40.
McAdam, J. (2006). *The Refugee Convention as a Rights Blueprint for Persons in Need of International Protection. New Issues in Refugee Research*. New York: UNHCR.
Moore, P.S., Marfin, A.A., Quenemoen, L.E., Gessner, B.D., Ayub, Y.S., and Sullivan, K.M. (1993). Mortality rates in displaced and resident populations of Central Somalia during the famine of 1992. *The Lancet*, **341**, 935–8.
Moore, W. and Shellman, S. (2004). Fear of persecution: forced migration 1952–1995. *Journal of Conflict Resolution*, **48**, 723–45.

Myers, N. and Kent, J. (1995). *Environmental Exodus: An Emergent Crisis in the Global Arena*. Washington, DC: Climate Institute.
Organization of African Unity (OAU) (1969). *OAU Convention Governing the Specific Aspects of the Refugee Problems in Africa*. Addis-Ababa: OAU.
Prince, M., et al. (2007). No health without mental health. *The Lancet*, **370**, 859–77.
Prinzo, W. and Benoist, B. (2002). Meeting the challenges of micronutrient deficiencies in emergency affected populations. *Nutritional Society*, **2002**, 251–7.
Regional Refugee Instruments & Related (1984). *Cartagena Declaration on Refugees. Colloquium on the International Protection of Refugees in Central America, Mexico and Panama*. [Online] Available at: http://www.oas.org/dil/1984_Cartagena_Declaration_on_Refugees.pdf
Salama, P., Spiegel, P., and Brennan, R. (2001). No less vulnerable: the internally displaced in humanitarian emergencies. *The Lancet*, **357**, 1430–2.
Salama, P., Spiegel, P., Talley, L., and Waldman, R. (2004). Lessons learned from complex emergencies over past decade. *The Lancet*, **364**, 1801–13.
Siddique, A.K., et al. (1995). Why treatment centres failed to prevent cholera deaths among Rwandan refugees in Goma, Zaire. *The Lancet*, **345**, 359.
Silove, D., Austin, P., and Steel, Z. (2007). No refuge from terror: the impact of detention on the mental health of trauma-affected refugees seeking asylum in Australia. *Transcultural Psychiatry*, **44**, 359–93.
Silove, D., Ekblad, S., and Mollica, R. (2000a). The rights of the severely mentally ill in post-conflict societies. *The Lancet*, **355**, 1448–549.
Silove, D., Steel, Z., and Watters, C. (2000b). Policies of deterrence and the mental health of asylum seekers. *Journal of the American Medical Association*, **284**, 604–11.
Spiegel, P. (2004). HIV/AIDS among conflict-affected and displaced populations: dispelling myths and taking action. *Disasters*, **28**, 322–39.
Tarantola, D., Byrnes, A., Johnson, M., Kemp, L., Zwi, A.B., and Gruskin, S. (2008). Human rights, health and development. *Australian Journal of Human Rights*, **13**, 2.
Toole, M. and Waldman, R. (1990). Preventing excess mortality in refugee and displaced populations in developing countries. *Journal of the American Medical Association*, **263**, 3296–302.
Toole, M. and Waldman, R. (1997). The public health aspects of complex emergencies and refugee situations. *Annual Review of Public Health*, **18**, 283–312.
Turner, S. (2004). New opportunities: angry young men in a Tanzanian refugee camp. In: Essed, P., Frerks, G., and Schrijvers, J. (eds.) *Refugees and the Transformation of Societies. Agency, Policies, Ethics and Politics*. New York: Berghahn Books, pp. 94–106.
United Nations (1997). *Glossary of Environment Statistics, Studies in Methods. Series F*. New York: UN.
United Nations (2000). *Protocol Against the Smuggling of Migrants by Land, Sea and Air*. New York: UN.
United Nations (2003). *Protocol to Prevent, Suppress and Punish Trafficking in Persons, Especially Women and Children, Supplementing the United Nations Convention Against Transnational Organized Crime, G.A. Res. 25, annex II, U.N. GAOR*, 55th Sess. New York: UN.
United Nations High Commissioner for Refugees (2006). *Looking to the Future. The State of The World's Refugees 2006: Human Displacement in the New Millennium*. Geneva: UNHCR.
United Nations High Commissioner for Refugees (2009). *Protecting Refugees and the Role of UNHCR*. [Online] Available at: http://unhcr.org.au/unhcr/images/protecting%20refugees.pdf
United Nations High Commissioner for Refugees (2011). *UNHCR Global Trends*. Geneva: UNHCR.
World Health Organization (WHO) (2014). Measles deaths reach record lows with fragile gains towards global elimination. [Online]. Available at: http://www.who.int/mediacentre/news/notes/2014/measles-20140206/en/.

Zwi, A.B. and Alvarez-Castillo, F. (2003). Forced migration globalization and public health: getting the big picture into focus. In: Allotey, P. (ed.) *The Health of Refugees*. Melbourne: Oxford University Press, pp. 14–34.

Zwi, A.B. and Cabral, A.J. (1991). Identifying 'high risk situations' for preventing AIDS. *British Medical Journal*, **303**, 1527–9.

Zwi, A.B., et al. (2006). Placing ethics in the centre: negotiating new spaces for ethical research in conflict situations. *Global Public Health*, **1**, 264–77.

10.10

Prisoners

A wicked problem for public health

Tony G. Butler and Peter W. Schofield

Background

The relationship between incarceration and population and individual health outcomes is influenced by a vast array of contextual factors. Broad determinants include national or regional differences in criminal justice system policies and their application—which influence how many individuals are incarcerated—or in the nature of the healthcare system as it functions for the general public and in prison. At the level of the individual, factors such as pre-offending medical history, indigenous or minority status, gender, education and occupational status, financial resources, and family history may each influence the impact of criminal justice system interventions on health outcomes. Incarceration can be evaluated for its immediate impact on health (i.e. among prison inmates) as well as for the consequences for health upon release. Furthermore, from a public health perspective, it is appropriate to consider the health consequences of incarceration not only upon prisoners themselves, but on their family and friends. The challenge in confronting the health needs of those incarcerated indeed constitutes a ‘wicked problem’, embodying huge social complexities, including striking health inequities, for which optimal data are frequently lacking (Plamondon and Pemberton 2019).

Historical perspective on prison public health

Prison health reformers such as John Howard and Sir George Onesiphorous Paul in the 18th century recognized that infectious conditions could easily

spread from those detained in gaols to the wider community, with dire consequences (Howard 1973). The main point of contact between the incarcerated population and the community occurred when those in detention were brought up from the bowels of squalid prisons to attend court. Well-documented cases exist whereby infectious diseases spread from prison to the community during so-called assizes (courts set up by judges who travelled across various circuits in England and Wales). Notable outbreaks (dubbed ‘Black Assizes’), included the Oxford Black Assize outbreak of 1577 resulting in over 300 deaths. The Black Assize of Exeter 1586 was caused by the filthy condition of Exeter Prison situated underneath the court. By the time the assize occurred, some individuals had been in gaol for a long period; so weak and thin, they had to be carried into the courtroom and subsequently spread disease to the judges and locals in attendance, and infected the population of the Westcountry for many months after. Typhus, known as ‘gaol fever’, was responsible for these deaths.

In 1759, an English authority estimated that each year one-quarter of prisoners died from gaol fever and that ‘Imprisonment until the next term of court’ was often equivalent to a death sentence. This concept of prisons as a public health threat and the need for interventions to improve health is eloquently expressed by the eighteenth-century reformer and philanthropist, Sir George Onesiphorous Paul, who in 1784 stated that prison hygiene procedures were:

> ‘Not for the benefit of the prisoners but rather to increase the 'salutary humiliation' of prison life and to prevent the spread of epidemic disease beyond the prison walls and to the general citizenry’.

The spread of disease from prisons to the wider community can take place prior to release because of the substantial contact that occurs between prisoners and the general community during incarceration from legal interactions such as court appearances, via the vast number of prison guards, teachers, administrators, health staff, and welfare workers employed in prisons, and by relatives and friends during visits and day-release. All provide a potential bridge between prison and the community. Furthermore, the cycle of incarceration ensures that many individuals relentlessly cycle back and forth between prison and the community (Andolsek 1997).

This notion of the prison as an incubator of disease originates from these historic examples but persists today in the consciousness of prison public health officials. While ‘gaol fever’ is not the focus of concern, new threats

such as tuberculosis, influenza, Ebola, HIV, viral hepatitis, and COVID-19 have replaced it. A striking example of prisoners' health impacting on the general community occurred in the countries of the former Soviet Union around the turn of the twentieth century where the epidemic of tuberculosis (TB) among the prisoner population was linked to its re-emergence in the community (Farmer et al. 1999; Stern 1999, 2000; Kenyon 2009). In 1997, an amnesty of 40,000 prisoners occurred, of whom 4,000 were estimated to have TB with a further amnesty releasing up to 120,000 prisoners to relieve chronic prison overcrowding (Gentleman 2000). This example illustrates two salient factors in prisoner healthcare—the need to ensure that appropriate infection control measures are in place in prison during incarceration to prevent the spread of infection within the prisoner population, and secondly that prison health services should be aligned with those of the community to ensure the standard of healthcare in prison is the same as that afforded to those in the community. This latter point is referred to as the 'principle of equivalence' and is underpinned by the United Nations' Principles for the Treatment of Prisoners (United Nations 2015).

Prisoner populations: an international perspective

Offender populations, most notably prisoners, comprise some of the most stigmatized and socially excluded groups in society, and are characterized by extreme socioeconomic, health, and psychological disadvantage. Those exposed to the criminal justice system are typically poorly educated, unemployed, socially isolated, and financially dependent. Epidemiological surveys of prisoners consistently find high levels of physical and psychiatric ill health, exposure to communicable diseases such as bloodborne viruses and sexually transmissible infections, and engagement in health risk behaviours including tobacco use, violence, harmful alcohol consumption, and illicit drug use. It is unsurprising therefore that studies have found increased mortality compared with community peers, particularly in the areas of drug overdose and suicide (Kariminia et al. 2007a, 2007b). Over the past 15 years, numerous systematic reviews and meta-analyses have unambiguously highlighted the extreme health needs of the offender population across all health domains (Fazel and Baillargeon 2011; Dolan et al. 2016; Fazel and Seewald 2018). While this information provides an important platform for health policy formulation it does not supplant the

need for better, more effective interventions to prevent events such as suicide and drug overdose. Without question, enormous challenges face prison medical services and those involved in prison health provision and research with respect to prioritizing the use of limited resources.

Prisoner populations

Overall, the world's prisoner population increased in absolute terms between 2000 and 2015 to over 10 million (Coyle et al. 2016), this 20% increase in the world's total prison population is broadly in line with the increase in the overall population (18%) over this period (see Figs. 10.10.1 and 10.10.2 from Coyle et al. 2016). However, at regional level, the prisoner populations of the Americas, Asia and Oceania all outstripped natural population growth whereas in Africa and Europe the prisoner population grew at a lesser rate than the overall population. In percentage terms, Europe is the only region showing a trend towards decarceration.

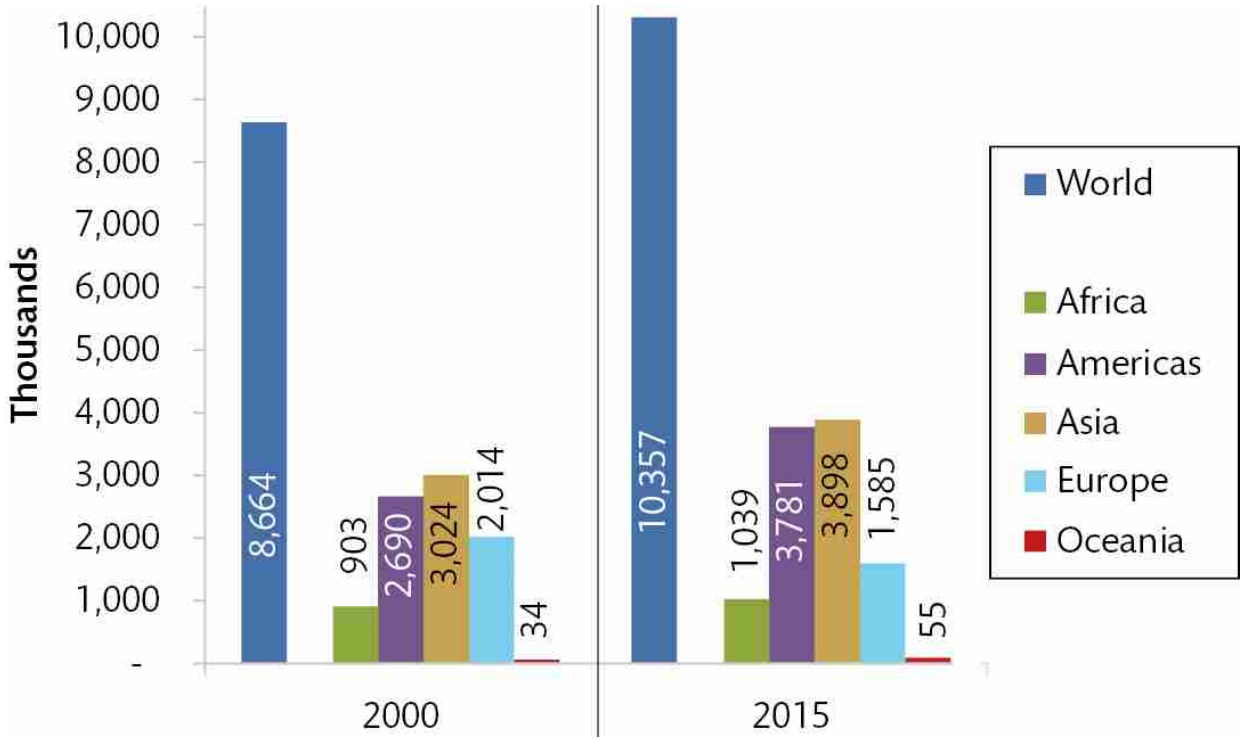


Fig. 10.10.1 Prison population totals, 2000 and 2015*: entire world and by region.

*Data are aggregated national prisoner population totals for the year 2000 or nearest available year. Figures for 2015 are based on World Prison Brief data as of November 2015 and vary as to the specific dates to which they refer. Source: data from Coyle A, Fair H, Jacobson J, Walmsley R. Imprisonment worldwide. The current situation and an alternative future. Policy Press, Bristol. 2016.

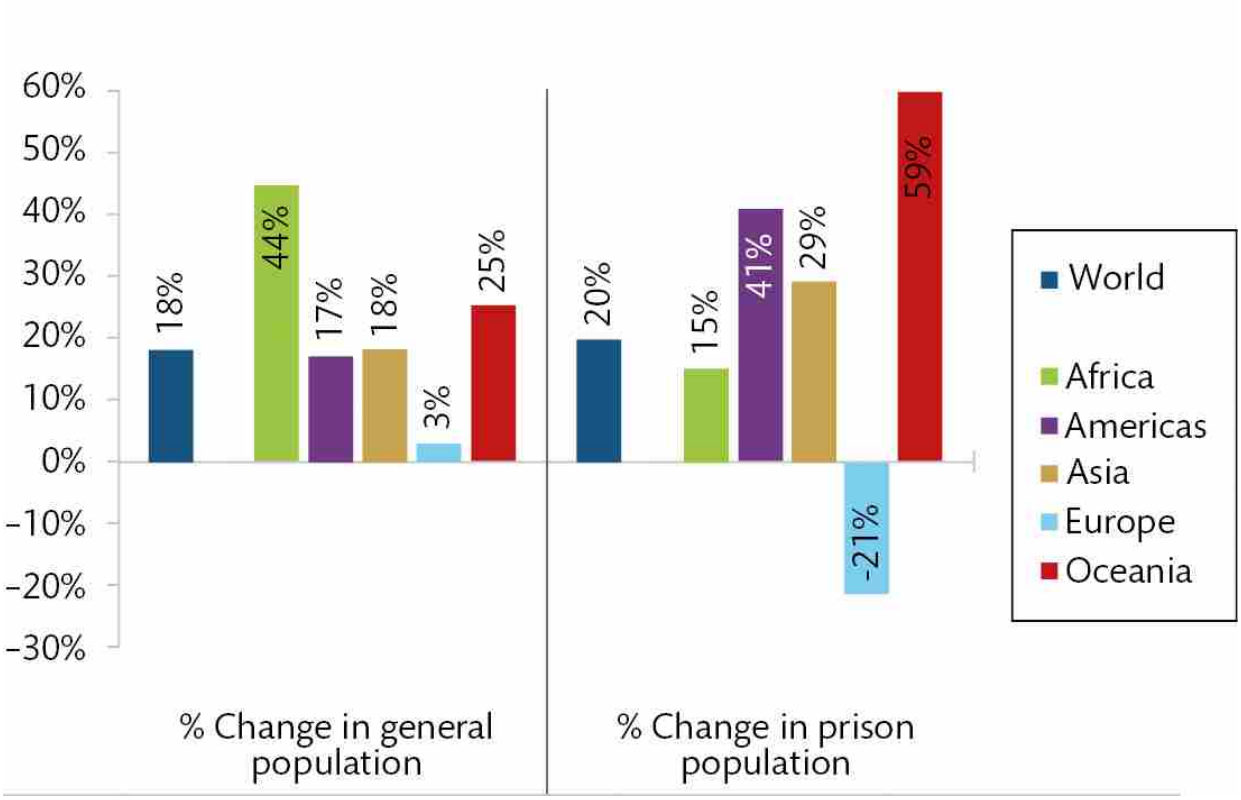


Fig. 10.10.2 Percentage change in general population total and prison population totals, 2000 to 2015: entire world and by region.

Source: data from Coyle A, Fair H, Jacobson J, Walmsley R. Imprisonment worldwide. The current situation and an alternative future. Policy Press, Bristol. 2016.

The United States of America (USA) is the world's largest incarcerator with over two million individuals currently behind bars (Drucker 2011). Although crime rates have fallen in recent times in many countries, the prisoner population has increased due to myriad changes in justice approaches such as: 'three strikes and you're out' (in some states in the USA), mandatory minimum prison terms, better police detection methods allowing cold-case prosecutions, commissions of inquiry into historic sexual abuse of those in care, and a trend towards bail refusal (in the case of Australia) (Coyle et al. 2016).

Internally, prisoner populations are far from static with significant prisoners movements occurring due to security reasons including splitting up gangs, for a prisoner's safety and protection, moving prisoners to and from courts, attending medical appointments either in the community or at other correctional facilities better equipped to address certain health problems, moving individuals closer to their family in the community, to attend

rehabilitation programmes at other sites, as a punishment by moving a prisoner to a facility with a harsher regime or different security classification, and moving closer to a city or town in anticipation of release.

While the static prisoner population represents a cross-sectional count on a census date, the throughput (or churn) of prisoners is large and significant. In the USA for example, the cross-sectional population is around 2 million, but an estimated 12 million individuals cycle through USA correctional facilities each year. Globally, while around 11 million individually are incarcerated at any one time, an estimated 30 million people are released from prisons annually (Coyle et al. 2016). By further example in New South Wales (Australia's most populous state and the jurisdiction with the highest prisoner population), the static population is around 13,000 with an estimated 180,000 prisoner movements per year. This presents challenges to prison health services in ensuring continuity of treatment within the prison system.

In Australia, the number of adult former prisoners returning to prison within two years of their release increased from 33% in 2000 to 51% in 2015 (Corrective Services 2016). Over two-thirds of state prisoners in the USA were arrested at least once within 3 years of their release in 2005 (Durose et al. 2014), and in the United Kingdom (UK) the re-offending rate is approximately 46% after 1 year of release from prison (Ministry of Justice 2016). These figures indicate the challenges faced by departments of corrective services in preventing recidivism suggesting the need for innovation and multipronged approaches involving a range of government agencies, including the health, welfare, and justice sectors.

Indigenous incarceration

Beyond the socioeconomic characteristics that define prisoner populations as over-representing the poor and disadvantaged, closer examination reveals that key ethnic groups are also over-represented in custodial settings. In the USA in 2017, Blacks represented 12% of the general adult population but 33% of the sentenced prisoner population while Whites accounted for 64% of adults in the community but 30% of prisoners (Bronson and Carson 2019). Expressed as a rate, there were 1,549 Black prisoners for every 100,000 Black adults but for Whites the rate is 272 per 100,000. Similarly, in the UK, 27% of prisoners identified as coming from an ethnic minority,

compared with 13% in the general population (Ministry of Justice 2019). Black or British Black people comprised 13% of the UK prisoner population (Ministry of Justice 2019) but approximately 3% of the UK general population (UK Office for National Statistics 2018).

Indigenous or first nations people have also been identified as over-represented in the prison systems of Canada, New Zealand, and Australia. In Canada, 'In 2016/2017, Aboriginal adults accounted for 28% of admissions to provincial/territorial correctional services and 27% for federal correctional services, while representing 4.1% of the Canadian adult population' (Malakieh 2018). While forming 12.5% of the general population aged 15 and over in New Zealand, 50% of prisoners identify as Māori, with Māori women accounting for 60% of the female prisoner population (New Zealand Department of Corrections 2007). In Australia, the pattern is similar with Aboriginal people comprising around 2% of the general population aged 18 years and over but 28% of the adult prisoner population (Australian Bureau of Statistics 2018). The overall crude incarceration rate in Australia is 164 per 100,000 but for Aboriginal people the rate is 2,481 per 100,000 (Australian Bureau of Statistics 2018). The rate varies considerably across Australia from 672 per 100,000 in Tasmania to 4,257 per 100,000 in Western Australia. It is believed that Australia has the highest incarceration rate of Aboriginal people in the world.

Explanations for these high rates among Aboriginal people include offence criminalization with Aboriginal people more likely to be imprisoned for certain, often petty, crimes than a non-Aboriginal person, poverty, unemployment, inadequate legal representation, a lack of diversionary options for Aboriginal people, intergenerational offending, and colonization and dispossession. A comprehensive set of recommendations from a socio-legal perspective have been developed by the Australian Law Reform Commission covering Justice Reinvestment, bail, sentencing, use of community-based sentences, mandatory sentencing, prison programmes and parole, access to justice, fines and driver licences, alcohol, police response, and child protection and adult incarceration (Australian Law Reform Commission 2017). All have been identified as important in issues in the overincarceration problem.

Why is the prisoner population so unhealthy?

Wilkinson and Pickett have drawn attention to a consistent, positive association of national or regional levels of inequality with poor health and adverse social outcomes (such as homicide or rates of incarceration) (Wilkinson and Pickett 2009) citing evidence for the potency—for its negative impact on those experiencing social inequality—of 'social evaluative stress' related to individuals' perceptions of their relatively lower positions within the social hierarchy. Chronic stress can have direct effects, mediated by the autonomic nervous and endocrine systems as well as, most importantly, the brain directly, such that early experiences and exposures become biologically embedded with lifelong consequences with respect to health, behaviour and resilience to future life challenges (see next).

Another influential theory, that of 'fundamental causes' posits that higher socioeconomic status (SES) is associated with better health because '*SES embodies an array of resources, such as money, knowledge, prestige, power, and beneficial social connections that protect health no matter what mechanism are relevant at any given time*' (Phelan et al. 2010). Thus, relative to those of higher SES, individuals with *less* privilege are less likely to be informed about health practices, less likely to adopt them, and to have poorer access to health or preventive healthcare leading to later detection and treatment of illness when it occurs. The fundamental theory has some explanatory relevance for the disproportionate rates of incarceration—representing an unwanted outcome through contact with the criminal justice system—among those least privileged in the community.

Finally, it is apposite to reference one of the most influential theories of crime causation, which considers low self-control as a fundamental mechanism and for which there is very considerable empirical support (Hirschi and Gottfredson 1995). The theory claims to explain more than just why individuals break the law. Poor self-control is suggested as a key cause of '*practically everything that (…) provides short-term benefit at the risk of long-term cost*' (Hirschi 2004). For this reason, gambling, smoking, heavy drinking, or misusing prescription medication—examples of conduct conceivable as analogous to crime—are part of the behavioural package that the theory seeks to explain (Moffitt et al. 2011; Wills and Stoolmiller 2002; Hirtenlehner 2019). Consistent with this theory, *incarcerated* populations display high rates of substance abuse, illnesses related to risky sexual practices, and traumatic brain injury related to other risk taking or impulsive behaviour. These theories each contribute a perspective on why offenders,

and the incarcerated in particular, are some of the most socially disadvantaged and unhealthy individuals in society.

Enormous advances in our understanding of how early life experiences may influence later life health and behaviour have come from the study of epigenetics (Moore 2015). The extent to which specific genes are accessible to the machinery by which they are transcribed (leading to protein synthesis and the attendant biological effects) can be significantly influenced by specific events or exposures, especially in early life. For example, epigenetic changes adjacent to the gene encoding the glucocorticoid receptor (GR) implicated in the hypothalamic–pituitary–adrenocortical (HPA) axis which plays a major role in the stress response have been shown to occur in the context of adverse exposures, leading to reductions in cortisol which have been shown to be associated with impaired stress response and poor self-regulation (Conradt 2017). Highlighting how early exposures of relevance may occur, in one study, intimate partner violence during pregnancy was shown to be associated with changes in the expression of the GR in the children born to that pregnancy (Radtke et al. 2011).

The data deriving from these sophisticated genetic techniques, together with that from structural and functional brain imaging which have demonstrated important adverse biological correlates of low SES and of criminality, serve to emphasize the importance of optimizing early childhood development and reducing toxic adversity as likely to be the most effective long-term measures to reduce the individual and social harms associated with offending behaviour, often the product of individuals exposed to early life disadvantages (Hackman et al. 2010; Lange et al. 2010; Lawson et al. 2013).

A snapshot of mental and physical ill health in prisoners

Prisoner populations are characterized by very high rates of mental morbidity and illness with rates of major mental illness such as psychosis and depression substantially exceeding community rates (Butler et al. 2006; Torrey et al. 2010; Fazel and Seewald 2018). Increased rates are seen also for post-traumatic stress disorder (PTSD), personality disorder, ADHD and substance abuse (Fazel and Seewald 2018). The evidence for major depression and psychotic illnesses is the strongest (Fazel et al. 2016; Fazel and Seewald 2018). Despite the high level of need, these disorders are

frequently underdiagnosed and poorly treated (Wilper et al. 2009). The extent to which, at the prisoner population level, incarceration itself is a cause of mental ill health is not clear but in some individuals it has an adverse impact with the development of depression with or without suicidality (Baillargeon et al. 2009; Mears and Cochran 2012).

Mental illness has been shown to be a risk factor for suicide, premature mortality on release from prison (Kariminia et al. 2007b) and recidivism (Baillargeon et al. 2009; Fazel and Yu 2011). Issues that naturally arise in relation to this high burden of psychiatric morbidity include its optimal and efficient detection, appropriate management, and its impact on the incarceration experience and on release (Wilper et al. 2009). Upon entry into prison, immediate risks related to existing psychiatric conditions include suicidality, the health consequences of withdrawal from substances (e.g. alcohol, amphetamine, or opiates), or the sudden discontinuation of medication for which adverse consequences might ensue (e.g. anticonvulsants, antidepressants, antipsychotics). Approaches to screening for significant mental illness vary considerably both within and between countries, but as with other illnesses, incarceration presents an opportunity to identify and treat mental illnesses that may have been either unrecognized or untreated previously. The stigma and practical adverse consequences in terms of securing employment, relationship/family breakdown, or resumption of addictive behaviours can make the post-release period particularly challenging psychologically.

Of the psychiatric illnesses prevalent among the incarcerated, there is evidence to suggest that depression and psychosis are risk factors for offending and that adequate treatment may reduce offending risk (Fazel et al. 2014; Albalawi et al. 2019). Substance abuse is also associated with an increased offending risk. The situation with respect to acquired brain injury, in many instances due to trauma, is complex. A consistent finding throughout the international literature is of a very high prevalence of self-reported traumatic brain injury (TBI) within populations of offenders and prisoners, ranging up to 85% of male prisoners in some studies (Schofield et al. 2006, 2011; Durand et al. 2017). The frontal lobes of the brain, important for regulating behaviour, are known to be especially susceptible to damage by trauma, and therefore the possibility that TBI may be causally implicated in subsequent offending behaviour arises. Several retrospective cohort studies using record linkage have identified an increased risk of offending

following a TBI (Fazel et al. 2011; Schofield et al. 2015); however a study in which non-head trauma cases served as controls failed to find evidence of offending risk associated with TBI, although in that study only mild TBI was evaluated (Kennedy et al. 2017). Indeed, many of the TBIs reported by offenders may be mild, following which the likelihood of significant brain damage impacting brain function is low. In some unknown proportion of offenders who report a TBI which preceded their offending, the TBI may have been a consequence of an underlying predisposition towards impulsive or risky behaviour (Schofield et al. 2019).

With respect to physical health, prisoners have a substantial excess of morbidity relative to the general community. One study from the USA, in which adjustment was made for age, gender, race, education, employment, the USA as birthplace, marital status, and alcohol consumption found elevated rates of hypertension, asthma, arthritis, cervical cancer, hepatitis but no increased odds of diabetes, angina, myocardial infarction and a lower odds of obesity (Binswanger et al. 2009).

For some individuals, incarceration may be associated with health benefits, related to the provision of increased access to healthcare compared with the community, abstinence from illicit drugs, alcohol and tobacco, and reduced risk of accidents or violence. The benefits may differ according to racial/ethnic group and by country (Fazel and Benning 2006; Kariminia et al. 2007a; Spaulding et al. 2011). Thus, in the USA, compared to community controls, African American (but not white) men were reported to have substantially reduced mortality for the period of time in prison although the effect largely disappeared when adjustment was made for the compassionate release of prisoners with terminal illnesses (Spaulding et al. 2011). By contrast, in the UK and France, in prison mortality was reduced overall among prisoners, although deaths by pneumonia and other infective illnesses were higher (Clavel et al. 1987). More consistent are the findings of increased mortality in the months after release from prison, related to factors including suicide, overdose either deliberate or unintentional, violence or liver disease (Kariminia et al. 2007b; Spaulding et al. 2011).

Drug use in prisoner populations

Alcohol and other drug (AoD) use is a defining characteristic of prisoner populations (Butler et al. 2006; Fazel et al. 2006), and the association

between AoD use and offending behaviour, including violent offending, is well established. In New South Wales, Australia, 60% of prisoners surveyed as part of the highly regarded New South Wales' Inmate Health Survey report they were drunk, 'high', or 'stoned' at the time of committing the offence for which they were currently imprisoned, most commonly on amphetamine or methamphetamine (49.1%), alcohol (43.8%), and cannabis (27.1%) (Justice Health & Forensic Mental Health Network 2015). Methamphetamine use in the 12 months prior to incarceration has now surpassed cannabis use in Australia (Butler and Simpson 2017).

Approximately half of all prisoners report histories of injecting drug use (Australian Institute of Health and Welfare 2015; Justice Health & Forensic Mental Health Network 2015; Butler and Simpson 2017). While most injecting drug users (IDUs) suspend this behaviour during incarceration, approximately half of IDU continue to do so during incarceration (Fetherston et al. 2013). Only a small proportion of individuals report commencing injecting drug use during imprisonment (Butler et al. 2003). Drug use in prison is fairly common with over half of participants (56%) reporting that drugs were either 'quite easy' or 'very easy' to obtain in NSW prisons (Justice Health & Forensic Mental Health Network 2015). Injecting drug use is the overwhelming risk factor associated with hepatitis C transmission (Luciani et al. 2014). However, in the prison context other risk behaviours implicated in the transmission of infectious diseases include: amateur tattooing, sharing barbering equipment, and blood-to-blood contact during fights (Luciani et al. 2014).

Problematic alcohol use receives far less attention than illicit drugs in the prison context, possibly due to its legal status, yet two-thirds of those reporting alcohol use in the 12 months prior to incarceration (67%) had consumed hazardous or harmful quantities of alcohol (Justice Health & Forensic Mental Health Network 2015). Despite the widespread and problematic use of alcohol among those in prison when in the community, a lack of evidence-based prison-based alcohol treatment programmes is apparent in the peer reviewed literature (de Andrade et al. 2018; Doyle et al. 2019). This is important given the link between substance use and re-offending and suggests a need for more prison-based drug treatment options for this population (de Andrade et al. 2018).

Apropos the high rates of drug use in this population, histories of injecting drug use, prison drug use, and also injecting drugs in prison, it is notable that

few jurisdictions internationally have allowed the provision of clean injecting equipment despite its proven effectiveness in reducing the spread of bloodborne viruses in the community (Stöver and Hariga 2016). Reasons for this include arguments that it will effectively arm prisoners with a weapon to be used against prison personnel. However, to our knowledge only one case of a prison officer being stabbed with a needle has been reported in the literature. This occurred in Australia and tragically resulted in the officer dying from AIDS. This incident has harmed efforts to consider the provision of clean injecting equipment both in Australia and beyond. This day is still remembered by custodial officers in New South Wales over 20 years since the incident occurred. It is also argued that introducing needle exchanges in prison would be viewed as condoning drug use and this should be avoided. Interestingly, similar arguments are used by opponents of condom provision to prisoners, i.e. if filled with liquid and frozen they can be used as a cudgel, and symbolically they are seen as condoning and promoting sex between prisoners (even providing sexual predators with protection from sexually transmissible infections). However, evidence suggests that condom introduction in the Australian context did not increase sex between prisoners but, when sex did occur, promoted safe sex between prisoners (Butler et al. 2013).

Smoking

Tobacco smoking is recognized as one of the deadliest health risk behaviours, responsible for killing up to half of its users and more than seven million deaths each year (WHO n.d.). More than six million of these deaths are the result of direct tobacco use while around 890,000 are the result of non-smokers being exposed to second-hand smoke (WHO n.d.). Successful public health campaigns have contributed to reducing daily smoking rates in the general population in many developed countries, but these reductions are generally not reflected among prisoners. For example, smoking rates fell from 19% in the general Australian population in 2001 to 12% in 2016 (Australian Institute of Health and Welfare n.d.) However, among prisoners, there has been no corresponding downward trend and the rate of current tobacco smoking has remained stubbornly high at around 84%. This is due to high rates of tobacco use in key marginalized groups who are over-

represented in prisoner populations—IDUs, those with mental illness, and Aboriginal people (in the case of Australia) (Baker et al. 2006).

The main response to tobacco control in prisons has been to ban smoking altogether (Spaulding et al. 2018). Whether this is in response to concerns for smokers' health or a fear of litigation by non-smoker prisoners and staff is unclear. While banning smoking undeniably improves air quality (Proescholdbell et al. 2008) and reduces the health risk posed to non-smokers, this measure is not without consequences such as the creation of another black market borne out of tobacco scarcity and the problems associated with this, and smoking other noxious substances as a substitute (Puljevic et al. 2018). The issue of smoking bans can present a dilemma for public health officials who by nature are opposed to tobacco smoking yet believe that the human rights of one of the most vulnerable groups in society ought to be respected. The effectiveness of smoking bans as a cessation strategy is questionable with several studies indicating that many of those released to freedom go on to relapse. One USA study found that 63% of former prisoners relapsed on their first day of release, 82% after 1 week, 86% after 4 weeks, and 97% had resumed smoking at 6 months (Lincoln et al. 2009). These figures illustrate several important points for those working in public health in prison and the extreme challenge posed by this population. Firstly, the scale of the problem is staggering with few, if any, other population group in the community having such high levels of tobacco smoking. Secondly, the high rates suggest an absence of appropriately tailored and effective interventions specific to the needs of this population to achieve sustained abstinence. Thirdly, going from total bans that promote drastically reduced tobacco smoking during incarceration to almost all relapsing after fairly short periods in the community suggests existing interventions are ineffective for this population or don't reach them, or that continuity of care is lacking with respect to health promotion activities aimed at tobacco smoking in the transition from prison to the community.

Women prisoners

The number of women and girls in prison worldwide has been reported to have increased by 53% since 2000, against a global population growth (from 2000 to mid-2016) of 21% (ICPR 2017). Consistent with the situation in other countries, in Australia women represent a substantially smaller

proportion of the total prisoner population than men (9% vs. 91% in Australia), but there was an 85% increase in women prison numbers between 2008 and 2018 (*vs.* 55% in men) (Australian Bureau of Statistics 2018). The increase has been particularly marked among Aboriginal women with a 117% increase between 2008 and 2018 (versus a 74% increase in non-Aboriginal women in Australia). Many of these women have endured significant abuse and trauma and much of their substance use, mental health issues, and violence likely has its origins in past traumatic events (Wilson et al. 2017). Higher levels of psychiatric morbidity and substance use are found among women prisoners relative to men (Fazel et al. 2016). In Australia, 2 in 5 (38%) women coming into prison reported they had children in the community who were dependent on them for their basic needs (Australian Institute of Health and Welfare 2015). The rate of women incarcerated for violence offences is increasing (Australian Bureau of Statistics 2018). Recognizing that '*women prisoners are one of the vulnerable groups that have specific needs and requirements*', and that prison system worldwide have been developed around men, in 2010 the 'United Nations Rules for the Treatment of Women Prisoners and Non-custodial Measures for Women Offenders'—the Bangkok Rules (UNODC 2010)—were passed by the UN. This resolution encourages member states '*to adopt legislation to establish alternatives to imprisonment and to give priority to the financing of such systems, as well as to the development of the mechanisms needed for their implementation*'.

Family members

There is a growing literature attesting to the adverse health effects of incarceration upon family members. Significant increases in the frequency of adverse physical and mental health problems, as well as increased mortality, have been found in the children of prisoners (Wildeman 2014; Wildeman et al. 2014; Wildeman and Wang 2017) and rates of obesity and cardiovascular disease may be elevated in women who have a family member incarcerated, even after controlling for socioeconomic status (Roettger and Boardman 2012; Lee et al. 2014). Because African Americans in the USA, and those of indigenous background in the USA and other countries, are far more likely to be imprisoned, the burden of such health consequences is more likely to affect women and children of the same racial or ethnic background,

compounding pre-existing levels of disadvantage. The impact on children of having a parent incarcerated can be extensive, prolonged, and insidious, including developmental problems affecting learning behaviour, and psychological issues (Lee et al. 2013).

The ageing offender

In the USA, the incarceration epidemic has been associated with a substantial increase within the prison system of older inmates, and other countries are experiencing similar developments, albeit in lower absolute numbers. In comparison with the non-offender population, related at least in part to the greater burden of chronic disease, prisoners as a group display features of accelerated ageing due to poor health, mental illness and lifestyle factors, including drug and alcohol misuse, and other social determinants such as trauma, homelessness and victimization (Angus 2015). In the USA, from 1996 to 2008, the number of older adults in jail who were aged 55 or older increased 278% compared to a 53% increase in the overall jail population. Data from Canada indicate that 25% of the federal prison population are aged 50+, a 50% increase over the preceding decade (Darrell and Beck 1997; Beck and Berzofsky 2010; OCI 2019).

Based on figures from the Australian Bureau of Statistics, the number of prisoners aged 65 years and over has increased by 348% between 2001 and 2017 (Ginnivan et al. 2018). The increase was over 700% in the case of women prisoners in the same age group. The World Health Organization (WHO) has recommended restructuring prisons to accommodate the environmental needs of elderly prisoners (WHO 2014). Recommended provisions include grip rails, seats in showers, bunk bed ladders, ramps and widened doorways to accommodate wheelchairs, and better access to bathroom facilities. The WHO has also urged institutions to adopt appropriate screening for conditions associated with ageing, such as sensory impairment, mobility problems and arthritis (WHO 2014).

The serious nature of some older prisoner’s crimes (e.g. historic sexual offences or murders) often precludes their diversion into the community, release at the end of their allotted sentence, or compassionate release on diagnosis of a terminal illness. Those released at the conclusion of their prison sentence face difficulties finding accommodation in aged care

facilities who are often reluctant or unable to accommodate ex-offenders, placing this group at an increased risk of homelessness.

In a recent USA study using the Montreal Cognitive Assessment (MoCA) to evaluate cognition, 70% of 310 individuals who had been incarcerated for at least 48 hours aged 55 years and older met criteria for cognitive impairment (defined by a MoCA score of <24) (Ahalt et al. 2018). Those scoring within this range were more likely to be non-white, report fair or poor health and, on six month follow-up, were more likely to make multiple visits to the emergency room and to be rearrested.

Some of the most difficult and controversial issues with respect to the application of criminal justice system procedures, including the appropriateness of incarceration, arise in relation to violent or other serious illegal acts by individuals who have sustained brain damage or disease affecting parts of the brain that mediate or regulate behaviour. Such individuals include those suffering from frontotemporal dementia (FTD), a condition associated specifically with behavioural dysregulation and loss of insight (Liljegren et al. 2015). The growing recognition that cognitive impairment, like psychiatric problems, when present in individuals who come to the attention of the criminal justice system, may better be managed through the health system than the criminal justice system has, in some settings, been reflected in changes in legislation. For example, in New South Wales, Australia, the Mental Health Act was recently changed to allow for the possible diversion of individuals with cognitive impairment away from courts and legal sanction to healthcare settings for further assessment and management. The operational consequences of such legislation, however, include practical challenges with respect to how best to screen for cognitive impairment. One distinct challenge relates to the fact that impairments of self-control and insight of major relevance to offending behaviour (so-called executive impairment), may not be accompanied by any formal test abnormalities but are typically detected based upon history from an informant.

Research challenges in the prison setting

An extensive literature describes the myriad health problems facing prisoners, but a dearth of evidence exists regarding interventions for this population. Information regarding health status is important but it needs to

be complemented by culturally sensitive treatments and interventions that are customized to this population and the unique environment. A search of PubMed in June 2019 covering the period 1980 to 2019 using the terms ‘prisoners’ + ‘randomized trial’ returned only 34 results while ‘prisoners’ + ‘review’ or ‘meta-analysis’ yields 1,641 results. The small number of randomized trials in this setting reflects the poor evidence base in this area and also the considerable challenges of undertaking research in the prison environment.

Difficulties undertaking research in the prison setting and challenges associated with the population may explain this trend towards descriptive rather than applied research and the employment of study designs such as randomized trials. Challenges include a protracted ethics approval processes and a culture among ethics committees of being overly protective of this population who they characterize as inherently vulnerable and incapable of providing informed consent. This perception of prisoners has likely been formed on the basis of well-documented abuses by researchers of this population in the past (Hornblum 1997). However, these are unlikely to apply today but nonetheless are still projected onto this population by zealous ethics committee members. As Beecher (1966) comments ‘*the harms that are apt to result from social scientific research are rarely of the same magnitude as those produced by research in the medical sciences, which have placed the lives of unwitting research subjects at risk by infecting them with dysentery and malaria and using them as guinea pigs for LSD research*’ (Beecher 1966).

Another challenge is a tendency for the prison system not to like being examined by an external organization, including researchers. Researchers cannot simply design a study and approach participants via telephone or methodologies that rely on direct recruitment or mail-out surveys as might occur in the community. Any of these research approaches in the correctional setting requires a rigorous ethics approval process which can sometimes take several years to reach a resolution. Even with ethics approval, site-specific concerns from local prison management may prevent the research from progressing should it be deemed disruptive to the smooth running of the prison.

Issues raised by ethics committees include whether or not prisoners should be paid for participation in research, and what form of payment (Moreno 2001). Payment is often cited by ethics committees as an inducement despite

its widespread use in the community to compensate research participants. However, as Moreno states ‘none of the landmark scandals that have marred the history of medical ethics— the concentration camp experiments, the radiation experiments, the Jewish Chronic Disease Hospital case, the Willowbrook hepatitis studies, the Tuskegee syphilis study—did financial inducement figure’ (Moreno 2001).

‘Vulnerability’ of prisoner research participants is also a widely used concept to stymie research but is often poorly articulated by ethics committees in adopting a protectionist view of this group. Fitness to provide fully informed consent is often raised as a concern. However, it is our experience across numerous surveys that most prisoners can comprehend the research being undertaken, the reason it is being conducted, and hence make an informed decision whether or not they wish to participate in a study. Recent work has established that prisoners feel insulted at the thought that a third party can decide on their behalf that they are incapable of making such a decision.

Numerous physical barriers exist in the prison setting which present a challenge to the conduct of research including: locked doors, X-ray scanners, biometric scanners, sniffer dogs, and a reliance on custodial officers to locate research participants. However, once inside the prison walls, prisoners are mostly willing to participate in research often for altruistic reasons of genuinely wanting to improve the lot of prisoners and as a form of civic duty (Lafferty et al. 2016). Response rates in the order of 80% are not uncommon in the prison setting if participants are approached with respect and the research is explained properly. Contrary to popular perception, prisoners are reliable survey respondents. We validated self-reports of TBI in a sample of 200 men and found we could locate evidence of the reported head injury in over 80% of cases using hospital medical records (Schofield et al. 2011).

Many individuals in prison have never been involved in surveys or invited to express opinions and often relish the opportunity to participate; they often report involvement in research as a cathartic experience. For many, involvement in research is often the first time they have ever told anyone about their background, upbringing, etc. and feel valued at being consulted. It is important that treatment pathways are put in place by researchers as part of a duty of care to participants including counselling for those who may be distressed, follow-up testing for bloodborne viruses and sexually

transmissible diseases, access to hepatitis treatments, vaccinations, and other forms of treatment and care.

The process whereby ethics committees make research on prisoners more difficult has been referred to as ‘ethics creep’ and results in research being hampered or even abandoned (Gill 2019). As Haggerty (2004) states ‘The range of potential research related harms envisioned by REBs at times seems to be limited only by the imagination of different reviewers’ (Haggerty 2004).

While we recognize the important and necessary role that ethics committees play, a more realistic approach should be adopted to facilitate a better understanding of the population they serve, and greater recognition be given to the positive benefits that can accrue from research in this field.

Cost of incarceration

Taken in its broader context, the vast sums of money spent on incarceration indirectly impact on public health as the criminal justice system absorbs funding that could otherwise be spent on health and education (Settumba et al. 2018). Justice Reinvestment is one approach that has been touted as a solution to this conundrum and posits that funding used for incarceration and running the justice apparatus might be better spent on programmes and infrastructure that keep people out of prison (Willis and Kapira 2018). Despite the enthusiasm of Justice Reinvestment advocates, the evidence base is small which could explain the lack of uptake by jurisdictions internationally.

Synergizing health services between prisons and the community is important to ensure continuity of healthcare on reception into prison and on release from prison into the community. The provision of health services to prisoners is most effective when it meets the same standards as those of the broader community in terms of both the scope of services provided and the standard of healthcare. During discussions on the provision of health services to prisoners in the UK, it was commented by the Inspectorate of Prisons for England and Wales that:

> Prisoners were members of the wider community before their reception into prison and their health care was the responsibility of the NHS (National Health Service); the vast majority will return to that community and to the NHS on release. It is illogical that, during the time they are prisoners, their health care is provided through separate channels. (Inspectorate of Prisons for England and Wales 1996)

Public health opportunity

Contact with the justice system via the police, courts, and prisons provides an important point of contact to connect and engage with this underserved population to identify health and social problems and initiate treatments to improve health, social, and justice outcomes. It has been suggested that imprisonment, while unfortunate, represents an important but largely overlooked public health opportunity to improve the health of this group (Glaser and Greifinger 1993). Evidence for this comes from studies showing that some prisoners are more likely to use health services when in prison than in the community (Butler and Milner 2003; Indig et al. 2009). Australia is unique in terms of having a national prisoner health data collection that collects indicators on various aspects of prisoner health and health service use on entry to prison, during incarceration, and at the time of release (Australian Institute of Health and Welfare 2015). Indicators from this survey show that almost all those entering prison received a health assessment, one-fifth referral to mental health services, few reporting needing to see a health professional but did not; almost all could see a health professional in prison if they wanted to; and eight out of ten prisoners reported their health either improved or stayed the same while in prison (Australian Institute of Health and Welfare 2015). See Table 10.10.1.

Table 10.10.1 Prison health indicators, Australia

Health indicator	Proportion
Proportion of prison dischargees who received a health assessment upon entry to prison.	92%
Proportion of prison entrants who, at reception, were referred to mental health services for observation and further assessment.	22%
Proportion of prison entrants who, in the last 12 months, consulted a medical professional for their own health within the community.	66%
Proportion of prison entrants who were in prison in the last 12 months and consulted a medical professional for their own health in prison.	57%
Proportion of prison entrants who, in the last 12 months, needed to consult with a health professional in the community but did not.	34%
Proportion of prison entrants who were in prison in the last 12 months and needed to consult with a health professional while in prison but did not.	15%
Proportion of prison dischargees who were tested for bloodborne viruses or a sexually transmissible infection.	69%
Proportion of prison dischargees who could easily see a medical professional (GP or nurse) in prison if they had a health problem.	91%
Proportion of prison dischargees who rated the healthcare they received in the prison clinic as excellent.	27%
Proportion of prison dischargees reporting that their physical health improved or stayed the same in prisons.	83%

Source: data from The health of Australia's prisoners. AIHW Catalogue no PHE 123. Canberra: Australian Institute of Health and Welfare. 2015.

Using the incarceration opportunity to tackle a major public health problem is currently being trialled in a world-first study in New South Wales (Australia) as part of efforts to eliminate hepatitis C (HCV). The Surveillance and Treatment of Prisoners with Hepatitis C (STOP-C) project aims to halt the transmission of HCV in prisons and hence the community when otherwise infected individuals are released from prison back to the community (Hajarizadeh et al. 2017). As the global burden of HCV infection continues to rise with an estimated 71 million people affected worldwide interventions that target hard-to-access populations are needed (The Polaris Observatory HCV Collaborators 2017). People who inject drugs represent

the core of the HCV epidemic in developed countries, accounting for most new (80%) and existing (60%) cases (Shepard et al. 2005). HCV is particularly prevalent among prisoners with around half reporting a history of injecting drug use and one-fifth exposed to hepatitis C (Sarah et al. 2013; Butler and Simpson 2017). Significant interest exists in whether treating those in prisons with HCV can impact on the wider HCV epidemic in Australia. The treatment-as-prevention approach adopted by the STOP-C trial comes from the HIV field where it has been shown that using antiretroviral therapy can prevent HIV transmission (Cohen et al. 2011). STOP-C leverages off the new treatments for HCV which have 90–95% cure rates, few side effects, and taken for only 8–12 weeks. If successful, the possibility exists to scale-up HCV treatment thus representing one of the first attempts to use the incarceration opportunity to impact on a major public health problem affecting the wider community. If successful, this approach could be adopted for other public health problems prevalent in this population.

International covenants and protections for prisoners' health

Numerous statements and covenants exist that *theoretically* protect the health of prisoners and those held in detention; many of these have an HIV prevention focus. The most famous is the United Nations' Standard Minimum Rules for the Treatment of Prisoners which was first adopted in 1955 and revised following consultation as the Nelson Mandela Rules in 2015 (United Nations 1955, 2015). While not legally binding, the Mandela Rules provide broad guidelines for best practice for citizens held in prisons and other forms of custody covering physical and mental health, disability, confidentiality of healthcare, accommodation, clothing, exercise, nutrition, etc. (United Nations 2015). Although not enforceable, they do provide principles against which countries can be judged in terms of their treatment of prisoners. Similarly, the Optional Protocol to the Convention against Torture (OPCAT) provides a framework under the United Nations' human rights system for oversight of places of detention and closed environments (United Nations 2002). It requires signatories to establish an independent national preventive mechanism to conduct inspections of all places of detention. As of June 2018, 88 countries have ratified the protocol (Wikipedia n.d.). However, three of the world's largest incarcerators—the

USA, Russia, and China—who together account for over 4 million prisoners—have not ratified the protocol. Brazil, the world's third largest incarcerator *has* signed the protocol.

Recent statements by the cities of San Francisco and London calling for public health approaches to be adopted to tackle incarceration (San Francisco) and violence (London) give rise to the view that traditional approaches to crime prevention have not been effective and new thinking is required to address these problems. In specifying how this approach might operate to fight violence, the public health approach is spelled out as one which:

takes a population approach, not one which just focuses on high risk individuals;
is preventive: by tackling 'upstream' risk factors, it aims to lessen 'downstream' consequences;
takes a system wide multiagency approach including involving business and volunteers;
takes brave decisions that require a long-term commitment;
recognizes the complexity of the issue, and seeks to build an evidence base that reflects that (Catch 22 2018).

While this [preventive] thinking is the bread and butter of public health, it represents an important conceptual shift in thinking in the fight against crime which has traditionally been downstream focused and tackled by justice agencies once contact with criminal justice system has occurred.

The public health approach to violence outlined here is important in highlighting the role of upstream factors in crime prevention which includes genetic and biological drivers of offending behaviour which to date have received little attention given the focus on sociological approaches. Traditional approaches to offending have often relied on sociological theories. Indeed, as Raine (2013) states in 'The Anatomy of Violence', '*The dominant model for understanding criminal behaviour has been, for most of the twentieth century, one built on social and sociological models. ….. sole reliance on these social perspectives is fundamentally flawed*' (Raine 2013).

The father of criminology, Cesare Lombroso, sought to identify criminal characteristics in physiognomy (Lombroso 2006). His work has long since been discredited, but advances in techniques such as brain imaging and genetics have resulted in a resurgence of interest into the biological

influences on offending. The emergent discipline of ‘biosocial criminology’ seeks to explain crime and antisocial behaviour by exploring both biological and environmental factors and their interactions (Beaver 2009).

Clearly, ‘brave decisions’ need to be made in addressing crime. Politicians frequently adhere to a perceived punitiveness on behalf of the community adopting a ‘tough on crime’ stance, driven by a populist press motivated by delivering sensationalist headlines to their readership. However, recent deliberative work that takes a more considered approach than that offered by snap opinion polls challenges this view of the community as punitive and driven by a desire for revenge and retribution. This important work, incorporating Citizens Juries and deliberative roundtables, suggests that the public, when provided with relevant information regarding the costs of incarceration and its limited deterrent and therapeutic qualities, favours the development and implementation of effective interventions to treat this population (Simpson et al. 2014). Public health can contribute to this area by promoting the use of rigorous research methodologies such as randomized control trials, cohort studies, case-control studies, and systematic reviews in the criminal justice area.

Multi-agency approaches will be required to tackle crime and hence reduce the numbers held in prison. This approach will necessarily have to harmonize biological and social factors for optimal success. To this end, the approaches of these two cities (San Francisco and London) is both refreshing and forward thinking.

Future challenges

Consistent with our intention of preserving relevance for an international readership, we recognize that prioritization, in terms of advancing health outcomes of individuals who come in contact with the criminal justice system will differ by country, or region. Thus, a major priority within the USA concerns the imperative to halt and reverse the epidemic of incarceration, particularly among Blacks (Drucker 2011). In relatively less developed regions, such as Latin America, improvement in the physical conditions of prisons and the better recognition and treatment of psychiatric conditions may be paramount. Setting region-specific considerations aside, we consider that the following challenges are likely to have broad relevance regardless of setting: creating and operating diversionary programmes,

reducing the impact of incarceration on families, dealing with cognitive impairment and other ageing-related illnesses, optimizing transition to the community and health maintenance on release from prison and, of crucial importance, more rigorous research as mentioned in the previous section (Adams 2018).

Prisons represent individual, societal, and institutional failures at multiple levels. They are costly and have limited efficacy in terms of their rehabilitative properties. To this end, they represent one of society's greatest challenges and one requiring a concerted effort by agencies spanning health, welfare, justice, law, and education. The prisoner population represents one of the most unhealthy in the community across all health metrics and one that presents numerous challenges, including the attitude of the community towards a group that has transgressed the rules of society. What has not occurred to date is the 'ownership' of this population by mainstream public health agencies, although recent moves have sought to remedy this, calling on public health professionals to apply the arsenal of methodologies and evidenced-based approaches to tackle offending behaviour. This will require a 'genes to gaol' approach that will not come quickly and will require effort, funding, innovation, and dedication.

REFERENCES

Adams, B.L. (2018). Paternal incarceration and the family: fifteen years in review. *Sociology Compass*, 12(3), e12567.

Ahalt, C., et al. (2018). Cognition and incarceration: cognitive impairment and its associated outcomes in older adults in jail. *Journal of the American Geriatric So*ciety, 66(11), 2065–71.

Albalawi, O., et al. (2019). Court diversion for those with psychosis and its impact on re-offending rates: results from a longitudinal data-linkage study. *BJPsych Open*, 5(1), e9.

Andolsek, K.M. (1997). Meningococcal transmission from jails to the community. *American Family Physician*, **55**, 290.

Angus, C. (2015). *Older Prisoners: Trends and Challenges*. [Online] Available at: https://www.parliament.nsw.gov.au/researchpapers/Documents/older-prisoners-trends-and-challenges/Older%20prisoners%20-%20trends%20and%20challenges.pdf

Australian Bureau of Statistics (2018). 4517.0— Prisoners in Australia, 2018. Canberra. [Online] Available at: https://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/4517.02018?OpenDocument

Australian Institute of Health and Welfare (2015).The health of Australia's prisoners. *AIHW Catalogue no PHE 123*. Canberra: Australian Institute of Health and Welfare.

Australian Institute of Health and Welfare (n.d.). *Alcohol, Tobacco & Other Drugs in Australia*. Available at: https://www.aihw.gov.au/reports/alcohol/alcohol-tobacco-other-drugs-australia/contents/interactive-data/tobacco

Australian Law Reform Commission (2017). *Pathways to Justice—Inquiry into the Incarceration Rate of Aboriginal and Torres Strait Islander Peoples, Final Report No 133*. [Online] Available at: https://www.alrc.gov.au/wp-content/uploads/2019/08/final_report_133_amended1.pdf

Baillargeon, J., Binswanger, I.A., Penn, J.V., Williams, B.A., and Murray, O.J. (2009). Psychiatric disorders and repeat incarcerations: the revolving prison door. *American Journal of Psychiatry*, 166(1), 103–9.

Baker, A., et al. (2006). Where there's smoke, there's fire: high prevalence of smoking among some sub-populations. *Drug & Alcohol Review*, **25**, 85–96.

Beaver, K. (2009). *Biosocial Criminology: A Primer*. Ken Hunt Publishing Company.

Beck, A. and Berzofsky, M. (2010). *Sexual Victimization in Prison and Jails Reported by Inmates, 2008–2009. NCJ 231169*. Washington DC: Department of Justice. Office of Justice Programs, Bureau of Justice Statistics.

Beecher H. (1966). Ethics and Clinical Research. *New England Journal of Medicine*, **274**, 1354–60.

Binswanger, I.A., Krueger, P.M., and Steiner, J.F. (2009). Prevalence of chronic medical conditions among jail and prison inmates in the USA compared with the general population. *Journal of Epidemiology & Community Health*, 63(11), 912–19.

Bronson, J. and Carson, A. (2019). *Prisoners in 2017. U.S. Department of Justice. Office of Justice Programs*, NCJ 252156. [Online] Available at: https://www.bjs.gov/content/pub/pdf/p17.pdf

Butler, T., et al. (2006). Mental disorders in Australian prisoners: a comparison with a community sample. *Australian & New Zealand Journal of Psychiatry*, 40(3), 272–6.

Butler, T., Levy, M., Dolan, K., and Kaldor, J. (2003). Drug use and its correlates in an Australian prisoner population. *Addiction Research & Theory*, 11(2), 89–101.

Butler, T. and Milner, L. (2003). *The 2001 Inmate Health Survey*. Sydney: NSW Corrections Health Service.

Butler, T., Richters, J., Yap, L., and Donovan, B. (2013). Condoms for prisoners: no evidence that they increase sex in prison, but they increase safe sex. *Sexually Transmitted Infections*, 89(5), 377–9.

Butler, T. and Simpson, M. (2017). *National Prison Entrants' Blood-borne Virus Survey Report 2004, 2007, 2010, 2013, and 2016.* Sydney: Kirby Institute. [Online] Available at: https://kirby.unsw.edu.au/report/national-prison-entrants-bloodborne-virus-and-risk-behaviour-survey-report-2004-2007-2010

Catch 22 (2018). *What does a 'Public Health' Approach to Violence Really Mean?* [Online] Available at: https://www.catch-22.org.uk/news/public-health-approach-to-violence/

Clavel, F., Benhamou, S., and Flamant, R. (1987). Decreased mortality among male prisoners. *The Lancet*, 2(8566), 1012–14.

Cohen, M.S., et al. (2011). Prevention of HIV-1 infection with early antiretroviral therapy. *New England Journal of Medicine*, 365(6), 493–505.

Conradt, E. (2017). Using principles of behavioral epigenetics to advance research on early-life stress. *Child Development Perspectives*, 11(2), 107–12.

Corrective Services (2016). *Report on Government Services. Productivity Commission. Canberra*. Available at: https://www.pc.gov.au/research/ongoing/report-on-government-services/2016/justice/corrective-services

Coyle, A., Fair, H., Jacobson, J., and Walmsley, R. (2016). *Imprisonment Worldwide. The Current Situation and an Alternative Future*. Bristol: Policy Press.

Darrell, G. and Beck, A. (1997). *Prison and Jail Inmates at Midyear* 1996. *NCJ 162843*. Washington, DC: U.S. Department of Justice Programs. Bureau of Justice Statistics.

de Andrade, D., Ritchie, J., Rowlands, M., Mann, E., and Hides, L. (2018). Substance use and recidivism outcomes for prison-based drug and alcohol interventions. *Epidemiologic Reviews*, 40(1), 121–33.

Dolan, K., et al. (2016). Global burden of HIV, viral hepatitis, and tuberculosis in prisoners and detainees. *The Lancet*, 388(10049), 1089–102.

Doyle, M.F., Shakeshaft A, Guthrie J, Snijder M, and Butler T. (2019). A systematic review of evaluations of prison-based alcohol and other drug use behavioural treatment for men. *Australian & New Zealand Journal of Public Health*, 43(2), 120–30.

Drucker, E. (2011). *A Plague of Prisons: The Epidemiology of Mass Incarceration* in America. New York: The New Press.

Durand, E., et al. (2017). History of traumatic brain injury in prison populations: a systematic review. *Annals of Physical and Rehabilitation Medicine*, 60(2), 95–101.

Durose, M., Cooper, A., and Snyder, H. (2014). Recidivism of prisoners released in 30 states in 2005: patterns from 2005 to 2010 – update. Bureau of Justice Statistics. NCJ 244205. [Online] Available at: https://www.bjs.gov/index.cfm?ty=pbdetail&iid=4986#

Farmer, P.E., et al. (1999). Recrudescent tuberculosis in the Russian Federation. In: *The Global Impact of Drug Resistant Tuberculosis*. Boston MA: Harvard Medical School/Open Society Institute.

Fazel, S. and Baillargeon, J. (2011). The health of prisoners. *The Lancet*, 377(9769), 956–65.

Fazel, S., Bains, P., and Doll, H. (2006). Substance abuse and dependence in prisoners: a systematic review. *Addiction*, 101(2), 181–91.

Fazel, S. and Benning, R. (2006). Natural deaths in male prisoners: a 20-year mortality study. *European Journal of Public Health*, 16(4), 441–4.

Fazel, S., Hayes, A.J., Bartellas, K., Clerici, M., and Trestman, R. (2016). Mental health of prisoners: prevalence, adverse outcomes, and interventions. *The Lancet Psychiatry*, 3(9), 871–81.

Fazel, S., Lichtenstein, P., Grann, M., and Långström, N. (2011). Risk of violent crime in individuals with epilepsy and traumatic brain injury: a 35-year Swedish population study. *PLoS Medicine*, 8(12), e1001150.

Fazel, S. and Seewald, K. (2018). Severe mental illness in 33 588 prisoners worldwide: systematic review and meta-regression analysis. *British Journal of Psychiatry*, 200(5), 364–73.

Fazel, S. and Yu, R. (2011). Psychotic disorders and repeat offending: systematic review and meta-analysis. *Schizophrenia Bulletin*, 37(4), 800–10.

Fazel, S., Zetterqvist, J., Larsson, H., Långström, N., and Lichtenstein, P. (2014). Antipsychotics, mood stabilisers, and risk of violent crime. *The Lancet*, 384(9949), 1206–14.

Fetherston, J., Carruthers, S., Butler, T., Wilson, D., and Sindicich, N. (2013). Rates of injection in prison in a sample of Australian-injecting drug users. *Journal of Substance Use*, 18(1), 65–73.

Gentleman, A. (2000). Russia to release 120,000 from packed prisons. *The Guardian*. Date: 27 May. [Online] Available at: https://www.theguardian.com/world/2000/may/27/russia.ameliagentleman

Gill, J. (2019). Prison research project is throttled by red tape. Mental-health services study abandoned after a 14-month bureaucratic nightmare. *The Times Higher Education*. [Online] Available at: https://www.timeshighereducation.com/news/prison-research-project-is-throttled-by-red-tape/405434.article

Ginnivan, N.A., Butler, T.G., and Withall, A.N. (2018). The rising health, social and economic costs of Australia's ageing prisoner population. *Medical Journal of Australia*, 209(10), 422–4.e421.

Glaser, J.B. and Greifinger, R.B. (1993). Correctional health care: a public health opportunity. *Annals of Internal Medicine*, 118(2), 139–45.

Hackman, D.A., Farah, M.J., and Meaney, M.J. (2010). Socioeconomic status and the brain: mechanistic insights from human and animal research. *Nature Reviews & Neuroscience*, 11(9), 651–9.

Haggerty, K.D. (2004). Ethics creep: governing social science research in the name of ethics. *Qualitative Sociology*, 27(4), 391–414.

Hajarizadeh, B., et al. (2017). Incidence of hepatitis C virus infection in two maximum-security prisons in New South Wales, Australia: the SToP-C study. *Journal of Hepatology*, 66(1), S274.
Hirschi, T. and Gottfredson, M.R. (1995). Control theory and the life-course perspective. *Studies on Crime & Crime Prevention*, 4(2), 131–42.
Hirschi, T. (2004). Self-control and crime. In: Baumeister, R., and Vohs, K. (eds.). *Handbook of Self-Regulation: Research, Theory, and Applications*. New York: Guilford Press, pp. 537–52.
Hirtenlehner, H. (2019). Self-control and offending in late adulthood. Investigating self-control's interaction with opportunities and criminal associations in advanced age. *Journal of Criminal Justice*, **60**, 117–29.
Hornblum, A.M. (1997). They were cheap and available: prisoners as research subjects in twentieth century America. *British Medical Journal*, **315**, 1437–41.
Indig, D., et al. (2009). *2009 NSW Inmate Health Survey*. Justice Health NSW. [Online] Available at: https://www.justicehealth.nsw.gov.au/about-us/publications/inmate-health-survey-aboriginal-health-report.pdf
Inspectorate of Prisons for England and Wales (1996). *Patients or Prisoners? A New Strategy for Health Care in Prisons*. London: Home Office.
Institute for Criminal Policy Research (ICPR) (2017). *World Female Imprisonment List* (4th ed.). [Online] Available at: https://www.prisonstudies.org/sites/default/files/resources/downloads/world_female_prison_4th_edn_v4_web.pdf
John, H. (1973). *Prisons and Lazarettos: The State of the Prisons in England and Wales*. First published 1777. Glen Ridge, NJ: Patterson Smith.
Kariminia, A., et al. (2007a). Extreme cause-specific mortality in a cohort of adult prisoners-1988 to 2002: a data-linkage study. *International Journal of Epidemiology*, **36**, 310–16.
Kariminia, A., et al. (2007b). Suicide risk among recently released prisoners in New South Wales, Australia. *Medical Journal of Australia*, **187**, 387–90.
Kennedy, E., Heron, J., and Munafo, M. (2017). Substance use, criminal behaviour and psychiatric symptoms following childhood traumatic brain injury: findings from the ALSPAC cohort. *European Child & Adolescent Psychiatry*, 26(10), 1197–206.
Kenyon, G. (2009). Russia's prisons fuel drug-resistant tuberculosis. *The Lancet Infectious Diseases*, 9(10), 594.
Lafferty, L., Treloar, C., Butler, T., Guthrie, J., and Chambers, G.M. (2016). Unlocking dimensions of social capital in the prison setting. *Health & Justice*, **4**, 9.
Lange, N., Froimowitz, M.P., Bigler, E.D., and Lainhart, J.E. (2010). Associations between IQ, total and regional brain volumes, and demography in a large normative sample of healthy children and adolescents. *Developmental Neuropsychology*, 35(3), 296–317.
Lawson, G.M., Duda, J.T., Avants, B.B., Wu, J., and Farah, M.J. (2013). Associations between children's socioeconomic status and prefrontal cortical thickness. *Developmental Science*, 16(5), 641–52.
Lee, H.,. Porter, L.C., and Comfort, M. (2014). Consequences of family member incarceration: impacts on civic participation and perceptions of the legitimacy and fairness of government. *The Annals of the American Academy of Political and Social Science*, 651(1), 44–73.
Lee, R.D., Fang, X., and Luo, F. (2013). The impact of parental incarceration on the physical and mental health of young adults. *Pediatrics*, 131(4), e1188–95.
Liljegren, M., et al. (2015). Criminal behavior in frontotemporal dementia and Alzheimer disease. *JAMA Neurology*, 72(3), 295–300.
Lincoln, T., et al. (2009). Resumption of smoking after release from a tobacco-free correctional facility. *Journal of Correctional Health Care*, 15(3), 190–6.
Lombroso, C. (2006). *Criminal Man*. Durham, NC: Duke University Press.
Luciani, F, et al. (2014). A prospective study of hepatitis C incidence in Australian prisoners. *Addiction*, 109(10), 1695–706.

Malakieh, J. (2018). *Adult and Youth Correctional Statistics in Canada, 2016/2017*. Statistics Canada. [Online] Available at: https://www150.statcan.gc.ca/n1/pub/85-002-x/2018001/article/54972-eng.htm

Mears, D.P. and Cochran, J.C. (2012). U.S. prisoner reentry health care policy in international perspective: service gaps and the moral and public health implications. *The Prison Journal*, 92(2), 175–202.

Ministry of Justice (2016). *Proven Reoffending Statistics Quarterly Bulletin July 2013 to June 2014, England and Wales. Ministry of Justice Statistics Bulletin. 28 April 2016*. [Online] Available at: https://www.gov.uk/government/statistics/proven-reoffending-statistics-quarterly-july-2013-to-june-2014

Ministry of Justice (2019). *Offender Management Statistics Quarterly: October to December 2018. Prison Population: 31 March 2019. UK Ministry of Justice.* [Online] Available at: https://www.gov.uk/government/statistics/offender-management-statistics-quarterly-october-to-december-2018

Moffitt, T.E., et al. (2011). A gradient of childhood self-control predicts health, wealth, and public safety. *Proceedings of the National Academies of Science U S A*, 108(7), 2693–8.

Moore, D. (2015). *The Developing Genome: An Introduction to Behavioral Epigenetics*. Oxford: Oxford University Press.

Moreno, J.D. (2001). It's not about the money. *The American Journal of Bioethics*, 1(2), 46–7.

Justice Health & Forensic Mental Health Network (2015). Network Patient Health Survey Report [Online] Available at: https://www.justicehealth.nsw.gov.au/publications/2015_NHPS_FINALREPORT.pdf

New Zealand Department of Corrections (2007). *Over-representation of Māori in the Criminal Justice System: An Exploratory Report. NZ Department of Corrections*. [Online] Available at: https://www.corrections.govt.nz/resources/research_and_statistics/over-representation-of-maori-in-the-criminal-justice-system.html

Office of the Correctional Investigator (OCI) (2019). *Aging and Dying in Prison: An Investigation into the Experiences of Older Individuals in Federal Custody.* Government of Canada. [Online] Available at: https://www.oci-bec.gc.ca/cnt/rpt/oth-aut/oth-aut20190228-eng.aspx

Phelan, J.C., Link, B.G., and Tehranifar, P. (2010). Social conditions as fundamental causes of health inequalities: theory, evidence, and policy implications. *Journal of Health & Social Behavior*, 51 Suppl, S28–40.

Plamondon, K. and Pemberton, J. (2019). Blending integrated knowledge translation with global health governance: an approach for advancing action on a wicked problem. *Health Research Policy and Systems*, 17(1), 24.

Proescholdbell, S.K., Foley, K.L., Johnson, J., and Malek, S.H. (2008). Indoor air quality in prisons before and after implementation of a smoking ban law. *Tobacco Control*, 17(2), 123–7.

Puljevic C, et al. (2018). 'Teabacco': smoking of nicotine-infused tea as an unintended consequence of prison smoking bans. *Drug and Alcohol Review*, 37(7), 912–21.

Radtke, K.M., et al. (2011). Transgenerational impact of intimate partner violence on methylation in the promoter of the glucocorticoid receptor. *Translational Psychiatry*, 1(7), e21.

Raine, A. (2013). *The Anatomy of Violence: The Biological Roots of Crime*. New York: Pantheon/Random House.

Roettger, M.E. and Boardman, J.D. (2012). Parental incarceration and gender-based risks for increased body mass index: evidence from the National Longitudinal Study of Adolescent Health in the United States. *American Journal of Epidemiology*, 175(7), 636–44.

Sarah, L., et al. (2013). Incidence and prevalence of hepatitis C in prisons and other closed settings: results of a systematic review and meta-analysis. *Hepatology*, 58(4), 1215–24.

Schofield, P., Butler, T., Hollis, S., and D'Este, C. (2011). Are prisoners reliable survey respondents? A validation of self-reported traumatic brain injury (TBI) against hospital medical records. *Brain Injury*, 25(1), 74–82.

Schofield, P., et al. (2006). Neuropsychiatric correlates of traumatic brain injury (TBI) among Australian prison entrants. *Brain Injury*, **20**, 1409–18.
Schofield, P., Butler, T., Perkes, I., and Hollis, S. (2011). Traumatic brain injury rates and sequelae: a comparison of prisoners with a matched community sample in Australia. *Brain Injury*, **25**, 131–41.
Schofield, P.W., et al. (2015). Does traumatic brain injury lead to criminality? A whole-population retrospective cohort study using linked data. *PLoS One*, 10(7), e0132558.
Schofield, P.W., Mason, R., Nelson, P.K., Kenny, D., and Butler, T. (2019). Traumatic brain injury is highly associated with self-reported childhood trauma within a juvenile offender cohort. *Brain Injury*, 33(4), 412–18.
Settumba, S., Chambers, G., Shanahan, M., Schofield, P., and Butler, T. (2018). Are we getting value for money from behavioral interventions for offenders? A research note reviewing the economic evaluation literature. *American Journal of Criminal Justice*, **43**, 411–31.
Shepard, C.W., Finelli, L., and Alter, M.J. (2005). Global epidemiology of hepatitis C virus infection. *Lancet Infect Diseases*, 5(9), 558–67.
Simpson, P., Guthrie, J., Lovell, M., Walsh, C., and Butler, T. (2014). *Views on Alternatives to Imprisonment: A Citizens Jury Approach*. Melbourne: The Lowitja Institute.
Spaulding, A.C., et al. (2018). Smoking in correctional settings worldwide: prevalence, bans, and interventions. *Epidemiological Reviews*, 40(1), 82–95.
Spaulding, A.C., et al. (2011). Prisoner survival inside and outside of the institution: implications for health-care planning. *American Journal of Epidemiology*, 173(5), 479–87.
Stern, V. (2000). Prison reform and public health: the case of tuberculosis in the former Soviet Union. *European Journal of Public Health*, **10**, 4.
Stern, V. (1999). *Sentenced to Die? The Problem of TB in Prisons in Eastern Europe and Central Asia*. London: International Centre for Prison Studies.
Stöver, H. and Hariga, F. (2016). Prison-based needle and syringe programmes (PNSP)—still highly controversial after all these years. *Drugs: Education, Prevention and Policy*, **23**, 103–12.
The Polaris Observatory HCV Collaborators (2017). Global prevalence and genotype distribution of hepatitis C virus infection in 2015: a modelling study. *Lancet Gastroenterol Hepatol*, 2(3), 161–76.
Torrey, E., Kennard, A., Eslinger, D., Lamb, R., and Pavle, J. (2010). More mentally ill persons are in jails and prisons than hospitals: a survey of the states treatment advocacy. Treatment Advocacy Center. [Online] Available at: http://www.treatmentadvocacycenter.org/storage/documents/final_jails_v_hospitals_study.pdf
UK Office for National Statistics (2018). Population of England and Wales. [Online] Available at: https://www.ethnicity-facts-figures.service.gov.uk/uk-population-by-ethnicity/national-and-regional-populations/population-of-england-and-wales/latest
United Nations (1955). *Standard Minimum Rules for the Treatment of Prisoners. United Nations Congress on the Prevention of Crime and the Treatment of Offenders*. Geneva: United Nations.
United Nations (2002). *Optional Protocol to the Convention Against Torture and Other Cruel, Inhuman or Degrading Treatment or Punishment*. [Online] Available at: https://treaties.un.org/doc/source/RecentTexts/iv_9bE.pdf
United Nations (2015). *The United Nations Standard Minimum Rules for the Treatment of Prisoners (The Nelson Mandela Rules)*. Geneva: United Nations Office on Drugs and Crime.
United Nations Office on Drugs and Crime (UNODC) (2010). *United Nations Rules for the Treatment of Women Prisoners and Non-custodial Measures for Women Offenders with their Commentary (Bangkok Rules)*. [Online] Available at: https://www.unodc.org/documents/justice-and-prison-reform/Bangkok_Rules_ENG_22032015.pdf
Wikipedia (n.d.). *Optional Protocol to the Convention against Torture*. [Online] Available at: https://en.wikipedia.org/wiki/Optional_Protocol_to_the_Convention_against_Torture

Wildeman, C., Andersen, S.H., Lee, H., and Karlson, K.B. (2014). Parental incarceration and child mortality in Denmark. *American Journal of Public Health*, 104(3), 428–33.
Wildeman, C. and Wang, E.A. (2017). Mass incarceration, public health, and widening inequality in the USA. *The Lancet*, 389(10077), 1464–74.
Wildeman, C. (2014). Imprisonment and Infant Mortality. *Social Problems*, 59(2), 228–57.
Wilkinson, R. and Pickett, K. (2009). *The Spirit Level: Why More Equal Societies Almost Always Do Better*. London: Bloomsbury Publishing.
Willis, M. and Kapira, M. (2018). *Justice Reinvestment in Australia: A Review of the Literature*. Canberra: Australian Institute of Criminology.
Wills, T.A. and Stoolmiller, M. (2002). The role of self-control in early escalation of substance use: a time-varying analysis. *Journal of Consulting & Clinical Psychology*, 70(4), 986–97.
Wilper, A.P., et al. (2009). The health and health care of US prisoners: results of a nationwide survey. *American Journal of Public Health*, 99(4), 666–72.
Wilson, M., et al. (2017). Violence in the lives of incarcerated aboriginal mothers in Western Australia. *Sage Open*, 7(1), 2158244016686814.
World Health Organization (WHO) (2014). *Prisons and Health.* [Online] Available at: http://www.euro.who.int/__data/assets/pdf_file/0005/249188/Prisons-and-Health.pdf
World Health Organization (WHO) (n.d.). *Tobacco: Key Facts.* Available at: https://www.who.int/news-room/fact-sheets/detail/tobacco

SECTION 11

Public health functions

11.1 Health needs assessment

11.2 *Michael P. Kelly, Jane E. Powell, and Natalie Bartle* *The political economy of non-communicable diseases: lessons for prevention*

11.3 *Anne Marie Thow, Raphael Lencucha, and K. Srinath Reddy* *Immunization and vaccination*
Eleonora A.M.L. Mutsaerts and Shabir A. Madhi

11.4 Principles of infectious disease control
Robert J. Kim-Farley

11.5 Medical screening: theories, methods, and effectiveness
Tang Jin-ling and Li Li-ming

11.6 The practice of environmental health in an era of sustainable development
Yasmin E.R. von Schirnding and Lynn R. Goldman

11.7 Strategies and structures for public health interventions
Sian Griffiths and Kevin A. Fenton

11.8 Strategies for health services
Chien Earn Lee and Fran Baum

11.9 Training of public health professionals in developing countries
San Hone and Roger Detels

11.10 Transformative learning for health professionals in the twenty-first century for the future health workforce
Wanicha Chuenkongkaew and Suwit Wibulpolprasert

11.11 Humanitarian emergencies
Craig Spencer and Les Roberts

11.12 Principles of public health emergency response for acute environmental, chemical, and radiation incidents
Virginia Murray, Thomas Waite, and Paul Sutton

11.13 Private support of public health
Quarraisha Abdool Karim and Roger Detels

11.14 Global health in the era of sustainable development
Fiona Fleck

11.1

Health needs assessment

Michael P. Kelly, Jane E. Powell, and Natalie Bartle

Introduction

The idea of health needs assessment is, on the face of it, straightforward—estimate the health needs of a population or subgroup in a population and organize services accordingly. The English National Institute for Health and Care Excellence (NICE n.d.) defines Heath Needs Assessment as ‘a systematic process used by NHS organizations and local authorities to assess the health problems facing a population. This includes determining whether certain groups appear more prone to illness than others and pinpointing any inequalities in terms of service provision. It results in an agreed list of priorities to improve healthcare in a particular area’ (NICE n.d.). In the United Kingdom health needs assessment has been used in a wide variety of settings. These have included Immigration Removal Centres (NHS England 2015), local NHS organizations (NHS Kent and Medway 2013); and by local authorities (Horsley and Hollingsworth 2014; Derby City Public Health—Knowledge, Intelligence and Strategic Planning Council 2017). In an uncomplicated world the assessment of health need and healthcare need should be a tool for tackling inequalities, a prerequisite for the allocation of resources and the first step in planning and evaluating care (Stevens and Raftery 1997). In practice however, measuring health needs is not straightforward. It raises difficult questions about healthcare, equity, equality, fairness, and justice (Gillam et al. 2012).

Needs assessment embraces a variety of methods and techniques to gather information about populations (Health Development Agency 2003). In the United Kingdom traditionally it has meant taking ‘a population-based, epidemiological and public health approach to the planning of health interventions’. That means using data about population patterns of health to

do rational planning in order to meet the particular needs of all, or parts, of the community, taking account of equity, efficiency, and affordability. In principle, this approach to needs assessment might be used as the basis for planning in any health system (Clarke et al. 2009). It aspires to 'maximize the appropriate delivery of effective health interventions or care …in an evidence-based way … [to] maximize equity' (Powell 2006).

Globally however, the term health needs assessment is more widely used and deployed to mean a range of different things. Sometimes the approach is highly specific focusing on particular risks or specific diseases (Cox et al. 2013; Cicero et al. 2014; Nacul et al. 2014; Allchorne and Green 2016). Sometimes it is aimed at special subgroups and populations (Ash and Mackereth 2013; Vyas et al. 2013) or for specific services (Ingold and Hicks 2015), and particular populations at particular points in the life course (Pretty 2014). In the United States its currency includes linkages to charity hospitals and the needs of disadvantaged groups who might be eligible to use the hospital facilities or to determine what sorts of services ought to be provided (Apatu et al. 2018; Van Gelderen et al. 2018). The US 2010 Patient Protection and Affordable Care Act created a new legal requirement for non-profit hospitals to conduct Community Health Needs Assessments (CHNA) every three years to maintain not-for-profit status with the Internal Revenue Service (Rosenbaum 2013; Fischer et al. 2018) and various techniques have been used to do this (Mathews et al. 2015; Pennel et al. 2015; Franz et al. 2017; Stone et al. 2018). Practice varies widely (Alfano-Sobsey et al. 2014; Becker 2015).

The website of the US Centers for Disease Control and Prevention (CDC) defines community health assessment as a systematic examination of the health status indicators for a given population that is used to identify key problems and assets in a community. The ultimate goal of this kind of community health assessment is to develop strategies to address the community's health needs and identified issues. A variety of tools and processes may be used according to CDC to conduct a community health assessment; the essential ingredients of which are community engagement and collaborative participation. https://www.cdc.gov/stltpublichealth/cha/plan.html. Health needs assessment has also been used in the United States in Medicaid enrolment (Leininger et al. 2014).

In disaster zones the term denotes ways of getting emergency provision to the population in the most acute of circumstances (Malilay et al. 2014; Hong et al. 2017) or assessing need after disaster has struck (Subaiya et al. 2014). There is no single method or methodology applied, and the term health needs assessment or just needs assessment has come to refer to any way of making an assessment of need.

The UK population-based, epidemiological approach to health needs assessment

We begin by focusing on the approach to needs assessment which originally drew heavily on epidemiology and health economics. In this view assessment of need must include an assessment of the effectiveness of interventions to meet identified health needs (Mooney 1992). The assumption is that the relative total need can be measured sufficiently by just a few factors such as mortality or morbidity rates. It emphasizes quantification and objective comparative measurement. This type of needs assessment is usually dependent on existing or available data, it seldom involves generating new information from primary research because this would be too expensive and take too long. Health needs assessment tends therefore to be based pragmatically on the routine data sources that are there already (Clarke et al. 2009).

There are a number of distinct steps. It begins with clear problem definition. To do this, two questions need to be considered. What do we want to find out about the population, and how can we go about finding it out? A series of further questions can then be asked to help clarify the process: why is this assessment needed now, who will be affected, what would the consequences be of doing nothing, how much time is available, how can the results and the recommendations be presented to maximal effect, are sufficient resources available, and how will the needs assessment be itself assessed and evaluated? (Clarke et al. 2009).

A detailed project plan identifying the component parts of the exercise should be prepared. Central to the exercise will be an epidemiological assessment, which will determine how many people in the population need care and at what level and what services are available for them. It is important to identify the denominator population, that is the total population

or relevant sub- population because otherwise it is very difficult to interpret prevalence (total number of cases) and incidence (new cases in a given time period) (Clarke et al. 2009). Collecting data like these tends to be more straightforward in developed countries. In less developed countries data systems tend not to be so reliable, although where data systems are less than optimal other options may be available (Bonnefoy et al. 2007).

Details of the structure of the population need to be built into needs assessment; the age, ethnic, occupational, and geographical contours of the local population and the patterns of health inequalities will be assessed and described (Kelly 2010a). The level and severity of the diseases of interest need to be examined at this stage too, along with an assessment of the prevalence of relevant risks—for example levels of smoking, physical activity, and alcohol consumption. In a population with a significant number of Afro Caribbean people an assessment of sickle cell disease might be a particular focus of interest, in a mining community chest disease might command attention. The absolute number of people suffering from the condition, and the degree of severity can then be calculated (Clarke et al. 2009). If assumptions have to be made in the absence of data and proxy variables are used, these must be made explicit.

The next stage is sometimes to develop an assessment of the clinical effectiveness of interventions for the condition or conditions of interest. There are a variety of ways of assessing effectiveness. However, if we are concerned to determine the clinical effectiveness of a medicine, or some other kind of intervention, we will get a more accurate assessment if we use evidence in which a control is used and potential biases are minimized. If this is so, we can be reasonably sure that the observed size of the effect is a consequence of the intervention not some other biasing factor. Some commentators suggest that using only evidence drawn from the top of the evidence hierarchy to assess clinical effectiveness is the way to do this (Clarke et al. 2009). The evidence hierarchy is a device which categorizes studies according to the methods they have used and the degree of bias which is associated with the methods. Randomized controlled trials (RCTs) and the meta analyses of such trials rest at the top of the hierarchy as these methods are deliberately designed to reduce bias and offer the greatest certainty that the observed relationship between the independent and the dependent variable is the consequence of that relationship, not some other factor.

It is possible to make an assessment of the effectiveness of interventions by examining clinical trial data for its quality or bias. So it would be quite appropriate, but very time consuming, for someone conducting a needs assessment, who is interested in the effectiveness of particular treatments, to examine RCT findings directly. An easier route is to use the evidence of effectiveness which has already been appraised for its quality in for example Cochrane reviews. These are produced by Cochrane (formerly the Cochrane Collaboration), a world-wide network of reviewers conducting quality appraisal of primary intervention studies. Alternatively, data examined by the National Institute for Health and Care Excellence (NICE) who conduct clinical and cost-effectiveness analyses of new technologies, treatment pathways, and preventive interventions in England can also be consulted. The assessment of effectiveness is based on a set of principles which collectively are known as health technology assessment (Kelly and Moore 2012).

After the effectiveness analysis is complete a synthesis of the evidence is then undertaken comprising the epidemiology in the local population, incidence and prevalence, underlying risk factors, treatments and interventions ranked according to effectiveness, along with evidence of cost-effectiveness and actual costs. These will, of necessity, be imprecise, but if assumptions are made explicit they will suffice as a starting point for needs assessment (Clarke et al. 2009).

Once data have been collected on the local population of interest, the next stage is to compare locally derived data with data from other places in a *comparative needs* assessment. This allows an appraisal of the degree to which local provision is consistent with what might be expected on the basis of the comparisons. This is sometimes difficult because rates of interventions vary both within and between countries, but the process is about putative differences between the observed and the expected values rather than exactness. It is important to consult with stakeholders at this point and genuine community participation is important. Rapid needs assessment may come into play where statistical and other data are unavailable (Clarke et al. 2009). All of this is brought together in a ‘case for change’ document which will outline what is in place, what ought to be done, what stakeholders believe should be done and what the community want done. Costed options are essential to inform redistribution of

resources. The final step is to act on the needs assessment and implement a plan.

Health systems are complex sociotechnical arrangements consisting of people, cultures and practices, organizational structures, equipment, and technologies. Health systems are constrained, because whatever the system, the resources to fund them are finite. Because of finite resources it is not possible to do everything, to fulfil every need or to adopt every new drug, procedure, or device. At the same time demand for health services rises because expectations of what medicine can achieve increase. People, especially in advanced societies, frequently think that they deserve or have some entitlement to—that they *need*—the best and most up to date medicines and procedures. The availability of a medicine therefore often creates the demand *and* the need for it—patients want it and doctors want to use it. All of this adds to the demands on the resources in the system.

Health systems have to adapt to constantly evolving inputs in the form of new medicines, medical technologies, and approaches to medical and surgical interventions. Decisions have to be made about the adoption and use of these new things. In addition, the health needs of populations are complex and change. For example, the age and ethnicity distributions of populations alter. Not only do populations change, but health systems themselves also evolve. Needs assessment, is one means of deciding how to allocate scarce resources to prioritize and tackle need in the face of these complexities. A needs assessment provides a basis for decision making, which will include local knowledge and understanding of the nature of communities, their sociodemographics, and physical environment. Health needs assessment consists of a range of techniques to assess and meet need in a rational and systematic way against this background of complex and changing systems, population diversity, evolution, and increasing expectations and demand. An understanding of the social environment and the social relationships within those environments is essential to successful implementation of needs assessment.

Later in this chapter we explore some of the difficulties attached to this approach. However, it is very important to remember that applying rational principles in this way arose from an understanding that the operation of an unmanaged system in which historical patterns of supply, the generosity of benefactors, the preference of medical practitioners to live in attractive areas, and their wish to provide services that were consistent with *their*

interests led to very unequal and inefficient distribution of resources. So whatever the imperfections of the approach just described, it was born out of a desire to do things better and more fairly.

Other techniques for doing needs assessment

Other techniques may be and have been used to assess need. It has been recommended in the United States in the context of the Affordable Care Act that community health assessments use such principles as: multisector collaborations that support shared ownership of all phases of community health improvement, including assessment, planning, investment, implementation, and evaluation; proactive, broad, and diverse community engagement to improve results; a definition of community that encompasses both a significant enough area to allow for population-wide interventions and measurable results, and includes a targeted focus to address disparities among subpopulations; maximum transparency to improve community engagement and accountability; the use of evidence-based interventions and encouragement of innovative practices with thorough evaluation to inform a continuous improvement process; and, use of the highest quality data pooled from, and shared among, diverse public and private sources (Rosenbaum 2013).

A very wide range of methods have been pressed into service in the United States. Becker notes that ‘Research designs vary across states and agencies, and little is known about the reliability or representativeness of results’ (Becker 2015, p. 15). Telephone and door to door surveys have been used to ascertain people’s self-reported health needs (Stone et al. 2018). Routine data and records (Alfano-Sobsey et al. 2014; Franz et al. 2017), community health partnerships and community surveys (Mathews et al. 2015), key informant interviews, focus groups, mixed methods designs (Becker 2015, p. 15) and opinion surveys (Alfano-Sobsey et al. 2014) have all been used.

Community participation and community engagement are particularly emphasized ‘community participation and mobilization in CHA processes includes the following: recognizing the community as a unit of identity; building on strengths and resources; facilitating collaborative partnerships; emphasizing locally relevant problems and an ecological perspective;

promoting power sharing, co-learning, and capacity building; improving cultural sensitivity, reliability, and validity through quality community participation; increasing community trust and ownership; developing community systems through a cyclical and interactive process' (Pennel et al. 2015, p. 104). Pennel et al. (2015) note that the following should be included 'gathering and analysing quantitative and qualitative data; using data to identify health issues; using broad social determinants of health to identify influences on health issues, including environment, behaviour, socioeconomics, and culture; identifying resources and resource gaps; identifying health disparities; engaging and mobilizing the community; organizing and sharing findings; setting health priorities; developing an action plan to address health priorities; implementing action plans; and providing opportunities for continual feedback with community members' (Pennel et al. 2015, p. 103). But they also note that this broad public health approach, may be at odds with the way that hospitals apply more biomedical principles.

In England, Joint Strategic Needs Assessment (JSNA) develops analysis and strategy for current and future health and care needs of local populations to inform and guide the planning and commissioning of health, well-being, and social care services within a local authority area. Health in All Policies (HiAP) aims to identify interactions between strategic targets, policies and strategies and population health and wellbeing (Public Health England and the Local Government Association 2016) utilizing 'Health Lens' assessment (Kickbusch et al. 2014). Health Lens Analysis has five essential elements that make the most of opportunities to implement 'Health for All' actions for potential population health impact and value. These are:

To engage in order to innovate and establish and maintain strong collaborative relationships across sectors and disciplines.
To determine agreed policy focus among stakeholders to achieve take-up, reach, and delivery of programmes and services. To gather evidence and data: to establish insight on population health need in the policy areas of focus. The use of insights and inference drawn from big, digital routine data sources within JSNA, is increasingly common.
To generate policy recommendations and a jointly owned final strategy report.
To navigate to help steer recommendations through the decision making process.
To evaluate to determine the effectiveness of the health lens.

Health economics and needs assessment

Over the last several decades, two important techniques have come into general use to assist the allocation of resources in the face of scarcity in health systems; health technology assessment which helps to assess effectiveness of interventions and cost utility analysis in health economics, to help assess cost-effectiveness (Kelly et al. 2010). We have already discussed health technology assessment previously in the context of assessments of effectiveness. We now turn our attention to the contribution of health economics.

As a starting point we look to utilitarianism—which is the ethical basis of the discipline of economics. Utilitarian ideas, theories and techniques are concerned with the common good or the greatest happiness of the greatest number (Bentham 1834). Utilitarianism in allocation of resources for the common good of society has much to commend it. But the tools and techniques do tend to favour the status quo. They do not challenge or on the whole try to amend existing resource allocations that might be considered to be unfair or unequal. Utilitarian tools and techniques if applied without critical reflection implicitly approve existing resource allocation and aim to improve social welfare starting from that point. They make what might be viewed as conservative changes to reallocate resources, so that at least one person is made better off and nobody else worse off (Pareto 1935).

Utilitarianism, or the greatest utility of the greatest number, is the ethic that continues to guide economic thinking, tools, and techniques and is also central to the ethic of health needs assessment. However, at the end of the nineteenth century, the future direction of economics as a discipline was influenced by the introduction into mainstream economic theories of mathematical thinking from the physical sciences and engineering.

Economics moved away from broad questions of how resources should best be applied for the good of society—a mix of value judgement on social welfare and observation of facts, to empirical testing of hypotheses using data. For example, economists became very interested in the relationship between unemployment and inflation and used empirical testing of data to test hypotheses about these variables.

A new more scientific language and a range of tools and techniques emerged from economic theories which tended to distance themselves from the messy details of complex human behaviour. The legacy of these changes still survives within economics. A quick perusal of the main economics journals today reveals numerous articles containing strings of equations to test hypotheses, as well as outlining the intricacies and new derivatives of different economic techniques and tools and how to apply them rigorously. In addition, the need within the economics discipline to find mathematical solutions to applied problems has guaranteed survival of the basic assumptions of the economic way of thinking, i.e. people, firms, and institutions are perfectly rational and behave predictably, as if, they have perfect information and control over environment and events. In economics, people are assumed to be equal and social/cultural and environmental/place variations between them are not acknowledged explicitly. So when health needs assessment is viewed as rational, this is the underlying rationality.

Health economics has developed particular techniques and tools, in particular the quality-adjusted life year (QALY) (Cohen and Henderson 1988). This technique is steeped in utilitarianism—allocating resources efficiently for the common good—in this instance fair allocation of health technologies and other interventions in the face of finite resources. Economic thinking is predicated on the notion that most resources are scarce and have limits, but demand on those resources is potentially infinite. This creates choice and opportunity cost meaning that once resources have been allocated and used, they cannot be reallocated and used again in

another way (Powell 2007). Health has been characterized as a multidimensional, dynamic concept subject to changing human expectations and revision over time (McGuire et al. 1988). The economics of health has therefore also developed over time to reflect these changes. Since the Second World War, successive generations of theorists have sought to embrace new ideas about health. For example, economic, social, and psychological aspects of health have been included to reflect changing expectations, the nature of health systems, stage of economic development and passing time (WHO 1947, 1948, 1958). From the late 1980s in developed countries, definitions of health broadened considerably to reflect the notion that wellbeing is inseparable from and crucial to health, and in turn, both are influenced by society as a whole (Bowling 2001).

Cost utility analysis and the QALY have been widely used around the world in the assessment of the cost-effectiveness of new drugs and of other types of clinical interventions. But it has proved less well suited to assessing preventive public health type interventions. The literature identifies a number of conceptual, methodological, and practical difficulties in evaluating the effectiveness and cost-effectiveness of primary prevention interventions (Lorgelly et al. 2010). There are difficulties of producing evidence of cost-effectiveness for public health guidance and methodological difficulties associated with applying health economic techniques to public health interventions (Kelly et al. 2005). These difficulties and issues spring from the focus on efficiency within economic evaluation techniques to the exclusion of the determinants of health and inequalities in health, which are central preoccupations in public health (Barendregt 2006). This is important because as we will show, health needs assessment is itself premised on principles of economic efficiency as well as the principle of equity. And so the same problem that applies to the application of simple cost utility analyses to preventive interventions also applies more broadly to health needs assessment.

Equity, efficiency, and health economics

Health systems may be characterized by the way they organize, finance, and deliver resources to attempt to balance competing health and well-being objectives. Equity in the distribution of resources is the main objective of a

health system where citizens have entitlement and equal access to the means of achieving good health and well-being according to health need. Efficiency in the allocation of resources—where the cost of good health and well-being is minimized and the benefits maximized—is often prioritized by governments over equity in order that the overall cost of a health system as a proportion of GDP, is controlled. Efficiency and equity are generally viewed as competing objectives that most health systems fail to reconcile because both objectives cannot be optimized at the same time—creating an equity-efficiency trade-off. This in turn raises concerns about social justice and generates philosophical arguments about where the balance between efficient resource allocation and equitable distribution of health resources to meet health need, should lie. Early forms of health needs assessment and similar approaches—social audit and rapid appraisal—attempted to establish the size of different health needs and to prioritize and allocate resources pro rata on that basis.

Health economic techniques and tools were directed at inefficiencies in the allocation of healthcare resources and ‘the need for healthcare’ defined as ‘capacity to benefit’ from healthcare became recognized as a key aspect of health status. As a consequence, the health intervention needed in any given circumstance is a function of factors such as the level of prevailing resources, the availability and effectiveness of health interventions, and the perspective and values of those making the assessment. Over time notions of health need have come to be defined as a result of momentum in economic discourse and debates and the development and proliferation of common currency quality of life measures that reflect the outcome and efficiency of interventions to improve health and well-being.

Health systems and the equity–efficiency trade-off

We next examine how economic conditions influence the inter-relationship between equity, efficiency, and cost, describing debates about resource allocation in Beveridge and Bismarckian health systems. These systems take their names respectively from William Beveridge whose report was used as the basis for the policies which helped establish the Welfare State in Britain in the 1940s (Ross 1952) and Prince Otto Von Bismarck who set up the first schemes of national insurance in Germany in the 1880s (Taylor

1955; Thomson 1957). We outline the ways theories of justice underpin equity-efficiency trade-offs, explaining the background to the emergence of new policy concepts and techniques of assessing, measuring, evaluating, and prioritizing efficient and equitable allocation of resources. The nature of the determinants of health and the role of physical and social environment in the improvement of community health outcomes are considered with respect to equity, need, and efficiency.

Healthcare is financed and provided free at the point of use by government through taxation in Beveridge-type health systems. Beveridge health systems differ from Bismarck-type systems in which healthcare is financed by multiple employer-based insurance schemes and taxation in which providers are privately rather than publicly owned. Both types of system however face the same problem of increasing demand in the face of finite resources to fund the systems. In recent decades, some reformers of health systems and health policies have attempted to create market incentives within the system to control overall healthcare expenditure, increase quality, bring down prices and increase choice (Bevan et al. 2010) —for example, providing a choice of provider for healthcare treatments and services as a way of taking some people off waiting lists. However, the economic and social trade-off is that this comes at a price—leaving those who cannot afford to move to languish on the longer waiting list.

Alternatively, other systems that prioritize access and give universal coverage to meet health need provide a narrower range of healthcare services because price signals to reallocate resources within the system are absent. An equity-efficiency trade-off occurs and this can create a dash from central decision making to localism where the allocation of resources is determined by local decision makers. Much of the tension in the balance between efficiency and equity, however, can be traced to expectations and changes over time in the common understanding of what constitutes good health and well-being in communities and countries. Inevitably, these understandings are related to stage of economic development, economic performance, and historical gross national product (GNP).

Some political and philosophical considerations.

At the heart of the approaches to needs assessment outlined so far is a relatively one dimensional idea of need. There is an altogether different approach in the philosophical literature which defines need as a dynamic, politically, and ideologically value-laden idea. In the political and philosophical view, one focus is *unmet* health needs in the form of morbidity in individuals and populations. It is argued that unmet need creates unfair deficits in the distribution of health and wellbeing among communities and populations—deficits that society should take steps to address. In this view, need is described as something that is likely to be dynamic over time as technologies, expectations, and availability of services change. Measurement of need will therefore vary in different contexts, such as, in the clinical setting and at the population level. Needs assessment planning would therefore involve considerations of allocative efficiency along with social values (McIntyre et al. 2009).

Unfortunately, morbidity turns out not to be a good guide to unmet need! We may posit that one person's need is greater than another's because there is a greater degree of morbidity—they are sicker. This in turn implies that one health state is more deserving than another because of the extent of the disease. It also assumes we can accurately measure the differences in disease states. This however may be a false premise because degrees of illness in a strictly biological or pathological sense, even if they can be measured accurately and meaningfully compared, tell us nothing about an individual's quality of life with different degrees of morbidity and therefore their needs in a social or psychological sense. Some people cope extremely well with illness and have a relatively good quality of life while others seem to be rendered incapable of normal social functioning with relatively minor ailments. The subjective experience of illness and an individual's response to it significantly affects their expression of need. It is difficult to measure need solely by assessing biological morbidity in any absolute sense (Anderson and Bury 1988).

There is a still more difficult issue relating to health differences in populations. All health systems, and by definition all health needs assessment, have to confront the systematic differences in the pattern of heath in populations. The health of individuals varies, and the health of groups of individuals also varies; so men and women, age groups, and ethnic groups show *average* differences in life expectancy and patterns of disease. The health of one country varies compared to other countries;

within countries there are differences between individuals and groups. These variations occur because of biological inheritance, because of differential exposure to factors which cause disease and because health services are not uniformly or equally spread between individuals and populations. In any case, individuals make widely different use of available services especially preventive ones. Some people die relatively young, while others live to a ripe old age. Some people live life with multiple health problems and disabilities, others lead lives of a good quality and die peacefully in their beds in their nineties. So as a starting point the demands placed on health systems of these different individuals and groups vary. If we are trying in some way to meet the health needs or demands of various individuals and populations on what basis should we do it?

The straightforward answer at the heart of most approaches to health needs assessment is that we should seek to measure the needs of individuals or populations and we should allocate provision to meet the need accordingly, as that would be the fairest way to do it. But that in turn raises several questions—how exactly would we measure need? Could we do so in a way that was accurate and precise and more importantly could we do it in a way that was fair? And fair exactly to whom? To the people in need or to the people who are not in need and whom we might call upon to fund the needs of others in greater need than them. But why should those not in need, subsidize those who are in need? Is that fair? The answer as to whether one thinks it is fair will be dictated by a political preference as well as general ideas about fairness which would in turn be influenced by other ideas about ethics, morality, duty, responsibility, and liberty for example. Is it fair to have one's income taken away to fund the needs of others? Not forgetting that some people's health needs are in part generated by the choices that they themselves have made in how they have lived their lives, whether they have chosen to smoke, to excessively consume alcohol and food and to take drugs or engage in sexual practices which might expose them to risk of infection. These people have health needs, but whose responsibility is it to meet those needs? Again the answer will be influenced by other value positions. It must also be acknowledged that in response to the international financial crisis of 2008 austerity policies have been pursued in many countries. Consequently, the recent economic environmental context is less likely to trigger positive change in population health impact and value through strategic needs assessment. In response to

austerity in England, local authorities have developed Sustainability and Transformation Plans (STPs) alongside JSNA. These aim to move resources from acute care to prevention by targeting those most in need.

Patterning of health differences

There is another very important dimension at play here which, we have not so far explored in detail; this is the *patterning* of health differences. This removes the discussion from the level of the individual to the level of the social or population. The differences in health experience and health outcomes, the differences in access to services, the differences in exposures to risk and the differences in activities which are health damaging are not distributed randomly or evenly in the population. Health experience and health outcomes are strongly patterned by social position. On whatever measure used to assess health be it mortality, morbidity or self-defined health, the measures follow a social gradient. Those who are better off *on average* enjoy better health, live longer, and make more use of available services. This applies under whatever arrangements for the funding of services operate.

It holds true in market systems like the United States, it holds true in social insurance systems as found in much of Western Europe and it holds true in societies where care is free at the point of use like the United Kingdom. This is called the health gradient—and is one of the most enduring and vexatious characteristics of contemporary health systems (Graham and Kelly 2004).

Comparative data for the United Kingdom and the United States illustrate the gradient. They show similar patterns of graded health differences measured by income for, among other things, self-assessed health, diabetes, heart disease, and lung disease (Banks et al. 2006). The gradient is described in many texts, reports, and papers (Bleich et al. 2012). The shape of the gradient varies; it tends to be steeper in societies with very heterogeneous populations and rather gentler in societies which are more socially homogeneous; but only by degree. Health differences linked to social position remain an enduring and structural feature of contemporary developed societies. In both developing and low income countries the same features of health inequities apply though the shape of the gradient tends to

be more curvilinear. In some developing and low income countries a small affluent elite enjoy good health outcomes and the majority of the population are in a less desirable situation (Bonnefoy et al. 2007).

In broad terms the state of the contemporary world is one where health differences, however measured, prevail across all societies and also between societies. There are absolute health differences between rich and poor societies and between rich and poor people and those not so poor, in all societies (Kelly and Doohan 2012). So here we encounter a major problem with the apparently rational calculus of health needs assessment. Whether health need is expressed in terms of health status or outcome, need (including subjective expression of need) varies systematically across the population and there are considerable health inequalities. It can be, and often is argued that the needs of the most or the relatively disadvantaged are paramount and therefore resources should be deployed in such a way that they meet those needs first. All that would remain to be done would be to find a technical solution to the measurement of need.

In addition to the health gradient there are two other ways of looking at health differences called health disadvantage and health gaps (Graham and Kelly 2004). Health disadvantage simply focuses on differences, acknowledging that there are differences between individuals, distinct segments of the population, or between societies. It is a descriptive and non-judgemental approach. The health gaps approach, in contrast, focuses on the differences between the worst off and everybody else, often inviting the conclusion that those who are not the worst off enjoy uniformly good health. The gap approach also invites the conclusion that this state of affairs is wrong or unfair.

A gap approach and a gradient approach lead to rather different solutions. Conceptually, narrowing health gaps look to actions which will improve the health of the poorest regardless of the rest of society. Such an approach would be one which achieved both an absolute and a relative improvement in the health of the poorest groups. The health gradient approach takes, as its starting point in contrast, the acknowledgement that the penalties of inequities in health affect the whole social hierarchy even though they increase from the top to the bottom.

Gaps and gradient approaches lead to quite different assessment of need and how to meet those needs. It may seem counterintuitive, but if policies or actions only attempt to target the needs of people at the bottom of the

social hierarchy, there will be little or no impact on health inequalities across the rest of society (Rose 1985, 1992). This is because inequities in health will still exist, the social determinants continue to exert their malign influence and the health needs of the majority of the population remain unattended to. The alternative approach advocated by the Marmot reviews for example and World Health Organization (WHO) (Bonnefoy et al. 2007) involves a consideration of the whole gradient in health inequities rather than only focusing on the health of the most disadvantaged. An effective policy in this regard is one that meets two criteria. It is associated with (a) improvements in health (or a positive change in its underlying determinants) for all socioeconomic groups up to the highest; and (b) a rate of improvement which increases at each step down the socioeconomic ladder. In other words, a differential rate of improvement is required: greatest for the poorest groups, with the rate of gain progressively decreasing for higher socioeconomic groups. It locates the causes of health inequity, not in the disadvantaged circumstances and health-damaging behaviours of the poorest groups, but in the systematic differences in life chances, living standards and lifestyles associated with people's unequal positions in the socioeconomic hierarchy (Graham 2004a, 2004b, 2006). The significant caveat is that where the health gap is both large and the population numbers in the extreme circumstances are high, a process of prioritizing action by beginning with the most disadvantaged would be the immediate concern.

From health inequality to health equity

The gradient leads to some other questions. Does it matter that there are patterns of inequalities in health? Why should we seek to remedy this state of affairs? Why use the instruments of health service provision to deal with this problem? Why conduct needs assessment to try to change the gradient? After all, all societies demonstrate a gradient to some extent. Even though health differences are clear, a case may be made that over the last century and a half, things have improved for nearly everyone—at least in the developed West. Rates of infant mortality are at an historic low and life expectancy has never been greater. Furthermore, while *average* patterns of morbidity and mortality in different social groups and populations may be

clear, there are wide variations between individuals. In short, not all well to do people live to a healthy ripe old age, and not everyone who is disadvantaged dies young. To be human is to know that we will eventually die and that the manner and timing of our death is in almost all circumstances beyond our control whether we are rich or poor. So is it the case that patterned average differences in *early* and largely *preventable* death and suffering are unfair and unjust just because they are theoretically preventable—particularly when there is no necessary inevitability that relatively disadvantaged people should die earlier than anyone else? There is no straightforward answer.

A number of writers have explored the theme of injustice in this context. The WHO has had a long-standing interest in the matter. The WHO's Commission on the Social Determinants of Health used a series of definitions to clarify things which were based on the work of Whitehead (Whitehead 1992; Whitehead and Dahlgren 2006) and Solar and Irwin (Solar and Irwin 2010). A distinction is made between health inequality and health inequity. Health inequality is defined as *health differences which are not avoidable or preventable, are not the consequence of human actions and activities, but are based on genetic or constitutional individual differences, age, or biological sex.* These are sometimes also referred to as variations (Kelly et al. 2007). Health inequity, in contrast, is defined as *unfair and avoidable or remediable differences*. Health equity in turn is defined as *the absence of unfair and avoidable or remediable differences in health among social groups* (Solar and Irwin 2010). The italics highlight the difference in definitions. It is particularly important to note that the difference in definition between inequity and inequality is not used universally and many writers and commentators employ the two terms as synonyms. Also the distinctions between individual differences which are based on human biology and differences arising from interaction between an organism and an external man-made hazard are in reality difficult to draw in anything other than an analytic sense. Empirically the divides are much fuzzier than these definitions suggest. However, as a way of beginning to find some clarity the distinction is helpful. Equity and inequity are not products of nature they are the products of human actions and, as they are socially, economically, and/or politically produced they are, theoretically at least, modifiable. The defining characteristics of equity are

fairness and justice; the defining characteristics of inequity are unfairness and injustice (Whitehead and Dahlgren 2006)

Fairness and unfairness can be conceptualized as absolutes; something or some state of affairs is either fair or unfair—it cannot be both at the same time. But thinking like this in absolute terms misses the point that fairness and unfairness are not properties, things, or states of affairs, but are about *relationships* between people. Fairness and unfairness arise as a consequence of the nature of the *relationships* between people and the ebb and flow of human affairs. So too justice is not a simple measure of equitable distribution of resources according to need, but is about the nature of relationships in society.

Justice may be understood in terms of the properties of people, their conduct, the rules that govern their affairs and the characteristics of institutions—an absolute definition The alternative is a *relational* concept of justice which concerns itself with the 'justness' of relations between people—X is unjust to Y (Pogge 2003). Such relations involve human agency, purpose, and motive. From a relational point of view justice should not be about fair distribution; it should be about seeking to identify the agents responsible for the social arrangements that determine the shape of human relationships (Pogge 2003).

Health equity may be conceptualised in distributional or relational terms in the same way that justice can be. The distributional approach is focused on the goal of equalizing good health across society—which is really the underlying value position of most health needs assessment and much of the discourse about health inequity. The relational view is about the balance between the harms inflicted and harms mitigated or prevented. In the relational view social institutions and those responsible for them should have more concern to prevent and mitigate those things they cause themselves and for which they are responsible, rather than those things which are outside of their control (Pogge 2003). So the real issue in relational terms, with which equity should be concerned, is not the fact that health is differentially distributed, but that social systems contribute to the differential distribution of health and disease. Poverty is the greatest contributor of all to ill health. Organizing economic arrangements so that they do not generate the conditions of poverty which generate ill health is thus, it may be argued, a moral duty. Therefore, focusing on assessing need without looking at the wider determinants of the causes of inequity in the

first place, is misguided. The global economic order is responsible for the generation of ill health and health inequities (Pogge 2003), and the social arrangements that have negative effects on health are unjust (Venkatapuram 2011).

In a celebrated paper called '*What is the point of equality*?' Anderson outlined a number of the problems associated with the distributional concept of equality. She warns against the notion that we can construct institutions to make them more equal and fairer, not least because this tends towards greater and greater state interference in the lives of the citizenry. She notes that one of the dangers associated with egalitarianism and the distributional approach is that it pushes the limits of the state further into the lives of ordinary people. This is a particular danger if states become concerned with equities of health rather than oppressions of the powerful against the powerless. In other words, focusing on rearranging healthcare resources to make the patterning of health more equitable is rather like moving deck chairs around on the deck of the Titanic after she hit the iceberg—largely irrelevant and pointless when the ship is sinking. 'Recent egalitarian writing has become dominated by the view that the fundamental aim of equality is to compensate people for undeserved bad luck—being born with poor native endowments, bad parents and disagreeable personalities, [and] suffering from accidents and illnesses' (Anderson 1999). This she says is not really the issue. We should instead be preoccupied with ending oppressive social relations rather than trying to ensure that everyone gets what they morally deserve. It is about creating a community in which people have equality in relationships with one another (Anderson 1999)—a community built on cooperation between members who see themselves as equals. She argues for democratic equality which means that all law abiding citizens are allowed effective access to the social conditions of their freedom. Anderson criticizes the view that the purpose of distributive justice is to compensate people for their misfortune. Her argument strikes at the heart of the idea that health inequities are simply unfair; her position is that while life may be unfair, bad luck is not at the heart of it—it is the relations between people and the way that they treat each other that is much more fundamental.

Anderson draws our attention to the fact that the concept of equality can mean a number of different things depending on the underlying political value position and the epistemological assumptions of the theory. So a

utilitarian seeking to maximize the greatest happiness of the greatest number would see things differently to a Marxist seeking absolute parity in access to wealth for example. She demonstrates that equality is a rationalist rather than an empiricist concept, meaning that the discourse about distributional equality and heath equity is premised on the manipulation of ideas and the contest between different ideas and political philosophies, rather than being something that can be demonstrated by empirical methods (Millican 2007). True, protagonists will appeal to empirical evidence about poverty and about wealth to justify their arguments, but in the end, much of the discussion about equality is grounded in ideal discussions of future desired states and institutions and the manipulation of ideas to justify that view, rather than the case being made that empirically things could be changed by doing x, y, or z. Where x, y, or z are suggested, they are selected on the basis of ideological preference, not empirical science.

Bernard Williams argued that the proper grounds for the distribution of healthcare are health need (Williams 1962). This he saw as a fundamental truth (Wolff 2007). Others have argued that this is not a fundamental truth at all. Nozick has suggested that the focus on need, common among egalitarian thinkers, is to define people quintessentially as consumers. The task then becomes one of finding the best way to ensure the fair distribution of available goods to consumers (Nozick 1974). The problem of course with this approach, is that this casts humanity into a fundamentally passive role and it doesn't consider individuals as active producers (Wolff 2007). Williams also argued that there is nothing about need itself intrinsically that should motivate action on equity (Williams 1962). What society decides to do for people, particularly in health need, depends upon a moral and ethical stance.

Another argument is that the goal of fair distribution should be treating the worst off as well as possible rather than flat equity (Wolff 2007). John Rawls argued that difference in the distribution of primary goods such as health is tolerable, so long as the welfare of the most disadvantaged is looked after (Rawls 1971). Dworkin takes this argument a stage further. He suggests that we need to determine why the worst off are in that position. Dworkin asserts that some may not be able to work because they are unable to find work; but others may decide they do not want to work. Dworkin also notes that some people's needs are greater than others (someone who is severely disabled has greater need than someone who is able bodied for

example—although as we noted earlier, this is a highly questionable first premise). But needs, Dworkin notes, are compounded by taste. Some people's needs are determined by expensive tastes say for drugs or wine or jewellery. Dworkin therefore distinguishes between brute bad luck and circumstance from optional bad luck involving some degree of culpability for being in need. For Dworkin, this distinction is all important, as it is his view that the state has a duty to deal with the former, but not the latter (Dworkin 2000).

Some further philosophical reflections

Health needs assessment is not value neutral it is premised on a number of philosophical and political concepts. It is helpful to explore the arguments relating to these ideas not least because although they are fundamental to health needs assessment the conventional literature about health needs assessment seldom considers the underlying assumptions and issues in detail. This is particularly so with respect to the central importance of utilitarianism and its contested place in the philosophical canon. The discussion is made all the more confusing because many of the relevant ideas like justice, equality, fairness and so on are in common as well as technical philosophical usage and the meanings attached to common sense understanding of the words and the technical vocabulary usage are seldom commensurate.

Another very intriguing characteristic of the literature—a literature after all which readily uses terms like justice, fairness, equality, is that with the exception of Sen (2009) (more of whom to follow) the origins of the arguments about justice and need are conducted almost entirely without reference to their antecedents in Christianity or the teachings of the other major world religions. The secular attempts to define the equal society, justice, and fairness struggle in this regard, to find a moral or ethical point of reference. The other very odd thing is that conventional political philosophy has, down the years, actually paid scant attention to health and health inequalities and issues of distribution of healthcare resources (Venkatapuram 2011).

At the heart of the debates about equity and health inequity in contemporary society is, as we saw earlier, social justice. Health inequities

are considered by many commentators to be unjust and unfair and this is advanced as a reason for allocating resources following needs assessment, more fairly. Just as utilitarianism is central to health needs assessment, it has also been central to theories of justice too viz. what produced the greatest happiness in the greatest number was considered to be socially just (Venkatapuram and Marmot 2009). This absolute or distributional approach to justice was based on a highly individualistic model and was premised in turn on the individual as the unit of analysis and not the alternative which is the relational conception of social life. The solution in the utilitarian view of the world is fair distribution and systems which can efficiently distribute resources, a principle at the very heart of health needs assessment. The utilitarian argument overlooks, by virtue of its basis in the maximization of utility, not only the fact that humans are motivated by a variety of things, not just maximizing utility, like love, social conflict, human venality, and sheer and utter evil, but also it doesn't seem to have a concept of the social (Etzioni 1988). In other words, individual utility maximizing motivated agents are the focus of the argument, not social relations between people. This individualistic approach chimes with medicine.

Medicine has as its principal focus, pathology in the individual human body. This means that, with only a few historical exceptions, the intellectual interests have been oriented to phenomena located in individual human bodies or minds with pathology measurable in ways that reflect the individual and individual variation from some notion of what is normal or healthy (Antonovsky 1985, 1987; Carter 2003). This approach received an enormous boost with the rise of germ theory and with the fantastic successes of the isolation of pathogenic microbes and then antibiotics to combat them (Kelly and Russo 2018). Of course medicine has other foci too; especially body systems, but the variables and phenomena of particular interests are individual. The pathogenic paradigm, in which the fundamental rationale is isolating the specific cause (pathogens) of specific outcomes (pathologies), reinforces the individualistic approach (Antonovsky 1987). Psychology has likewise made its main focus the individual so measures of intelligence, personality, quality of life, all reside in the individual or are characteristics or properties of individuals. The approach is about the degree to which things go wrong in individual bodies and minds and the preceding causes of the pathology. Need as an individual property is an obvious next step in this way of thinking.

Economics, and especially health economics has bought into this individualistic paradigm in a big way. The QALY and the application of cost utility analysis and health technology assessment are deeply infused with an individualistic orientation. The greatest happiness of the greatest number is the utilitarian philosophical view which sits very comfortably with the individualistic paradigm because it conceptualizes the notion of the good as the aggregate of lots of different individual utilities. The ontological consequence of this—i.e. the assumptions made about what constitutes human life and how and why it is the way that it is—is the idea that the essence and meaning of human existence can be captured by isolating these individual characteristics and seeing how they connect to each other. So individual characteristics as different as height, weight, blood pressure, bone density, hair colour, IQ, biological sex and size of tumour for example can each in turn be linked to the presence of other characteristics in the individual like health of mother, poor nutritional status in childhood, genetic coding, parental heredity, age, chromosome structure, and exposure to tobacco smoke. The causal link is from one individual characteristic to another.

Now all of this is intuitively meaningful—in the modern Western world the individual and the individuated self are touchstones of the way we live our lives and the way the state regulates our lives—we have for example individual national insurance numbers, passport numbers, birth certificates, tax codes, and genomic structures. Individual variables seem natural. Further in medical terms there have been some remarkable successes in isolating disease mechanisms and offering curative technologies using the individual approach– although far less than the popular imagination often supposes—using these principles.

However, there are two fundamental objections which suggest that an additional way of viewing things might be helpful. First the individualist approaches dehumanize people—people are reduced to some subhuman characteristic—a number, a genetic code, a pathological organism, a utility. But second, and more importantly, and notwithstanding the advances that have followed in the wake of some of the individualistic connections which have been made, above all it ignores or relegates the fact that humans live in groups and that those groups are in relations with each other and that membership of those groups is a defining characteristic of identity and of profound importance to most people—one’s family, class, tribe, caste,

gender, ethnicity, nationality are all paramount social markers in life. Moreover, the relationships within which we live our lives, the relations with other people, the relationships within and between different groups, shapes the nature of our human selves, our experiences, and our behaviour (Kriznik et al. 2018). The defining characteristic of human life is belonging —to be a member of multiple groups and communities. We live our lives in a network of interlinked and overlapping relations with others. Knowingly or not, the desire to belong has far reaching consequences on the types of behaviours that we adopt and the choices that we make. Not surprisingly therefore, the effects of social exclusion or isolation from social groups are of paramount importance when trying to explain the health of an individual.

What this makes us pause and reflect on is not that the individual understanding of human affairs is unimportant—that would clearly be an absurd position to adopt. Rather it is that a full understanding of the human condition requires additionally another set of concepts—which capture human relationships—relational concepts—in order to develop a rounded account of human life (Kriznik et al. 2018). And moreover inequity and injustice are best understood in relational not individualistic terms.

Social theorists have grasped the idea of a well-rounded approach very neatly in the conception of the dynamic interaction between agency and structure. The idea of agency is that we are all unique biological, psychological, and physical individuals. We all engage in individually motivated actions and behaviours which are in part the results of our unique individuality. But the sum of all human behaviours is the social structure which is the product of the millions and billions of human relationships that are the medium for individual actions. Those structures or webs of human relationships are relational and in turn they constrain, drive, and facilitate individual human behaviour. So we have individual behaviour, the medium for its expression which are human relationships, and social structures which are the sum of all those relationships which in turn impinge on and delimit the possibilities of individual behaviour (Giddens 1979, 1984; Elder-Vass 2010).

This rather abstract approach allows us in turn to consider some of the thorny problems at the heart of health needs assessment. Thinking of the dynamic interaction between agency and structure moves us beyond the methodological individualism of traditional epidemiology and therefore of health needs assessment and provides an escape from the individual level of

explanation (Frohlich et al. 2001). This in turn allows for a fuller understanding of the dynamics of the development of health inequalities. (Abel and Frohlich 2012),

This way of thinking has been linked to what is called capability theory (Sen 2009). Under austerity policies, the notion of capabilities has started to enter into JSNA, HiAP (see earlier section on some political and philosophical reflections). The core characteristic of capability theory is its focus on what people are effectively able to do within relationships with each other (Abel and Frohlich 2012). Individuals being able to engage effectively in what they really want to do, is the core idea. These engagements include being active, healthy, and being able to work but the list is not limited to these activities. In the capability approach, resources, and their fair distribution to individuals (the focus of the traditional utilitarian approach) are not the central interest. Resources are not ends in themselves—they are means to ends. And neither are resources reducible to monetary utilities. People's abilities to realize their life goals and plans are the focal point. In the capability approach, the issue of justice, fairness, and need does not apply to resources *per se* but to the range of options for agency—the capabilities (Abel and Frohlich 2012).

Sen is a key thinker in relation to these arguments. He makes the case for a dynamic approach to social justice. He has argued that to think about justice like this is a major shift away from the traditional ways that philosophers have thought about justice. This is because they were locked into the utilitarian/ distributional tradition (Sen 2009). This is important in the context of needs assessment which is fundamentally utilitarian and implies that through its rationality and market allocative efficiency it will be possible to deliver fairness by market redistribution. This follows a tradition which goes back, according to Sen, to the Enlightenment. The utilitarian position is a rationalist position writ large, i.e. a position which is about the manipulation of ideas, rather than the observation of empirical facts (Millican 2007). The ideas about social justice which are being manipulated are that it is possible to design institutional arrangements that will deliver justice and this will be the basis of a perfect society. For Sen, justice is a relative concept and is about relations between people. Justice is a process, an aspiration; it is about advancing justice or reducing injustice. It is not about finding or describing the perfectly just society. Because social structures and their properties emerge out of human relationships, it means

that to try to legislate to change social systems to make them more just or fair can only ever be partially effective. In a sociological sense justice and injustice are properties of social systems not abstract transcendental things that can be made by social actors. Sen's approach is comparative, plural, iterative, dynamic, and acknowledging of alternatives.

Justice in this view is not a given—a rationally derived static universal principle. It is about relations between people and arises as a consequence of social action and social structure. Therefore, injustice will also arise socially in social interaction and is decided upon morally or metaphysically. The judgement about whether the relations between people, such as differences in health, are just or unjust is a value judgement. In short, seeking to bring about equality in relation to health by the utilitarian redistribution of resources is never likely to work, or to produce fairness and justice because justice is not a quality of individuals or institutions. Equality is about human relationships and the utilitarian approach at the heart of health needs assessment is based on an alternative individualistic ontology.

Sen's argument is that theories of justice major on something; it could be happiness as in the case of classical utilitarianism, and it could be resources or income. Sen argues that in contrast to these utility-based or resource-based lines of thinking, individual advantage is judged in the capability approach by a person's capability to do things they have a reason to value. A person's advantage in terms of opportunities is judged to be lower than that of another if they have less capability—less real opportunity—to achieve those things that they have reason to value. The focus is on the freedom that a person actually has to do this or be that—things that they value doing or being (Sen 2009). Sen's approach is about human life and the opportunities for living—living as against just existing. Justice for Sen cannot be created by social institutions. Justice should aim to reduce injustice, i.e. change the nature of the relationships between people, rather than aiming to produce a perfectly just society.

Conclusion

Health needs assessment can be seen as a highly rational and straightforward means of identifying health needs, of linking needs to

patterns of health inequities, of marshalling appropriate resources to match those needs and then deploying resources accordingly. And of course that is what it aspires to do. But that aspiration is based on the traditional rationalist approach of utilitarianism and in turn on an individualist ontology. Need, as we have shown, is not a static, objective thing. It is therefore intrinsically difficult to measure and as soon as one starts to try to capture the idea fundamental questions about fairness and justice are raised.

The objective and measurable concept of need is located in an individualist ontology. When we think instead in relational terms about equity, justice, and human capabilities a different perspective is possible. Capabilities theory captures the idea of equity in a far more nuanced way than individualistic utilitarian accounts. It also offers a more complete way of understanding how we might rethink health needs assessment.

So rather than seeking to measure individualistic objective variables and then seeking to apply resources accordingly, the relational capabilities approach bids us to think about the relations between service providers and users in a novel way. It requires us to not to try to match resources to a completely slippery and spuriously objective concept of need, but instead makes us consider the nature of the relationships between people and services.

Maximizing health outcomes, a utilitarian fundamental principle, may not be what social justice requires and alleviating injustice may require more than maximizing efficiency (Venkatapuram 2011). The argument then hinges on several fundamental things: the degree to which the structure and organization of services permits people to meet their own capabilities, the degree to which it allows human potentials to be realized, the degree to which the relationships with services do not distort or alienate people from themselves or from others. The capability approach privileges human functioning, not maximizing utility or achieving an idealized and biologically improbable disease free state (Venkatapuram 2011).

This means that service design should reflect and respect human dignity; this means respect for an individual person, but at the same time acknowledging the limitations which disease and disability genuinely impose on people while recognizing that there is not a direct cause and effect linear relationship between disease state and social functioning. It should acknowledge the fact that illness and disease by their very nature tend to exacerbate disconnectedness between the person and their normal

role responsibilities, their primary social attachments, and their desired capabilities. The experience of disease and disability can also produce a profound separation of the individual from their sense of self or their sense of who and what they want to be. The capability approach therefore requires us to respect the lifeworlds of ordinary people and work with them to build their skills, assets, and capabilities (rather than focusing on their deficits and trying to correct them via resource redistribution). These skills, assets, and capabilities allow them to manage their lifeworlds with minimal interference. The kinds of skills which enhance capabilities include interpersonal relationships, technical skills to manage the routine aspects of social and economic and domestic life, skills to develop emotional and psychological resistance (often referred to as coping or resilience mechanisms) and an ability to make life seem meaningful. These skills enable people to manage the routine travails of ordinary living as well as the more significant life events which engulf everybody from time to time (Kelly 2010b). It also allows people to manage the material lifeworld they inhabit. This is important because the material and psychological lifeworlds mediate the stressors—physical, psychological, and biological—which assault the human body periodically. The greater the ability people have to control their lifeworlds, the greater the resilience they will have. Skills to control the lifeworld are quintessentially capabilities. The inability to exert that control forms the basis of the patterning of health inequities because the ability to exert control is not spread uniformly through the population. Following Sen's prescription, we should seek not to measure need in a potentially spuriously scientific way, but rather acknowledge that the *total* population, all of us, are in need of strengthening our capabilities. Further it is relatively easy to predict where the need to develop more capabilities and skills is greatest and that is among the poor, the disadvantaged, subgroups, and minorities, and people with disabilities. Efforts to develop and build the skills and capabilities should be proportionately, but not exclusively, focused on these groups. The level of self-empowerment to help to realize these capabilities needs to be appropriate for the individual. This is an important consideration when deciding where to deploy resources. When this happens, it almost always involves active public participation to determine what it is that people value.

Moreover, upstream efforts need to embrace a public health preventive approach; an education in basic skills for living, appropriate role modelling

as the basis of skills for human interaction, proper deportment, manners, respect for others, management of emotions, and the development of a balanced sense of self and identity—are the basics of human socialization. This means that the early years are particularly important and that taking the edge off encounters between people that are destructive and harmful is paramount.

While all of this plays out on life's grand stage, it must also apply to the way health services are provided and delivered. If services are organized around allocative efficiency alone and neglect the sense of self, the identity, the skills required to negotiate the system, and they alienate patients and the public from and systematically deskill the recipients of care, they meet no one's needs at all, and serve only to make the patterns of inequities worse. They become, in other words, a contributor to the problem, rather than a solution to it. Unfortunately, the utilitarian approach with its individualistic ontology, and its emphasis on ideal rationality, is focused on a question, which on the face of it is not unreasonable, of how to match resource deployment to need. This is in fact the wrong question, so it often ends up doing exactly the opposite of what it is trying to do and making matters worse. A purely deficits approach to health needs assessment may well be neat and tidy, however it only tells one side of the story. Widening the approach to encompass assets starts to get complex and messy. Where Sen's capability theory is involved there is no ready-to-go formula. Instead of avoiding what is hard to do we should embrace the challenge. It is after all, the challenge of understanding what it is to be human. Fortunately, there are many people who have risen to this challenge and remarkable things are happening across the Globe. While helping people in need is a virtue we mustn't lose sight of the fact that we are all part of the human species and we all live on the same planet. None of us are equal until all of us are equal.

We should understand needs assessment in terms of building assets and capabilities rather than focusing on deficits. In the context of the aftermath of the global financial crisis, austerity, the ageing population, rising levels of obesity, and costly healthcare systems there are clear implications for the nature of health needs assessment. What is required is greater capacity to harness insight from routine digital data, more acceptance of ideas about creating agency among those most in need and acknowledging the impact of the structural determinants of health on capability. Population

approaches, targeting those in greatest need allow the harder edge of human existence to be ameliorated.

It does mean though that the temptation to try to find purely rationalist solutions which can only ever work in theory, or indeed in Utopia—nowhere—rather like transcendent theories of justice, must be resisted. It must be acknowledged that it is about human relationships and allowing those relationships to be as humanly meaningful and fulfilling and permitting of enhancing human capabilities as possible. The most important task in a good and just society is to prevent the erosion of those things which protect and maintain human capabilities. It is to protect the virtuous and the virtues and to enhance them where possible and to protect them by supporting social arrangements which facilitate that. Relationships cannot be legislated for, nor made to happen by complex processes of resource allocation. Justice and the social arrangements that maintain relationships are in the end the emergent properties of social relations. Those relations must be cultivated in a humane, and as far as possible, just way that protects and enhances people's capabilities. This must be the prerequisite for any state wanting to deploy its healthcare resources in a way that is both effective and meaningful. Building capability into service design is a fundamental prerequisite.

REFERENCES

Abel, T. and Frohlich, K.L. (2012). Capitals and capabilities: linking structure and agency to reduce health inequalities. *Social Science and Medicine*, **74**, 236–44.

Alfano-Sobsey, E., Ledford, S.L., Decosimo, K., and Horney, J.A. (2014). Community health needs assessment in Wake County, North Carolina: partnership of public health, hospitals, academia, and other stakeholders. *North Carolina Medical Journal*, 75(6), 376–83.

Allchorne, P. and Green, J. (2016). Identifying unmet care needs of patients with prostate cancer to assist with their success in coping. *Urologic Nursing*, 36(5), 224–32.

Anderson, E.S. (1999). What is the point of equality? *Ethics*, **109**, 288–337.

Anderson, R. and Bury, M. (eds.) (1988). *Living with Chronic Illness: The Experience of Patients and their Families*. London: Unwin Hyman.

Antonovsky, A. (1985). *Health Stress and Coping*. San Francisco, CA: Jossey Bass.

Antonovsky, A. (1987). *Unravelling the Mystery of Health: How People Manage Stress and Stay Well*. San Francisco, CA: Jossey Bass.

Apatu, E., Hamadi, H., Abrams, J.A., Metzger, I.W., and Spaulding, A. (2018). Association between hospital community services and county population health in the USA. *Journal of Public Health*, 41(1), 71–9.

Ash, M. and Mackereth, C. (2013). Assessing the mental health and wellbeing of the lesbian, gay, bisexual and transgender population. *Community Practitioner*, 86(3), 24–7.

Banks, J., Marmot, M., Oldfield, Z., and Smith J.P. (2006). *The SES Health Gradient on Both Sides of the Atlantic*. Cambridge, MA: National Bureau of Economic Research.

Barendregt, J. (2006). Economics and public health: an arranged marriage. *European Journal of Public Health*, 17(2), 124.

Becker, K.L. (2015). Conducting community health needs assessments in rural communities: lessons learned. *Health Promotion Practice*, 16(1), 15–19.

Bentham, J. (1834). Deontology or the science of morality. In: Bowring, J. (ed.) *The Works of Jeremy Bentham: Two Volumes*. London: Longman.

Bevan, G., Helderman, J., and Wilsford, D. (2010). Changing choices in healthcare: implications for equity, efficiency and cost. *Health Economics Policy and Law*, **5**, 251–67.

Bleich, S.N., Jarlenski, M.P., Bell, C.N., and LaVeist, T.A. (2012). Health inequalities: trends, progress and policy. *Annual Review of Public Health*, **33**, 7–40.

Bonnefoy, J., et al. (2007). *Constructing the Evidence Base on the Social Determinants of Health: A Guide*. [Online] Available at: http://www.who.int/social_determinants/knowledge_networks/add_documents/mekn_final_guide_112007.pdf

Bowling, A. (2001). *Measuring Disease* (2nd ed.). Buckingham: Open University Press.

Carter, K.C. (2003). *The Rise of Causal Concepts of Disease: Case Histories*. Aldershot: Ashgate.

Cicero, A.F., et al. (2014). Population health needs assessment and healthcare services use in a 3 years follow-up on administrative and clinical data: results from the Brisighella Heart Study. *High Blood Pressure & Cardiovascular Prevention*, 21(1), 45–51.

Clarke, A., Powell, J., and Lansang, M.A. (2009). Needs assessment: a practical approach. In: Detels, R., Beaglehole, R., Lansang, M.A., and Gulliford, M. (2009). *Oxford Textbook of Public Health*. Three Volumes (5th ed.). Oxford: Oxford University Press, p. 1549.

Cohen, D.R. and Henderson, J.B. (1988). *Health Prevention and Economics*. Oxford: Oxford Medical Publications.

Cox, C.L., et al. (2013). Development of a comprehensive health-related needs assessment for adult survivors of childhood cancer. *Journal of Cancer Survivorship*, 7(1), 1–19.

Derby City Public Health—Knowledge, Intelligence and Strategic Planning (2017). *Children and Young People's Mental Health and Emotional Wellbeing Health Needs Assessment*. South Derbyshire, North Derbyshire, Erewash, Hardwick Clinical Commissioning Groups, Derbyshire County Council and Derby City Council.

Dworkin, R. (2000). *Sovereign Virtue*. Cambridge MA: Harvard University Press.

Elder-Vass, D. (2010). *The Causal Power of Social Structures: Emergence, Structure and Agency*. Cambridge: Cambridge University Press.

Etzioni, A. (1988). *The Moral Dimension: Towards a New Economics*. New York: Free Press.

Fischer, K.R., et al. (2018). A content analysis of hospitals' community health needs assessments in the most violent U.S. cities. *Journal of Community Health*, 43(2), 259–62.

Franz, B., Skinner, D., and Kelleher, K. (2017). The impact of the affordable care act on hospital-led community health evaluation in the U.S. Appalachian Ohio region. *Journal of Evaluation in Clinical Practice*, 23(4), 882–7.

Frohlich, K.L., Corin, H., and Potvin, L. (2001). A theoretical proposal for the relationship between context and disease. *Sociology of Health & Illness*, 23(6), 776–7.

Giddens, A. (1979). *Central Problems in Social Theory: Action, Structure and Contradiction in Social Analysis*. Berkeley, CA: University of California Press.

Giddens, A. (1984). *The Constitution of Society: Outline of the Theory of Structuration*. Berkeley, CA: University of California Press.

Gillam, S., Yates, J., and Badrinath, P. (2012). *Essential Public Health: Theory and Practice*. Cambridge: Cambridge University Press.

Graham, H. (2004a). Social determinants and their unequal distribution: clarifying policy understandings. *Milbank Quarterly*, **82**, 101–24.
Graham, H. (2004b). Tackling health inequalities in England: remedying health disadvantages, narrowing gaps or reducing health gradients. *Journal of Social Policy*, **33**, 115–31.
Graham, H. (2006). Socioeconomic inequalities in health: evidence on patterns and determinants. *Benefits*, **14**, 77–90.
Graham, H. and Kelly M.P. (2004). *Health Inequalities: Concepts, Frameworks and Policy*. London: Health Development Agency.
Health Development Agency (2003). *The Working Partnership*. London: Health Development Agency.
Hong, K.J., et al. (2017). Rapid health needs assessment after typhoons Bolaven and Tembin using the public health assessment for emergency response toolkit in Paju and Jeju, Korea 2012. *Journal of Korean Medical Science*, 32(8), 1367–73.
Horsley, S. and Hollingworth, E. (2014). *Children and Young People's Emotional Wellbeing and Mental Health: Health Needs Assessment*. Sheffield: Sheffield City Council.
Ingold, K. and Hicks, F. (2015). Using a public health approach to improve end-of-life care: results and discussion of a health needs assessment undertaken in a large city in northern England. *BMJ Supportive & Palliative Care*, 5(2), 200–2.
Kelly, M.P. (2010a). The axes of social differentiation and the evidence base on health equity. *Journal of the Royal Society of Medicine*, **103**, 266–72.
Kelly, M.P. (2010b). A theoretical model of assets: the link between biology and the social structure. In: Morgan A. and Ziglio E. (eds.) *International Health and Development: Investing in the Assets of Individuals, Communities and Organisations*. New York: Springer.
Kelly, M.P., McDaid D., Ludbrook A., and Powell J. (2005). Economic appraisal of public health interventions. Health Development Agency. [Online] Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.476.3820&rep=rep1&type=pdf
Kelly, M.P., Morgan, A., Ellis, S., Younger, T., Huntley, J., and Swann, C. (2010). Evidence based public health: a review of the experience of the National Institute of Health and Clinical Excellence (NICE) of developing public health guidance in England. *Social Science and Medicine*, **71**, 1056–62.
Kelly, M.P., et al. (2007). The social determinants of health: developing an evidence base for political action. Final Report to the World Health Organization Commission on the Social Determinants of Health. [Online] Available at: http://www.who.int/social_determinants/resources/mekn_report_10oct07.pdf
Kelly, M.P. and Moore, T.A. (2012). The judgement process in evidence based medicine and health technology assessment. *Social Theory and Health*, **10**, 1–19.
Kelly, M.P. and Doohan, E. (2012). The social determinants of health. In: Merson, M.H., Black, R.E., and Mills, A.J. (eds.) *Global Health: Diseases, Programs, Systems and Policies*. (3rd ed.) Burlington, MA: Jones & Bartlett, pp. 75–113.
Kelly, M.P. and Russo, F. (2018). Causal narratives in public health: the difference between mechanisms of aetiology and mechanisms of prevention in non-communicable diseases. *Sociology of Health and Illness*, **40** (1), 82–99.
Kickbusch, I., Williams, C., and Lawless, A. (2014). Making the most of open windows: establishing Health in All Policies in South Australia. *International Journal of Health Services*, **44**, 185–94.
Kriznik, N.M., Kinmonth, A.L., Ling, T., and Kelly, M.P. (2018). Moving beyond individual choice in policies to reduce health inequalities: the integration of dynamic with individual explanations, *Journal of Public Health*, 40(4), 764–75.
Leininger, L.J., Friedsam, D., Voskuil, K., and DeLeire, T. (2014). Predicting high-need cases among new Medicaid enrolees. *American Journal of Managed Care*, 20(9), e399–407.

Lorgelly, P., Lawson, K., Fenwick, E., and Briggs A. (2010). Outcome measures in economic evaluations of public health interventions: a role for the capacity approach? *International Journal of Environmental Research and Public Health*, 7(5), 2274–89.

McGuire, A., Henderson, J., and Mooney, G. (1988). *The Economics of Healthcare*. London: Routledge.

McIntyre, D., Mooney, G., and Jan, S. (2009). What is need and how do we measure it? In Detels, R., Beaglehole, R., Lansang, M.A., and Gulliford, M. *Oxford Textbook of Public Health*. Three Volumes (5th ed.). Oxford: Oxford University Press, pp. 1535–48.

Malilay, J., et al. (2014). The role of applied epidemiology methods in the disaster management cycle. *American Journal of Public Health*, 104(11), 2092–102.

Mathews, A.L., Coyle, B.S., and Deegan, M.M. (2015). Building community while complying with the Affordable Care Act in the Lehigh Valley of Pennsylvania. *Progress in Community Health Partnerships*, 9(1), 101–12.

Millican, P. (ed.) (2007). *Introduction to Hume D. An Enquiry Concerning Human Understanding*. Oxford: Oxford University Press.

Mooney, G.H. (1992). *Economics Medicine and Healthcare* (2nd ed.). Brighton: Wheatsheaf,

NICE (n.d.). *Glossary*. [Online] Available at: https://www.nice.org.uk/Glossary?letter=H#Health%20needs%20assessment

NHS England (2015). *Health and Wellbeing Health Needs Assessment Programme: Immigration Removal Centres and Residential Short Term Holding Facilities, National Summary Report*. London: NHS England.

NHS Kent and Medway (2013). *Breastfeeding Services in West Kent and Dartford Gravesham and Swanley: Health Needs Assessment, Kent and Medway Public Health Observatory*. Maidstone: NHS Kent and Medway.

Nacul, L.C., et al. (2014). A Toolkit to assess health needs for congenital disorders in low- and middle-income countries: an instrument for public health action. *Journal of Public Health*, 36(2), 243–50.

Nozick, R. (1974). *Anarchy, State and Utopia*. Oxford: Blackwell.

Pareto, V. (1935). *Mind and Society*. New York: Harcourt Brace Jovanovich.

Pennel, C.L., McLeroy, K.R., Burdine, J.N., and Matarrita-Cascante, D. (2015). Nonprofit hospitals' approach to community health needs assessment. *American Journal of Public Health*, 105(3), e103–13.

Pogge, T.W. (2003). Relational conceptions of justice: responsibilities for health outcomes. In: Anand S., Fabienne P., and Sen A. (eds.) *Health, Ethics and Equity*. Oxford: Clarendon, pp. 1–10.

Powell, J. (2006). *Health Needs Assessment: A Systematic Approach*. National Library for Health: Health Management Specialist Library. [Online] Available at: https://www.library.nhs.uk/healthmanagement/ViewResource.aspx?resID=29549

Powell, J. E. (2007). Health economics in public health. In: Orme J., Powell J., Taylor P., and Grey M. (eds.) *Public Health in the 21st Century, New Perspectives on Policy Participation and Practice*. Maidenhead: Open University Press.

Pretty, I.A. (2014). The life course, care pathways and elements of vulnerability: a picture of health needs in a vulnerable population. *Gerodontology*, 31 Suppl 1, 1–8.

Public Health England, Local Government Association (2016). *Local Wellbeing, Local Growth: Implementing Health in all Policies at Local Level: Practical Examples*. London: Public Health England.

Rawls, J. (1971). *A Theory of Justice*. Cambridge MA: Harvard University Press.

Rose, G. (1985). Sick individuals and sick populations. *International Journal of Epidemiology*, **14**, 32–8.

Rose, G. (1992). *The Strategy of Preventive Medicine*. Oxford: Oxford University Press.

Rosenbaum, S. (2013). *Principles to Consider in the Implementation of a Community Health Needs Assessment Process*. Washington DC: Department of Health Policy School of Public Health and Health Services, George Washington University. [Online] Available at: http://nnphi.org/wp-content/uploads/2015/08/PrinciplesToConsiderForTheImplementationOfACHNAProcess_GWU_20130604.pdf

Ross, J.R. (1952). *The National Health Service in Great Britain: An Historical and Descriptive Study*. London: Oxford University Press.

Sen, A. (2009). *The Idea of Justice*. London: Allen Lane.

Stevens, A. and Raftery, J. (eds.) (1997). *Health Needs Assessment: The Epidemiologically Needs Based Reviews*. Oxford: Radcliffe.

Stone, K., Sierocki, A., Shah, V., Ylitalo, K.R., and Horney, J.A. (2018). Conducting community health needs assessments in the local public health department: a comparison of random digit dialing and the Community Assessment for Public Health Emergency Response. *Journal of Public Health Management and Practice*, 24(2), 155–63.

Solar, O. and Irwin, A. (2010). *A Conceptual Framework for Action on the Social Determinants of Health. Social Determinants of Health Discussion Paper 2, Policy and Practice*. Geneva: WHO.

Subaiya, S. Moussavi, C. Velasquez, A., and Stillman, J. (2014). A rapid needs assessment of the Rockaway Peninsula in New York City after Hurricane Sandy and the relationship of socioeconomic status to recovery. *American Journal of Public Health*, 104(4), 632–8.

Taylor, A.J.P. (1955). *Bismarck: The Man and the Statesman*. London: Hamish Hamilton, pp. 202–8.

Thomson D. (1957). *Europe Since Napoleon*. London: Longman, pp. 356–9.

Van Gelderen, S., Krumwiede, K.A., Krumwiede, N.K., and Fenske, C. (2018). Trialing the community-based collaborative action research framework: supporting rural health through a community health needs assessment. *Health Promotion Practice*, 19(5), 673–83.

Venkatapuram, S. (2011). *Health Justice: An Argument from the Capabilities Approach*. Cambridge: Polity.

Venkatapuram S. and Marmot, M. (2009). Epidemiology and social justice in the light of social determinants of health research. *Bioethics*, **23**, 79–89.

Vyas A.N., Chaudhary, N., Ramiah, K., and Landry, M. (2013). Addressing a growing community's health needs: project SAHNA (South Asian Health Needs Assessment). *Journal of Immigrant & Minority Health*, 15(3), 577–83.

Whitehead, M. (1992). Perspectives in health inequity. *International Journal of Health Services*, **22**, 429–45.

Whitehead, M. and Dahlgren, G. (2006). *Levelling Up. Part 1. A Discussion Paper on Concepts and Principles for Tackling Inequities in Health*. Copenhagen: WHO.

Williams, B. (1962). The idea of equality. In: Laslett. P. and Runciman, W.G (eds.) *Philosophy, Politics and Society.* Second Series. Oxford: Blackwell, pp. 110–31.

Wolff, J. (2007). Equality: the recent history of an idea. *Journal of Moral Philosophy*, **4**, 126.

World Health Organization (WHO) (1947). *Constitution of the World Health Organization.* Geneva: WHO.

World Health Organization (WHO) (1948). *Official Records of the World Health Organization*. Geneva: WHO.

World Health Organization (WHO) (1958). *The First Ten Years: The World Health Organization*. Geneva: WHO.

11.2

The political economy of non-communicable diseases

Lessons for prevention

Anne Marie Thow, Raphael Lencucha, and K. Srinath Reddy

Introduction to non-communicable diseases

Non-communicable diseases (NCDs) are the major cause of death and disability globally, and the prevalence of these diseases continues to rise, particularly in low and middle-income countries (LMIC) (WHO 2013). Annual deaths from NCDs increased by 14.1% between 2005 and 2015 (Wang et al. 2016). Cardiovascular disease—mainly ischaemic heart disease and stroke—is the largest contributor to global mortality, accounting for 17.9 million deaths in 2015 (Sacco et al. 2016). The number of adults with diabetes, worldwide, has quadrupled since 1980 (NCD Risk Factor Collaboration 2016), and the number of deaths from cancer increased by 17% between 2005 and 2015, with tracheal, bronchus, and lung cancer the leading causes of cancer deaths (Wang et al. 2016). Obesity prevalence doubled between 1980 and 2008, from 6.4% to 12%, with half of this rise occurring between 2000 and 2008 (Stevens et al. 2012; WHO 2016). In 2014, one in three adults aged over 18 years were overweight (Stevens et al. 2012; WHO 2016).

A large proportion of NCD deaths are premature (over 40% occur before the age of 70), and 80% of these premature deaths occur in LMIC (WHO 2014b). The four major NCDs (cardiovascular disease, cancer, chronic respiratory disease, and diabetes) are responsible for over 80% of NCD deaths (WHO 2014b). Although NCDs are still often thought of as diseases of the wealthy, the burden of these diseases in LMIC has escalated rapidly over the past few decades and people with these diseases in LMIC are far more likely to die than those in high-income countries (Sommer et al.

2015). The World Health Organization's (WHO) South East Asia and Western Pacific Regions (including India and China respectively), account for approximately half the diabetes cases in the world (WHO 2016). Between 1970 and 2008, the proportion of newly reported cancers occurring in LMIC increased from 15% to 56%; by 2030, the proportion is expected to be 70% (Wagner and Brath 2012). By 2020, it is estimated that chronic respiratory disease will be the fourth leading cause of lost 'disability-adjusted life years' in LMIC (Checkley et al. 2014).

These diseases have a devastating impact at the personal and household level; in addition to causing death, chronic diseases have a high burden of morbidity. For example, the complications of diabetes include blindness, kidney failure, and lower limb amputation—all of which impact significantly on quality of life as well as participation in employment (WHO 2016). As a result of their associated premature mortality and morbidity, NCDs pose a substantial household and national economic burden, both from the cumulative healthcare costs associated with chronic conditions and lost economic productivity. The economic losses due to cardiovascular disease, chronic respiratory disease, cancer, diabetes, and mental health between 2005 and 2025 are projected at US$ 47 trillion, representing 75% of global GDP in 2010 (US$ 63 trillion) (Bloom et al. 2011). Not only this, but the investment required to prevent NCDs is significantly less than the costs currently associated with these diseases: a 10% reduction in mortality from ischaemic heart disease and stroke would reduce economic losses in LMICs by an estimated US$25 billion per year (three times greater than the investment required) (WHO 2011). Recent estimates suggest that investment of an additional $1.50 per capita for prevention in the 20 countries with the highest NCD burden would avert 15 million deaths, 8 million incidents of ischaemic heart disease, and 13 million incidents of stroke in the 20 countries (Bertram et al. 2018).

Despite the rising tide of NCDs, it is widely accepted that these diseases are largely preventable. For example, at least one-third of the 12 million cancer cases that occur annually are preventable (Boyle and Levin 2008). In this chapter, we focus on five major modifiable risk factors for NCDs: tobacco use, alcohol use, unhealthy diets, physical inactivity, and air pollution (Lim et al. 2013; WHO 2017). In 2016, tobacco use was responsible for 7.1 million deaths, mainly from cancer and respiratory disease, with most of these deaths occurring in China, India, the United

States, and Russia (Global Burden of Disease 2017). In 2010, dietary risk factors (particularly diets low in fruits and vegetables and those high in sodium, but also including diets high in fats and sugars) were the largest contributors to the global disease burden—and particularly risk for cardiovascular disease, diabetes, and some cancers Lim et al. 2013). Alcohol use/consumption is the seventh leading risk factor for NCDs, in terms of the morbidity burden as measured by disability-adjusted life years —and globally, alcohol exposure has increased by 15.2% since 1990 (Global Burden of Disease 2017). Physical inactivity is associated with an increased risk of cancer (breast and colon), diabetes, ischaemic heart disease and stroke, and even low amounts of activity are beneficial in reducing risk (Kyu et al. 2016). In 2014, the prevalence of physical inactivity in high-income countries was almost double that of LMIC. However, with changes to work and leisure, the prevalence of physical inactivity is increasing in LMICs (WHO 2016). In 2017, air pollution was estimated to cause one-quarter of all adult deaths from heart disease and stroke, 43% from chronic obstructive pulmonary disease and nearly a third from lung cancer, and is also associated with diabetes (WHO 2016; Global Burden of Disease 2017; Bowe et al. 2018).

In this chapter, we discuss how these four major risk factors emerged in a context whereby economic interests converge with public policy to produce environments of risk. The important thread running through this chapter is that the global rise in NCDs is conditioned by modifiable political and economic patterns. This chapter concludes by presenting a number of strategies that public health practitioners can take to remedy this situation.

Policy action on risk factors is essential for NCD prevention

The World Health Organization (WHO) *Global Action Plan for the Prevention and Control of Noncommunicable Diseases 2013–2020* was agreed by member states at the World Health Assembly in 2013 (WHO 2013). The NCD global action plan sets out a clear vision and targets for addressing the heavy global burden of NCDs, including nine voluntary global targets with a strong focus on all of these modifiable risk factors:

A 25% relative reduction in the overall mortality from cardiovascular diseases, cancer, diabetes, or chronic respiratory diseases
At least 10% relative reduction in the harmful use of alcohol, as appropriate, within the national context
A 10% relative reduction in prevalence of insufficient physical activity
A 30% relative reduction in mean population intake of salt/sodium
A 30% relative reduction in prevalence of current tobacco use in persons aged 15+ years
A 25% relative reduction in the prevalence of raised blood pressure or contain the prevalence of raised blood pressure, according to national circumstances
Halt the rise in diabetes and obesity
At least 50% of eligible people receive drug therapy and counselling (including glycaemic control) to prevent heart attacks and strokes
An 80% availability of the affordable basic technologies and essential medicines, including generics, required to treat major noncommunicable diseases in both public and private facilities

The member states of the WHO have also agreed on a comprehensive package of effective policy options to support achievement of these targets by reducing the prevalence of all four major risk factors for NCDs (WHO 2013):

Policy options for tobacco control focus on implementation of the WHO Framework Convention on Tobacco Control (FCTC), which includes mandating smoke-free environments, graphic health warning labels on tobacco products, bans on advertising, promotion and sponsorship, and taxes on tobacco products.
Policy options for promoting a healthy diet include breastfeeding promotion, restrictions on the food marketing, creating incentives for reformulation of food to reduce sugar, salt, and fat content, promoting healthy food environments, social marketing, nutrition labelling, and taxes and subsidies.
Policy options for reducing the harmful use of alcohol include strengthening health services' response and community action, addressing drink–driving and availability of alcohol, restrictions on marketing, and taxation. A critical component of global recommendations for all four risk factors is protection of policy making from influence by vested interests.
Policy options identified for promoting physical activity include promotion of 'active transport', recreational activity and sport, including through urban planning and transport policies, creation of healthy built environments which support physical activity, and public campaigns to promote activity.

Growing recognition of the link between environmental risk factors—and, in particular, air pollution—has resulted in emerging global consensus regarding the need for policy action in this sphere for effective NCD prevention (Lelieveld et al. 2015; WHO 2017). This will require policy change across health, transport, housing, and energy sectors to reduce air pollution from particulate matter. Innovation by countries has included support for clean household energy, cleaner vehicle standards and support for public transport (WHO 2018).

Achieving the Sustainable Development Goal targets related to NCDs and the global NCD targets will require structural, policy-led change that addresses environments in which people live, work, and play, with attention to contextual and regional differences (Reddy 2015). These environments have a significant impact on peoples' capacity to engage in healthy behaviours, and can either support and incentivize the healthier choices, or the opposite. For example, redesigning streets to form squares and one-way streets, implementing traffic calming measures and creating bicycle lanes, creating or renovating playgrounds, improving lighting, and enhancing

aesthetics are associated with increased physical activity, as measured by an increase in the percentage of people engaging in active transport and other measures of activity, and will also have benefits in reducing traffic-related air pollution (Heath et al. 2006). Similarly, there is a growing body of evidence that indicates that food environments have a major influence on food choice and consumption, such that shifts in access, pricing, portions, marketing, and other powerful drivers of eating can have major effects on weight (Wansink and Sobal 2007; Brownell et al. 2010).

However, there has not yet been widespread system policy implementation to create supportive environments for health promotion and NCD prevention. This lack of action stems from policy tensions that arise when government attempts to control health-harming products come up against powerful economic interests. Companies that produce alcohol are notorious for resisting any type of regulation. This resistance often utilizes arguments and corresponding tactics that locate the cause of NCDs at the individual, rather than environmental level (Roberto et al. 2015). These tensions are also seen with industries with interests contrary to reducing car use, in relation to interventions aimed to reduce physical inactivity and pollution. As we will discuss, this emphasis on individual responsibility has important implications for how governments and the public health community addresses NCD prevention. The message of personal responsibility often translates into public health approaches such as health communication campaigns and voluntary standards to guide industry practices. Alone, these approaches are largely ineffective when environments continue to make the healthy choice the more difficult choice (Golden et al. 2015).

The structural, environment-focused policies required to prevent NCDs —including labelling, taxation, restrictions on marketing, and reform of transport and planning policy—require change in finance, commerce, trade, planning, and communications policies. Achieving health oriented policy change in these sectors, especially those that govern unhealthy commodities and physical environments, has proved challenging because of the overt or covert influence of industry over government decision-making or more nuanced misalignments between the mandates of government sectors and health objectives (Miller and Harkins 2010; Lencucha et al. 2015; Schram 2017). Governments that do implement strong environmentally focused NCD prevention policy consistently face challenges in national and

international economic forums; for example, tobacco, alcohol and nutrition labelling measures have all been challenged at the World Trade Organization (WTO) (Lencucha and Drope 2015; Global Burden of Disease 2017; Mitchell and Casben 2017; Thow et al. 2017). It is for this reason, that a political economy perspective is indispensable in understanding and acting on product and environmental risk.

Global political economy and environments of risk

A political economy perspective on health is one that links individual health risk (e.g. choice or biology) with political and economic structures and practices. This perspective asks: 'How do economic interests intersect with political decisions to shape the environments in which individuals live?'. In its simplest form, this perspective brings together and analyses political and economic structures and how these structures shape the conditions for health. For example, as we discuss in more depth, political decisions driven by pressures to free the commercial sector from regulation have led to product environments where healthy choices are difficult.

Political structures refer to public policy, what a government decides to do or not do, and the institutions, namely rules, norms, and strategies established by government, that form and enact policy (Dye 1972; Ostrom 1986). Importantly, observers note that the contemporary political economy is also marked by a deliberate retreat of the state from public life, whereby non-state actors, whether powerful individuals, companies, or other groups, play an important role in shaping the economic patterns of society (Strange 1996; Zohlnhöfer et al. 2018). By considering the political economy of NCDs prevention one can examine how structural factors can impact the incidence and prevalence of disease. This perspective serves as a backdrop to understand why health-harming products continue to dominate the consumer environment despite their known risk for NCDs. The same perspective can elucidate how political decisions and economic interests shape how we move about the built environment (e.g. by foot, by bike, or by car) and how this movement is conditioned by the existence of road systems, forms of public transportation, and access to safe alternative transportation (e.g. curbed bike lanes on busy streets).

When ‘environments of risk’ are considered from a political economy perspective, it enables public health actors to critically assess sociopolitical factors that generate product environments characterized by health-harming products, built environments that condition physical inactivity over physical activity, and marginalization of pollution reduction. This type of analysis positions these five modifiable risk factors, not as ‘behavioural’ risk factors, but in a context wherein individual choice is bounded by political economic structures that often make the healthy choice the difficult choice.

To begin, we examine these global political economic patterns and the underlying (or at least purported) rationality informing these patterns. Neoliberalism has emerged as a dominant way of thinking about the relationship between governments, markets, and society. Benatar and colleagues note that ‘neoliberalism has triumphed as the prevailing and dominant paradigm of global capitalism’ (Benatar et al. 2018). Early proponents of neoliberalism were driven by the goal of economic freedom, both in terms of individual freedom to make economic choices from a wide range of options, and the ability of companies to provide a wide range of appealing options available to consumers (e.g. Friedman 2002). As Gill notes ‘these (neoliberal) policies are cast within a neoliberal discourse of governance that stresses the efficiency, welfare, and freedom of the market, and self-actualization through the process of consumption’ (Gill 1995, p. 401). In this view, economic freedom has intrinsic value, as an expression of the human aspiration and capacity to express one’s agency, and instrumental value, a means of upholding political and economic freedom (Friedman 2009). This neoliberal vision has been largely, if not uniformly, institutionalized since the 1980s (Labonté and Stuckler 2016). David Harvey notes that neoliberalism is ‘in the first instance a theory of political economic practices that proposes that human well-being can best be advanced by liberating individual entrepreneurial freedoms and skills within an institutional framework characterized by strong private property rights, free markets, and free trade. The role of the state is to create and preserve an institutional framework appropriate to such practices’ (Harvey 2007).

As indicated, a key tenet of neoliberalism is the separation of market from state control, with the idea that consumer decisions, not government control, will correct market failures. The ongoing result of the neoliberal paradigm is an acute tension between those who believe that government

should take an active role in protection and promotion of health by intervening in the market, and those who seek a clearer separation between government and market whereby responsibility is left to individuals to make healthy choices (or not) and thus shape the market through these decisions. In sum, the dominant neoliberal paradigm has created a situation where governments struggle, or neglect, to regulate in the public interest because of a dominant paradigm that argues for the virtue of separating market and state, and a legacy of pressure from powerful institutions to conform to a set of policies that drives a separation of market and state. Opponents of the current neoliberal orthodoxy argue that 'We need to work harder to reclaim the narrative: we do not have a fiscal crisis. We have a crisis of inadequate taxation. We are not living in conditions of economic scarcity. We are living in conditions of extreme inequality' (Labonté and Stuckler 2016, p. 316). This argument is tied to the widespread recognition that although neoliberalism is thought of as a theory and practice involving freeing markets from state control, in reality neoliberalism is largely perpetuated by state decisions and often brings seemingly paradoxical state interventions to protect market activity (e.g. government bailouts to large companies) (Brenner et al. 2010).

In this context, the intent of public health regulation to reduce consumption of commodities associated with NCD risk gives rise to a very real tension for governments, with a perceived choice between profits and public health interests. The companies that benefit from the continuation of behaviours associated with NCD risk are all major contributors to GDP and employment in many countries. In addition to this, risky commodities are often more profitable than healthier alternatives. For example, highly processed foods generally have higher profit margins due to their long shelf life when compared to fresh, unprocessed foods, and there is minimal profit to be had from active transport. However, it is important to note that profitability does not directly translate to economic benefits for government or society, and often leads to wealth concentration and heightened economic inequality (Goda and Lysandrou 2013; Saez and Zucman 2016). An example of this tension is seen in the public discourse regarding taxation of tobacco, alcohol, pollution, and unhealthy foods and beverages. If finance ministers see this as a revenue measure with diminishing returns (as consumption/pollution falls), neoliberal opponents can sway them with the 'freedom from nanny state' and 'regressive impact on poor' arguments. If

taxation is projected as a public health imperative, with a particular focus on the socioeconomic aspect of the NCD burden, broader public support can be built up (Sassi et al. 2018; Thow et al. 2018a).

Tobacco, food, and alcohol companies, as well as companies with a vested interest in individually-oriented transport, are strongly transnationally integrated, both vertically and horizontally, and are valued in the billions. The world's top ten food companies, which produce primarily highly processed foods, own almost all the familiar international brands and in 2017 had combined sales of over 400 billion dollars (Top 100 food and beverage companies n.d.). This is more than the GDP of Thailand, and equivalent to half of Indonesia's GDP. In 2016, seven of the top 10 biggest companies in the world were petroleum and car companies, including China National Petroleum (#3), Royal Dutch Shell (#5), Exxon Mobil (#6) and Volkswagen (#7) (Fortune Global 500 List 2017).

Despite a global decline in tobacco consumption, transnational tobacco companies continue to generate increasing profits. For example, Philip Morris International saw a rise in profits from roughly 7 billion Canadian dollars (CAD) to over 9 billion CAD from 2006 to 2012. The same increases can be seen with British American Tobacco and Japan Tobacco (Physicians for a Smoke Free Canada n.d.). The total profits for the big six tobacco companies rose from 35 billion USD in 2012 to over 60 billion USD in 2015 (Bowers 2012). It is this massive wealth that drives the tobacco industry to employ every means to fight tobacco control efforts.

This perception of profit doesn't take into account the negative externalities that arise from consumption of tobacco, alcohol and unhealthy foods, or that are associated with living in built environments that disincentivize physical activity or generate air pollution—namely poor health, social costs, and lost productivity—but it is nevertheless an influential perception (Kankeu Tchewonpi et al. 2013; Lencucha et al. 2018b). The economic paradigm just described creates a framework for decision-making in which costs that are harder to quantify are less likely to be considered in decision-making; for example, costs in the form of quality of life, social capital, or rising inequality (Stiglitz 2017). The vertical integration of these industries can also result in significant inequality in the way that profits accrue, geographically. The profits from value adding often accrue to companies based in high-income countries, with a strong extractive element evident in the GDP and employment gains in LMICs.

For example, although tobacco is a significant contributor to GDP for many countries in Africa, such as Malawi, the economic gains to tobacco farmers is dwarfed by the huge profits made by transnational tobacco companies (Makoka et al. 2017).

Industry power and the political economy of non-communicable disease risk

The enormous size and reach of these 'industries of risk' translates into significant political power. In a global political economy largely driven by the ideas and corresponding policies of neoliberalism, industry is positioned as creator of wealth and the voice of industry is often given preference. Thus, when the health sector talks about policy change in economic sectors that will reduce consumption of profitable commodities, their voice is often drowned out by others with a vested interest in these commodities. In this section, we explore three key ways in which industry exercises power in politics and policy making, to the detriment of governments efforts to reduce consumption of harmful commodities and create built environments that foster activity. Drawing on Lukes' conception of power, we consider ways in which industry actively exerts power in formal decision-making forums ('decision-making power'), shapes the agenda in political debates ('non-decision-making power'), and exercises ideological power (Lukes 2005).

Decision-making power focuses on policy preferences revealed through political action.

Industry actors invest heavily in shaping government policy through establishing relationships, and direct lobbying efforts. The power of these relationships is widely recognized in the area of tobacco control. The tobacco industry is notorious for employing different tactics to aggressively influence policy; namely to prevent the regulation of their products (Saloojee and Dagli 2000). In the United States, industry actors are formally consulted on the negotiating text of trade agreements by the Office of the US Trade Representative, while public health actors are not. It is common for governments to consult with industry on economic policy decisions and this practice often conflicts with the aims of health sectors to protect health policy from industry influence. For example, the Philippines Inter-Agency Committee on Tobacco was established to coordinate tobacco control policy

within government but is chaired by a representative from the trade ministry who views the tobacco industry as a legitimate stakeholder and thus the industry has a reserved seat at the agency (Lencucha et al. 2015).

The influence of relationship building and policy lobbying has also been used effectively by the alcohol industry (McCambridge et al. 2018). For example, the alcohol industry in sub-Saharan Africa—taking the mantle of 'technical expert'—has supported governments in the drafting of their national alcohol policies (Bakke and Endal 2010). Evidence from Lesotho, Malawi, Uganda, and Botswana shows that the resulting policies minimized a public health approach to alcohol and maintained a narrow focus on the economic benefits from trade and alcohol. In terms of addressing alcohol problems, the policy content focused on individual drinkers, with no mention of effective environmental interventions (Bakke and Endal 2010). Similarly, in Thailand, there has been little action on environment-focused alcohol policy—in part because alcohol company executives sit in the national policymaking body (Treerutkuarkul 2017).

Industry also exerts direct power in decision-making through involvement in the negotiation of trade and investment agreements, and through using dispute settlement mechanisms in these agreements to influence policy. For example, the Government of Thailand proposed graphic warning labels on alcohol that highlighted the associated harms, similar to approaches used in tobacco control (Cecchini and Belloni 2015). This policy proposal was withdrawn in 2010 after Specific Trade Concerns were raised at the World Trade Organization (O'Brien 2013). Similarly, despite neoliberal commitment to free trade arguably increasing access to medicines and health technologies in developing country markets, commercial interests prevent affordable access to these products through a variety of trade provisions (United Nations 2016). In particular, strict patent protection that benefits international pharmaceutical companies mean that affordability of NCD treatments remains a challenge in LMICs.

Investor-State Dispute Settlement mechanisms are included in many international trade and investment agreements and enable foreign investors to directly contest government policy (this is in contrast to traditional 'state-state' mechanisms of dispute settlement, where only governments can bring challenges regarding national policies that inhibit trade or investment). These mechanisms create avenues for industry to contest regulation that undermines 'reasonable expectations' of profits—which resonates with the

tension between risky products and public health regulation described earlier (Hawkins and Holden 2016). For example, the tobacco multinational Philip Morris brought a claim against Australia under an International Investment Agreement (IIA) between that country and Hong Kong Special Administrative Region in China, in 2011. This claim challenged Australia's mandatory 'plain packaging' of tobacco products as an unreasonable measure that would expropriate trademarks and undermine good will indirectly (Allens Arthur Robinson 2011). Philip Morris International filed a similar claim in 2010 against Uruguay's decision to expand graphic health warning labels on tobacco products (Lencucha 2010; Porterfield and Byrnes 2011).

Non-decision-making power is that which sets the agenda in debates and makes certain issues unacceptable for discussion in 'legitimate' public forums. A key element of exercising non-decision-making power is framing discourse and public debate regarding an issue. The tobacco industry has consistently attempted to shape the discourse on government regulation by drawing on the powerful value of personal freedom. Early efforts to dissuade governments from regulating tobacco products, were supported by the message that individuals should have the freedom to choose the products that they want to consume. That the government should not intervene with the personal decisions of its citizens. The following statement by the Institute of Economic Alternatives, an ardent supporter of the tobacco industry, on tobacco product regulation reflects the logic used to oppose public regulation: 'We argue that the interests of consumers are nearly always better advanced by the provision of accurate information and free choice than by prohibitions and restrictions on commercial speech. Markets can correct themselves more quickly than governments … Restrictions on product development and marketing in the nicotine industry have led to … failures of both government and the market. Buttressed by political grandstanding, corporate rent-seeking, rigid ideology, and old-fashioned puritanism, the result has been a disaster for the only group of people whose interests are of direct relevance: smokers' (Snowdon 2013, p. 41). These frames, however, don't acknowledge the disproportionately high experience of NCD risk factors, and concomitant NCD morbidity and mortality, among people of low socioeconomic status. The narrow definitions and frames of 'benefit' ignore the significant impact of NCDs on economic productivity and on inequalities (Nugent et al. 2018).

Another of the powerful messages used by industry in opposition to government regulation of their products is that such regulation will create economic ripples that will harm employment throughout the supply chain. In LMICs this message is tied to economic development. Industry argues that regulation will result in loss of employment, with particular emphasis to the most vulnerable, the farmers. The tobacco industry has long argued that tobacco regulation will hurt tobacco farmers and have convinced government officials of this point. We see this argument being made by industry representatives in the media (Karombo 2018; Khan 2018) and by government officials in high-level policy venues such as the World Trade Organization (Lencucha et al. 2016). These arguments harken back to the ideas of economic freedom and growth that underpin the neoliberal paradigm.

However, recent research has questioned the evidence base for the claims that farmers are prosperous from crops such as tobacco. A study conducted in Malawi in 2014 found that smallholder tobacco farmers who were on contract with tobacco leaf buying companies, after accounting for family labour costs, were making roughly 220 USD per year, less than one dollar per day (Makoka et al. 2017). In the same study independent tobacco farmers were finding themselves in debt at the end of the season. This is in a context where the CEO of British American Tobacco tops 20 million USD in salary and bonuses in one year (Treanor 2012). Similar arguments are evident in the food industry's opposition to taxes on sugar-sweetened beverages, which are linked to increased risk of diabetes and obesity, on the basis of their potentially regressive fiscal impact on poor households (Thow et al. 2018a). For example, in the Philippines, the industry argued consistently that the proposed sugar-sweetened beverage tax should be reduced because it was anti-poor, and would be 'crucial to the fate of millions of low-income Filipino consumers' (PhilStar.com 2017). This argument has been found to ignore the disproportional health benefits that accrue to low-income consumers, as their higher price sensitivity leads them to decrease their consumption more than their higher income counterparts (Thow et al. 2018a).

Strong frames that create tension between profits and public health regulation are also evident in public industry submissions to government. For example, food industry submissions regarding the proposed TransPacific Partnership Agreement assumed that trade liberalization would

raise living standards and benefit the economy and country; food was framed exclusively as an economic commodity with implications for revenue generation, with no connection to nutritional health (Friel et al. 2016).

Ideological power shapes thoughts and behaviour by invoking strong values, emotions, and ethical principles. Marketing and other strategies that seek to create brand loyalty and influence habit formation in children can be a source of ideological power. Powell has argued that the food industry has used the school system to integrate and normalize messages of personal responsibility over diet and the idea of energy balance (i.e. emphasizing physical activity as a remedy for unhealthy diet). He shows how Coca-Cola systematically worked to shape the discourse towards healthy choice and consumer responsibility by implementing over 250 'anti-obesity' programmes in schools. Powell and Gard interrogate a number of programmes that Coca-Cola designed to stimulate physical activity among the children. Underlying these types of programmes is a message that there is no bad food but bad choices. For example, messages that consumption of a 'high energy' product like Coca-Cola, must be followed by the decision to exercise in order to balance the energy in with energy out. Food industry interests have pushed these types of messages at all levels. Brownell and Warner provide the example of how the Sugar Association in the United States enlisted the support of US government officials to oppose the WHO draft strategy to address issues of diet and physical activity. The Association pushed the message that there should be greater emphasis on personal responsibility, a focus on physical activity, and the oft made assertion that there are no good and bad foods just bad choice (Brownell and Warner 2009).

Ewing and Cervero (2010) conducted a meta-analysis of over 200 studies that examined the relationship between transportation and the built environment. One prominent finding is that Vehicle Miles Travelled (VMTs) was positively associated with the accessibility of the destination, such as work. They also found that mixed-use settings, such as retail outlets, schools, grocery stores, had a negative relationship with VMTs. For example, they also found that having a bus stop nearby was positively associated with public transportation usage. These factors are examples of how the way that the built environment is designed shapes modes of transportation. If we take one step back, we can think about how policy can

shape these environments and one step further to see how certain economic interests can shape this policy. For example, the rise of car culture in the United States was partly driven by ideas of personal freedom tied to convenient mobility (Seiler 2009). The rise of affordable personal use vehicles was supported by public policy that lead to car-friendly infrastructure developments in cities, such as roads and parking lots, and policies that facilitated access to cheaper fuel sources, through subsidized fuel extraction, refinement, and transport to urban centres.

How public health can support strong policy for non-communicable disease prevention

Despite the power wielded by industry interests in the policy process, there has been innovation and meaningful policy change for NCD prevention. In this section, we examine cases of public health success—where governments have successfully regulated in the face of strong economic interests. In many cases, this innovation is led by LMIC; in part due to the devastating burden of NCD morbidity and mortality in these countries. These case studies provide the opportunity to draw lessons from experience and consider how public health actors can protect policy space and support regulation to promote healthy product and built environments.

Strong, local evidence

Assembling strong, local evidence to underpin policy is essential to convince policy makers of the need to take action, and of the solutions needed—even though alone it is insufficient for policy change. The power of evidence of effectiveness is clear from the decision by the Government of Peru to scale up strong environmentally focused regulation regarding the hours alcohol can be sold. Data on the impact of a local ban in the district municipality of La Victoria in Lima on the sale of alcohol (Sundays to Wednesdays up to midnight and Thursdays to Sundays up to 03:00) over 2 years showed that, compared to a district without a ban, violent incidents had halved. Subsequently, a similar ban was established in the entire Metropolitan area of Lima, and the national police began an intensive action campaign for zero tolerance to driving under the influence of

alcohol. A year later, the number of violent deaths had decreased by 11%, homicides by 19%, and the number of victims of alcohol-attributable traffic crashes by 28% (WHO 2014a).

It is not just quantitative evidence that it required to underpin policy change. A turning point in tobacco control legislation was brought about by evidence that the tobacco industry had acted unethically in promoting their products, undermining the assumption by policy makers that consumers had full access to information about the harms and benefits of the product. Availability of adequate, accurate information regarding products is an important assumption in a market-based economy—and asymmetry of information is a recognized aspect of market failure. For example, a free choice may involve knowing the risk and benefits of consuming a medicine whereby the consumer can decide if the benefits outweigh the risks. In the United States in the mid-1990s, a series of lawsuits were filed against the tobacco industry in an attempt to retrieve damages caused by tobacco to consumers and the pressing burden of tobacco-related illness on the health system (Schroeder 2004). One of the important outcomes of these suits, and particularly the Master Settlement Agreement, was the requirement that the tobacco companies release internal documents for public scrutiny. At the time, the CEOs of the major tobacco companies were testifying that tobacco was not addictive and posed no, or at least minimal, harm to users. Contrary to this testimony the internal documents revealed that the tobacco industry was well aware of the harms of tobacco use and more importantly, despite this knowledge, had been aggressively and deceptively marketing their products to consumers. This is a history of the most sinister cases of corporate deception where ‘The tobacco industry knowingly and successfully deceived the public for 50 years about the connection between smoking and cancer’ (Leahy 2014). The example of tobacco industry deception illustrates how economic interests can shape consumer decisions by controlling the information available to consumers.

As discussed earlier, the frames and norms surrounding a product have dramatic implications on the extent to which the companies producing the product have access to the policy making arena. For example, there is now no question as to the risk posed by tobacco products, to the extent that Philip Morris, one of the largest tobacco companies, acknowledges the harms of tobacco consumption on their website. The National Cancer Institute in the United States asserts that ‘there is no safe tobacco product,

and there is no safe level of exposure to tobacco smoke' (National Cancer Institute n.d.). This categorical position informs decisions to exclude tobacco interests from accessing arenas where health policy, and now even other social policy, is being made. For example, the World Bank instituted a policy in 1991 that prohibits the organization from any involvement with tobacco (World Bank n.d.). Article 5.3 of the FCTC implores governments to protect health policy from tobacco industry interference (WHO 2003), given the aforementioned efforts of the tobacco industry to shape policy in their favour (Bialous and Peeters 2012; Gilmore et al. 2015). In contrast, we still see companies like Pepsi Co partnering with health groups in the fight against NCDs (Yach et al. 2010). Evidence regarding both the scale of the problem and the influence of industry in policy making will help to change norms and frames regarding other NCD risk factors.

Coalitions

Theories of the policy process highlight the influence of strong coalitions of actors with shared interests, in shaping policy outcomes (Sabatier 1987; Weible et al. 2012; Shiffman 2016). Key features of strong actor coalitions in a health context include their ability to construct compelling frames of the issue and convincing reasons to act. It is also critical that they engage with the politics of the issue, and develop relationships outside of the health sector (Shiffman 2016). An example in relation to policy action on NCDs is CLAS (the Healthy Latin America Coalition) (NCD Alliance 2017). CLAS is a coalition with 250 member organizations who has developed a rapid-response system for nutrition policy advocacy, based on consensus position statements. When there is a nutrition policy crisis in a Latin American country then the Coalition can coordinate a strong public health voice to protect good policies—they don't have to gain consensus every time, as they can speak on the basis of the agreed statements. For example, when the Mexican health commission wanted to lower the rate of the sugar-sweetened beverage tax, the coalition organized a coordinated lobby to retain the tax.

Strong coalitions across sectors have also been effective in supporting urban planning policies that promote and support physical activity. Cities that encourage physical activity through active transport are now seen as a priority from both sustainability and health perspectives (Giles-Corti et al.

2015). In Australia, draft urban planning legislation incorporated health considerations because of concerted lobbying by health actors (Harris et al. 2017). Health was explicitly identified as a relevant issue for planning policy in submissions by health-focused agencies, who also connected health to wider planning concerns including economic issues, transport, public open space, and environmental sustainability.

Prepare for push-back from industry

As described in detail earlier, effective policies to prevent NCDs will be contested by the industries that stand to lose with decreased consumption or use of their products. Pro-active consideration of these arguments and development of thoughtful responses is a critical component of protecting policy space for NCD prevention. Industry is predictable in their focus on the economic and profit implications of policies to prevent NCDs, and strategic counterarguments to industry lobbying can help to minimize their impact. For example, when South Africa proposed their soft drink tax, industry argued strongly against it on the basis of job losses: they stated that 3,400 jobs would be lost in the sugar-sweetened drinks production; 25,200 jobs would be lost in 'upstream' jobs; 19,200–29,000 jobs in retail and 15,400 job losses directly and indirectly as a result of reduced demand for drinks (Wilkinson and Bhardwaj 2017). Data from an independent think tank showed that these estimates were vastly overestimated, as the industry analysis assumed that producers would not shift to producing healthier, untaxed drinks (which would result in more jobs) and did not consider that with the funds raised from the tax, treasury spending would also result in some jobs being created. This balanced data helped to keep the policymakers focused on the health concerns, and the tax was passed (Cullinan 2017). Economic interests have the ability to wield massive resources to move such arguments into the public discourse. In the case of tobacco, health proponents have recognized that it is crucial to intercept such messages by instituting rules within government that either prevent industry access to government decision-makers or ensure that such access is transparent and the information about any such consultations is open to public scrutiny.

Healthier economic policy

Finally, it is essential for public health actors to identify specific opportunities for economic policy change that creates healthy environments for NCD prevention. Using tools such as Health Impact Assessment can be vital in identifying specific points at which economic policy can negatively impact on health and NCDs (Hirono et al. 2016). At a practical level, it is important that health policy actors strategically frame health policy objectives in ways that ensures that they are able to defend necessity and non-discrimination as key principles for trade compliance. For example, analysis of concerns regarding nutrition labelling for NCD prevention at the World Trade Organization revealed a lack of clarity regarding the necessity of the measures, as well as practical concerns, such as the use of stickers to label imports (Thow et al. 2017). However, it is also important to go beyond simply health focused policy measures to consider broader, whole-of-government objectives that determine how governments balance competing objectives. It is essential for governments to clearly identify national interests in relation to NCDs and, more broadly, public health, to set the context for economic policymaking—including (and in some ways, especially) the negotiation of international trade and investment agreements. For example, policy incoherence with respect to nutrition in South Africa has its roots in whole-of-government priorities that preference economic indicators over nutrition (Thow et al. 2018b). It is also important to recognize that government has numerous tools that can be used to foster healthy product environments. The regulatory role of government can be complimented by targeted interventions to ensure that healthy products, such as fruits and vegetables are accessible and affordable. Governments can leverage economic resources to subsidize or incentivize healthy food production, including subsidies to small retail outlets to carry products like fruits and vegetables (Story et al. 2008; Lencucha et al. 2018a).

Despite the dominant neoliberal economic paradigm that often preferences industry interests over public health, significant changes to economic policy can and have been made in the interests of public health. After the Phillip Morris challenge to plain packaging, the Government of Australia made a decision not to agree to investor-state dispute settlement (Voon and Mitchell 2016). Similarly, the Governments of Brazil, South Africa, and India have terminated a number of their bilateral investment

agreements, due to the constraints they place on domestic policy space (UNCTAD 2016). Strategic engagement by public health actors with the sectors that govern NCD risk—as well as recognition of the political economy of NCD risk—can support meaningful policy change for environments that promote health.

REFERENCES

Allens Arthur Robinson (2011). *Notice of Arbitration, Australia/Hong Kong Agreement for the Promotion and Protection of Investments, Philip Morris (Asia) Limited, 21 November 2011*. Australian Government.

Bakke, Ø. and Endal, D. (2010). Vested Interests in addiction research and policy: alcohol policies out of context: drinks industry supplanting government role in alcohol policies in sub-Saharan Africa. *Addiction*, 105(1), 22–8.

Benatar, S., Upshur, R., and Gill, S. (2018). Understanding the relationship between ethics, neoliberalism and power as a step towards improving the health of people and our planet. *The Anthropocene Review*, 5(2), 155–76.

Bertram, M.Y., et al. (2018). Investing in non-communicable diseases: an estimation of the return on investment for prevention and treatment services. *The Lancet*, 391(10134), 2071–8.

Bialous, S.A. and Peeters, S. (2012). A brief overview of the tobacco industry in the last 20 years. *Tobacco Control*, 21(2), 92–4.

Bloom, D.E., et al. (2011). *The Global Economic Burden of Non-Communicable Diseases*. World Economic Forum. [Online] Available at : http://www3.weforum.org/docs/WEF_Harvard_HE_GlobalEconomicBurdenNonCommunicableDiseases_2011.pdf

Bowe, B., Xie, Y., Li, T., Yan, Y., Xian, H., and Al-Aly, Z. (2018). The 2016 global and national burden of diabetes mellitus attributable to PM2·5 air pollution. *The Lancet Planetary Health*, 2(7), e301–e312.

Bowers, S. (2012). Global profits for tobacco trade total $35bn as smoking deaths top 6 million. *The Guardian*. [Online] Available at: https://www.theguardian.com/business/2012/mar/22/tobacco-profits-deaths-6-million#:~:text=8%20years%20old-,Global%20profits%20for%20tobacco%20trade%20total%20%2435,smoking%20deaths%20top%206%20million&text=Revenues%20from%20global%20tobacco%20sales,more%20than%20%241%2C100%20a%20second

Boyle, P. and Levin, B. (2008). *World Cancer Report*. Lyon: International Agency for Research on Cancer.

Brenner, N., Peck, J., and Theodore, N. (2010). After neoliberalization? *Globalizations*, 7(3), 327–45.

Brownell, K., et al. (2010). Personal responsibility and obesity: a constructive approach to a controversial issue. *Health Affairs*, 29(3), 379–87.

Brownell K, Warner K (2009). The perils of ignoring history: big tobacco played dirty and millions died. how similar is big food? *The Milbank Quarterly*, 87(1), 259–94.

Cecchini, M. and Belloni, A. (2015). *Tackling Alcohol-Related Harms: What Policy Approaches?* Paris: OECD Publishing.

Checkley, W., et al. (2014). *Management of NCD in Low- and Middle-Income Countries*. *Global Heart*, 9(4), 431–43.

Cullinan, K. (2017). Parliament finally passes sugary drinks tax. *South Africa: health24*. [Online] Available at: https://www.health24.com/Diet-and-nutrition/Beverages/parliament-finally-passes-sugary-drinks-tax-20171205

Dye, T.R. (1972). Policy analysis and political science: some problems at the interface. *Policy Studies Journal*, 1(2), 103–7.

Ewing, R. and Cervero, R. (2010). Travel and the built environment. *Journal of the American Planning Association*, 76(3), 265–94.

Fortune Global 500 List (2017). *See Who Made It.* [Online] Available at: http://fortune.com/global500/

Friedman, M. (2002). *Capitalism and Freedom* (40th ed.). Chicago, IL: University of Chicago Press.

Friedman, M. (2009). *Capitalism and Freedom*. Chicago, IL: University of Chicago Press.

Friel, S., et al. (2016). Shaping the discourse: what has the food industry been lobbying for in the Trans Pacific Partnership trade agreement and what are the implications for dietary health? *Critical Public Health*, 26(5), 518–29.

Giles-Corti, B., Sallis, J.F., Sugiyama, T., Frank, L.D., Lowe, M., and Owen, N. (2015). Translating active living research into policy and practice: one important pathway to chronic disease prevention. *Journal of Public Health Policy*, 36(2), 231–43.

Gill, S. (1995). Globalisation, market civilisation, and disciplinary neoliberalism. *Millennium*, 1(24), 399–423.

Gilmore, A.B., Fooks, G., Drope, J., Bialous, S.A., and Jackson, R.R. (2015). Exposing and addressing tobacco industry conduct in low-income and middle-income countries. *The Lancet*, 385(9972), 1029–43.

Global Burden of Disease (2017). Global, regional, and national comparative risk assessment of 84 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *The Lancet*, 390(10100), 1345–422.

Goda, T. and Lysandrou, P. (2013). The contribution of wealth concentration to the subprime crisis: a quantitative estimation. *Cambridge Journal of Economics*, 38(2), 301–27.

Golden, S.D., McLeroy, K.R., Green, L.W., Earp, J.A.L., and Lieberman, L.D. (2015). Upending the social ecological model to guide health promotion efforts toward policy and environmental change. *Health Education & Behavior*, 42(1_suppl), 8S–14S.

Harris, P., et al. (2017). Creating 'healthy built environment' legislation in Australia; a policy analysis. *Health Promotion International*, 33(6), 1090–100.

Harvey, D. (2007). *A Brief History of Neoliberalism*. Oxford: Oxford University Press.

Hawkins, B. and Holden, C. (2016). A corporate veto on health policy? Global Constitutionalism and Investor-State Dispute Settlement. *Journal of Health Politics, Policy and Law*, 41(5), 969–95.

Heath, G., Brownson, R., Kruger, J., Miles, R., Powell, K., and Ramsey, L. (2006). The effectiveness of urban design and land use and transport policies and practices to increase physical activity: a systematic review. *Journal of Physical Activity and Health*, 3(s1), S55–76.

Hirono, K., Haigh, F., Gleeson, D., Harris, P., Thow, A.M., and Friel, S. (2016). Is health impact assessment useful in the context of trade negotiations? A case study of the Trans Pacific Partnership Agreement. *British Medical Journal Open*, 6(4), e010339.

Kankeu Tchewonpi, H., Saksena, P., Xu, K., and Evans, D.B. (2013). The financial burden from non-communicable diseases in low- and middle-income countries: a literature review. *Health Research Policy and Systems*, 11(31), 1–12.

Karombo, T. (2018). Zimbabwe is upbeat about its tobacco crop. [Online] Available at: https://www.iol.co.za/business-report/international/zimbabwe-is-upbeat-about-its-tobacco-crop-14080202

Khan, R. (2018). Tobacco cultivation creates jobs for 75,000 farmers' families. *The Nation*. [Online] Available at: https://nation.com.pk/31-Mar-2018/tobacco-cultivation-creates-jobs-for-75-000-farmers-families

Kyu, H.H., et al. (2016). Physical activity and risk of breast cancer, colon cancer, diabetes, ischemic heart disease, and ischemic stroke events: systematic review and dose-response meta-analysis for the Global Burden of Disease Study 2013. *British Medical Journal*, 354, i3857.

Labonté, R. and Stuckler, D. (2016). The rise of neoliberalism: how bad economics imperils health and what to do about it. *Journal of Epidemiology & Community Health*, 70(1), 312–18.

Leahy, S. (2014). *Merchants of Doubt* film exposes slick US industry behind climate denial. *The Guardian.* [Online] Available at: http://www.theguardian.com/environment/2014/nov/20/merchants-of-doubt-film-exposes-slick-us-industry-behind-climate-denial

Lelieveld, J., Evans, J.S., Fnais, M., Giannadaki, D., and Pozzer, A. (2015). The contribution of outdoor air pollution sources to premature mortality on a global scale. *Nature*, **525**, 367.

Lencucha, R. (2010). Philip Morris versus Uruguay: health governance challenged. *The Lancet*, 376(9744), 852–3.

Lencucha, R. and Drope, J. (2015). Plain packaging: an opportunity for improved international policy coherence? *Health Promotion International*, 30(2), 281–90.

Lencucha, R., Drope, J., and Chavez, J. (2015). Whole-of-government approaches to NCDs: the case of the Philippines Interagency Committee—Tobacco. *Health Policy and Planning*, 30(7), 844–52.

Lencucha, R., Drope, J., and Labonte, R. (2016). Rhetoric and the law, or the law of rhetoric: how countries oppose novel tobacco control measures at the World Trade Organization. *Social Science & Medicine*, **164**, 100–7.

Lencucha, R., Dubé, L., Blouin, C., Hennis, A., Pardon, M., and Drager, N. (2018a). Fostering the catalyst role of government in advancing healthy food environments. *International Journal of Health Policy and Management*, 7(6), 485–90.

Lencucha, R., et al. (2018b). Global tobacco control and economic norms: an analysis of normative commitments in Kenya, Malawi and Zambia. *Health Policy and Planning*, 33(3), 420–8.

Lim, S.S., et al. (2013). A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010. *The Lancet*, 380(9859), 2224–60.

Lukes, S. (2005). *Power: a radical view* (2nd edn). London: Palgrave Macmillan.

Makoka, D., et al. (2017). Costs, revenues and profits: an economic analysis of smallholder tobacco farmer livelihoods in Malawi. *Tobacco Control*, 26(1), 634–40.

McCambridge, J., Mialon, M., and Hawkins, B. (2018). Alcohol industry involvement in policy making: a systematic review. *Addiction*, 113(9), 1571–84.

Miller, D. and Harkins, C. (2010). Corporate strategy, corporate capture: food and alcohol industry lobbying and public health. *Critical Social Policy*, 30(4), 564–89.

Mitchell, A.D. and Casben, J. (2017). Trade law and alcohol regulation: what role for a global Alcohol Marketing Code? *Addiction*, 112(S1), 109–16.

NCD Alliance (2017). NCD Civil Society Atlas. In: *Brussels: Noncommunicable Disease Alliance*. [Online] Available at: https://ncdalliance.org/resources/ncd-atlas

NCD Risk Factor Collaboration (2016). Worldwide trends in diabetes since 1980: a pooled analysis of 751 population-based studies with 4·4 million participants. *The Lancet*, 387(10027), 1513–30.

National Cancer Institute (n.d.). *Cigar Smoking and Cancer: Facts Sheet.* [Online] Available at: http://www.cancer.gov/cancertopics/factsheet/Tobacco/cigars

Nugent, R., et al. (2018). Investing in non-communicable disease prevention and management to advance the Sustainable Development Goals. *The Lancet*, 391(10134), 2029–35.
O'Brien, P. (2013). Australia's double standard on Thailand's alcohol warning labels. *Drug and Alcohol Review*, 32(1), 5–10.
Ostrom, E. (1986). An agenda for the study of institutions. *Public Choice*, 48(1), 3–25.
PhilStar.com (2017). Sweetest compromise: cutting sweetened beverage taxes and its consequences. *The Philippine Star*, 17 November. Manila. [Online] Available at: https://www.philstar.com/other-sections/news-feature/2017/11/17/1759832/sweetest-compromise-cutting-sweetened-beverage-taxes-and-its-consequences
Physicians for a Smoke Free Canada (n.d.). *Tobacco Industry Profits: Revenue and profits of the world's largest tobacco companies*. [Online] Available at: https://www.fctc.org/member/physicians-for-a-smoke-free-canada-psc/
Porterfield, M.C. and Byrnes, C.R. (2011). Philip Morris v. Uruguay: will investor-state arbitration send restrictions on tobacco marketing up in smoke. *Investment Treaty News*, 1(4), 3–5.
Reddy, K.S. (2015). Regional roadmaps for reducing premature deaths from NCDs. *The Lancet Global Health*, 3(12), e725–6.
Roberto, C.A., et al. (2015). Patchy progress on obesity prevention: emerging examples, entrenched barriers, and new thinking. *The Lancet*, 385(9985), 2400–9.
Sabatier, P. (1987). Knowledge, policy-oriented learning and policy change: an advocacy coalition framework. *Science Communication*, **8**, 649–92.
Sacco, R.L., et al. (2016). The heart of 25 by 25: achieving the goal of reducing global and regional premature deaths from cardiovascular diseases and stroke: a modeling study from the American Heart Association and World Heart Federation. *Global Heart*, 11(2), 251–64.
Saez, E. and Zucman, G. (2016). Wealth inequality in the United States since 1913: evidence from capitalized income tax data. *The Quarterly Journal of Economics*, 131(2), 519–78.
Saloojee, Y. and Dagli, E. (2000). Tobacco industry tactics for resisting public policy on health. *Bulletin of the World Health Organization*, 78(7), 902–10.
Sassi, F., et al. (2018). Equity impacts of price policies to promote healthy behaviours. *The Lancet*, 391(10134), 2059–70.
Schram, A. (2017). When evidence isn't enough: ideological, institutional, and interest-based constraints on achieving trade and health policy coherence. *Global Social Policy*, 18(1), 62–80.
Schroeder, S. (2004). Tobacco control in the wake of the 1998 master settlement agreement. *New England Journal of Medicine*, 350(3), 293–301.
Seiler, C. (2009). *Republic of Drivers: A Cultural History of Automobility in America*. Chicago, IL: University of Chicago Press.
Shiffman, J. (2016). Networks and global health governance: introductory editorial for *Health Policy and Planning* supplement on the Emergence and Effectiveness of Global Health Networks. *Health Policy and Planning*, 31(Suppl 1), i1–i2.
Snowdon, C. (2013). *Free Market Solutions in Health: The Case of Nicotine*. In: *IEA Current Controversies Paper No 45*. Institute of Economic Affairs. [Online] Available at: http://www.iea.org.uk/sites/default/files/publications/files/Free%20Market%20Solutions%20in%20Health.pdf
Sommer, I., et al. (2015). Socioeconomic inequalities in non-communicable diseases and their risk factors: an overview of systematic reviews. *BMC Public Health*, 15(1), 914.
Stevens, G.A., et al. (2012). National, regional, and global trends in adult overweight and obesity prevalences. *Population Health Metrics*, 10(1), 22.
Stiglitz, J.E. (2017). Countering the power of vested interests: advancing rationality in public decision-making. *Journal of Economic Issues*, 51(2), 359–65.

Story, M., Kaphingst, K.M., Robinson-O'Brien, R., and Glanz, K. (2008). Creating healthy food and eating environments: policy and environmental approaches. *Annual Review of Public Health*, **29**, 253–72.

Strange, S. (1996). *The Retreat of the State: The Diffusion of Power in the World Economy*. Cambridge: Cambridge University Press.

Thow, A.M., et al. (2018a). Fiscal policy to improve diets and prevent noncommunicable diseases: from recommendations to action. *Bulletin of the World Health Organization*, 96(3), 201.

Thow, A.M., Greenberg, S., Hara, M., Friel, S., duToit, A., and Sanders, D. (2018b). Improving policy coherence for food security and nutrition in South Africa: a qualitative policy analysis. *Food Security*, **10**, 1105–30.

Thow, A.M., Jones, A., Hawkes, C., Ali, I., and Labonte, R. (2017). Nutrition labelling is a trade policy issue: lessons from an analysis of specific trade concerns at the World Trade Organization. *Health Promotion International*, 33(4), 561–71.

Top 100 food and beverage companies (n.d.). [Online] Available at: https://www.foodengineeringmag.com/2017-top-100-food-and-beverage-companies

Treanor, J. (2012). Boardroom pay packages close to £10m are on the rise. *The Guardian*. [Online] Available at: https://www.theguardian.com/business/blog/2012/mar/26/boardroom-pay-packages-10m-rise

Treerutkuarkul, A. (2017). Moving Thailand's mountain of alcohol-related harm. *World Health Organization Bulletin of the World Health Organization*, 95(7), 487.

UNCTAD (2016). *Taking Stock of IIA Reform.* Geneva: United Nations Conference on Trade and Development.

United Nations (2016). *High-Level Panel on Access to Health Technologies: The United Nations Secretary-General's High-Level Panel on Access to Medicines Report: Promoting Innovation and Access to Health Technologies*. Geneva: United Nations.

Voon, T.S. and Mitchell, A.D. (2016). *Philip Morris vs. Tobacco Control: Two Wins for Public Health but Uncertainty Remains*. [Online] Available at: https://www.researchgate.net/publication/309319497_Philip_Morris_vs_Tobacco_Control_Two_Wins_for_Public_Health_but_Uncertainty_Remains

Wagner, K.-H. and Brath, H. (2012). A global view on the development of non communicable diseases. *Preventive Medicine*, 54, S38–S41.

Wang, H., et al. (2016). Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980–2015: a systematic analysis for the Global Burden of Disease Study 2015. *The Lancet*, 388(10053), 1459–544.

Wansink, B. and Sobal, J. (2007). Mindless eating: the 200 daily food decisions we overlook. *Environment and Behavior*, 39(1), 106–13.

Weible, C.M., Heikkila, T., deLeon, P., and Sabatier, P.A. (2012). Understanding and influencing the policy process. *Policy Sciences*, 45(1), 1–21.

Wilkinson K, Bhardwaj V (2017). *South Africa's Proposed Sugar Tax: Claims About Calories & Job Losses Checked.* Africa Check. [Online] Available at: https://africacheck.org/reports/sas-proposed-sugar-tax-claims-calories-job-losses-checked/

World Bank (n.d.). *The World Bank and Tobacco Control: The Facts.* [Online] Available at: http://www.worldbank.org/en/topic/health/brief/world-bank-and-tobacco-control-the-facts

World Health Organization (WHO) (2003). *WHO Framework Convention on Tobacco Control*. Geneva: WHO.

World Health Organization (WHO) (2011). *From Burden to 'Best Buys'*. Geneva: WHO.

World Health Organization (WHO) (2013). *Follow-up to the Political Declaration of the High-level Meeting of the General Assembly on the Prevention and Control of Non-communicable Diseases. SIXTY-SIXTH WORLD HEALTH ASSEMBLY WHA66.10 Agenda item 13.1, 13.2, 27*

May 2013. Annex: Global Action Plan for the Prevention and Control of Noncommunicable diseases 2013–2020. Geneva: WHO.

World Health Organization (WHO) (2014a). *Global Status Report on Alcohol and Health*. Geneva: WHO.

World Health Organization (WHO) (2014b). *Global Status Report on Non Communicable Disease*. Geneva: WHO.

World Health Organization (WHO) (2016). *Global Report on Diabetes.* Geneva: WHO.

World Health Organization (WHO) (2017). *Preventing Noncommunicable Diseases by Reducing Environmental Risks*. Geneva: WHO.

World Health Organization (WHO) (2018). Media release: 9 out of 10 people worldwide breathe polluted air, but more countries are taking action (2 May 2018). Geneva: WHO. [Online] Available at: http://www.who.int/news-room/detail/02-05-2018-9-out-of-10-people-worldwide-breathe-polluted-air-but-more-countries-are-taking-action

Yach, D., Khan, M., Bradley, D., Hargrove, R., Kehoe, S., and Mensah, G. (2010). The role and challenges of the food industry in addressing chronic disease. *Globalization and Health*, **6**, 10.

Zohlnhöfer, R., Engler, F., and Dümig, K. (2018). The retreat of the interventionist state in advanced democracies. *British Journal of Political Science*, 48(2), 535–62.

11.3

Immunization and vaccination

Eleonora A.M.L. Mutsaerts and Shabir A. Madhi

Introduction

History of vaccination

The origins of vaccine-derived immunity date back to 900 AD when the Chinese observed that individuals surviving smallpox were immune to renewed illness. This led to the earliest documented experiments on vaccination by inoculating smallpox into scarified skin, a process called variolation. Although variolation was effective in conferring immunity to smallpox, it carried a significant risk of severe disease or even death in some recipients. The practice of smallpox inoculation spread to Africa and Turkey, before arriving in Europe and the Americas. In eighteenth-century Britain, Edward Jenner observed immunity to smallpox in cow-maids who developed cowpox after close exposure to cows. In 1796, he experimented inoculating a 13-year-old boy with vaccinia virus (cowpox) and challenged the child with smallpox virus to prove the concept of immunity. This led to the first smallpox vaccine being developed in 1798; ultimately resulting in eradication of smallpox 200 years later.

Subsequent to the success of developing a smallpox vaccine, Louis Pasteur developed the rabies vaccine in 1885, followed by antitoxins and vaccines against cholera, typhoid, plague, diphtheria, tetanus, anthrax, and tuberculosis through the 1930s. It was however only after the mid-1940s, that systematic immunization became common practice. The Expanded Programme on Immunization (EPI) was launched in 1974 by the World Health Organization (WHO). The goal of EPI was to immunize children worldwide with six core vaccines: diphtheria, tetanus, pertussis, measles, polio, and tuberculosis. Vaccination of children is now regarded as one of the

most successful public health interventions of the 20th century and continues to have vast impact on preventing infectious disease and death worldwide.

Challenges to immunization programmes

Following initiation of the EPI programme, only 5% of children in low- and middle-income countries (LMICs) received the full series of vaccines in the 1970s, which increased to around 80% by the late 1980s. However, global vaccine coverage reached a plateau and even declined in some places by the mid-1990s. Consequently, approximately 19.9–30 million children born each year would be underimmunized or unvaccinated. This is further accentuated by inequity in vaccine coverage, with children from the lowest economic quintile in many LMIC being 2–3 fold less likely to be fully immunized than those from the highest quintile. This paradox is manifest by the increased susceptibility of children from poverty-stricken backgrounds to developing vaccine-preventable diseases, as well as having less access to curative healthcare services.

Reasons for underimmunization of children are multifactorial and include lack of political commitment to prioritize immunization programmes; lack of access to health facilities in marginalized and poverty-stricken communities; vaccine hesitancy due to social, cultural, and religious issues and mis-information; logistical issues in vaccine supply such as maintaining cold-chain and sterile needle and syringe availability; as well as high pricing of newly developed vaccines. Furthermore, vaccine hesitancy, especially in high- and middle-income settings are fostered by antivaccine lobby groups, often based on unsubstantiated concerns.

To address some of these barriers, which prevent the full public health value of vaccines being realized, the Global Alliance for Vaccines and Immunization (Gavi) was established in 2000 as a partnership of governments, private sector, philanthropists, and civil society. The mission of Gavi is to reduce the number of unvaccinated children, as well as expedite the inclusion of new life-saving vaccines into public immunization programmes in low-income countries.

Public health impact of vaccination

Even at current vaccine coverage rates of approximately 83% of children being fully immunized, childhood vaccination is estimated to save 2 to 3 million lives annually. The benefit of immunization is evident from the decline in global mortality attributed to many pathogens for which vaccines have been deployed. See Figure 11.3.1.

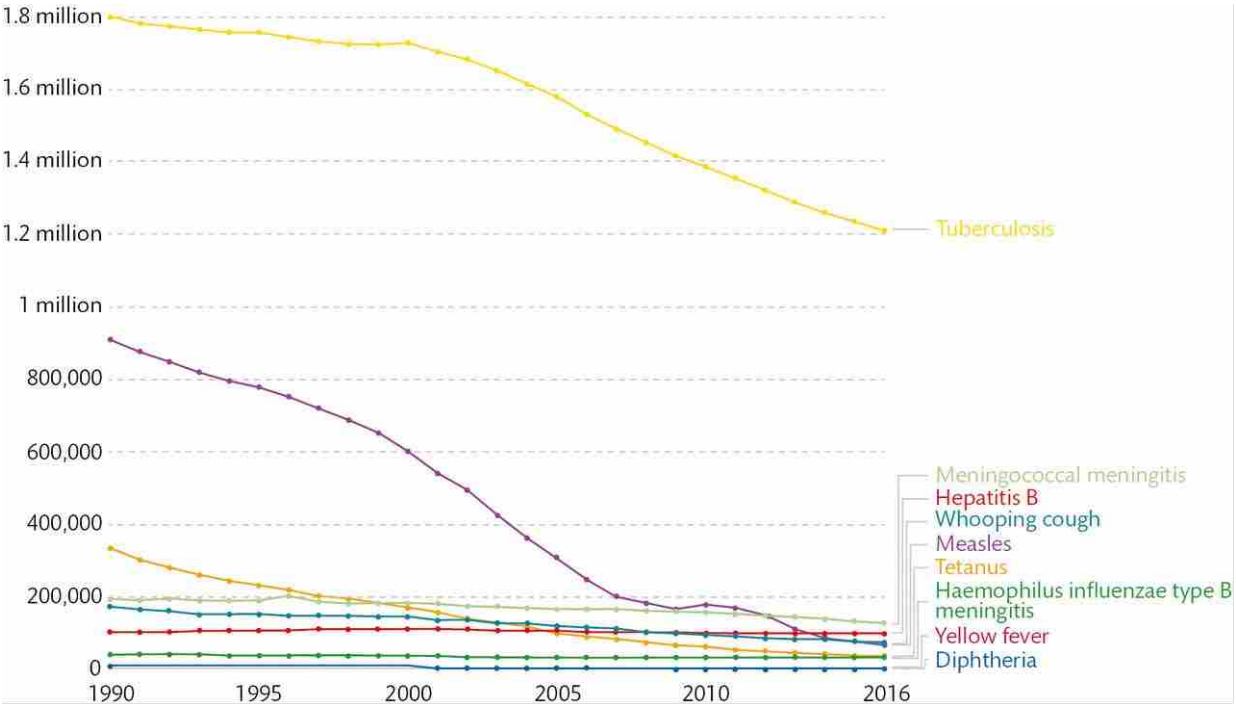


Fig. 11.3.1 Trends in deaths due to vaccine-preventable diseases from 1990 to 2016.
Source: Diphtheria, Tetanus, H influenzae type B meningitis, Hepatitis B, Measles, Meningococcal meningitis, Tuberculosis, Yellow fever, Whooping cough. OurWorldInData.org/vaccination. Licensed under CC BY-SA.

Further data illustrate the public health impact of immunization in the United States, where vaccine introduction was temporally associated with near-elimination of deaths from many vaccine-preventable diseases; see Table 11.3.1.

Table 11.3.1 Reduction of cases and deaths of vaccine-preventable diseases in the United States after introduction of vaccination

Disease	Pre-vaccine, cases per million per year (year)	Post-vaccine, cases per million per year (all post-vaccine cases refer to 2006)	Reduction cases, %	Pre-vaccine, deaths per million per year (year)	Post-vaccine, deaths per million per year (all post-vaccine cases refer to 2004)	Reduction deaths, %
Diphtheria	158 (1936–1945)	0	100%	13.7 (1936–1945)	0	100%
Measles	3,044 (1953–1962)	0.2	99.99%	2.5 (1953–1962)	0	100%
Mumps	830 (1963–1968)	22	97.4%	0.2 (1963–1968)	0	100%
Pertussis	1,534 (1934–1943)	52	96.6%	30.8 (1934–1943)	0.09	99.7%
Acute poliomyelitis	141 (1941–1950)	0	100%	10 (1941–1950)	0	100%
Paralytic poliomyelitis	103 (1951–1954)	0	100%	11.8 (1951–1954)	0	100%
Rubella	242 (1966–1968)	0.04	99.98%	0.09 (1966–1968)	0	100%
Congenital rubella syndrome	0.76 (1966–1969)	0.003	99.6%	No data	0	No data
Smallpox	250 (1900–1949)	0	100%	2.9 (1900–1949)	0	100%

Disease	Pre-vaccine, cases per million per year (year)	Post-vaccine, cases per million per year (all post-vaccine cases refer to 2006)	Reduction cases, %	Pre-vaccine, deaths per million per year (year)	Post-vaccine, deaths per million per year (all post-vaccine cases refer to 2004)	Reduction deaths, %
Tetanus	4 (1947–1949)	0.14	96.6%	3.2 (1947–1949)	0.01	99.6%
Hepatitis A	465 (1986–1995)	51	89%	0.5 (1986–1995)	0.06	88.7%
Acute hepatitis B	273 (1982–1991)	44	83.9%	1 (1982–1991)	0.16	83.6%
Haemophilus influenza type b	84 (1980s)	0.17	99.8%	No data	0.02	No data
Pneumococcal disease	233 (1997–1999)	139	40.5%	24 (1997–1999)	16.5	31.3%
Varicella	16,018 (1990–1994)	2046	87.2%	0.41 (1990–1994)	0.06	84.3%

Data source: Roush and Murphy (2007)—Historical comparisons of morbidity and mortality for vaccine-preventable diseases in the United States. In *The Journal of the American Medical Association*, 298, 18, 2155–63. OurWorldInData.org/vaccination. Licensed under CC BY-SA by the author Max Roser/data converted to table.

Notable successes of vaccination:

Eradication of smallpox by 1979 after a global immunization campaign. Until the 1960s, smallpox killed two million people per year.
Reduction of the number of polio cases from 350,000 per year in 1988 (endemic in 125 countries) to 74 cases in 2015 (endemic in three countries).
Eradication of neonatal tetanus in two-thirds of developing countries, with annual neonatal tetanus attributed deaths being reduced from 1.2 million in 1980 to less than 50,000 in 2015. The reduction in neonatal tetanus is a result of increase in tetanus toxoid vaccine coverage in pregnant women, coupled with changes in birthing and practices and prenatal care of newborns.
Reduction in the number of measles deaths by 79% between 2,000 (n = 580,000) and 2015 (<50,000), thereby preventing more than 20 million measles deaths during this period.
Eradication of rinderpest disease in animals.

Note: In addition to immunization, broader social and economic development, including improved sanitation, increased access to healthcare facilities, and improved infant nutrition are also likely to have played a role in the decline in childhood deaths due to vaccine-preventable disease.

Broader impact of vaccination

Immunization of infants is widely heralded as a cost-effective healthcare intervention, second only to access to potable water. It is estimated that the economic benefit of vaccination is approximately 44-fold greater than the up-front cost of immunization, inclusive of direct treatment cost and the life-time loss of productivity averted by prevention of disease. The benefits of immunization extend beyond the prevention of disease in the individual, with health benefits also potentially accruing to society at large by the interruption of transmission of some pathogens (indirect immunity) or person-to-person circulation of some live-attenuated vaccines (herd immunity). Furthermore, the societal valuation of vaccines could include broader outcomes, including the effects on comorbidities, nosocomial infections, educational attainment, cognitive development, labour productivity, income, savings, investment, and fertility.

Advances in the vaccine landscape

Since the 1970s, significant advances have occurred in the development of new vaccines. Vaccine development was fostered by progress in biotechnology and enhanced understanding of immunity against different pathogens, including newly designed adjuvant technology. The number of vaccines that have been licensed in the past two decades or are in the final stages of clinical development, exceeds the total number of vaccines that were developed in the previous two centuries; Table 11.3.2.

Table 11.3.2 Vaccine development timeline

1798–1899	1900–1949	1950–1979	1980–1999	2000s
1798 **Smallpox** 1885 **Rabies** 1891 **Anthrax** 1886 **Typhoid** 1896 **Cholera** 1897 **Plague**	1923 **Diphtheria** (toxoid) 1924 **Tetanus** (toxoid) 1926 **Pertussis** 1927 **Tuberculosis (BCG)** 1935 **Yellow fever** 1936 **Influenza** 1938 **Typhus**	1955 **Polio (IPV)** 1962 **Polio (OPV)** 1963 **Measles** 1967 **Mumps** 1969 **Rubella** 1974 **Meningococcal polysaccharides** 1977 **Pneumococcal polysaccharides**	1980 **Adenovirus**, **Rabies** (cell culture) 1981 **Hepatitis B** (plasma derived), **tick-borne encephalitis** 1983 **Maternal tetanus** 1985 ***Haemophilus influenzae* type B** (polysaccharide) 1986 **Hepatitis B** 1987 ***Haemophilus Influenzae*** (conjugate) 1989 **Typhoid** 1991 **Cholera (WC-rBC)** 1992 **Japanese encephalitis** (inactivated) 1995 **Varicella zoster** 1996 **Hepatitis A** (inactivated), **Pertussis** (acellular) 1998 **Rotavirus** 1999 **Meningitis C** (conjugate)	2000 **Pneumococcal conjugate** 2005 **Meningococcal quadrivalent** (A, C, W, Y) 2006 **Human papilloma virus, Rotavirus** (attenuated), **Varicella zoster** live) 2009 **Japanese encephalitis** (Vero cell culture), **Cholera** (whole cell) 2012 **Typhoid Vi conjugate vaccine** 2015 **Malaria falciparum (RTS-S)** 2015 **Shigella** 2016 **Dengue** 2018 **rBCG** Expected: **RSV vaccine for pregnant women and long-acting RSV monoclonal antibodies**

Expanded Programme on Immunization recommended vaccines

See Table 11.3.3

Table 11.3.3 Recommended routine immunizations for children

Antigen	Age of first dose	Doses in primary series	Booster dose	Considerations
BCG	Birth	1		
Hepatitis B	Birth	3 or 4; with DTPCV		
Polio	6 wks (bOPV)–8 wks (IPV)	4 (bOPV + IPV); 3–4 (1–2 IPV, 2 bOPV); 3 (IPV)		
DTP containing vaccine	6 wks	3	3 Boosters 12–23 mths (DTPCV); 4–7 yrs (Td); and 9–15 yrs (Td)	
Haemophilus influenzae type b	6 wks–59 mths	2–3		Not recommended for children >5 yrs
Pneumococcal (conjugate)	6 wks	2–3	After 2 doses: 9–15 months	
Rotavirus	6 wks with DTP1	2 (Rotarix); 3 (Rota Teq)		Not recommended if >24 mths old
Measles	9 or 12 mths (6 mths min)	2		HIV early vaccination
Rubella	9 or 12 mths with measles	1		Achieve and sustain 80% coverage
HPV	9 yrs of age (females only)	2		Target 9–14-yr-old girls; Older age ≥ 15 yrs 3 doses
Recommended for children living in certain regions				
Japanese encephalitis	6–9 mths	1 (live-attenuated and live recombinant); 2 (inactivated Vero-cell-derived)		

Antigen	Age of first dose	Doses in primary series	Booster dose	Considerations
Yellow fever	9–12 mths with measles	1		
Tick-borne encephalitis	≥1 yr (FSME-Immun and Encepur) ≥3 yrs (BE_Moscow and EnceVir)	3	At least 1 every 3 years	
Recommended for children in some high-risk populations				
Typhoid	>6 mths (Typhbar); 2 yrs (Vi PS); 5 yrs (Ty21a)	1 (Typhbar, Vi PS); 3 or 4 (Ty21a)	Every 3 yrs (Vi PS); Every 3–7 yrs (Ty21a)	
Cholera	1 yr (Shanchol, Euvchol, mORCVAX); 2 yrs (Dukoral)	2 (Shachol, Euvchol, and mORCVAX); 3 at 2–5 yrs or 2 ≥6 yrs (Dukoral)	Every 6 mths (Dukoral 2–5 yrs); Every 2 yrs (Dukoral ≥6 yrs); After 2 yrs (Shanchol, Euvchol, mORCVAX)	
Meningococcal	9–18 mths (Men A conjugate); 2–11 mths / ≥12 mths (MenC conjugate) 9–23 mths / ≥2 yrs (quadrivalent conjugate)	1 (MenA conjugate); 2 / 1 (MenC conjugate); 2 / 1 (quadrivalent conjugate)	After 1 yr (Men C conjugate)	
Hepatitis A	1 yr	At least 1		
Rabies	As required	2		PrEP vs. PEP
Dengue (CYD-TDV)	9 yrs	3		Seroprevalence
Recommended for children receiving vaccinations from immunization programmes				
Mumps	12–18 mths with measles	2		Coverage criteria > 80%;

Antigen	Age of first dose	Doses in primary series	Booster dose	Considerations
Seasonal influenza (inactivated)	6 mths	2 (<9 yrs); 1 (≥9 yrs)	Revaccinate annually: 1 dose	Priority risk groups: pregnant women, lower dosage for children 6–35 mths
Varicella	12–18 mths	1–2		Achieve & sustain ≥ 80% coverage

Bacillus Calmette–Guérin (BCG)

Tuberculosis (TB), caused by *Mycobacterium tuberculosis* (MTB), is now the leading vaccine-preventable cause of death worldwide. An estimated 1.7 billion people are infected with MTB globally, contributing to 10.4 million active cases in 2016, 10% of which is related to HIV-infection. There were an estimated 1.2 million deaths resulting from TB in 2016. The risk of developing TB following MTB infection is greatest in children under 2 years of age, with 30–40% of those infected with MTB progressing to disease. MTB infection during childhood manifests either as pulmonary TB (PTB) or as the more severe spectrum of extrapulmonary TB (30–40% of cases). Childhood TB serves as an indicator of ongoing *M. tuberculosis* transmission in a population.

Immunocompromised individuals are at heightened risk of progressing from MTB infection to disease, or having re-activation of underlying dormant MTB infection. HIV-infected individuals who are not treated with antiretroviral treatment (ART) have a 10% annual risk of TB disease, compared to a life-time risk of 10% among the general population. Although this risk is reduced if initiated on ART, it remains significantly greater in HIV-infected individuals than the general population. Other factors facilitating spread of and predisposition to TB include poverty, overcrowding, poor ventilation, tobacco use, and diabetes. The risk of TB is also increased in refugees, migrants, urban poor, and incarcerated individuals.

Despite availability of treatment, around 40% of TB cases remain undiagnosed. To achieve the United Nations Sustainable Development Goal of ending TB epidemics by 2030, there is an increased focus on early detection and treatment of TB cases; as well as provision of chemo-

prophylaxis in high-risk groups such as HIV-infected adults, children with MTB infection less than 5 years of age, and household contacts of active TB cases. In addition, vaccination with BCG vaccine of infants at birth, or shortly thereafter, remains a key component to achieve the End TB strategy; as does the development of new TB vaccines.

BCG was first introduced in humans in 1921, making it one of the oldest vaccines. It is a freeze-dried preparation of a live-attenuated strain of *M. bovis*. Currently, more than 10 manufacturing strains exist. Despite ongoing research for the development of new anti-TB vaccines, the absence of an established immunological correlate of protection is a major impediment to vaccine development. As such, BCG is the only available TB vaccine and also provides protection against leprosy. However, BCG vaccine effectiveness varies considerably between different manifestations of TB, vaccine effectiveness wanes over time and there is variability between different BCG vaccine strains. Neonatal BCG vaccination is reportedly associated with 59% (RR 0.41, 95% confidence interval [CI]: 0.29–0.58) protection against PTB, possibly being lower in countries closer to the tropics; and >80% against extrapulmonary TB during childhood. Recent evidence indicates that BCG vaccination might also protect against *M. tuberculosis* primary infection. The duration of protection following BCG vaccination at 10–15 years of age is uncertain. Currently, the additional benefit from booster doses is being re-investigated.

BCG vaccination is recommended in settings with a high TB incidence and/or high leprosy burden. A single dose is recommended at birth to all healthy neonates. The vaccine is contraindicated in individuals known to be allergic to vaccine components, immunocompromised individuals including symptomatic HIV-infected, and ART naïve with CD4% ≤25%. Children infected with HIV, who are receiving ART, and are clinically and immunologically stable should be vaccinated with BCG.

New TB candidate vaccines are currently being developed, prioritizing development of vaccines targeted at adolescents and adults, considering they contribute to >90% of TB cases and are the primary source of MTB transmission in communities.

Hepatitis B

Hepatitis B virus (HBV) causes acute hepatitis and chronic hepatitis; the latter could progress to hepatocellular carcinoma (HCC). Serological prevalence of the hepatitis B surface antigen (HBsAg) is used to measure endemicity of active HBV infection. HBsAg sero-positivity was estimated at 3.5% of the world population in 2015; including 257 million people living with chronic HBV. The prevalence of sero-positivity is highest in the Western Pacific (6.2%) and Africa (6.1%). Global mortality due to HBV infection was estimated at 887,220 in 2015 including deaths due to HCC, cirrhosis, and acute hepatitis. HBV infection in infancy has a higher risk of causing chronic infection than when acquired later in life. In HBV endemic areas, co-infection with other viral infections commonly occurs, including co-infection with HIV, hepatitis C virus, and hepatitis D virus.

Hepatitis B vaccines were first licensed in 1982 and contain the HBsAg surface protein. In 1986, recombinant vaccines were developed, which are now most widely used. Immunization with a complete primary series of three doses of HBV induces protective antibody levels (i.e. HBsAg antibody ≥10 mIU/ml) in >95% of infants; and protection is likely lifelong. Hence, no booster doses are recommended. Hepatitis B vaccines are available either as a monovalent vaccine or in combination with diphtheria-tetanus-pertussis (DTP) ± *Haemophilus influenzae* type b (Hib) and inactivated polio vaccine (IPV); only the monovalent vaccine should be used for immunization at birth. WHO recommends that hepatitis B vaccine should be given immediately after birth to prevent vertical acquisition from highly-infectious mothers (i.e. Hepatitis B antigen positive), which should be followed by two or three further doses to complete the primary series. As of 2018, 102 (53%) countries had introduced hepatitis B vaccination at birth. Neonatal vaccination provides partial (estimated 30–40%) prevention against perinatal infection.

Hepatitis B vaccination is recommended in persons at high risk of HBV infection in older age groups, including patients who frequently require blood or blood products. If resources are available, catch-up campaigns in unvaccinated cohorts are recommended. HIV-infected adults, in whom there is possibly waning of immunity, may need re-vaccination. Other measures to prevent HBV transmission include hepatitis B immunoglobulin at birth (costly and unavailable in most low-income countries) and screening of donated blood for transfusion.

Global coverage with three doses of hepatitis B vaccine was estimated at 84% in 2015. Routine HBV immunization has contributed to a reduction in prevalence of chronic hepatitis B infection in children aged 0–5 years from 4.7% in the pre-vaccination era to 1.3% in 2015. An estimated 14.2 million cases of chronic HBV infection were prevented by hepatitis B vaccine among children under 5 years of age worldwide in 2013.

Polio

Poliomyelitis is an acute communicable disease caused by three poliovirus serotypes (1, 2, or 3). Prior to routine introduction of vaccination, poliovirus was the major pathogen causing permanent disability in children; 1 in 200 susceptible individuals developed paralytic poliomyelitis.

Both oral, live-attenuated vaccines (OPV) and injectable, inactivated virus vaccines (IPV) are available. Live OPV vaccines contain attenuated virus strains, which may replicate and re-obtain their potential for transmission and neurovirulence, causing isolated cases or outbreaks of paralytic poliomyelitis. Since 2009, type 2 polio virus was removed from OPV, to eliminate vaccine-derived polio infections from this strain. The tri-valent OPV has since been replaced by the bivalent OPV (bOPV containing type 1 and 3 Sabin viruses) in immunization programmes still using OPV.

The EPI recommends three to four doses of bOPV and at least one dose of IPV. The advantages of bOPV compared with IPV include low cost, ease of administration, superior intestinal mucosal immunity, and potential for boosting immunity in household and community contacts. A sequential IPV-bOPV or IPV-only schedule can be used in countries with sustained high vaccination coverage and low importation risk from neighbouring countries. However, WHO does not recommend this switch in endemic countries and regions with high risk of wild poliovirus (WPV) importation.

In 1988, the World Health Assembly aimed at eradicating poliomyelitis by 2000 through initiation of the Global Polio Eradication Initiative. Since then, global incidence of poliomyelitis dropped by >99%. Wild poliovirus (WPV) type 2 was eradicated in 2015 and no detection of WPV type 3 occurred since November 2012. Countries with endemic polio have dropped from 125 in 1988 to 2 in 2015 (Afghanistan, Pakistan).

Diphtheria

Diphtheria is caused by *Corynebacterium diphtheriae*. In the past, diphtheria was one of the most devastating infectious diseases with case-fatality rates up to 50%. Diphtheria toxoid-based vaccines were developed in the late 1940s, but only became widely accessible worldwide in the 1970s when EPI was initiated. Introduction of the vaccine contributed to a >90% reduction of reported diphtheria cases from 1980 to 2000. Nevertheless, diphtheria remains a considerable health problem in countries with low vaccine coverage; including among displaced populations and refugees. Approximately 86% of children globally receive three doses of diphtheria-containing vaccines. Due to increased vaccination coverage in children, a shift in age distribution to adolescents and adults has occurred.

Diphtheria toxoid-containing vaccines contain a formaldehyde-inactivated preparation of diphtheria toxin, adsorbed onto alum to increase immunogenicity. Since the 1940s, a combination vaccine containing DTP has been widely used. Diphtheria toxoid, tetanus toxoid, and pertussis vaccine are also available in combination with HBsAg and *Haemophilus influenzae* type b conjugate vaccines (pentavalent vaccines), and IPV (hexavalent vaccines containing acellular pertussis vaccine). Routine infant immunization includes a three-dose primary series (starting at 6 weeks and 4 weeks apart) and a booster at 15–18 months. Presence of maternal diphtheria antibody interferes with immune responses following the first two doses during early infancy, however, protective antibody levels are reached after the third primary series dose. The diphtheria vaccine does not prevent infection, but inhibits diphtheria toxin's effects, thereby preventing systemic illness.

Immunity wanes over time after a three-dose primary vaccination schedule. In order to achieve protection, a booster dose is required. It is recommended that individuals ≥7 years of age should receive the low-dose diphtheria toxoid (Td) vaccine, to reduce risk of vaccine reactions.

Tetanus

Clostridium tetani is the causative bacterium of tetanus. It is acquired through exposure to *C. tetani* spores in contaminated skin wounds or tissue injury. All age groups are at risk; although neonates contributed to the highest burden of tetanus related deaths in the past. Tetanus infection is associated with case-fatality rates up to 100% in the absence of medical care.

In countries with suboptimal immunization programmes, including low coverage rates of tetanus vaccination during pregnancy, tetanus remains a considerable public health problem. Most cases of tetanus still occur in neonates in LMIC, as well as some cases in mothers. Poor hygiene during delivery/abortion, cultural practices, and cord care practices contribute to maternal and neonatal tetanus. Incidence rates are highest in rural and difficult to reach areas, where delivery predominantly takes place at home without sterile practices. In countries where maternal and neonatal tetanus is uncommon with high coverage rates in infants and pregnant women, the majority of cases occurs in adolescents and adults after injuries. Due to poor surveillance systems in LMIC, there is a lack of robust data on tetanus incidence. Nevertheless, it is estimated that approximately 34,000 neonates died from tetanus in 2015, a 96% reduction compared to 1988 (787,000 cases).

Tetanus-toxoid-containing vaccines (TTCV) are incorporated in routine immunization programmes around the world. Use of combination vaccines containing diphtheria toxoid is recommended. Prevention of neonatal tetanus is targeted by immunizing women of childbearing age with TTCV, either during or outside of pregnancy. By doing so, both the mother and the newborn (through transplacental transfer of tetanus antibodies) are protected. Administration of TTCV is recommended to women of childbearing age in countries with high risk for neonatal tetanus. In any country, tetanus vaccination history should be verified in all pregnant women when visiting the antenatal clinic for the first time and any missing doses should be administered. In all settings, clean procedures during delivery are important.

Recovery from tetanus disease does not induce natural immunity. Individuals who have been infected remain susceptible and thus need to be immunized. WHO recommends administration of three primary and three booster doses through routine childhood immunization (booster doses at 12–23 months, 4–7 years, and 9–15 yrs).

Pertussis

Pertussis, also known as whooping cough, is caused by *Bordetella pertussis* and is endemic globally. Prior to the vaccine era, pertussis was a common childhood disease with infections occurring in most individuals and >50% presenting with clinical symptoms. Young infants are at highest risk of death

due to pertussis infection. Pertussis can be prevented by vaccination and immunization programmes in children have been successful in reducing pertussis cases and deaths. Combined with diphtheria and tetanus toxoid, the pertussis vaccine was introduced in the EPI in 1974. Vaccination prevented >1.3 million pertussis related deaths worldwide in 2001 and 687,000 deaths in 2008.

Two type of pertussis vaccines are available. Whole-cell (wP) vaccines contain killed *Bordetella pertussis* bacteria and are commonly used in most LMIC. Acellular (aP) vaccines contain subunit pertussis toxin in combination with 1–4 other immunogenic pertussis proteins, including filamentous haemagglutinin, pertactin, and fimbriae. Many high-income countries (HICs) have replaced wP with aP vaccines in order to decrease adverse events following vaccination.

A rise in pertussis incidence in a number of countries using the aP vaccine has been reported. The increase in cases was partly attributed to cyclical patterns, improved surveillance, increased awareness, and more sensitive laboratory testing using polymerase chain reaction. However, aP vaccines have shorter duration of protection (4–5 years) than wP (10–12 years) and do not confer mucosal protection or prevent transmission as effective as wP. WHO recommends that countries currently using wP do not change to aP. In HICs, pertussis remains the least controlled vaccine-preventable disease; possibly contributed to by the transitioning from the wP to aP vaccines.

Pertussis vaccination is aimed at reducing severity of pertussis in infancy. The global target is to vaccinate >90% of infants with three doses of pertussis vaccine. In some countries a booster dose of TdaP is recommended in adolescents and adults to prevent pertussis in this population. There is insufficient evidence whether the booster dose is able to reduce severe pertussis in children; however, a booster dose of aP might be required decennially considering the duration of protection from aP. Furthermore, vaccination of pregnant women has been a recent focus of pertussis immunization, aimed at protecting young infants against pertussis disease during the first 2–3 months of life when risk of severe pertussis disease and death is greatest; and they are too young to have derived immunity from direct vaccination. Vaccination of pregnant women has been demonstrated to be safe and has >90% effectiveness in preventing pertussis in their infants. It is recommended that vaccination occurs from the second trimester onward, and at least 30 days before the anticipated birth of the child to optimize

transplacental transfer of antibody to the foetus. Maternal pertussis vaccination, however, has been shown to interfere with immune responses to aP and possibly pneumococcal conjugate vaccine (PCV) (due to diphtheria toxoid included in the dTap vaccine given to mothers—which is similar to the cross-reacting material (CRM) protein used to conjugate polysaccharide in PCV-CRM_{197} formulations). The clinical relevance of this interference in antibody responses in the infants of vaccinated women requires further investigation.

Haemophilus influenzae type b

Haemophilus influenzae type b (Hib) bacteria are responsible for severe pneumonia, meningitis, and other invasive diseases. The greatest burden of disease prior to routine Hib polysaccharide-protein conjugate vaccine (HibCV) immunization was in children aged 4–18 months, with younger children deriving immunity from maternal transfer of antibody and older children through natural immunity derived from Hib exposure. HIV-infected children have a 6-fold higher risk of Hib invasive disease compared to HIV-uninfected.

Hib disease is difficult to diagnose because it requires immediate laboratory testing in patients, before receiving antibiotics; and sampling of cerebrospinal fluid in cases of meningitis. Furthermore, the highest burden of severe disease and mortality from Hib in LMIC is due to pneumonia, for which the sensitivity of blood culture in diagnosing cases is <5%, and there are no other pragmatic diagnostic tools by which to make a pathogen specific diagnosis.

Prior to vaccine introduction in LMIC in 2000, Hib was estimated to cause 2–3 million cases of serious disease in children <5 and 386,000 deaths. In 2015, the number of Hib deaths dropped to around 30,000 (uncertainty range [UR] 18,400–40,700), a 90% reduction from 2000. HibCV is currently used in 191 (98%) WHO member states.

Although HibCV was introduced into high-income immunization programmes since the late 1980s, there was a lag of approximately a quarter of a century before the majority of low-income countries (LICs) were able to include the vaccine into their public immunization programme. This introduction was enabled by Gavi support, as well as a HibCV price reduction from approximately $20 per dose when first licensed to less than

$2 per dose nowadays as part of a combination vaccine with other vaccine epitopes. HibCV immunization has been associated with >90% reduction in invasive Hib disease (especially meningitis) in countries administering the vaccine in routine immunization programmes; as well as approximately 20% reduction in all-cause radiologically confirmed pneumonia.

The vaccine consists of proline-rich protein, a cell wall component of the Hib polysaccharide, conjugated to a protein to stimulate a T-cell dependent immune response. This induces immunity in young infants and establishes memory cell repertoire. The vaccine can be administered singularly, but is nowadays mainly combined with DTP ±HepB ±IPV. WHO recommends HibCV in all infant immunization programmes. The vaccine may be administered as a three primary dose schedule without a booster (3p+0); 2 dose primary schedule with a booster (2p+1); or a three-dose primary schedule with a booster (3p+1), depending on local epidemiology and the national schedule. Booster doses are currently used in most resource-rich countries, but are not generally recommended in the EPI schedule. Booster dosing might, however, be required for HIV-infected children in whom waning of immunity has been shown in the absence of treatment with ART.

Pneumococcal vaccines

Diseases caused by *Streptococcus pneumoniae* include pneumonia, meningitis, and bacteraemia, and other less severe but more frequent mucosal illnesses such as otitis media, sinusitis, and bronchitis. An estimated 294,000 (UR 192,000–366,000) HIV-uninfected children and 23,300 (15,300–29,700) HIV-infected children aged 1–59 months died of pneumococcal disease in 2015, predominantly in LMIC. The burden of pneumococcal disease is highest in children under 2 and the elderly.

Introduction of PCVs into immunization programmes has resulted in a dramatic decline in vaccine-serotype invasive pneumococcal disease (IPD) incidence, even in unvaccinated age groups due to the indirect effect of vaccination. This is achieved by PCV reducing nasopharyngeal acquisition of vaccine-serotypes among the young vaccinated children, who are also the main source of transmission in communities, and consequently interrupting the transmission of the vaccine-serotypes among unvaccinated age groups.

Two PCV vaccines are currently recommended: a 10-valent (PCV10) and 13-valent (PCV13) vaccine. The vaccines are designed to cover the

serotypes associated with severe pneumococcal disease in children. PCV10 includes protein D as the conjugating protein, which may provide some protection against non-encapsulated *Haemophilus influenzae* mucosal infection, a common cause of otitis media and possibly pneumonia. Similar to Hib conjugate vaccine, PCV also induces a T-cell dependent immune response resulting in memory responses; as well as protecting against nasopharyngeal colonization (50% efficacy). This indirect protection of unvaccinated individuals, including adults at high risk for pneumococcal disease, is achieved once infant vaccine coverage rates are approximately 50%.

Vaccine efficacy of PCV is >85% against vaccine-serotype IPD, approximately 50% against vaccine-serotype acute otitis media, and 30% against all-cause radiologically confirmed pneumonia hospitalization. Although there is replacement by non-vaccine-serotypes in the nasopharynx, this has not translated into substantial increase in non-vaccines serotype IPD incidence in children. However, recent data from the United Kingdom, indicate that the indirect protection against vaccine-serotype IPD in the elderly following childhood vaccination, has been offset by an increase in non-vaccine-serotype IPD. Substantial increase in non-vaccine-serotype IPD has also been observed in HIV-infected adults in South Africa, following childhood PCV13 immunization.

WHO recommends three doses, either as a three-dose primary series (3p + 0) or as a two-dose primary series and a booster dose at 9–15 months of age (2p+1 schedule), depending on locally relevant factors. A two-dose primary series at 6 and 14 weeks induces similar immunogenicity compared to a three-dose primary series schedule at 6, 10, and 14 weeks; with the advantage of using the third dose as a booster dose which is required in some settings to sustain durability of protection and might be more effective in preventing nasopharyngeal colonization.

By 2018, PCV had been introduced in the national immunization programme in 138 countries and a further three countries are planning on vaccine introduction in 2018. High costs of PCV ($20 per dose in middle-income countries and approximately $3.50 per dose in LICs) pose a significant burden on national health budgets. As a result, reduced dosing schedules are being evaluated in settings with low pneumococcal circulation and continued monitoring for prevalence of *S. pneumonia* colonization.

PCV vaccination is also licensed for use in high-risk adults, including HIV-infected adults and the elderly; for whom a single dose of vaccine appears to be adequate. In HIV-infected adults not on ART, PCV vaccination was associated with a 74% reduced risk for recurrent vaccine-serotype IPD; whereas in the elderly (>50 years of age), PCV immunization was associated with an efficacy of 46% (95%CI 22–63%) at preventing vaccine-serotype pneumonia. Currently, broader spectra for PCV sero-coverage are being developed, which would include 15–25 serotypes.

Pneumococcal polysaccharide vaccines (PPV) exist and contain 23 serotypes (PPV-23). PPV is not immunogenic for most serotypes in children <2 years of age and protective immunity in older individuals may be limited or short lived as it stimulates a T-cell independent response which does not result in memory induction. Furthermore, repeat dosing with PPV is associated with immune hyporesponsiveness. PPV-23 could be used 1–2 months after priming with PCV in specific groups at risk for severe infection such as sickle cell disease, chronic renal failure, immunosuppression, cerebrospinal fluid leaks, asplenia, diabetes mellitus, or chronic liver, heart, or lung disease. This could increase coverage for those serotypes not included in PCV. In resource-limited settings there is limited evidence to support routine immunization of the elderly and high-risk populations with PPV-23.

Rotavirus

Rotaviruses are responsible for most cases of severe diarrhoeal disease in young children globally in the absence of vaccination. Approximately 215,000 deaths per year were due to vaccine-preventable rotavirus infections in children <5 years of age, according to WHO estimates from 2013. Almost every child is infected by rotaviruses by 3–5 years of age. The median age of primary infection ranges from 6 to 9 months in LIC, and 2–5 years in HIC. Most (90%) cases of rotavirus-associated deaths occur in LIC and are associated with lack of healthcare facilities. Rotaviruses are responsible for up to a half of diarrhoea related hospital admissions and approximately 40% of all-cause diarrhoea deaths in LMIC in the absence of vaccination.

Rotavirus vaccine introduction has contributed to a 50% decrease in diarrhoeal deaths in children under 5 in Mexico and a significant decline in rotavirus gastroenteritis hospitalizations in the United States.

Currently, two oral, live, attenuated rotavirus vaccines are available: Rotarix™ (derived from a monovalent strain of human rotavirus) and RotaTeq™ (a re-assorted pentavalent bovine-human rotavirus). Both vaccines are safe and highly effective in preventing severe gastrointestinal disease. There is a small increased risk of intussusception (~1–2/100,000 infants) mainly after the first dose of vaccine, but this risk is far outweighed by the benefits of immunization.

In LIC, vaccine efficacy (45–70%) can be lower than in industrialized settings (>90%), however, rotavirus vaccination still results in large numbers of severe gastroenteritis cases and deaths being prevented, because of the high incidence of severe rotavirus disease.

WHO recommends rotavirus vaccine introduction in routine immunization programmes either as a two-dose course (Rotarix™) or a three-dose course (RotaTeq™). In India (Rotavac, Rotasiil), Vietnam (Rotavin-M1) and China (Lanzhou lamb) additional rotavirus vaccines are licensed and produced, however these are not available internationally. The manufacturers recommend that the last dose is given no later than 32 weeks of age, because of theoretical concerns of increased risk for intussusception in older children, however, WHO recommends that the rotavirus vaccines may be given with DTP vaccines until two years of age. In 2018, 94 countries worldwide have introduced rotavirus vaccines. Due to the less than desired vaccine efficacy of current live-attenuated formulations in LMIC, research is ongoing on improved formulation –including the development of subunit systemically administered vaccines.

Measles

The highly contagious measles virus caused epidemics every 2 to 3 years prior to vaccine introduction in 1963. Measles was responsible for 30 million cases of infection and >2 million deaths worldwide on an annual basis in the pre-vaccine era; with more than 95% of individuals infected before reaching 15 years of age. Despite availability of a vaccine since the 1960s, measles remained an important cause of death among children. In 2000 there were an estimated 850,000 deaths from measles, which decreased to an estimated 89,780 by 2016 due to a concerted effort at increasing vaccine coverage rates in routine immunization programmes, as well as undertaking periodic

community-wide supplementary immunization activities. This resulted in more than 20.4 million deaths being prevented between 2000 and 2016.

The age of measles acquisition varies with population immunity and birth rate. Infants and toddlers are at increased risk to acquire measles in LIC, refugee camps, countries with low population immunity, high birth rates, and high population density. Once vaccination coverage improves, a transition occurs from infection during childhood to adolescence or adulthood in cohorts that have not been vaccinated before.

As part of the Global Vaccine Action Plan, the WHO targets measles elimination in 5 of 6 WHO regions by 2020. The definition of elimination is the interruption of transmission in a specific geographical area for ≥12 months and is confirmed after sustaining this for ≥36 months. Once sufficient immunity in the population is reached, measles virus transmission can be interrupted. Due to its high infection rates (R_0), measles requires the highest coverage rates (approximately 90–95%) for herd protection to materialize. Global coverage of the first measles dose by 12 months was 85% in 2016, and even lower (78%) in sub-Saharan Africa and South Asia.

The measles vaccine is a live-attenuated virus and may be given as a monovalent vaccine or in combination with rubella, +/– mumps, +/– varicella (MR, MMR, MMRV). The timing of measles vaccine administration depends on coverage, measles incidence, and age at which visits take place in the immunization schedule. For example, countries with weaker immunization programmes reaching suboptimal coverage and high measles virus transmission may offer doses at younger ages (6–9 months) and may need to use supplemental vaccination campaigns. All children should receive two doses of measles vaccine. In countries with ongoing transmission, the first dose should be administered during infancy (at 9 months) and the second dose between 15 and 18 months of age. In areas of high measles and maternal HIV incidence, an addition dose should be given at 6 months to mitigate against the lower transplacental acquisition of measles antibody in children born to HIV-infected women, as well as more severe disease in HIV-infected children. If measles transmission is low, the first dose may be administered at 12 months.

Countries should reach ≥95% coverage with both doses in order to achieve measles elimination. During every contact with health services, children should be evaluated for missed doses and should be vaccinated if necessary.

Rubella

Rubella virus causes an acute, but mild infection and usually affects children and young adults. The importance for public health is related to its teratogenic outcomes. Rubella infection in pregnant women may result in miscarriage, foetal death, or congenital defects. Congenital rubella syndrome (CRS) is associated with ophthalmic (e.g. cataract, glaucoma, microphthalmia); auditory (e.g. sensorineural deafness), cardiac (e.g. patent ductus arteriosus, ventricular septal defects) and craniofacial (e.g. microcephaly) defects. Women of childbearing age who are susceptible to rubella are at highest risk of CRS. In countries that have not introduced rubella-containing vaccines (RCV), the burden of CRS is estimated to be similar as in 1996 when approximately 22,000 children were born with CRS in Africa (95%CI: 6,127–51,472), 46,000 (95%CI: 1,016–168,910) in South East Asia and 12,634 (95%CI: 1,545–21,396) in the Western Pacific Region. Many HIC and some LMIC have introduced rubella vaccination with high coverage rates and have almost reached rubella and CRS elimination.

Rubella vaccine, a live-attenuated vaccine, is mostly based on the RA 27/3 strain. The vaccine is often given as part of combination vaccine with measles (MR), measles and mumps (MMR), ± varicella (MMRV). The main purpose of rubella immunization is to decrease incidence of primary infection in pregnant women and thus protect against CRS in their offspring. When vaccine coverage is low (e.g. only offered in private sector), viral circulation may decline, and result in a shift of primary infection occurring during childhood to adolescence and adulthood. This may result in a paradoxical increase in CRS cases.

WHO recently recommended that countries that are providing two doses of measles vaccine should consider incorporating routine infant immunization against rubella when giving the second dose of measles vaccine. This approach (ideally accompanied by an initial immunization campaign targeting a wide range of adolescents and women of childbearing age) is preferred to the alternative strategy of immunizing adolescent girls and females of childbearing age, as this may be undermined by low vaccine coverage and ongoing circulation of the virus. Due to high response rates of a single dose of rubella vaccination and long-term persistence of protection, a booster dose is not required.

Human papilloma virus

Human papilloma virus (HPV) is a group of viruses, of which at least 13 types could cause precancerous lesions that may progress to cancer. HPV commonly causes infections in the reproductive tract, the majority resulting in asymptomatic infection with spontaneous resolution. HPV is transmitted through sexual contact and most people are infected shortly after the onset of sexual activity. HPV-16 and HPV-18 are responsible for 70% of cervical cancers and precancerous cervical lesions. HPV is also associated with anogenital and oropharyngeal cancers. In LMIC, cervical cancer is the second most common cancer in women. An estimated 630,000 new cases of HPV-related cancer occurred in 2012, of which 84% were cervical cancer. Most deaths due to cervical cancer (85%) were in LMIC.

Three HPV vaccines are licensed and available: bivalent (HPV types 16 and 18), quadrivalent (HPV types 16, 18, 6, and 11), and non-avalent (HPV types 6, 11, 16, 18, 31, 33, 45, 52, and 58) vaccines. All of the vaccines are recommended to be administered prior to onset of sexual activity and prevent cervical cancer. The quadrivalent and non-avalent HPV vaccines also protect against common genital wart-causing strains.

WHO recommends inclusion of HPV vaccination in all national immunization programmes. HPV is indicated primarily for girls aged 9–14 years with a two-dose schedule to prevent cervical cancer. Vaccination of females ≥15–25 years or males is only recommended if feasible, cost-effective, and affordable. The risk of HPV infection in boys is reduced if a vaccination coverage >80% is achieved in girls. A three-dose vaccine schedule in women 15–25 years of age with precancerous cervical lesions has recently been shown to be effective in preventing progression to cervical cancer; but not in older women.

By May 2018, 80 (41%) countries had introduced HPV vaccine in their national immunization programme. The implementation of HPV vaccine in immunization programmes is technically and logistically a departure from the services used for EPI vaccination; as the targeted age groups rarely visit childhood immunization centres. Introduction of HPV vaccine should be part of a larger comprehensive strategy for the prevention of cervical cancer and other HPV-related diseases. As part of this strategy, health personnel should be trained, people should be educated about behaviours leading to HPV acquisition, and women should be informed about screening, diagnosis, and treatment of precancerous lesions and cancer. Also, availability of cervical

cancer screening and treatment of invasive cancers should be addressed. It is important that HPV vaccination does not replace the need for screening during adulthood, because older age groups will not have been vaccinated and the vaccine does not prevent all high-risk HPV types.

Vaccines recommended when living in certain regions

Japanese encephalitis

Japanese encephalitis virus (JEV) is the main cause of viral encephalitis in Asia. Japanese encephalitis (JE) is a zoonotic viral disease. The virus amplifies in pigs and wading birds and is spread primarily by *Culex* mosquitoes. The most important vector, *Culex tritaeniorhynchus*, breeds in flooded rice fields and water pools and usually bites at night time. Elimination of JEV is not possible due to the animal reservoirs. However, control is possible by vaccination of humans in endemic regions. Because viraemia in humans is too low for transmission, they are a dead-end-host. The season for JEV transmission is from April/May to September/October in temperate climates. In the tropics or subtropics, transmission increases during the rainy season. However, during the dry season transmission may be possible if irrigation systems are in place.

The burden of JE is unclear due to lack of surveillance and difficulty of laboratory confirmation. An estimated 68,000 clinical cases occur every year, with approximately 13,600 deaths. Most cases of JE are mild, but case-fatality rates can be as high as 30% in case of encephalitis. Encephalitis may result in permanent neurologic or psychiatric sequelae in 30–50% of cases. 24 countries are endemic for JE, all of them in the WHO Southeast Asia and Western Pacific regions. People living in rural areas, close to rice fields, and in close proximity to pigs are at increased risk for JE.

Safe and effective vaccines can prevent JE. In all areas where JE disease is a public health concern, JE vaccination is recommend as part of the national immunization schedule. Three JE vaccines are currently available and recommended: inactivated Vero cell-derived vaccines, live-attenuated vaccines, and live recombinant (chimeric) vaccines. Inactivated Vero cell-derived vaccines are recommended as 2-dose (from 6 months of age), whereas live-attenuated (from 8 months of age) and live recombinant vaccines (from 9 months of age) are given as a single dose.

JE vaccination does not induce herd immunity, so high coverage is required in populations at risk. If this is sustained, elimination of viral circulation can occur. A common immunization strategy in JE endemic areas should encompass a one-time campaign in the primary target population (depending on local epidemiology, commonly <15 years), and inclusion of JE vaccine into the national routine childhood immunization schedule. If the burden of disease in older age groups is high, they may also be targeted for vaccination.

Yellow fever

Yellow fever (YF) is a viral disease transmitted by mosquitoes and affects humans and other primates. The virus is endemic in tropical areas in Africa (29 countries) and Central and South America (13 countries). YF can cause acute viral haemorrhage and is named ‘yellow’ due to jaundice being one of the clinical manifestations. The burden of YF is estimated at 84,000–170,000 cases of severe disease and 29,000–60,000 YF deaths in Africa in 2013. YF may be transmitted in sylvatic (jungle), intermediate, or urban cycles. Sylvatic YF is common in primates, other than humans. In humans it is predominantly seen in Central and South America in young men working in close proximity to the rain forest. Intermediate YF transmission occurs in humid settings in Africa infecting both humans and non-human primates. Urban YF transmission is associated with large epidemics, where *Aedes aegypti* transmits the virus from person-to-person in densely populated areas. Only a minor proportion of individuals who are infected with the virus develop severe symptoms. Among those with severe symptoms, around half die in 7–10 days. Most (>90%) of YF deaths occur in sub-Saharan Africa, with children being at highest risk.

YF can be prevented by a very effective live-attenuated vaccine. A single dose is adequate to confer sustained lifelong immunity against disease. Infant immunization at the same time as measles vaccination is recommended in YF-endemic countries. Additionally, mass immunization campaigns are recommended in settings with high risk of YF and low vaccination coverage.

Tick-borne encephalitis

Tick-borne encephalitis can cause viral infections of the central nervous system. It is endemic in in eastern, central, and northern Europe, northern China, Mongolia, and the Russian Federation. An estimated 10,000–12,000 clinical cases of tick-borne encephalitis are reported every year. In general, infections occur after tick bites during outdoor activities in the forest.

Vaccination using a formalin-inactivated strain of the virus could protect against tick-borne encephalitis if administered as a three-dose primary series. The need for public immunization varies between geographical areas, depending on the local incidence of tick-borne encephalitis. If the disease is highly endemic (incidence of clinical disease ≥5 cases/100,000 population per year prior to vaccination), WHO recommends vaccination in all age groups.

Vaccines recommended in high-risk populations

Typhoid

The enteric bacterium *Salmonella enterica* serovar Typhi (*S. Typhi*) causes typhoid fever, a life-threatening infection. Transmission occurs via the faecal-oral route through contaminated food and water. The only known reservoir of *S. typhi* are humans. *S. typhi* infection is associated with lack of safe water and sanitation, poor hygiene when handling food, high population density, overcrowding, low socioeconomic status, and low literacy rates.

LMICs have a substantial burden of typhoid disease, which has however decreased with improved sanitation, hygiene, and access to potable water. Globally, the burden ranges from 11 to 21 million cases and around 128,000–161,000 deaths annually. The incidence of typhoid fever varies significantly between countries in Asia and Africa. Children have the highest incidence of typhoid fever, in particular between 5 and 15 years of age.

Three types of typhoid vaccines are available: typhoid conjugate vaccine (TCV), unconjugated Vi polysaccharide (ViPS), and live-attenuated Ty21a vaccines. The licensed TCV vaccines are preferred at all ages due to their increased immunogenicity, use in younger children, and longer duration of protection. TCV should be prioritized in countries with a high burden of typhoid disease or antimicrobial resistance. Countries could also consider routine use of ViPS vaccine in individuals ≥2 years of age, and Ty21a

vaccine for individuals ≥6 years of age depending on costs and programmatic issues.

Cholera

Cholera is a diarrhoeal disease that may cause rapid dehydration and is caused by toxigenic serogroups of the *Vibrio cholera* bacterium. Risk factors for cholera include lack of clean drinking water, poor sanitation, and poverty. In the past, cholera outbreaks caused millions of cases and hundreds of thousands of deaths. Some LMICs still experience major epidemics and cholera remains an important public health issue in many countries; especially in areas with fragile infrastructure during times of unrest/wars, natural disaster, and in refugee camps. Cholera endemicity is defined as the detection of cholera due to local transmission in a specific area in the last 3 years. A country with one or more endemic regions is classified as a cholera-endemic country. A cholera outbreak is defined as the detection of at least one confirmed case of cholera and proof of local transmission. Outbreaks occurred in Haiti and Yemen between 2010 and 2017 and regions in Asia and sub-Saharan Africa have shown surges in endemic disease. The burden of cholera is highest is Africa and southern Asia (99% of all cases). Risk of cholera is increased in humanitarian emergencies and around 50% of the cases and deaths occur in children under 5.

Two types of oral cholera vaccines are available: WC-rBS, a killed whole cell monovalent vaccine with a recombinant B subunit of cholera toxin; and WC, a killed modified whole cell bivalent vaccine without the B subunit. WHO recommends that, along with other preventive measures, cholera vaccines should be used where disease is endemic and considered in areas at risk for outbreaks. The parenteral cholera vaccine has poor efficacy and is no longer recommended.

Meningococcal disease

Meningococcal meningitis is a serious bacterial infection affecting the membrane around the brain and spinal cord. If untreated, case-fatality rates may go up to 50% or could cause severe neurological sequelae (10%). The *Neisseria meningitidis* bacteria may be responsible for meningococcal meningitis. There are 12 *N. meningitidis* serogroups, of which A, B, C, W,

X, and Y can cause epidemics. The disease mainly affects children and young people and can occur as single cases, clusters, or epidemics.

Depending on the serogroup, the geographical burden of disease differs; however, due to inadequate surveillance, estimates of global burden of meningococcal disease are uncertain. The meningitis belt, an area in sub-Saharan Africa stretching from Senegal to Ethiopia (26 countries), has the highest burden of meningococcal disease. Risk of meningococcal disease is increased when the nasopharyngeal mucosa is damaged in the dry season due to dust winds, colds nights, and upper respiratory tract infections in combination with transmission facilitated by overcrowding.

Vaccines against specific serogroups are administered routinely or during outbreaks. There are three types of vaccines available: polysaccharide vaccines, conjugate vaccines, and protein-based vaccines. Polysaccharide vaccines can be bivalent (serogroups A, C), trivalent (A, C, W), or tetravalent (A, C, Y, W) and are usually given during outbreaks. However, protection decreases after several years and, apart from serogroup A, are poorly immunogenic in infants. These vaccines have largely been replaced by polysaccharide-protein conjugate vaccines (Men-CV)

Men-CVs confer longer-lasting immunity, induce herd immunity, and prevent nasopharyngeal colonization—hence also interrupting transmission of the bacteria. They are immunogenic in children from 1 year of age onward. Currently available vaccines include monovalent (A or C) or quadrivalent (A, C, Y, W) conjugate vaccines, also in combination with HibCV (HibMenC). Epitope specific protein based vaccines are available for serotype B.

Mass immunization campaigns using meningococcal A conjugate vaccine in the meningitis belt have contributed to a 58% decline in the meningococcal meningitis A cases and 60% reduction in the risk of epidemics since 2010. More than 280 million individuals have been vaccinated in 21 countries located in the belt as of November 2017. However, epidemics due to meningococcal serogroups W, X, and C still occur, with a burden of 30,000 cases of meningococcal disease reported annually.

Hepatitis A

Hepatitis A virus (HAV) can cause mild to severe liver disease. Most HAV infections are asymptomatic during childhood, but symptoms commonly occur when infected later in life. The virus is transmitted via the faecal-oral route by ingesting contaminated food or water, or through close contact with an infected person. The risk of hepatitis A infection is associated with socioeconomic factors such as income, availability of clean water, sanitation, and hygiene. The burden of hepatitis A is estimated to have increased from 117 million to 126 million in the period between 1990 and 2005, with deaths due to hepatitis A rising from 30,283 to 35,245 in the same period.

In most LICs, anti-HAV antibodies are present in 90% of children by the time they reach 10 years of age. Asymptomatic childhood infection induces immunity and few are susceptible to symptomatic disease during adolescence and adulthood. Yet, middle-income countries often have intermediate HAV endemicity. As sanitation and hygiene improve, a number of individuals will not be infected during childhood, but may be at risk of symptomatic infection during adult life. Hence, when HAV endemicity decreases, the number of clinical cases of hepatitis A will increase. In HIC, the prevalence of anti-HAV antibody is low, as is circulation of the virus.

Hepatitis A can be prevented by safe and effective vaccines. Inactivated and live-attenuated vaccines are available and result in long-lasting protection in adults and children. WHO recommends HAV vaccination to be integrated in the national routine immunization schedule for children ≥1 years of age in countries with high incidence of acute hepatitis A, who are transitioning from high to intermediate endemicity if it is cost-effective. Large scale vaccination in highly endemic countries is not recommended.

Rabies

Rabies is a viral disease transmitted from animals to humans through bites or scratches. Dogs are responsible for 99% of rabies transmissions and are the main source of human rabies deaths. Rabies can also be found in domestic animals and wildlife. Rabies is present in more than 150 countries and causes an estimated 59,000 human deaths and approximately 3.7 million disability-adjusted life years lost on an annual basis. Once clinical signs appear, it almost always results in death due to acute progressive encephalitis.

The majority of cases occur in Asia and Africa, with 40% in children under 15 years of age.

The principle strategy for rabies control consists of vaccination of dogs and prevention of dog bites.

Vaccination of humans and immunoglobulins exist, but they are usually not accessible to people living in remote areas. Vaccination is effective before or after exposure to rabies.

Dengue

Dengue virus is transmitted to humans by female mosquitoes. The primary vector is the *Aedes aegypti* species and to some extent the *Aedes albopictus* species. People with dengue infection develop flu-like illness, but infected individuals may develop severe fatal complications. Four dengue serotypes exist: DEN-1, DEN-2, DEN-3, and DEN-4. Infection will result in lifelong protection against the specific serotype, but there is limited and temporary cross-immunity to other serotypes. Instead, new infection by other serotypes increases the risk of developing severe dengue. In Asia and Latin America, severe dengue is a significant cause of serious illness and death among children.

In recent decades, dengue incidence has increased rapidly from 0.4–1.3 million in 1995–2005, to 2.2 million in 2020 and 3.2 million in 2015. Around 50% of the global population is at risk for dengue infection. Dengue is commonly found in the tropics and subtropics, predominantly in urban and semi-urban regions. Local risk factors include rainfall, temperature, and urbanization.

Prevention and control of dengue can be achieved by successful vector control. The first dengue vaccine (Dengvaxia, CYD-TDV) was licensed in 2015 for use in persons 9–45 years of age living in endemic settings. Vaccination is not recommended when seroprevalence is <50% in the targeted age group. Recent findings showed that vaccination resulted in more severe disease and hospitalizations from dengue in individuals previously seronegative at time of first vaccination compared to unvaccinated individuals. WHO now recommends a ‘pre-vaccination screening strategy’, in which only dengue-seropositive persons are vaccinated, the logistics of which might prove challenging.

Currently, other tetravalent live-attenuated vaccines are being developed in phase III clinical trials. Development of new vaccine candidates based on subunit DNA and purified inactivated virus platforms is also ongoing.

Vaccines recommended in certain immunization programmes

Mumps

Mumps is a viral infection spreading between humans via airborne droplet or direct contact. It is also known as infectious parotitis because of infection in the salivary (parotid) glands. Infection with mumps during childhood causes mild disease, primarily affecting children aged 5–9 years. Infection in adults may occur and may cause more serious complications such as meningitis (up to 15% of cases), orchitis, or deafness.

Mumps incidence infection rates ranges from 100 to 1,000 cases per 100,000 population in absence of vaccination. Vaccination against mumps resulted in a reduction of incidence. The vaccine has been incorporated in childhood immunization schedules of many middle- and high-income countries, usually combined with measles and rubella (MMR) immunization. WHO recommends mumps vaccination in countries with a high burden of mumps, an established routine vaccination programme and persistent high vaccination coverage with measles and rubella of more than 80%. Similarly to rubella vaccine, insufficient coverage of mumps vaccination during childhood can cause a shift in age distribution to infection at older ages, when the course of disease is more severe.

Seasonal influenza

Influenza viruses can cause seasonal influenza epidemics as well as sporadic cases and outbreaks outside of the influenza season. Symptomatic influenza attack rates of 5–10% have been reported in adults and 20–30% in children; with as many of 50% of individuals demonstrating sero-conversion following seasonal epidemics. Persons with increased risk of severe disease and complications include pregnant women, children under 5 years of age, the elderly, and individuals with chronic conditions or immunosuppression. Healthcare workers are also at increased risk of influenza acquisition due to increased exposure and higher risk of spreading disease. In tropical areas,

influenza occurs throughout the year, whereas temperate climates typically show a seasonal peak in incidence during winter.

WHO recommends protection of high-risk groups against severe influenza illness and death by annual influenza vaccination. These risk groups include pregnant women who are considered as the highest priority group, children aged 6 months to 5 years, elderly persons (>65 years), individuals with chronic medical conditions and healthcare workers. Vaccine effectiveness is dependent on the closeness of the genetic match between the vaccine and wild-type circulating strains, and could vary from 0% to 70% year-by-year. Influenza viruses continuously evolve over time. Therefore, it is important to monitor wild-type circulating viruses and update the vaccine formulation bi-annually.

Vaccination of pregnant women has been shown to protect the women (50% efficacy), as well as their infants <6 months of age (50% efficacy) against symptomatic influenza infection. Furthermore, a 20% reduction in all-cause pneumonia hospitalization has been reported in randomized controlled trials in the young offspring of vaccinated pregnant women.

A number of challenges exist with influenza vaccine development, including dependency on egg-culture for vaccine virus propagation, uncertainty of predicting the drifts in the upcoming wild-type strains; and short duration of protection. This is being addressed by the search for universal influenza vaccines, at least against group A viruses, which cause the most severe seasonal epidemics. Furthermore, the threat of virus shift resulting in a pandemic from a new novel strain of influenza virus, warrants countries to have a pandemic plan in place, which includes how they will procure and dispense vaccines should such a pandemic arise. Ideally, a ‘universal’ influenza virus vaccine should also confer cross-protection against pandemic virus; however, this is less likely to be realized over the next 5–10 years.

Varicella and Zoster vaccines

Varicella zoster virus (VZV) is a highly contagious virus that causes varicella (chickenpox), mostly during childhood, and herpes zoster (shingles) later in life. In absence of vaccination, almost all individuals are infected by the time they reach mid-adulthood. VZV epidemiology depends on climate, population density, and risk of exposure. The disease is generally

mild, but severe disease may develop (e.g. cellulitis, pneumonia, encephalitis, death), especially in infants and immunocompromised individuals. After primary infection, the virus may establish latency in sensory nerve ganglia and reactivate later in life, causing herpes zoster. Typically, herpes zoster affects elderly and immunocompromised persons.

Multiple formulations of live-attenuated varicella vaccine based on the OKA strain are available. It may be given as a monovalent dose or in combination with the MMR vaccine (MMRV). Routine immunization against VZV is employed in many resource-rich countries. Coverage of ≥80% is required to prevent a shift of infection to older age groups. Herpes zoster vaccination is currently not recommended due to limited availability of data. The recently licensed zoster vaccine in the United States showed efficacy for preventing herpes zoster of 70% (95%CI 54–81%).

Vaccination of special populations

Refugees, asylum-seekers, migrants

Refugees, asylum-seekers, and migrants may have increased risk for acquisition of infectious diseases because of mass population movement; limited access to clean water, shelter, and sanitation. Children in these populations also have increased risk of respiratory infections, diarrhoea, and skin infections. Despite high vaccination coverage in the countries of origin, young children may have missed vaccinations due to interruptions by civil unrest and war. Also, individuals may not have access to healthcare, health insurance, or registration in the health system.

Vaccination of refugees, asylum-seekers, and migrants may be challenging if populations are on the move. All vaccines should be administered according to the national routine immunization schedule of the country of stay. Priority vaccines include measles, mumps, rubella, and polio. Furthermore, refugees might benefit from early vaccination against cholera, especially if in or from settings where cholera is endemic.

HIV-infected

HIV-infected individuals require special consideration because of increased susceptibility to severe disease from infections, including TB, pneumococcal

disease, and measles. HIV-infected persons have reduced immunogenicity following vaccination when administered in advanced stages of AIDS, but HIV-positive adults and children are able to mount moderate to good vaccine responses. Waning of immunity may result in reduced protection over time for some vaccines. Individuals may require administration of booster doses, especially if children were vaccinated prior to initiation of antiretroviral therapy.

HIV-exposed-uninfected children and HIV-infected children, who were vaccinated when already on antiretroviral therapy, have similar responses to vaccination as HIV-unexposed children. Except for BCG, asymptomatic HIV-infected individuals should receive all EPI vaccines at the indicated ages. BCG vaccination should be given to children receiving ART, who are clinically well and immunologically stable. In neonates newly diagnosed with HIV, BCG vaccination should be delayed until antiretroviral therapy is initiated and the infant is confirmed immunologically stable. If an HIV-infected child has an acute severe illness, vaccination should be delayed, as for any severely ill child.

Healthcare workers

Healthcare workers have increased risk for acquisition of vaccine-preventable diseases. It is important that healthcare workers can maintain delivery of healthcare services during epidemics. They should also avoid spreading disease to vulnerable patient groups. Appropriate vaccination for BCG, hepatitis B, polio, diphtheria, measles, rubella, meningococcal, influenza, varicella, and pertussis is recommended, especially if at increased risk of exposure.

Emerging vector-borne diseases

Vector-borne diseases require vectors, such as mosquitoes, ticks, or sand flies for transmission between hosts or rely on other animals as reservoirs or carriers. They account for an estimated 17% of all infectious diseases and have the highest burden in tropical and subtropical areas. Since 2014, major outbreaks of chikungunya, dengue, malaria, Zika, and Ebola have affected populations and claimed many lives around the world.

Ebola

Ebola virus disease, previously called Ebola haemorrhagic fever, is a severe and often fatal illness in humans. The virus was first discovered in 1976, during outbreaks in South Sudan and the Democratic Republic of Congo. The largest outbreak occurred in 2014–2016 in West Africa. Five Ebola virus species have been identified, of which the Zaire species was responsible for the recent outbreak.

The virus is transmitted from wild animals to humans and can spread between humans. Case-fatality rates vary between 25% and 90%. Treatment consists of supportive care, including rehydration, but no cure is available.

An experimental vaccine (rVSV-ZEBOV) was shown to protect against the virus in a trial in Guinea. WHO supports Ebola vaccination in high-risk populations in the Democratic Republic of Congo using a ring vaccination strategy. This involves tracing contacts of recently confirmed cases as well as contacts of contacts. Other preventive measures should target reducing the risk of transmission and containment of outbreaks.

Maternal immunization

Despite considerable progress in the reduction of under-5 mortality, a significant number of deaths still occurs during the neonatal period and progress has been slower in this age group. Of 700,000 deaths in the first month of life, 28% is attributable to infectious diseases. It is therefore important to protect infants already in the first weeks of life. Yet, primary immunization schedules usually start at 6 weeks of age in most LMIC. Maternal immunization could reduce the morbidity and mortality from infectious diseases at birth. Transplacental transfer of maternal antibodies successfully protects newborns against tetanus, influenza, and pertussis after maternal immunization during pregnancy. New maternal vaccines currently being developed include immunization for group B streptococcus and respiratory syncytial virus.

Universal access to immunization

In 2012, the World Health Assembly adopted the Global Vaccine Action Plan (GVAP), which sought to deliver universal access to immunization. This was expected to avert 24.6–25.8 million vaccine-preventable deaths in this decade alone. Key to the GVAP agenda is ensuring equitable access to immunization services, regardless of geographic location, ethnic or religious beliefs, and socioeconomic background. Efforts to achieve this include Gavi's support for the introduction of new vaccines in the poorest countries by subsidizing vaccine procurement. As a result, the timelines by which PCV and rotavirus vaccines were introduced into public immunization programmes have been drastically reduced and 20 out of 25 countries within the meningitis belt have introduced the MenA conjugate vaccine. Furthermore, steady progress has been made in the field of polio and tetanus eradication/elimination.

However, immunization equity analyses still show that economic disparities, place of residence, and education status continue to determine a child's vaccination status. In many regions, vaccine supply remains a challenge with 16 countries reporting one or more stock-outs of vaccines lasting for more than one week in 2017. As a result, an estimated 19.9 million (15%) of the annual birth cohort of 132 million are either not vaccinated or underimmunized. The continued presence of vaccine-preventable deaths is primarily due to low immunization coverage rates. Understanding the mediators contributing to underimmunization and non-vaccination of children is, therefore, fundamental to derive the full benefits that vaccines stand to offer at an individual and societal level.

In this context, there is a need for community-engagement and a bottom-up approach that includes societal, cultural, and religious considerations to address the barriers contributing to underimmunization of children.

11.4

Principles of infectious disease control

Robert J. Kim-Farley

Principles of infectious disease control: overview

Infectious diseases remain a leading cause of morbidity, disability, and mortality worldwide. Lower respiratory infections result in the highest mortality for any category infectious diseases and comprise the fourth leading causes of death worldwide (WHO 2020). Their control is a constant challenge that faces health workers and public health officials in both industrialized and developing countries. Only one infectious disease, smallpox, was eradicated and stands as a landmark in the history of the control of infectious diseases. The international community is now well down the path towards eradication of poliomyelitis and dracunculiasis (Guinea worm infection). Other infectious diseases, like malaria and tuberculosis, foiled eradication attempts or control efforts and are re-emerging as increasing threats in many countries, including both developing and developed countries. Some infectious diseases, such as tetanus, will always be a threat if control measures are not maintained. Newer infectious diseases, like AIDS and COVID-19, demonstrate the truth of McNeill's statement that infectious disease will remain 'one of the fundamental parameters and determinants of human history'. The history of infectious diseases is an exciting story in itself and readers interested in the subject are referred to McNeill (1976) or the comprehensive work on the history of human diseases (Kiple 1993).

This chapter provides a global and comprehensive view of infectious disease control through examination of the magnitude of disease burden, the chain of infection (agent, transmission, and host) of infectious diseases, the varied approaches to their prevention and control, and the factors conducive to their eradication as well as emergence and re-emergence. Although this

chapter provides many examples of infectious diseases that illustrate modes of transmission and approaches to infectious disease control, this chapter does not attempt to be comprehensive in listing all infectious diseases. Detailed recommendations on control measures for any specific disease are outlined periodically in the updated reports of the American Public Health Association, *Control of Communicable Diseases Manual* (Heymann 2014), the comprehensive two-volume work *Mandell, Douglas and Bennett's Principles and Practice of Infectious Diseases* (Bennett et al. 2020), and the textbook *Infectious Diseases* (Gorbach et al. 2004). For readers specifically interested in paediatric infectious diseases there is the comprehensive two-volume *Feigin and Cherry's Textbook of Pediatric Infectious Diseases* (Cherry et al. 2014); for infectious diseases in emergency medicine settings there is the textbook *Infectious Diseases: Emergency Department Diagnosis & Management* (Red and White Emergency Medicine Series) (Slaven et al. 2006); and for tropical infectious diseases there is *Tropical Infectious Diseases: Principles, Pathogens, and Practice* (Guerrant et al. 2011). A comprehensive treatment of the worldwide distribution of infectious diseases is provided in *Atlas of Human Infectious Diseases* (Wertheim et al. 2012). The Centers for Disease Control and Prevention (CDC) publishes up-to-date disease surveillance information for the United States and recommendations for control measures in the *Morbidity and Mortality Weekly Reports* and provides annual summaries of notifiable infectious diseases in the *Summary of Notifiable Diseases, United States* (CDC 2016). Many other countries have similar types of publications. The World Health Organization (WHO) publishes worldwide surveillance information and recommendations for control measures in the *Weekly Epidemiological Record*. A more detailed background on infectious agents as determinants of health and disease is provided in Chapters 8.11–8.15.

Definitions of infectious diseases and their control

Infection occurs when an infectious agent enters a body and develops or multiplies. *Infectious agents* are organisms capable of producing inapparent infection or clinically manifest disease and include bacteria, rickettsia, chlamydiae, fungi, parasites, viruses, and prions. An *infectious disease*, or communicable disease, is an infection that results in clinically manifest disease. Infectious disease may also be due to the toxic product of an

infectious agent, such as the toxin produced by *Clostridium botulinum* causing classical botulism. As this is a textbook of public health, the infectious diseases considered are those that manifest in human hosts and are a result of the interaction of people, animals, and their environment. Infectious diseases may be due to infectious agents exclusively found in human hosts such as rubella virus, in the environment such as *Legionella pneumophila*, or primarily in animals such as *Brucella abortus*.

Control of infectious diseases refers to the actions and programmes directed towards reducing disease incidence (new infections), reducing disease prevalence (infections in the community at any given point in time), or completely eradicating the disease. Control aimed at reducing the incidence of infectious disease or their risk factors can be considered as *primary prevention* of infectious disease. Primary prevention protects health through individual and community-wide measures, including such actions as maintaining good nutritional status, keeping physically fit, immunizing against infectious diseases, providing safe water, and ensuring the proper disposal of faeces.

Control aimed at reducing the prevalence by shortening the duration of infectious disease can be considered as *secondary prevention* of infectious disease. Secondary prevention corrects departures from good health through individual and community-wide measures, including such actions as screening that result in early detection of disease, prompt antibiotic treatment, and ensuring adequate nutrition. It should be noted that such control efforts in secondary prevention in a group of infected individuals may also result in primary prevention in uninfected people. For example, prompt and specific drug therapy for tuberculosis patients resulting in sputum conversion to culture-negative status renders them no longer a source of infection to others and treatment of HIV-positive pregnant women reduces transmission of HIV to their newborns.

Control aimed at reducing or even eliminating long-term impairments of infectious disease can be considered as *tertiary prevention* of infectious disease. Tertiary prevention reduces or eliminates disabilities, minimizes suffering, and promotes adjustment to permanent disabilities through such actions as providing orthopaedic appliances and its associated rehabilitation for victims of poliomyelitis, counselling and vocational training, and prevention of opportunistic infections. For example, the prevention of

opportunistic infections in HIV infection can be considered as tertiary prevention (Osterholm and Hedberg 2005).

Global burden of infectious diseases

A WHO analysis of the global burden of disease estimates that infectious diseases caused 12.3 million deaths, accounting for 21% of total global mortality of 56.8 million deaths in 2008 (WHO 2008). Five diseases—lower respiratory infections, AIDS, diarrhoeal diseases, tuberculosis, and malaria—account for some 81% of the total infectious disease burden. Most infectious disease (some 97%) deaths occur in the economically developing group of countries where approximately a quarter of deaths are due to an infectious cause (WHO 2008). However, infectious diseases are also of importance in more developed countries. In the United States, for example, influenza still ranks as an important cause of death.

The current magnitude of morbidity and mortality due to infectious diseases worldwide is highlighted by the the Institute for Health Metrics and Evaluation (IHME) as follows (http://www.healthdata.org/results/gbd_summaries):

Acute lower respiratory infections, including pneumonia and influenza, cause some 489 million episodes of illness each year and result in 2.49 million deaths annually. These infections are the highest cause of infant and child mortality in developing countries.

HIV newly infects about 1.9 million people each year and causes an estimated 864,000 AIDS-related deaths annually, down from a peak of 2.2 million that occurred in the mid-2000s. There were some 36.8 million people living with HIV in 2019.

Diarrhoeal diseases are also a major cause of morbidity and mortality in infants and children in developing countries. Each year there are some 6.6 billion episodes with 1.5 million deaths due to diarrhoeal disease, of which the vast majority are among children under 5 years of age.

Tuberculosis, caused by *Mycobacterium tuberculosis*, infects some 1.82 billion people, about one-fourth of the world's population, and is the leading cause of death of a single pathogen. It was estimated that 8.5 million people who developed clinical disease and 1.18 million died of tuberculosis in 2019.

Malaria was estimated to cause some 231 million cases and 643,000 deaths in 2019, mainly in children under the age of 5 years.

Influenza epidemics, caused by influenza A and B viruses, are estimated to kill from 500,000 to 1 million people worldwide each year. Influenza pandemics, of which there were three in the last century, have killed as many as 40–50 million people (as occurred during the 1918–1919 'Spanish flu' pandemic) and, as with the current concerns regarding highly pathogenic avian influenza H5 viruses, continue to threaten to re-emerge as a major public health emergency.

• *The COVID-19 Pandemic*, ongoing at the time of writing this chapter, has become the most important infectious disease public health pandemic in the last 100 years. Deaths from COVID-19 are projected to exceed 3.5 million deaths worldwide by mid-2021.

Chain of infection: agent, transmission, and host

The chain of infection is the relationship between an infectious agent, its routes of transmission, and a susceptible host. The prevention and control of infectious diseases depend upon the interaction of these three factors that may result in the human host clinically manifesting disease.

Agent

The infectious agent is the first link in the chain of infection and is any microorganism whose presence or excessive presence is essential for the occurrence of an infectious disease. Examples of infectious agents include the following:

Bacteria: for example, spirochetes and curved bacteria such as *Borrelia burgdorferi* causing Lyme disease; Gram-negative rods such as *Yersinia pestis* causing plague; Gram-positive cocci such as *Streptococcus pyogenes*, group A causing erysipelas; and Gram-positive rods such as *Mycobacterium tuberculosis* causing tuberculosis.
Rickettsiae: for example, *Rickettsia ricketsii* causing Rocky Mountain spotted fever, and *Rickettsia prowazekii* causing epidemic louse-borne typhus fever.
Chlamydiae: for example, *Chlamydia psittaci* causing psittacosis (ornithosis), and *Chlamydia trachomatis* causing trachoma and genital infections.
Fungi: for example, *Trichophyton schoenleinii* and *Microsporum canis* causing tinea capitis (ringworm of the scalp), and *Trichophyton rubrum*, *Trichophyton mentagrophytes*, and *Epidermophyton floccosum* causing tinea pedis (ringworm of the foot, or athlete's foot).
Parasites: for example, helminths such as *Trichinella spiralis* causing trichinosis, filaria such as *Brugia malayi* causing filariasis (which may lead to the symptom of elephantiasis), nematodes such as *Enterobius vermicularis* causing enterobiasis (pinworm disease), trematodes such as *Clonorchis sinensis* causing clonorchiasis (oriental liver fluke disease), cestodes such as *Taenia solium* causing taeniasis (beef tapeworm disease), and protozoa such as *Trypanosoma cruzi* causing American trypanosomiasis (Chagas' disease).
Viruses: for example, Paramyxoviridae such as measles virus which causes measles, Togaviridae such as rubella virus which causes rubella, Picornaviridae such as the hepatovirus which causes hepatitis A, and arthropod-borne viruses (arboviruses) such as dengue viruses which cause dengue fever.
Prions: small proteinaceous infectious particles which cause such diseases as kuru, Creutzfeldt–Jakob disease (and its variant associated with exposure of humans to the bovine spongiform encephalopathy, or 'mad cow', agent), and the Gerstmann–Straussler–Scheinker syndrome.

There is increasing evidence that some infectious agents, often with cofactors, are associated with human tumours. Examples include: *Schistosoma haematobium* with bladder cancer, *Schistosoma japonicum* with colorectal cancer, *Clonorchis sinensis* with cholangiocarcinoma,

hepatitis B and C viruses with hepatocellular carcinoma, *Helicobacter pylori* with gastric cancer, and human papillomaviruses with cervical cancer. Some 2.2 million cancers cases are attributed by the International Agency for Research on Cancer (IARC) to infectious agents.

Agents can be described by their ability to infect (infectivity), be transmitted (infectiousness), and cause disease (pathogenicity) as well as their ability to cause serious disease (virulence). The *infectivity* of an infectious agent is the extent to which the agent can enter, survive, and multiply in a host and is measured by the ratio of the number of people who become infected to the total number exposed to the agent. Rhinoviruses that can cause the common cold are examples of infectious agents with high infectivity.

The *infectiousness*, or transmissibility, of an infectious agent is the extent to which the agent can be transmitted from one host to another. It is measured by the basic reproduction number (known as R_0) which is determined by the number of secondary infections in a susceptible population that result from a primary infection. Epidemic-prone diseases such as measles and pertussis caused by measles virus and *Bordatella pertussis* bacteria, respectively, are examples of diseases due to highly infectious agents with an R_0 of 10 or greater.

The *pathogenicity* of an infectious agent is the extent to which clinically manifest disease is produced in an infected population and is measured by the ratio of the number of people developing clinical illness to the total number infected. Examples of highly pathogenic infectious agents are the measles virus and the human (α) herpesvirus 3 (varicella-zoster) causing measles and chickenpox, respectively, in which most infected susceptible people will manifest disease.

The *virulence* of an infectious agent is the extent to which severe disease is produced in a population with clinically manifest disease. It is the ratio of the number of people with severe and fatal disease to the total number of people with disease. An example of a highly virulent infectious agent is HIV whereby nearly all untreated people with AIDS will die.

Characteristics of infectious agents that affect pathogenicity include their ability to invade tissues (invasiveness), produce toxins (intoxication), cause damaging hypersensitivity (allergic) reactions, undergo antigenic variation, and develop antibiotic resistance. An example of an infectious agent with high *invasiveness* is the *Shigella* organism which can invade the

submucosal tissue of the intestine and cause clinically manifest shigellosis (bacillary dysentery). An example of an infectious agent that has a high degree of ability to *produce toxins* is the *Clostridium botulinum* organism which can elaborate toxins in inadequately prepared food and cause classical botulism. An example of an infectious agent that is highly *allergenic* is the *Mycobacterium tuberculosis* organism which can cause tuberculosis. An example of an infectious agent that has a high degree of *antigenic variation* is the type A influenza virus which frequently experiences minor antigenic changes—antigenic 'drift'. Influenza A viruses, on an irregular basis, may also undergo a major antigenic change creating an entirely new subtype—antigenic 'shift'. Antigenic shifts that have the characteristic of high transmissibility between people may result in an influenza pandemic when large numbers of individuals are exposed to the new subtype for which they have no prior immunity. An example of an infectious agent that can develop *antibiotic resistance* that challenges control efforts is *Neisseria gonorrhoeae* that has both chromosomally mediated and resistance transfer plasmid-mediated genetic factors for antibiotic resistance.

The *infective dose* of an infectious agent is the number of organisms needed to cause an infection. The infective dose may vary depending upon the route of transmission and host susceptibility. An example of an infectious agent that needs only a very low infective dose (as few as ten organisms) is *Escherichia coli* O157:H7 which causes enterohaemorrhagic diarrhoea. An example of an infectious agent that needs a relatively higher infective dose is *Vibrio cholerae* which requires around 100 million bacteria to be ingested to cause cholera disease in a healthy adult.

Control measures for infectious diseases directed at inactivating the agent are designed according to the type of agent and its reservoirs and sources. For example, an agent like *Vibrio cholerae* can be inactivated through adequate chlorination of the water supply. This is a chemical method for provision of safe water to control cholera. An agent like hepatitis B virus can be inactivated through adequate autoclaving of injection and surgical equipment. This is a sterilization method to control hepatitis B. Details of these and other methods of control directed at the agent are provided in the sections in this chapter on control measures applied to the agent and the environment.

Routes of transmission

Control efforts are often designed to break the routes of transmission, the mechanisms by which infectious agents are spread from reservoirs or sources to human hosts. A *reservoir* of an infectious agent is any person, other living organism, or inanimate material in which the infectious agent normally lives and grows. The *source* of infection for a host is the person, other living organism, or inanimate material from which the infectious agent came. *Horizontal transmission* refers to transmission between individuals whereas *vertical transmission* refers to the specific situation of transmission between parent to offspring (e.g. transplacentally *in utero*, during passage through the birth canal, or through breast milk). The routes of transmission have been summarized in the *Control of Communicable Diseases Manual* (Heymann 2014) as follows:

Direct transmission

Direct and essentially immediate transfer of infectious agents to a receptive portal of entry through which human or animal infection may take place. This may be by direct contact such as touching, biting, kissing or sexual intercourse, or through direct projections (droplet spread) of droplet spray onto the conjunctiva or onto the mucous membranes of the eye, nose, or mouth during sneezing, coughing, spitting, singing or talking (risk of transmission in this manner is usually limited to a distance of about 1 metre or less from the source of infection). Direct transmission may also occur through direct exposure of susceptible tissue to an agent in soil, through the bite of a rabid animal, or transplacentally (Heymann 2014).

Indirect transmission

Vehicle-borne

Contaminated inanimate materials or objects (fomites) such as toys, handkerchiefs, soiled clothes, bedding, cooking, or eating utensils, surgical instruments or dressings; water, food, milk, and biological products including blood, serum, plasma, tissues or organs; or any substance serving as an intermediate means by which an infectious agent is transported and introduced into a susceptible host though a suitable portal of entry. The

agent may or may not have multiplied or developed in or on the vehicle before being transmitted (Heymann 2014).

Vector-borne

Mechanical: includes simple mechanical carriage by a crawling or flying insect through soiling of its feet or proboscis, or by passage of organisms through its gastrointestinal tract. This does not require multiplication or development of the organism.

Biological: propagation (multiplication), cyclic development, or a combination of these (cyclopropagative) is required before the arthropod can transmit the infective form of the agent to humans. An incubation period (extrinsic) is required following infection before the arthropod becomes infective. The infectious agent may be passed vertically to succeeding generations (*transovarian transmission*); *trans-stadial transmission* indicates its passage from one stage of the life cycle to another, as from nymph to adult (for example, ticks infected with *Rickettsia rickettsii* can remain infected with this agent as they pass from nymph to adult stages). Transmission may be by injection of salivary gland fluid during biting, or by regurgitation or deposition on the skin of faeces or other material capable of penetrating through the bite wound or through an area of trauma, often created by scratching or rubbing. This transmission is by an infected non-vertebrate host and not simple mechanical carriage by a vector as a vehicle. An arthropod in either role is termed a vector (Heymann 2014).

Airborne transmission

The dissemination of microbial aerosols to a suitable portal of entry, usually the respiratory tract. Microbial aerosols are suspensions of particles in the air consisting partially or wholly of microorganisms. They may remain suspended in the air for long periods of time, some retaining and others losing infectivity or virulence. Particles in the 1- to 5-micrometre range are easily drawn into the alveoli of the lungs and may be retained there. Not considered as airborne are droplets and other large particles that promptly settle out (see ‘Direct transmission’).

Droplet nuclei: usually the small residues that result from evaporation of fluid from droplets emitted by an infected host (see earlier). They also may be created purposely by a variety of atomizing devices, or accidentally as in microbiology laboratories, abattoirs, rendering plants, or autopsy rooms. They usually remain suspended in the air for long periods.

Dust: the small particles of widely varying size that may arise from soil (e.g. fungus spores), clothes, bedding, or contaminated floors (Heymann 2014).

Control measures for infectious diseases directed at interrupting transmission are designed according to the type of transmission for the agent. Direct transmission of an agent like *Neisseria gonorrhoeae*, for example, can be reduced by using condoms as a barrier method of control of gonorrhoea. Vector-borne transmission of an agent like *Plasmodium falciparum* can be reduced by using a residual insecticide against *Anopheles* mosquitoes as a chemical vector control method for malaria. Airborne transmission of an agent like *Mycobacterium tuberculosis* from sputum-positive pulmonary tuberculosis patients in hospital can be reduced using special ventilation to create a negative pressure in the patient's room as an environmental method of control of tuberculosis. It should be recognized that some infectious agents may have more than one route of transmission. Poliovirus, for example, can be spread via direct transmission through the faecal–oral route and pharyngeal spread, or indirect transmission through contaminated food or other materials. Details of these and other methods of control directed at interrupting transmission are provided in the section on 'Tools for control of infectious diseases' in this chapter.

Host

The human host is the final link in the chain of infection. The infectious agent may enter the host through the following portals of entry.

Respiratory tract: infectious agents can be inhaled into the respiratory tract and will be deposited at different levels of the pulmonary tree according to the size of the aerosol, droplet nuclei, or dust particles. For example, *Mycobacterium tuberculosis* in airborne droplet nuclei, 1–5 μm in diameter, may be inhaled into the alveoli of the lungs of a vulnerable host and result in tuberculosis.
Intact skin: some infectious agents, such as *Necator americanus* which causes hookworm disease, can penetrate the intact skin.
Gastrointestinal tract: an infectious agent, such as *Vibrio cholerae* which causes cholera, may enter via the gastrointestinal tract. People who have a compromised gastric function, such as gastric achlorhydria, may be at increased risk of disease.
Mucous membranes: infectious agents, such as measles viruses, may be deposited on mucous membranes, including the conjunctiva of the eye, by droplet spread or by direct contact with infected people or contaminated objects.
Genitourinary system: some infectious agents, such as *Escherichia coli* which causes urinary tract infections, can enter the urinary tract via an ascending route from the urethra colonized with the organism. Structural abnormalities of the urinary tract and procedures such as urinary catheterization may predispose the host to disease. Sexually transmitted infectious agents, such as *Neisseria gonorrhoeae* which causes gonorrhoea, may enter the male or female genital orifices.
Placenta: certain infectious agents, such as rubella virus which causes congenital rubella syndrome, can cross the placenta resulting in transplacental transmission, a direct route of transmission from the mother to the foetus that is a form of vertical transmission.

Infectious agents also enter the host though *mechanisms that get past the body's natural barriers*, including wounds that break the integrity of the skin or mucous membranes; invasive procedures, parenteral injections, parenteral infusions, blood transfusions, or organ transplants that may introduce an agent into the body; or insect vectors that may inject agents through the skin.

The most important *host factors* regarding developing clinically manifest disease and the severity of disease are immune status and age. Infants, young children, and the elderly are at generally higher risk from more

severe disease due to immaturity or deterioration of their immune systems, respectively. Host co-infections or other health conditions that diminish the immune status, for example, HIV infection or malnutrition, respectively, are also risk factors.

Many *host defence mechanisms* help prevent infection or disease. *Non-specific* host defence mechanisms include the intact skin (by providing a natural barrier), sinus and nasal cilia (by moving mucus and particles towards the nasal cavity), tears, saliva, and mucus (by preventing drying and containing lysozymes and other components with antimicrobial properties), and gastric acid (by the antimicrobial and antiparasitic properties of hydrochloric acid). *Specific* host defence mechanisms include naturally acquired immunity from previous infection, transplacentally acquired passive immunity in the newborn from the mother, artificially acquired active immunity from immunization, and artificially acquired passive immunity from immunoglobulins and antitoxins.

Host responses to infection that prevent or reduce the severity of infectious disease include: (1) polymorphonuclear leucocytosis stimulated by some bacterial infections that increases the number of phagocytic white blood cells; (2) fever that may slow the multiplication of some infectious agents; (3) antibody production that may neutralize some infectious agents or their toxins; (4) interferon production that may block intracellular replication of viruses; and (5) cytotoxic immune cell responses that kill cells infected with viruses.

The *manifestation of infection* in the host may vary from inapparent (subclinical) infection to severe disease that may even result in death. The interaction between an infectious agent, routes of transmission, and host factors determine the spectrum of signs and symptoms. Sometimes the host may become an asymptomatic carrier of the infectious agent and be a source of infection for others.

Control measures for infectious diseases directed at the host are designed according to the immune status of the host and the likelihood of host exposure to certain infectious agents. For example, measles disease can be prevented by active immunization with measles vaccine to develop host immunity. Pneumonic plague can be prevented in those in close contact with patients with plague pneumonia by chemoprophylaxis using antibiotics such as tetracycline or sulphonamide. Details of these and other methods of

control directed at the host are provided in the section on control measures applied to the host in this chapter.

Tools for control of infectious diseases

The primary concern of infectious disease control in public health, whether in developing or industrialized countries, is the reduction, elimination, or even eradication of infectious disease. This is accomplished by directing control measures to the agent, the routes of transmission, or the host. Such control measures include: (1) identifying and then reducing or eliminating infectious agents at their sources and reservoirs, (2) breaking or interfering with the routes of transmission of infectious agents, and (3) identifying susceptible populations and then reducing or eliminating their susceptibility.

The tools for control of infectious diseases are related to recognition and evaluation of the patterns of diseases and the results of interventions to control them. In infectious disease control, the most important tool for recognition and evaluation is *surveillance* of disease defined as:

the process of systematic collection, orderly consolidation and analysis, and evaluation of pertinent data with prompt dissemination of the results to those who need to know them, and particularly those who are in a position to take action. It includes the systematic collection and evaluation of:

Morbidity and mortality reports
Special reports of field investigations of epidemics and of individual cases
Isolation and identification of infectious agents by laboratories
Data concerning the availability, use, and untoward effects of vaccines and toxoids, immune globulins, insecticides, and other substances used in control
Information regarding immunity levels in segments of the population
Other relevant epidemiologic data

A report summarizing the aforementioned data should be prepared and distributed to all cooperating persons and others with a need to know the results of the surveillance activities. The procedure applies to all jurisdictional levels of public health from local to international (Heymann 2014).

Surveillance, therefore, is ‘information for action’. More detailed information on surveillance or on field investigations is given in Chapter 5.19.

Tools for control related to *interventions* include:

Control measures applied to the host: active immunization, passive immunization, chemoprophylaxis, surgical intervention, behavioural change, reverse isolation, barriers, and improving host resistance.
Control measures applied to vectors: chemical, environmental, and biological.
Control measures applied to infected humans: chemotherapy, surgical intervention, isolation, quarantine, restriction of activities, and behavioural change.
Control measures applied to animals: active immunization, isolation, quarantine, restriction or reduction, chemoprophylaxis, and chemotherapy.
Control measures applied to the environment: provision of safe water, proper disposal of faeces, food and milk sanitation, and design of facilities and equipment.
Control measures applied to infectious agents: cleaning, cooling, pasteurization, disinfection, and sterilization.

Achieving maximum impact on control of a specific infectious disease may involve more than one of these interventions. For example, the control of hepatitis A infection can be achieved through interventions that may include: active immunization, passive immunization, food preparation, and hand-washing behaviours, provision of safe water, food sanitation, and proper disposal of faeces.

The tools for control can also be considered according to the level at which they are applied: individual, institutional, or community based. At the *individual level*, control measures, usually initiated by a clinician, are directed towards the specific infectious disease threats to the specific person. Examples include chemoprophylaxis to prevent wound infection, pre-exposure prophylactic immunization against rabies for a veterinarian, and use of diphtheria antitoxin in a patient with diphtheria.

At the *institutional level*, control measures, usually initiated by the officials of the institution, are directed to a group of people who are in close contacts with each other, such as people in day-care centres, schools,

military barracks, hospital wards, nursing homes, and correctional facilities. Examples include: (1) administering amantadine hydrochloride or rimantadine for chemoprophylaxis or chemotherapy of influenza A in a high-risk institutional population; (2) quarantining institutionalized young children during a measles outbreak; and (3) hepatitis B immunization of staff and clients of institutions for the developmentally disabled.

At the *community level*, control measures, usually initiated by local, state, or national public health agencies, are directed to the community at large. Examples include childhood immunization programmes, provision of safe water, regulation of food supplies, and recall of contaminated food products.

It should be noted that some control measures, such as immunization, may take place at all levels while others, such as the provision of safe water to a community, are more specifically applied at a particular level.

The tools for the control of infectious diseases and their relationship to the chain of infection are the focus of the remainder of this chapter.

Control measures applied to the host

Control measures applied to the host range from relatively easily administered immunization to behavioural changes that may be extremely difficult for an individual to accept and practise. This section details the types of control measures applied to susceptible hosts and gives examples of their application in the control of selected infectious diseases.

Active immunization

One of the most efficient control measures applied to a host is one that renders the host immune from infectious disease by an infectious agent. Active immunization is a cornerstone of public health measures for the control of many infectious diseases and is considered one of the most cost-effective methods of individual, institutional, and community protection for vaccine-preventable infectious diseases. The most powerful example of the potential impact of active immunization against an infectious disease is that of smallpox vaccination. Mobilization of political will on a worldwide basis, coupled with full application of the strategies of active surveillance

and containment immunization against smallpox, ultimately resulted in the complete global eradication of the disease and cessation of transmission of the infectious agent, variola virus.

Active immunization is usually considered synonymous with the term vaccination, and is the process of administration of an antigen that can induce a specific immune response that protects a susceptible host from an infectious disease. Some draw a distinction between the two terms. Narrowly defined, *vaccination* is the process of administration of an antigen and *immunization* is the development of a specific immune response. Administering an antigen without evoking an immune response is possible, since no vaccine is 100% effective. Conversely, someone can become immunized even if they were not administered an antigen. For example, the live, attenuated oral polio vaccine viruses can be transmitted from the recipient to other close contacts.

Active immunization can be accomplished through different types of antigens, including the following:

Inactivated toxins: diphtheria toxoid is an example of a formaldehyde-inactivated preparation of diphtheria toxin that protects against clinically manifest disease, although the immunized person may still become infected with toxin-producing strains of *Corynebacterium diphtheria*. Tetanus toxoid and *Clostridium perfringens* toxoid (pig bel vaccine) are other examples of inactivated toxin preparations.
Inactivated complex antigens: whole-cell pertussis vaccine is an example of a heat or chemically treated preparation of killed whole pertussis bacteria that protects against clinically manifest disease, although the immunized person may still become infected with *Bordetella pertussis*. Inactivated polio vaccine and inactivated influenza vaccine are other examples of inactivated vaccines.
Purified antigens: acellular pertussis vaccine is an example of a vaccine composed of isolated and purified immunogenic pertussis antigens. Other vaccines with purified components include polyvalent capsular polysaccharide pneumococcal, polysaccharide meningococcal, protein-polysaccharide conjugate *Haemophilus influenzae* type b, and plasma-derived hepatitis B vaccines.
Recombinant antigens: hepatitis B recombinant vaccine is an example of a vaccine composed of hepatitis B surface antigen subunits made through recombinant DNA technology. Human papillomavirus vaccine is another example of a DNA recombinant vaccine.
Live, attenuated vaccines: measles vaccine is an example of a vaccine containing live infectious agents that are of reduced virulence, but induce protective antibodies against measles viruses. Other live, attenuated vaccines include oral polio, mumps, rubella, yellow fever, and bacille Calmette–Guérin (BCG) vaccines.
• *Messenger RNA (mRNA) vaccines*: some of the COVID-19 vaccines are examples of a new vaccine technology that utilizes mRNA to deliver one or more of the SARS-CoV-2 genes into the body’s cells that then cause those cells to make spike proteins that are subsequently released from the cell and provoke an immune response.

Protective *antibody responses* usually take 7–21 days to develop. Although most vaccines must be given before exposure to be effective, some vaccines may protect even if administered after exposure to an infectious agent. For example, measles vaccine may provide protection against measles disease if

given within 72 hours of exposure. This occurs since the percutaneous route of administration of the antigen evokes an immune response faster than the measles virus results in disease through the respiratory route of natural exposure.

Duration of protection varies from only months, such as with killed whole-cell cholera vaccine, to years or even life with some live attenuated vaccines, such as measles vaccine. Some inactivated toxoids and vaccines, such as tetanus toxoid, may require a priming series of doses to be optimally effective and additional booster doses to maintain protective antibodies. Many new technologies are being explored that may increase the number and efficacy of vaccines available against infectious disease, including immune stimulating complexes, live viral or bacterial vector vaccines, and timed-release vaccines.

It should be recognized that vaccines vary in their efficacy and no vaccine is 100% effective. *Vaccine efficacies* vary with type of vaccine, manufacturing techniques, storage, and handling conditions, skill of administration, age of vaccination, and other host factors. Vaccines for routine use are safe. However, no vaccine is 100% safe. Potential vaccinees, or their parents or guardians, should be screened for contraindications and be informed of potential side effects.

Immunization schedules for the routine control of infectious diseases preventable by immunization vary between countries and are usually based on expert advice to governments and physicians. For example, general and specific recommended policies for immunization in the United States are provided by the Advisory Committee on Immunization Practices and are published online (ACIP 2018). In addition, the American Academy of Pediatrics periodically publishes comprehensive immunization recommendations in its *Report of the Committee on Infectious Diseases* (Committee on Infectious Diseases 2018). At the global level, the WHO publishes recommended immunization schedules and control strategies for vaccine-preventable diseases that are periodically updated by expert advisory groups in the WHO *Weekly Epidemiological Record*.

In *outbreak settings*, immunization schedules may be modified. For example, the age of immunization for measles may be lowered to 6 months of age during a measles outbreak. In such situations, children receiving vaccine before the routinely recommended age of immunization should be

immunized again at the recommended age since immunization at the earlier age may not have been optimally effective.

Immunization programmes include vaccines for routine child, routine adult, travel, selected high-risk populations, and occupational settings. For example, tetanus toxoid is universally recommended; yellow fever vaccine is only recommended in geographic areas of epidemiological risk (i.e. in certain areas of Africa and South America); typhoid fever vaccine is only recommended for individuals subject to unusual exposure to typhoid, including people living in the same household as known carriers; and anthrax vaccine is only recommended for veterinarians and people occupationally exposed to possibly contaminated industrial raw materials.

Beyond protection of the individual, vaccination may also provide a degree of community protection. This phenomenon of community immunity is known as herd immunity. *Herd immunity* is the relative protection of a population group achieved by reducing or breaking the chains of transmission of an infectious agent because a sufficient percentage of the population is immune, that is, resistant to infection through immunization or prior natural infection. When the immunity level in a population becomes such that the effective reproduction number of the disease in that population, R, is less than 1, the disease will eventually die out. Herd immunity is a complex phenomenon and varies according to the infectious agent, its routes of transmission, the degree to which immunization protects against infection versus only clinically manifest disease, and the distribution of immunization in the population. Typically, for childhood vaccine-preventable diseases, the immunity levels in the population to achieve herd immunity range from around 84% (for rubella) to over 90% (for pertussis). The mechanisms of herd immunity are several, including ‘direct protection of vaccinees against disease or transmissible infection and indirect protection of non-recipients by virtue of surreptitious vaccination, passive antibody, or just reduced sources of transmission and, hence, risks of infection in the community’ (Fine 1993).

A particularly difficult problem for vaccine-preventable infectious disease control programmes is complacency by the population that can result from the very successes of the programmes and highlights the continuing need to educate the public and decision-makers. Low rates of vaccine-preventable infectious disease may mistakenly lead parents to consider that vaccination is no longer important for maintaining their

children's health and may result in political leaders reducing funding for immunization programmes. Low disease rates may also focus undue attention on the relatively rare serious side effects of vaccination in relation to current rates of disease. Such side effects should only be compared in relation to rates of disease and its complications that would occur without immunization programmes (Kim-Farley 2017).

A comprehensive treatment of active immunization is given by Plotkin and Orenstein (2012).

Passive immunization

Passive immunization is a temporary immunity in a host due to the protection afforded by antibody produced in another host or artificially as in the use of laboratory-made monoclonal antibodies against SARS=CoV-2 that mimic the body's immune system to fight the virus. Passive immunity may be acquired either naturally or artificially.

Naturally acquired passive immunity is achieved through transfer of maternal antibodies via the placenta. It is the way that newborn infants are provided with a temporary immunity against many infectious diseases for which the mother is immune. This immunity wanes over time and eventually leaves the infant susceptible to these diseases.

An important use of transplacental immunity as a control measure is in the prevention of tetanus neonatorum (neonatal tetanus) by immunization of women before or during pregnancy with tetanus toxoid. This is especially important in developing countries where the disease typically occurs when the umbilical cord is cut with an unclean instrument contaminated with tetanus spores or when substances contaminated with tetanus spores are placed on the umbilical stump after delivery. Control by active immunization of the infant cannot be achieved in sufficient time since the average incubation period is only 6 days (with a range from 3 to 28 days). An adequately immunized mother, however, will usually effectively transfer maternal antibodies against tetanus across the placenta to her newborn and prevent tetanus neonatorum.

Another example of naturally acquired passive immunity is the relative protection against measles disease in a young infant born to a mother who previously had the disease. Typically, such infants are immune for approximately 6–9 months or more after birth, depending upon the level of

residual maternal antibodies present at the time of pregnancy. Other diseases for which there is usually an effective transplacental immunity in infants, for variable amounts of time, include diphtheria, mumps, poliomyelitis, rubella, and varicella (chickenpox). It should be noted that if the mother is not immune, or if residual maternal antibodies have significantly waned, then the infant may be susceptible to disease.

Research is ongoing as to other infectious diseases that may be preventable in the neonate or infant through immunization of the mother before or during pregnancy. Examples include *Haemophilus influenzae* type b, group B streptococcal and meningococcal diseases (Insel et al. 1994). Many diseases, however, are not prevented by transplacental immunity (e.g. transplacental immunity to pertussis in infants has not been demonstrated even though antibodies cross the placenta).

Breastfeeding is a form of naturally acquired passive antibody transfer to neonates and infants. Breast milk and colostrum contain secretory immunoglobulin A (IgA) antibodies that may play a protective role in the prevention of infections with such agents as respiratory syncytial virus, rotavirus, and *Haemophilus influenzae* type b.

Artificially acquired passive immunity is acquired through administration of an antibody-containing preparation, antiserum, or immune globulin. It has a place in the control of certain infectious diseases in special situations. This immunity also wanes over a relatively short period of time.

Examples of the use of artificially acquired passive immunity to control infectious disease include the following:

Rabies: natural immunity to rabies in humans does not exist. Susceptible individuals bitten by an animal known or suspected to be rabid should receive rabies immune globulin to neutralize the rabies virus in the wound. It should be noted that, besides passive immunization with rabies immune globulin, such individuals should also receive active immunization with rabies vaccine.

Hepatitis A: in areas where sanitation is poor, hepatitis A infection commonly occurs at an early age and therefore most adults in developing countries are already immune. However, outbreaks may occur in industrialized countries. Passive immunization with immune globulin may be given to: (1) all household and sexual contacts of patients with hepatitis A; (2) other food handlers in an establishment where hepatitis A has occurred in a food handler; (3) all individuals in an institution where a focal outbreak of hepatitis A has occurred; and (4) persons from industrialized countries travelling to highly endemic areas. It should be noted that vaccines for active immunization for hepatitis A are now available and may be a preferred approach.

Diphtheria: treatment of this disease is an example of the use of an antibody-containing product (diphtheria antitoxin) produced in an animal (only diphtheria antitoxin from horses is available) administered as part of the treatment regimen for secondary prevention of disease. In suspected cases of diphtheria, the antitoxin must be given as soon as possible because it is only effective in neutralizing diphtheria toxins not yet bound to cells.

Other important infectious diseases, including hepatitis B, measles, tetanus, varicella, COVID-19: depending upon the circumstances of exposure, susceptibility of the host, and status of the host's general immune system there are circumstances under which hepatitis B immune globulin, tetanus immune globulin, varicella-zoster immune globulin, SARS-CoV-2 monoclonal antibodies, or immune globulin may be warranted.

Chemoprophylaxis

Chemoprophylaxis is the prevention of infection or its progression to clinically manifest disease through the administration of chemical substances, including antibiotics. Chemoprophylaxis can also consist of the treatment of a disease to prevent complications of that disease. Chemoprophylaxis may be specifically directed against a particular

infectious agent or it may be non-specifically directed against many infectious agents. The use of antibiotics before surgical procedures is an example of non-specific chemoprophylaxis to prevent wound infections in the postoperative period. Examples of specific chemoprophylaxis are given next.

The use of chemoprophylaxis to *prevent development of infection* is illustrated by using chloroquine to prevent malarial parasitaemia caused by *Plasmodium vivax*, *Plasmodium ovale*, *Plasmodium malariae*, and chloroquine-sensitive strains of *Plasmodium falciparum*. For some chloroquine-resistant strains of *Plasmodium falciparum*, alternative regimens include mefloquine alone, doxycycline alone, primaquine alone, or an atovaquone/proguanil combination. Primaquine may be given to reduce the risk of a relapse from intrahepatic forms of *Plasmodium vivax* and *Plasmodium ovale* after discontinuation of chemoprophylaxis with any chemosuppressive drugs other than primaquine. Determination of a specific malaria chemoprophylactic regimen is complex. It must consider the geographic area, the possibility of pregnancy, the weight of an individual (dose size for children is determined by body weight), and the risks of adverse reactions to the chemoprophylactic regimen.

Other examples of prevention of development of infection include the following:

The use of silver nitrate, erythromycin, or tetracycline instilled into the eyes of a newborn to prevent gonococcal ophthalmia by transmission of *Neisseria gonorrhoeae* from an infected mother during birth.

The use of tetracycline, sulphonamides (including sulfadiazine and trimethoprim-sulfamethoxazole), chloramphenicol, or streptomycin in close contacts of confirmed or suspected cases of plague pneumonia to prevent plague pneumonia by transmission of *Yersinia pestis*.

The use of benzathine penicillin in those in sexual contact with confirmed cases of early syphilis to prevent syphilis by transmission of *Treponema pallidum*.

An example of the use of chemoprophylaxis to *prevent the progression of an infection to active manifest disease* is the use of isoniazid (INH) to prevent the progression of latent infection with *Mycobacterium tuberculosis* to clinical tuberculosis. People less than 35 years of age who are

tuberculosis test positive should receive INH to prevent clinical tuberculosis. The decision to use INH, especially in individuals more than 35 years of age (who are at higher risk of clinical hepatitis from the use of the drug), must be determined based on such information as length of infection, closeness of association with a current case, status of the immune system, presence of acute liver disease, possibilities of pregnancy, and risks of adverse reactions.

Other examples of prevention of progression of an infection to active manifest disease through the use of chemoprophylaxis include the following:

Co-trimoxazole or pentamidine to prevent subclinical latent infection with *Pneumocystis jirovecii* (formerly *carinii*) from progressing to clinically manifest *Pneumocystis* pneumonia in immunosuppressed people such as HIV-infected individuals.

Mebendazole, albendazole, or pyrantel pamoate to prevent infection with *Necator americanus*, *Ancylostoma duodenale*, and *Ancylostoma ceylanicum* from progressing to the clinically manifest anaemia of hookworm disease.

Pyrimethamine combined with sulfadiazine and folinic acid (to avoid possible bone marrow depression) to prevent asymptomatic infants congenitally infected with *Toxoplasma gondii* from progressing to clinically manifest chorioretinitis and other sequelae of congenital toxoplasmosis.

In some situations, establishing *screening* programmes to detect and treat asymptomatic infections or unrecognized disease in defined populations is useful. An example is the screening for *Chlamydia trachomatis* in sexual partners of people infected with *Chlamydia trachomatis*, women with mucopurulent cervicitis, sexually active women 25 years of age or younger, and women older than 25 years of age with risk factors for chlamydia. A more detailed background on screening as a public health function is given in Chapter 11.4.

An example of the use of chemoprophylaxis to *treat an infectious disease to prevent complications of the disease* is the use of penicillin (or erythromycin in penicillin-sensitive patients) to treat streptococcal sore throats caused by *Streptococcus pyogenes* group A to prevent acute rheumatic fever.

Other examples of prevention of complications of an infectious disease through the use of chemoprophylaxis include the following:

Tetracycline for adults, or penicillin for children, for treatment of Lyme disease caused by *Borrelia burgdorferi* in the erythema chronicum migrans stage to prevent or reduce the severity of late cardiac, arthritic, or neurological complications.
Benzathine penicillin for treatment of syphilis in its primary, secondary, or early latency period to prevent late manifestations of the disease such as cardiovascular syphilis.
Ketoconazole for treatment of blastomycosis caused by *Blastomyces dermatitidis* in its early stages to prevent progression of chronic pulmonary or disseminated blastomycosis that may lead to death.

Potential problems with the use of chemoprophylaxis may include compromise of the host's own non-specific defence mechanisms, other replacement infectious agents causing disease by growing in the place of the infectious agent affected by the specific chemoprophylactic regimen (e.g. severe diarrhoea from intestinal infections with *Clostridium difficile* may occur when the normal bacteria flora in the intestines have been lost due to antibiotics), and emergence of resistant strains of the infectious agent. The development of antibiotic resistance can be reduced by using antibiotics only when needed, selecting the proper antibiotic (or, in some situations, the appropriate multidrug therapy) for the infectious agent, and ensuring compliance with the appropriate regimen for the duration of treatment.

Control by surgical intervention

Surgical intervention is an additional method of infectious disease control. Particularly in conjunction with chemoprophylactic prevention measures and as primary prevention, some communicable diseases of public health importance can have improved control through addition of surgical intervention.

An example of the use of surgical intervention for infectious disease control is its use as an effective measure for HIV primary prevention. There is sufficient evidence that male circumcision, the surgical removal of the

foreskin of the penis, reduces the risk of HIV infection. Three randomized controlled trials have determined a 50% to 60% reduction in the incidence of HIV among heterosexual men who had circumcision compared to those who were not circumcised (Smith et al. 2010). Though male circumcision should be conducted with a comprehensive HIV prevention package, including promotion of safer sex practices, provision of condoms and appropriate use of pre-exposure prophylaxis (PreP). WHO/UNAIDS recommends male circumcision particularly in countries and regions with heterosexual epidemics, high HIV, and low male circumcision prevalence (WHO 2017). It should be noted that the impact of circumcision on the transmission of non-HIV sexually transmitted infections is still controversial (MacNeily 2011).

Additional examples of surgery as a primary prevention measure include tonsillectomy, tympanostomy, and prophylactic appendectomy. Tonsillectomy (removal of the tonsils), though performed less frequently than in the past, is most commonly used to prevent recurrent tonsillitis and obstructive sleep apnoea (Patel et al. 2014). Tympanostomy tube placement, a procedure involving a small incision in the eardrum (myringotomy) and placement of a tube, can allow drainage of fluid and is used as a preventive measure for otitis media (middle ear infection) when infections are recurrent (Lieberthal et al. 2013). Regarding prophylactic appendectomy, or removal of an apparently normal appendix during abdominal surgery for other indications, while a clear benefit has not been established by randomized trials, there appears to be very little increased morbidity and so the decision to remove during laparoscopic procedures should be made on a case by case basis (Occhionorelli et al. 2014). Also, there is the rare circumstance of prophylactic appendectomy for a sole physician wintering over in Antarctica to avoid the possibility of having to perform an emergency auto-appendectomy as occurred in a Russian physician in 1961 (Rogozov and Bermel 2009).

Behavioural change

Perhaps the most challenging tool for the control of infectious diseases, and sometimes one of the most powerful and cost-effective, is behaviour change in the host that reduces or eliminates risk of exposure to an agent. Everyone has developed habits of living (lifestyles) that are not easily changed. Some

of these behaviours are protective against infectious diseases. Others render the individual at higher risk of infection.

Examples of higher risk of exposure to infectious agents through behaviour, and behaviour changes that can have an impact on the chain of transmission, include the following.

Sexual behaviour

Many infectious agents are transmitted by the direct transmission route through sexual contact, including *Chlamydia trachomatis* causing chlamydial genital infections, *Neisseria gonorrhoeae* causing gonorrhoea, *Treponema pallidum* causing venereal syphilis, *Calymmatobacterium granulomatis* causing granuloma inguinale, *Haemophilus ducreyi* causing chancroid, herpes simplex virus causing herpes simplex, *Trichomonas vaginalis* causing trichomoniasis, human papillomaviruses causing condyloma acuminata, and HIV causing AIDS.

Abstinence behaviour, that is, refraining from sexual activity with other persons, eliminates the risk of transmission of these agents through sexual contact. The delaying of age of first sexual activity avoids the risk of transmission of these agents at an early age. Restricting sexual contact to only between two uninfected people who do not have sexual activity with any other people virtually eliminates the risk of transmission of these agents through sexual behaviours. The exceptions are due to other routes of transmission of some of these agents (e.g. HIV acquired through intravenous drug use in one partner being transmitted through sexual contact to the other partner). Limiting the number of sexual partners, and limiting those sexual partners to people who also have few sexual partners, reduces the risk of exposure. However, at the individual level, if one of these sexual partners has an infectious agent transmissible by sexual contact, the risk of transmission may still be high. Finally, condom use during sexual activity in high-risk situations will markedly reduce, but not eliminate, transmission. A more detailed background on sexually transmitted infections is provided in Chapter 8.12.

Intravenous drug use behaviour

Injection of drugs using non-sterile needles or syringes previously used by other intravenous drug users may transmit infectious agents in blood

through the vehicle-borne route of indirect transmission, including HIV causing AIDS; hepatitis B virus causing viral hepatitis B; and *Plasmodium vivax*, *Plasmodium malariae*, and *Plasmodium ovale* causing malaria.

Abstinence behaviour, that is, refraining from intravenous drug use, eliminates the risk of transmission of such agents through contaminated needles and syringes. Using a sterile needle and sterile syringe for intravenous drug use will break the chain of transmission of these infectious agents through this route. Some community public health programmes, in addition to promoting drug abstinence and drug rehabilitation, conduct needle and syringe exchanges and education regarding methods of decontamination to help promote the use of sterile injection equipment among intravenous drug users (see Chapter 9.2).

Eating behaviour

Eating certain foods may result in exposure to infectious agents through the vehicle-borne route of indirect transmission. These behaviours include consuming raw molluscs by which an infectious agent like the hepatitis A virus can cause viral hepatitis A, eating raw eggs by which an infectious agent like a *Salmonella* serotype can cause salmonellosis, and consuming raw beef by which an infectious agent like *Taenia saginata* can cause beef tapeworm infection.

Although food and diet are strongly ingrained behaviours, modification of dietary patterns is possible. Cooking foods like beef, pork, and eggs can markedly reduce risk of transmission of infectious agents. In addition, reducing risks by elimination of infectious agents from the food may be possible (see the section on control methods applied to the environment). Eating shortly after cooking so that foods are not left standing at ambient temperatures for extended times and paying attention to food recalls are also important behaviours. Hand washing before eating also reduces risk of transmission of many infectious agents that are spread through direct or indirect routes of faecal–oral transmission, such as *Shigella dysenteriae*, *Shigella flexneri*, *Shigella boydii*, and *Shigella sonnei* which may cause shigellosis (bacillary dysentery), which is estimated to cause 600,000 deaths per year with two-thirds of the cases in children under 10 years of age (Heymann 2014).

Working behaviour

In certain occupations, many behaviours may result in exposure to infectious agents and should be targets for control programmes in occupational safety and health settings. Specific examples include the following:

Dental workers performing procedures with bare hands may result in exposure to hepatitis B viruses from infected patients.
Health workers improperly handling used needles may result in needle-stick injuries leading to exposure to HIV from infected patients.
Hospital laboratory workers improperly processing specimens containing infectious agents without appropriate glove or eyewear protection may result in exposure to these agents.
Veterinarians who do not properly handle animals may result in brucellosis (undulant fever) due to exposure to *Brucella abortus*, *Brucella melitensis*, *Brucella suis*, or *Brucella canis*.

Occupational hazards related to non-infectious materials may predispose an individual to increased risk of infectious diseases. For example, working conditions and behaviours in industrial plants and mines that lead to silicosis due to long-term inhalation of free crystalline silica dust will greatly increase the risk of developing tuberculosis.

Working behaviours appropriate for the particular occupational setting may include wearing protective clothing, eyewear, and gloves; not eating food at the worksite; hand washing, and changing clothes after work; receiving appropriate vaccinations for the working environment; meticulous adherence to needle disposal and equipment sterilization procedures; and using hooded laboratory benches when handling infectious specimens that can become aerosolized.

Other behaviours

Other behaviours that may reduce the transmission of infectious agents include the following:

Scheduling outdoor activities at periods of low vector activity, applying insect repellents, and sleeping under bednets reduce the indirect transmission of vector-borne agents of infectious diseases like malaria.
Searching oneself or a young child for attached ticks every 3 to 4 hours when playing or working in tick-infested areas reduces the indirect transmission of vector-borne agents of infectious diseases like Rocky Mountain spotted fever.
Avoiding sharing of utensils, cups, toothbrushes, or towels reduces the indirect transmission of vehicle-borne agents of infectious diseases like mononucleosis.
Wearing of shoes reduces the direct transmission of infectious agents like those causing hookworm disease.
Frequently bathing and regular washing of clothes in hot soapy water controls body lice.
Breastfeeding reduces diarrhoeal diseases in the infant, although it may transmit HIV from HIV-infected mothers.
Hand washing after defecation or touching potentially contaminated surfaces, people, or animals (including those found in petting zoos) prior to preparing food or touching one's own eyes, mucous membranes, or mouth reduces the risk of direct or indirect transmission of a wide variety of infectious agents.
Large family sizes and crowding may facilitate airborne transmission of infectious agents in droplet nuclei for infectious diseases like tuberculosis.

Some of these other behaviours, like crowding, are conditioned by circumstances such as poverty that are not easily or directly amenable to programmes promoting behavioural change.

Reverse isolation

Certain rare circumstances exist where a means of avoiding transmission of an infectious disease to a highly susceptible host is to provide reverse, or protective, isolation. Such isolation attempts to protect infection-prone patients from potentially harmful infectious agents. Reverse isolation procedures range from provision of a private room with the use of masks, gloves, and gowns by all people entering the room, to elaborate facilities with laminar airflow rooms and sterilization of all food. Protective isolation

is usually conducted for a limited time until the normal immune system recovers, a regimen of passive immunization is begun, or a bone marrow transplant is successful.

Examples of people who may need periods of reverse isolation include those who have such diseases as X-linked agammaglobulinaemia, DiGeorge's syndrome, and severe combined immunodeficiency; or those who have received therapies, such as some forms of cancer chemotherapy, that have severely compromised the person's immune system in combating many infectious diseases.

Barriers

One tool of control that can be applied to the host is the use of barriers between the host and the infectious agent. The effectiveness of such barriers, however, may be dependent on the behaviour of the host to use them consistently and education to use them correctly. Examples of barriers include the following:

Screens, bednets (including bednets impregnated with pyrethroid insecticides such as permethrin), long-sleeved shirts and trousers (with the cuffs tucked into boots as a mechanical barrier), and repellents (such as N,N-diethyl-meta-toluamide, also known as DEET) to prevent transmission of malaria through the bite of infected female *Anopheles* mosquitoes or West Nile virus through *Culix* mosquitoes.
Condoms to prevent transmission of HIV and other sexually transmitted infectious agents through sexual intercourse.
Masks (air-purifying respirators) to prevent transmission of tuberculosis through airborne droplet nuclei from patients with sputum-positive pulmonary tuberculosis.

General improvement in host resistance

Improving host resistance though general improvement of the immune system is a non-specific approach, but may be important in certain settings. Kwashiorkor, marasmus, and other forms of malnutrition debilitate the host's immune system and may make an individual more susceptible to infectious diseases. Moreover, people who are malnourished and contract an

infectious disease are at higher risk of the disease being of greater severity and leading to other complications.

Malnutrition also encompasses micronutrient deficiencies. Vitamin A deficiency, for example, has been linked to higher rates of mortality associated with measles disease. Correcting vitamin A deficiency, through programmes of supplementation, fortification, and dietary modification in high-risk populations, can reduce mortality rates due to measles.

A complex interaction exists between infectious diseases, such as diarrhoeal diseases, and malnutrition. A downward spiral of infection may lead to malnutrition that, in turn, leads to more infections, and so on. If unchecked, especially in developing countries, this downward spiral can ultimately result in death.

Further, infectious diseases themselves can impact host resistance. Measles' proposed "immune amnesia" has been shown to have long-term effects on child mortality (Mina et al. 2015).

International travel

The special situation of international travel combines many control measures applied to the host already mentioned. The increase in the numbers of travellers, the speed of travel, and the ability to reach areas previously infrequently visited have reduced the effectiveness of surveillance for infectious diseases at ports of arrival and increased infectious disease risks. Advice for prevention against infectious diseases must be both general and specific.

General advice includes such issues as avoidance of eating and drinking potentially contaminated food or drink (including ice) and swimming or bathing in polluted water. Specific advice provided by health professionals should be provided based on information about the area to be visited and may include such measures as active immunization against yellow fever, active or passive immunization against hepatitis A, chemoprophylaxis against malaria, repellents against potentially infected mosquitoes, and not walking barefoot in areas of risk for infection with hookworms *Strongyloides stercoralis* and *Strongyloides fuelleborni*. A more detailed background on international travel and health is provided in the annually updated WHO publication on *International Travel and Health* (WHO 2012b).

Control measures applied to vectors

Vector-borne transmission is the only or main route of transmission for many infectious diseases. There exist more than 100 arthropod-borne viruses that may produce clinically manifest diseases in humans. Control of vector-borne diseases includes measures to: (1) change behaviour and create barriers to the susceptible host, (2) reduce or break the chain of transmission of the infectious agent from an infected host to the vector, and (3) directly control the vector population itself. Chemical, environmental, and biological controls are the primary means of directly controlling the vector population.

Chemical control

Chemicals used in the control of vectors that act as digestive poisons, contact poisons, or fumigants include minerals, natural plant products (botanicals), chlorinated hydrocarbons, organophosphates, and carbamates. Chemical control measures include the following public health interventions:

Spraying chemical insecticides such as organochlorine insecticides (e.g. dichlorodiphenyltrichloroethane, or DDT, and dieldrin), organophosphorus insecticides (e.g. malathion and fenitrothion), and carbonate insecticides (e.g. propoxur and carbaryl) to prevent malaria through control of mosquitoes.

Spraying chemical biodegradable insecticides such as temephos (Abate®) to prevent onchocerciasis through control of *Simulium* fly vectors.

Using traps impregnated with decamethrin to prevent African trypanosomiasis (sleeping sickness) through reduction of the population of infective species of *Glossina* (tsetse fly) vectors.

Treating snail-breeding places with chemical molluscicides to prevent schistosomiasis due to the free-swimming cercariae (larval forms) of *Schistosoma mansoni, Schistosoma haematobium, and Schistosoma japonicum* that develop in snails.

Treating step-wells and ponds with chemical insecticides such as temephos (Abate®) to prevent dracunculiasis due to infected cyclops (a crustacean copepod).

Suppressing rat populations by poisoning, preceded, or accompanied by measures to control fleas, as an additional measure to supplement environmental sanitation to control rodent populations to prevent human plague.

The use of spraying for control of mosquitoes is complicated due to concerns about environmental contamination by chemicals such as DDT and dieldrin which have led to them being banned in many countries. In addition, the emergence of mosquito vectors resistant to the insecticides diminishes their effectiveness in many areas. New methods of application, such as ultra-low-volume spraying of malathion, reduce the amounts of insecticide used.

Environmental control

Environmental control of vectors includes the following public health interventions:

Eliminating breeding sites of mosquito larvae by filling and draining areas of stagnant water and removing objects around houses that may collect water.
Destroying the habitats of the tsetse fly vector.
Properly implementing landfill procedures, placing lids on rubbish bins, covering food for human consumption, screening privies, cleaning up spilled food, and appropriately storing food.
Placing roach and fly traps.
Constructing rat-proof houses.
Eliminating rodent habitats.

It is also important to note that certain development projects may have an impact on the environment that facilitates the growth of vector or intermediate-host populations and results in increased infectious diseases. Construction of artificial waterways may serve as breeding sites for *Simulium* fly vectors that can transmit *Onchocerca volvulus* resulting in onchocerciasis. Irrigation schemes can foster the growth of snail intermediate hosts required for the transmission of species of *Schistosoma* resulting in schistosomiasis. Carefully conducted environmental and health impact studies that consider the impact of a construction project on the vector and intermediate-host populations, and ways to modify the project to reduce such populations, are important environmental control measures.

Biological control

Biological control of vectors includes the following public health interventions:

Introduction of predators and parasites: the introductions of *Gambusia affinis*, a small fish that feeds on mosquito larvae, and of *Coelomomyces*, a fungal parasite, are examples of control measures that are effective against *Aëdes* mosquitoes.
Insect growth regulators: the use of such regulators may result in death or sterility of vectors by interfering with normal insect development. An example is the use of methoprene (Altosid®) to control flood water mosquitoes.
Genetic modification: although still at an experimental phase, researchers have developed transgenic, or genetically modified, mosquitoes that are malaria resistant with higher survival rates that could eventually replace mosquitoes that carry malaria parasites.

Control measures applied to infected humans

Control measures may be applied to infected people at the individual level, in the institutional or hospital setting, and at the community level.

Hospital infection control

The hospital setting is a unique situation that requires special efforts to prevent and control nosocomial infections, or healthcare-associated infections (HAIs), which are infections that originate or occur in a hospital or other healthcare setting. HAIs are a major problem worldwide. In the United States alone, some 2 million HAIs occur annually resulting in an estimated 90,000 deaths (McKibben et al. 2005). It has been estimated that HAIs result in some US$35.7–45 billion in annual direct medical hospital patient costs based on using the Consumer Price Index for inpatient hospital services (Scott 2009).

Infection control programmes for hospitals should ideally include the following elements:

An infection control committee responsible for overall coordination of infection control activities.
One or more infection control practitioners (Infection Preventionists) responsible for nosocomial disease surveillance, analysis of data, consultation, and training of hospital staff.
A hospital epidemiologist to supervise the infection control practitioners, oversee data collection and analysis, and implement any needed emergency infection control measures.
An engineer to direct engineering and preventive maintenance operations, especially ventilation equipment.
A sanitarian to develop procedures for proper disposal of liquid and solid wastes; and sanitation of water, ice, and food.
Effective guidelines for patient care practices.
Surveillance of patient care practices, patient infections, and environmental contamination by infectious agents.
Coordination with other departments (microbiology laboratory, central services, housekeeping, food service, and laundry).
Vector control.
Thorough investigation of problems.

Examples of specific control measures that may be applied to infected humans at the individual, institutional, and community levels are detailed next.

Chemotherapy

Treatment of people with infectious diseases or subclinical infections may be a control tool for some infectious diseases. Such treatment may or may not have an impact on disease progression in the patient. It should be noted that rapid case detection and prompt application of appropriate chemotherapeutic agents are needed to limit infectivity.

Some important examples of control by chemotherapy include the following:

Treatment of patients with sputum-positive pulmonary *tuberculosis* with appropriate multidrug therapy will usually result in sputum conversion rendering them non-infectious to others within a few weeks. Recommended treatment regimens include isoniazid (INH) combined with one or more of the following antibiotics: rifampin, streptomycin, ethambutol, and pyrazinamide. The WHO has recommended that adherence to a complete course of multidrug therapy be directly observed by another responsible person as part of the DOTS (directly observed treatment, short-course) global strategy for the control of tuberculosis.
Patients with *leprosy* treated with appropriate multidrug therapy are considered no longer infectious within 3 months of regular and continued treatment. Recommended treatment regimens for multibacillary leprosy include the following antibiotics: rifampicin, dapsone, and clofazimine.
Treatment of patients with *streptococcal sore throats* with penicillin (or erythromycin for penicillin-sensitive patients) will usually render them no longer infectious after 24–48 hours.
Patients with *pertussis* treated with antibiotics such as erythromycin or trimethoprim-sulfamethoxazole, although they may not affect the patient's symptoms, will usually result in the patient no longer being infectious after 5–7 days.

Of special note is the situation of treatment of people who are carriers:

> A person or animal that harbors a specific infectious agent without discernible clinical disease, and which serves as a potential source of infection. The carrier state may exist in an individual with an infection that is unapparent throughout its course (such an individual is commonly known as *healthy or asymptomatic carrier*), or during the incubation period, convalescence, and post-convalescence of a person with a clinically recognizable disease (commonly known as *incubatory or convalescent carrier*). Under either circumstance the carrier state may be of short or long duration (*temporary or transient carrier, or chronic carrier*). (Heymann 2014)

A chronic carrier of diphtheria, for example, may shed the infectious agent *Corynebacterium diphtheriae* for 6 months or more, but appropriate antibiotic therapy will usually promptly stop the carrier state. Another example is that of untreated patients with typhoid fever due to *Salmonella typhi*. Between 2 and 5% of such patients will become permanent carriers. Treatment with appropriate antibiotics may be effective in ending the carrier state.

Antibiotic treatment may not always eliminate a carrier state for some infectious agents. For example, the treatment of people with salmonellosis with an antibiotic may not terminate the period of communicability and can even result in emergence of antibiotic-resistant strains. However, antibiotic therapy may be still warranted under certain circumstances.

In some situations, establishing screening programmes in defined target populations for identification of asymptomatic or unrecognized infections that could be transmitted to others may be appropriate. Such screening should include the necessary follow-up with appropriate chemotherapy and counselling. An example would be screening close contacts of diphtheria patients with nose and throat cultures for the presence of *Corynebacterium diphtheriae*. Identified carriers with positive cultures should be treated with appropriate antibiotic therapy.

Control by surgical intervention

Surgical intervention is an additional method of infectious disease control in an infected person. Particularly in conjunction with chemotherapeutic control measures, some communicable diseases of public health importance can be better controlled through the addition of surgical therapy.

In the first part of the 20th century, surgery played a major role in control of tuberculosis (TB) until the availability of effective medications mid-century (Sihoe et al. 2009). With the increase of multidrug-resistant (MDR) and extensively drug-resistant (XDR) TB, surgical intervention is again being considered as an adjunct treatment option (Kempker et al. 2012). The American Thoracic Society, CDC and Infectious Diseases Society of America released a 2003 report communicating that surgical therapy should be used only as an adjunct to medical treatment (Blumber et al. 2003). Recent research has shown early pulmonary resection surgery, removing either part or all of a lung, as an adjunct to chemotherapy may result in high rates of cure for both MDR and XDR TB (Kang et al. 2010). The WHO additionally recommends that any adjunct surgical therapy may be warranted earlier (rather than later) in disease progression (WHO 2008a). It should be noted that treating TB would constitute primary prevention to stop transmission to others.

Surgical therapy can also play a role in treatment of other infections such as Methicillin-Resistant *Staphylococcus aureus* (MRSA). Surgical

debridement, or the removal of dead or contaminated tissue from a wound to prevent further infection, is recommended in complicated MRSA infections, or when there is bone or joint involvement (Liu et al. 2011). Another example is that of gangrene, a condition in which cell death occurs due to lack of sufficient blood supply caused by infection and/or noncommunicable conditions such as diabetes. Gangrene can be caused by only a non-infectious condition manifested as dry gangrene, or by a variety of bacteria, manifested as wet gangrene, gas gangrene, internal gangrene, and necrotizing fasciitis (an infection that results in the death of soft tissue).

Isolation

Isolation is the 'separation, for a period at least equal to the *period of communicability*, of infected persons or animals from others, in such places and under such conditions as to prevent or limit the direct or indirect transmission of the infectious agent from those infected to those who are susceptible to infection or who may spread the agent to others' (Heymann 2014).

The US CDC and the Hospital Infection Control Practices Advisory Committee (HICPAC) have provided guidelines for isolation precautions in hospital settings (Siegel et al. 2007). There are two levels of isolation precautions, namely: (1) a standard precautions level designed for the care of all hospitalized patients; and (2) a transmission-based precautions level designed for the care of hospitalized patients that are suspected or confirmed to be infected by agents spread by contact, droplet, or airborne routes of transmission. These are summarized from guidelines as follows.

Standard precautions are universally applied precautions designed to reduce the risk of transmission by infectious agents from blood; body fluids, secretions, and excretions; non-intact skin; and mucous membranes. The essential elements of standard precautions include hand washing; personal protective equipment (use of gloves, appropriate application of mask and eye protection or a face shield, and utilization of gowns); proper handling of patient care equipment; adequate environmental control measures for routine care, cleaning, and disinfection of frequently touched surfaces; appropriate handling, transporting, and processing of used linen; proper handling and disposal of needles, scalpels, and other sharp instruments; adequate protection when undertaking patient resuscitation;

respiratory hygiene/cough etiquette; and placement of patients who contaminate the environment in private rooms.

Airborne precautions are used, in addition to standard precautions, in settings where patients are suspected or confirmed to be infected by agents transmitted by airborne droplet nuclei. The essential elements of airborne precautions include preferably placing patients in an airborne infection isolation room (AIIR) that has monitored negative air pressure (if necessary, it is possible to use cohorting of patients with the same active infections); use of mask respirators (N-95 air-purifying respirators); and limiting patient movement and transport from the room (placing a surgical mask on the patient if they are being moved for an essential purpose). An example of an infectious disease for which patients are recommended to be placed under airborne precautions is a patient in hospital with measles through the fourth day of rash. Although isolation of patients with measles not in hospital is not practical in the general population, schoolchildren should remain out of school until at least the fourth day of rash.

Droplet precautions are used, in addition to standard precautions, in settings where patients are suspected or confirmed to be infected by agents transmitted by droplets. The essential elements of droplet precautions include placement of patients in a private room (if necessary, it is possible to use cohorting of patients with the same active infections or maintaining a spatial separation of at least 3 feet between the infected patient and other patients and visitors); use of a mask when working within 3 feet of the patient; and limiting patient movement and transport from the room (placing a surgical mask on the patient if they are being moved for an essential purpose). Examples of infectious diseases for which patients are recommended to be placed under droplet precautions include pharyngeal diphtheria caused by *Corynebacterium diphtheriae* and pneumonic plague caused by *Yersinia pestis.*

Contact precautions are used, in addition to standard precautions, in settings where patients are suspected or confirmed to be infected or colonized by agents transmitted by direct or indirect contact. The essential elements of contact precautions include placement of the patient in a private room (if necessary, it is possible to use cohorting of patients with the same active infections); use of gloves when entering the patient’s room and removing them before leaving the room; wearing a gown when entering the patient’s room and removing the gown before leaving the room; limiting

patient movement and transport to essential purposes only; and, when possible, dedicate the use of patient care equipment to a single patient (if necessary, it is possible to use such equipment on a cohort of patients with the same active or colonized infections). Examples of infectious diseases for which patients are recommended to be placed under contact isolation precautions include cutaneous diphtheria caused by *Corynebacterium diphtheriae*, *C. diff* infections caused by the bacterium *Clostridium difficile*, and disseminated herpes simplex caused by herpes simplex virus.

Quarantine of potentially infected persons

Quarantine is the 'Restriction of activities for well persons or animals who have been exposed (or are considered to be at high risk of exposure) to a case of communicable disease during its period of communicability (i.e. *contacts*) to prevent disease transmission during the incubation period if infection should occur' (Heymann 2014).

Two categories of quarantine are as follows (Heymann 2014):

> *Absolute or complete quarantine*: the limitation of freedom of movement of those exposed to a communicable disease for a period of time not longer than the longest usual incubation period of that disease, in such a manner as to prevent effective contact with those not so exposed.
>
> *Modified quarantine*: a selective, partial limitation of freedom of movement of contacts, commonly on the basis of known or presumed differences in susceptibility and related to the assessed risk of disease transmission. It may be designed to accommodate particular situations. Examples are exclusion of children from school, exemption of immune persons from provisions applicable to susceptible persons, or restriction of military populations to the post or to quarters. Modified quarantine includes: *personal surveillance*, the practice of close medical or other supervision of contacts to permit prompt recognition of infection or illness but without restricting their movements; and *segregation*, the separation of some part of a group of persons or domestic animals from the others for special consideration, control or observation; removal of susceptible children to homes of immune persons; or establishment of a sanitary boundary to protect uninfected from infected portions of a population. (Heymann 2014)

Examples of diseases where quarantine may be considered include the following.

Pneumonic plague: people who have been in the same household or in face-to-face contact with patients with pneumonic plague and who do not accept chemoprophylaxis should be placed under *absolute quarantine* with strict isolation, including careful surveillance, for 7 days.
Measles: although absolute quarantine is impractical, a *modified quarantine* is recommended in settings where young children are living in dormitories, wards, or institutions. When measles occurs in such institutional settings, strict *segregation* of infants is recommended.
Lassa fever: close *personal surveillance* of all close contacts is recommended. Such people include those who live or are in close contact with Lassa fever patients as well as laboratory personnel testing specimens from such patients.

Restriction of activities

Controlling infectious disease transmission by restriction of the activities of people in the community who are potentially infectious to others may be appropriate in certain circumstances. Examples include the following:

Individuals with a diarrhoeal disease should be excluded from handling food and caring for patients in hospital, children, and elderly people.
Known carriers of *Salmonella typhi* should be excluded from food-handling and care of patients.
People with staphylococcal disease should avoid contact with debilitated people and infants.
People with rubella should be excluded from school or work for 7 days after the onset of rash and from contact with pregnant women.

Behavioural change

Behaviour change in an infected person to protect others may be difficult to accomplish and often requires continuing education and counselling. However, this should be considered in preventing the transmission of infectious agents in the following situations.

Sexual behaviour

Examples of infectious agents transmitted through sexual activities are discussed in the earlier section on control measures applied to the host and in more detail in Chapter 8.12. Individuals who suspect that they may have a sexually transmitted infection should be encouraged to have health-seeking behaviours. Persons with a sexually transmissible infectious agent should be treated and asked to cooperate with health officials to trace their sexual contacts. Patients with diseases such as lymphogranuloma venereum and syphilis, for example, should refrain from sexual contact until all lesions are healed. HIV-infected individuals should be counselled to treat genital ulcer disease promptly since such disease may increase transmissibility of HIV. Also, HIV-infected people should avoid sexual intercourse with HIV-negative individuals or, if having sexual intercourse with HIV-negative individuals, use methods such as condoms to reduce the risk of transmission. For a more detailed overview of HIV and AIDS see Chapter 8.13.

Intravenous drug use behaviour

In addition to counselling to abstain from intravenous drug use and establishing drug rehabilitation programmes to help individuals who wish to abstain, promoting behaviour change in the use of injection equipment is important. Discouraging the sharing of injection equipment and education on methods for the decontamination of needles and syringes for intravenous drug use reduce risks of transmission of infectious agents through contaminated injection equipment.

Food preparation behaviour

Individuals who should be restricted from handling food (e.g. carriers of *Salmonella typhi*) should be counselled regarding their condition and potential to infect others if they handle food. Food handlers who have an infectious disease that is potentially transmissible through the vehicle-borne means of food should be discouraged from handling food for others. The importance of hand washing, especially after defecation and before handling food, should be stressed.

Other behaviours

Other behaviours that may reduce risk of transmission of infectious agents to other people include the following:

Cough and sneeze behaviour: patients with infectious diseases directly transmitted by droplet spread or airborne transmitted by droplet nuclei (e.g. patients with sputum-positive tuberculosis) should cover their mouth and nose when coughing or sneezing.
Avoidance of contaminated drinking water: people suffering from dracunculiasis should avoid entering a source of drinking water if they have an active ulcer or blister.
Avoidance of vector bites: patients with the vector-borne disease of African trypanosomiasis (sleeping sickness) with trypanosomes in their blood should prevent tsetse flies from biting.
Avoidance of donating organs or bodily fluid by certain people: individuals who are infected with HIV or who have sexual and other behaviours that have placed them at increased risk for HIV infection should not donate blood, plasma, tissues, cells, semen for artificial insemination, or organs for transplantation.

Control measures applied to animals

A zoonosis is any infectious agent or infectious disease that may be transmitted under natural conditions from vertebrate animals, both wild and domestic, to humans. A detailed approach to zoonoses is given in the comprehensive work *CRC Handbook Series in Zoonoses* (Beran and Steele 1994). In the control of zoonoses many approaches are used that are applied to animals, including the following.

Active immunization

Active immunization, or vaccination, of selected animals may protect susceptible animal hosts from certain infectious diseases. This protection of animals, in turn, prevents susceptible humans from exposure to the infectious agent of those diseases from animals. An example of an infectious disease in animals in which some control can be achieved through immunization in selected animal populations is rabies. The

reservoir of the rabies virus is varied and includes dogs, foxes, wolves, skunks, raccoons, and bats. Preventive measures include efforts to vaccinate all dogs.

Other examples of immunization of animals under certain conditions include: (1) immunization of young goats and sheep using a live attenuated strain of *Brucella melitensis* and calves using a strain of *Brucella abortus* in areas of high endemicity for brucellosis; and (2) immunization of animals at risk for acquiring infection with *Bacillus anthracis* that could be transmitted to man causing anthrax.

Restriction or reduction

Restriction is the limiting of the movement of animals and includes isolation and quarantine. Reduction is the killing, known as culling, of selected animals. Selective use of restriction of animals or reduction of animal populations that are infected, or potentially infected, with a zoonotic infectious agent are methods used to decrease or eliminate the opportunity of exposure of susceptible humans, or other animals, to such animals.

The example of rabies can also be used to illustrate the use of restriction or reduction of an animal population to help control an infectious disease. Heymann (2014) recommends the following measures:

Preventive measures

Register, licence, and vaccinate all owned dogs, and other pets when feasible, in enzootic countries; control ownerless animals and strays. Educate pet owners and the public on the importance of local community responsibilities (e.g. pets should be leashed in congested areas when not confined on the owner's premises; strange-acting or sick animals of any species,—domestic or wild—should be avoided and not handled; animals that have bitten a person or another animal should be reported to relevant authorities, such as the police/local health departments; if possible, such animals should be confined and observed as a preventive measure; and wildlife should be appreciated in nature and not be kept as pets). Where animal population reduction is impractical, animal contraception and repetitive vaccination campaigns may prove effective.

Detain and observe for 10 days any healthy-appearing dog or cat known to have bitten a person (stray or ownerless dogs and cats may be euthanized and examined for rabies by fluorescent microscopy); dogs and cats showing suspicious clinical signs of rabies should be euthanized and tested for rabies.

Euthanize unvaccinated domestic animals bitten by known rabid animals; if detention is elected, hold the animal in a secure facility for at least 6 months under veterinary supervision, and vaccinate against rabies 30 days before release. If previously vaccinated, booster immediately with rabies vaccine, and detain for at least 45 days.

Cooperate with wildlife conservation authorities in programmes to reduce the carrying capacity of wildlife hosts of sylvatic rabies, and to reduce exposures to domestic animals and human populations—such as in circumscribed enzootic areas near campsites, and in areas of dense human habitation (Heymann 2014).

Epidemic (epizootic) measures

In urban areas of industrialized countries, strict enforcement of regulations requiring collection, detention and euthanasia of ownerless and stray dogs, and of non-immunized dogs found off owners' premises; control of the dog population by castration, spaying or drugs have been effective in breaking transmission cycles (Heymann 2014).

Other measures

Other examples of restricting or reducing of animal populations include the following:

Rat-proofing dwellings and reduction of the rat population to prevent rat bites that may transmit the infectious agents *Streptobacillus moniliformis* and *Spirillum minus* causing the rat-bite fevers of streptobacillosis and spirillosis, respectively.
Rat suppression by poisoning (after achieving flea control) in rodent populations with a high potential for epizootic plague.
Elimination of animals infected with *Brucella abortus*, *Brucella melitensis*, *Brucella suis*, and *Brucella canis* by segregation or slaughter to prevent brucellosis.
Slaughtering dairy cattle that test positive for infection with *Mycobacterium bovis*, the infectious agent of bovine tuberculosis.

Chemoprophylaxis and chemotherapy

Chemoprophylaxis of an animal is using chemical substances (e.g. antibiotics) that prevent infection or its progression to clinically manifest infectious disease in the animal. Chemotherapy of an animal is using these chemical substances to treat an infectious disease in an animal. Both chemoprophylaxis and chemotherapy are control measures that may be used to reduce or prevent the opportunity of an infectious agent from being transmitted from an animal to susceptible humans. However, caution should be applied in the routine use of chemoprophylaxis in cattle, feed lots, and poultry farms that can promote the emerging of drug resistance and its associated problems in humans as well.

Psittacosis is an example of a zoonosis controlled by chemoprophylaxis or chemotherapy in selected animal populations. The infectious agent, *Chlamydia psittaci*, may be directly transmitted to humans from infected birds when the dried droppings, secretions, or dust from the feathers of such infected birds are inhaled. Imported psittacine species of birds should be placed under quarantine and receive an appropriate antibiotic chemotherapeutic regimen such as chlortetracycline administered in their feed for 30 days.

Another example is chemoprophylaxis in selected dogs at high risk of infection with *Echinococcus granulosus*. This infectious agent can be

transmitted to humans through hand-to-mouth transmission of the tapeworm eggs from dog faeces causing echinococcosis due to *Echinoccus granulosus*, or cystic hydatid disease. Such high-risk dogs should periodically receive antihelminth treatment with a chemotherapeutic agent such as praziquantel (Biltricide®).

Control measures applied to the environment

Control measures applied to the environment are designed to interrupt the routes of transmission by which an infectious agent may be spread through the environment. Just as the routes of transmission are varied, so, too, are the control methods that can be applied. Control measures that affect transmission that can be applied to the host, agents, vectors, infected humans, and other animals are reviewed elsewhere in this chapter. Environmental factors may also have a direct impact on the host, agent, or vector. For example, low humidity may predispose to certain infections due to a greater permeability of mucous membranes in the host; cold, dry climates inhibit development of the infective larvae agent of hookworm disease; and higher altitudes and colder climates limit habitats of the mosquito vector.

The recognition of the relationship between disease and filth led to a sanitary revolution in industrialized countries that markedly reduced infectious diseases even before the arrival of the antibiotic era. Improved methods for storing and preserving food, better housing, and smaller families with a resultant decrease in the risk of infections at an early age all contributed to reductions in infant and child mortality rates.

This section focuses on general environmental control measures not mentioned elsewhere. Some of these methods, such as provision of safe water, have the potential to prevent several different infectious diseases and significantly reduce rates of disease in the community.

Provision of safe water

It has been estimated that some 663 million still lack safe drinking water in 2015 (WHO and United Nations Children’s Fund 2015). Contaminated drinking water, sometimes the result of poorly designed or maintained

systems of sewerage, may lead to the water-borne indirect transmission of such infectious agents as *Giardia lamblia* causing giardiasis, pathogenic serotypes of *Salmonella* causing salmonellosis, and *Cryptosporidium* species causing cryptosporidiosis.

Purification of water can occur though natural methods or human intervention. Examples of natural methods that contribute to water purification include the processes of evaporation and condensation, filtration through the earth, plant growth, aeration, and reduction and oxidation of organic material by bacteria. Purification of water for public consumption is conventionally done through such processes as coagulation of colloids by aluminium salts (such as aluminum sulfate, or alum) or with other techniques; filtration through such materials as coal, sand, or diatomaceous earth; and disinfection with such chemicals as chlorine derivatives. In special situations, boiling and distillation can be used for purification (Solomon et al. 2009).

Proper disposal of faeces

It has been estimated that some 2.4 billion people in the developing world do not have an adequate system for proper disposal of faeces with some 946 million people practising open defecation (WHO and United Nations Children's Fund 2015). Infectious agents in faeces that may result in infectious diseases include poliovirus causing poliomyelitis; *Shigella dysenteriae*, *Shigella flexneri*, *Shigella boydii*, and *Shigella sonnei* causing shigellosis; and *Entamoeba histolytica* causing amoebiasis.

Infectious agents in faeces may be transmitted by the direct transmission route (including the faecal–oral mode), the vehicle-borne route (including water as noted in the previous section), and the vector-borne route (including the simple mechanism of flies carrying infected faeces on their feet). Public health environmental control measures to interrupt these routes of transmission by ensuring the proper disposal of faeces include the following:

Appropriate on-site disposal through such means as properly constructed sanitary privies in rural areas with no sewerage systems.
On-site disposal of domestic wastewater (such as use of septic tanks or cesspools).
Sewerage systems with treatment of wastewater. Such treatment may include preliminary treatment, sedimentation, chemical coagulation and flocculation, biological treatment (such as activated sludge units and trickling filters), stabilization ponds, sludge management, and disinfection (usually with chlorine) of effluents discharged into drinking, bathing, or shellfish-growing waters.

The importance of personal health-promoting behaviours of using toilets, keeping toilets clean, and hand washing after defecation are a part of control efforts aimed at the proper disposal of faeces.

Food sanitation

Food-borne infectious diseases remain a problem in both industrialized and developing countries. In the United States in 2011 alone, it is estimated that some 48 million people became ill, 128,000 were hospitalized, and 3,000 died of food-borne-related diseases (CDC 2012a). Significant food-borne outbreaks and sporadic cases continue to occur due to such factors as the following:

Contamination of meat, poultry, and eggs with infectious agents, including pathogenic serotypes of *Salmonella*, *Yersinia pseudotuberculosis* and *Yersinia entercolitica*, and *Listeria monocytogenes.*
Contamination of vegetables, especially lettuce and leafy green vegetables, with the infectious enterohaemorrhagic strain of *Escherichia coli* O157:H7. Other outbreaks in fruits, juices, and vegetables have been due to contamination with hepatitis A and pathogenic serotypes of salmonella.
Problems in food storage, handling, and preparation in commercial eating places and in homes.
Larger and more centralized production and processing facilities, coupled with increasingly extensive distribution networks, which may result in transmission of infectious agents to many people if a commercial product becomes contaminated.

Industrialized countries have significantly reduced the transmission of some infectious agents through major public health programmes in food sanitation, including: (1) inspecting eating and drinking establishments, (2) inspecting meat and poultry, (3) improving shellfish sanitation, and (4) promoting adequate cooking, canning, and refrigeration methods. Some cities and counties have instituted restaurant grading systems based on inspection reports, including required public display of the restaurant’s grade as a guide for consumers.

Examples of vehicle-borne indirect transmission of infectious agents through food that can be controlled though a comprehensive public health food sanitation programme include the following:

Pathogenic serotypes of *Salmonella* transmitted by ingesting food made from infected animals or contaminated by the infectious agent in faeces that may cause salmonellosis. Methods of control—preventive measures: ‘a) handwashing before, during and after food preparation; b) refrigerating prepared foods in small containers; c) thoroughly cooking all foodstuffs derived from animal sources, particularly poultry, pork, egg products, and meat dishes; d) avoiding recontamination within the kitchen after cooking is completed; and e) maintaining a sanitary kitchen and protecting prepared foods against rodent and insect contamination’ (Heymann 2014).

Staphylococcus aureus causing staphylococcal food intoxication by ingesting food containing the staphylococcal enterotoxin. Methods of control—preventive measures: ‘Reduce food-handling time (from initial preparation to service) to a minimum, no more than 4 hours at ambient temperature. If foods are to be stored for more than 2 hours, keep those that are perishable hot (above 60°C/140°F) or cold (below 5°C/41°F), in shallow containers, and covered’, ‘Temporarily exclude people with boils, abscesses and other purulent lesions of hands, face or nose from food handling’ (Heymann 2014).

Trichinella spiralis transmitted by ingesting raw or improperly cooked meat or meat products, mainly pork, containing infectious encysted larvae that may cause trichinosis. Methods of control—preventive measures:

1. Educate the public on the need to cook all fresh pork and pork products and meat from wild animals at a temperature and for a time sufficient to allow all parts to reach at least 71°C (160°F), or until meat changes from pink to grey, which allows a sufficient margin of safety. This should be done unless it has been established that these meat products have been processed either by heating, curing, freezing, or irradiation adequate to kill trichinae.
2. Grind pork in a separate grinder, or clean the grinder thoroughly before and after processing other meats.
3. Adopt regulations to encourage commercial irradiation processing of pork products. Testing carcasses for infection with a digestion technique is useful, as is immunodiagnosis of pigs with an approved ELISA test.
4. Adopt and enforce regulations that allow only certified trichinae-free pork to be used in raw pork products that have a cooked appearance, or in products that are traditionally not heated sufficiently in final preparation to kill trichinae.
5. Adopt laws and regulations to require and enforce the cooking of garbage and offal before feeding to swine (Heymann 2014).

Milk sanitation

Milk may be a vehicle for indirect transmission of such infectious agents as: *Mycobacterium bovis* causing tuberculosis, *Corynebacterium diphtheriae* causing diphtheria, *Listeria monocytogenes* causing listeriosis, and *Campylobacter jejuni* and *Campylobacter coli* causing *Campylobacter* enteritis.

Public health control measures to break the chain of transmission of infectious agents through milk include:

Mechanization and sanitization of milking processes.
Refrigeration of milk to inhibit bacterial growth.
Pasteurization of milk through high-temperature short-time, batch, ultra-pasteurization, or ultra-high-temperature methods that help to kill any bacteria from the cow or others that may have come into direct contact with milk during the milking and handling process.
Monitoring milk quality by testing for bacteria using a standard bacterial plate count, by testing for density of coliform organisms, and by use of the phosphatase test to assay for pasteurization.
Periodically testing cows for tuberculosis and brucellosis.

The use of raw milk for human consumption may result in outbreaks. A study showed that some 60% of dairy-related outbreaks reported to the CDC between 1993 and 2006 were associated with raw milk products (CDC 2012b). It should be noted that post-pasteurization contamination of milk may also result in outbreaks of milk-borne diseases.

Design of facilities and equipment

The design and proper maintenance of buildings, rooms, and equipment can help break the chain of transmission of infectious agents. Laminar airflow hoods in laboratory workbenches, ventilation systems in hospitals, and disposable intravenous equipment are examples of systems designed to reduce risk of transmission. Routine maintenance needed to retain the original design standards for control of transmission of infectious agents include: (1) replacement of air filters, (2) cleaning of cooling towers, (3) monitoring of negative and positive pressure rooms and their airlocks, and (4) replacement of in-dwelling peripheral venous catheters.

Examples of infectious agents whose transmission can be reduced through proper design and maintenance include the following:

Legionella species, the infectious agents responsible for legionellosis, are usually transmitted through airborne transmission via aerosol production. Transmission of the agent from cooling towers can be reduced by periodically cleaning off any scale or sediment, routinely using biocides to kill slime-forming organisms, and draining such towers when not in use.
Staphylococcus aureus, the infectious agent responsible for staphylococcal disease in medical and surgical wards of hospitals, can be controlled by enforcing strict aseptic technique, including procedures to change intravenous infusion sites at least every 48 hours and replace in-dwelling peripheral venous catheters every 72 hours.
Bacillus anthracis, the infectious agent responsible for anthrax, can be transmitted, among other ways, through inhalation of anthrax spores. Proper design of industrial plants that handle raw animal fibres includes providing facilities for adequate ventilation and control of dust, washing, and changing clothes after work, and eating away from the places of work.

Other methods

In addition to the environmental methods of control of transmission of infectious agents already mentioned, the following are other methods, some specific and some general, that should also be noted.

Improvement of housing conditions to *reduce crowding* (as measured by the number of people per room and not total population density). Reduction in crowding is a general measure that can reduce the transmission of infectious agents, especially direct transmission from direct contact or direct projection (droplet spray).

Improvement in working conditions can affect the risk of infectious disease. For example, control of particulate matter by proper ventilation in occupations such as textile mill workers, metal grinders, and pottery factory workers can reduce inflammation of the lungs and thus decrease the risk of developing tuberculosis. Excessive physical exertion and the stress of exhausting work can also increase the risk of tuberculosis.

Improved *irrigation* and agricultural practices and *removing vegetation* or *draining and filling* of snail-breeding sites can reduce or eliminate the freshwater snail hosts of such infectious agents as *Schistosoma mansoni*, *Schistosoma haematobium*, and *Schistosoma japonicum* that cause schistosomiasis in humans.

Adequate screening of blood, serum, plasma, tissues, or organs can break the chain of vehicle-borne transmission from such biological products. Examples include screening for hepatitis B surface antigen and HIV antibodies in donated blood to prevent transmission of hepatitis B and HIV, respectively.

Installation of *screened living and sleeping quarters* and use of *bednets*, including bednets impregnated with a synthetic pyrethroid such as permethrin, can reduce exposure to mosquitoes infected with the infectious agents of malaria.

• Erecting *plexiglass barriers* around student desks or between workers on a food processing line as a measure, in addition to masking, to reduce the risk of droplet spread of SARS-CoV-2 to prevent COVID-19.

Control measures applied to the agent

Control of some infectious diseases can be achieved through means that remove the infectious agents from the environment or inactivate the agents. Physical measures (such as heat, cold, ultraviolet light, and ionizing radiation) and chemical measures (such as liquid disinfectants and

antiseptics, gases, and chlorination) can be used. Examples of control measures applied to infectious agents include the following:

Cleaning is the removing of infectious agents from surfaces through such physical actions as vacuum cleaning or washing and scrubbing using soap or detergent and hot water. Cleaning also helps remove organic materials that might support the growth or survival of infectious agents.
Cooling may inhibit bacterial multiplication and some infectious agents, such as *Trichinella* cysts and *Taenia solium* larvae (cysticerci), can be killed by freezing temperatures.
Pasteurization is heating to a temperature of 75°C/167°F for 30 minutes to kill pathogenic vegetative bacteria. It does not inactivate bacterial spores. Pasteurization is a commonly used process to help ensure safety of milk and to prolong its storage quality.
Disinfection is the reduction or killing of vegetative harmful bacterial infectious agents outside the body or on objects. Disinfection may not inactivate all bacterial spores and viruses. *Disinfectants* are used to eliminate pathogenic bacteria from the skin surface and from contaminated inanimate surfaces and include: (1) alcohols, (2) halogens such as iodine and chlorine, (3) surface active compounds such as the quaternary ammonium compound benzalkonium chloride, (4) phenolics, and (5) alkylating agents such as glutaraldehyde and formaldehyde. *Antiseptics* are a class of disinfectant that can be applied on body surfaces; they have a lower toxicity than environmental disinfectants and are usually less effective in killing microorganisms.
Sterilization is the complete removal or killing of all infectious agents in, or on, an object. Sterilization of equipment for surgery and wound dressings; parenteral administration of drugs, vaccines, or nutrients; catheterization; and dental work are all important means of controlling infectious diseases by killing infectious agents. Sterilization can be accomplished through use of fire, pressurized steam (such as in an autoclave), heated air, certain gases (such as ethylene oxide), ultraviolet light, ionizing radiation, liquid chemicals, and filtration. The method of sterilization chosen depends on the type of equipment to be sterilized.

The use of sterilized *disposable equipment*, such as disposable needles, syringes, and catheters, has the potential to reduce the risk of transmission

of infectious agents. However, it must be assured that such equipment is disposed of properly and is not reused. For example, disposable syringes cannot be properly resterilized because the plastic from which they are made cannot withstand the heat necessary for sterilization. Technologies, such as the single-use disposable needle and syringe developed for immunization programmes, help assure such disposable equipment is not reused.

Control and prevention programmes

The preceding sections have considered the issues and given examples of control measures for infectious diseases at individual, institutional, and community levels and the tools for control directed at the host, routes of transmission, and the agent. Control and prevention programmes using these tools must be developed according to a number of factors including: (1) the risk of disease; (2) the magnitude of disease burden (as measured by mortality, degree of disability, morbidity, and economic costs); (3) the feasibility of control strategies; (4) the cost of control measures; (5) the effectiveness of such measures (on current levels of disease and impact on future cases or outbreaks); (6) the adverse effects or complications of the control measures; and (7) the availability of resources. Public health planning for the control of infectious diseases must consider these issues to design optimal, evidence-based control and prevention programmes.

The tools of disease surveillance for recognition and evaluation of the patterns of disease can provide information on the risk and magnitude of disease burden to individuals, people in institutions, subgroups of populations, and the community at large. Establishment and maintenance of the infrastructure for surveillance, including a system for the reporting of notifiable infectious diseases and unusual events, must be a high priority.

Feasibility of possible control and prevention strategies must be assessed through operational research, pilot projects, or from field experience. The fact that a particular measure can help control a disease does not mean it can be applied on a sufficient scale to have the desired impact. The cost of control activities (in both human and material resources) can be assessed through costing studies that can also provide the data needed to conduct more rigorous cost–benefit and cost-effectiveness analyses. A costly

measure, even if it provides a high degree of control for an infectious disease, may not be affordable to the society or reasonable to apply in the light of other less expensive alternative strategies. For example, use of sterile, disposable products may be costlier than reusable, sterilizable products. Effectiveness of control measures may be assessed through epidemiological studies to find out their impact on reduction in the incidence or prevalence of disease.

The availability of resources for prevention and control programmes forces public health planners to set priorities by taking into account all these factors and then designing programmes that have maximum impact within available resources. Planners have a responsibility to mobilize additional necessary resources by raising public awareness and generating political will. Effective communication of disease burden and the results achievable through well-managed and effective control programmes can be a powerful tool for advocacy. Ideally, communities or their representatives should actively participate in the planning, execution, and evaluation of public health programmes.

Prevention effectiveness is 'the measure of the impact on health (including effectiveness, safety and cost) of prevention policies, programmes, and practices. The assessment of prevention effectiveness is the ongoing process of applying evaluation tools to prevention practices' (CDC 1995). Recognizing that systems for assessing the effectiveness of prevention strategies (including prevention strategies for infectious diseases) are weak or non-existent in both developing and industrialized countries alike, the CDC has suggested the following objectives for prevention effectiveness activities: 'evaluate the impact of prevention, use results of evaluation research to establish programme priorities, and establish or apply standardized methods to compare the benefits and effectiveness of alternative prevention strategies' (CDC 1995).

The current situation of *international migration* of many people worldwide presents an additional complexity to the design of programmes for the control of infectious diseases. Pertinent issues include refugee camps, legal status of migrants in recipient countries, and temporary return migration. Public health officials must consider the most effective mix of combined control measures applied to the host, agent, and routes of transmission when designing suitable control and prevention programmes (Gellert 1993). Particularly in terms of humanitarian emergencies, whether

due to human violence or natural disaster, crowding can make diseases such as measles, acute respiratory infections (ARI)s and meningitis of particular concern for groups of displaced persons (Heymann 2014). A recent statement by the American Society of Tropical Medicine and Hygiene's Executive Committee calls for US migration policy to be evidence-based, compassionate, and maintain international human rights law. They further argue that investment in the health needs of the world's vulnerable populations is beneficial not only to these affected populations, but to the United States as well, as infectious diseases are a global issue (Rosenthal et al. 2017).

International commerce and transportation are important areas of concern for public health infectious disease control programmes, especially as the speed of travel has increased. The tools of control include such measures as:

Spraying insecticides effective against mosquito vectors of malaria in aircraft before departure, in transit, or on arrival.
Rat-proofing or periodic fumigation to control rats on ships, docks, and warehouses to prevent plague.

Specific international control measures relating to aircraft, ships, and land transportation for infectious diseases are detailed in the *International Health Regulations (2005)* (WHO 2005).

The challenge facing infectious disease control programmes is to design an optimal set of interventions at local, institutional, community, national, and international levels supported and accepted by the political leadership and the people to whom these measures are applied.

Eradication

A unique end point in the control of infectious diseases is that of eradication. Eradication is the cessation of all transmission of an infectious agent by extermination of that agent. To date, only one human infectious disease, *smallpox*, and one animal infectious disease, rinderpest, have been eradicated. The WHO World Health Assembly in May 1980 confirmed its global eradication some 3 years after the last naturally acquired case of

smallpox in October 1977 (Fenner et al. 1988). The magnitude of this accomplishment is appreciated when one realizes that, in the early 1950s, it was estimated that 50 million cases of smallpox still occurred each year in the world, some 150 years after Edward Jenner had performed the first vaccination and wrote: ‘it now becomes too manifest to admit of controversy, that the annihilation of the Small Pox, the most dreadful scourge of the human species, must be the final result of this practice’ (Fenner et al. 1988). Rinderpest (known as cattle plague or steppe murrain) was an infectious viral disease of cattle and some other even-toed ungulates, that was associated with a high mortality rate. In 2011, after widespread vaccination campaigns, the United Nations declared the disease eradicated.

The goal of global eradication has been set by the World Health Assembly for two other infectious diseases, poliomyelitis caused by wild poliovirus and dracunculiasis (Guinea worm infection), the latter caused by the infectious agent *Dracunculus medinensis*. A high level of sustained political will, aggressively applied disease surveillance, and effective control measures are the required elements to achieve eradication of the infectious agents for these diseases.

Impressive progress has been made towards the global *eradication of poliomyelitis* since the 1988 World Health Assembly set the goal for its eradication. Polio cases have decreased by over 99% since that date, from some 350,000 cases in over 125 countries to only 22 reported cases in 2017. In 2018, only parts of three countries in the world (Nigeria, Pakistan, and Afghanistan) remain endemic for the disease (WHO 2018). Poliomyelitis control measures that will lead to eradication include the following:

Achieving and maintaining high levels of routine coverage of infants with at least three doses of oral polio vaccine.
Mass application of oral polio vaccine in countries where poliomyelitis is endemic through national immunization days, usually by providing oral polio vaccine to every child less than 5 years of age twice each year, separated by 4–6 weeks, and conducted during the low season of poliovirus transmission.
‘Mopping-up’ operations after the use of national immunization days has reduced transmission of disease to defined focal geographic areas, usually by providing oral polio vaccine house-to-house to all children less than 5 years of age on two occasions separated by 4–6 weeks.
Aggressive action-oriented surveillance for acute flaccid paralysis. Such surveillance includes: (1) case investigation, (2) a laboratory network for isolation and characterization of polio-viruses in suspect cases of poliomyelitis and people in close contact with them, and (3) limited outbreak response immunization providing one house-to-house round of oral polio vaccine to children less than 5 years of age living in the same village or neighbourhood of the patient.

Significant strides in the *eradication of dracunculiasis* have also been made. Over the last three decades the total number of dracunculiasis cases has declined by more than 99%. In 2017, only 30 cases were reported from just two countries, Chad and Ethiopia (WHO 2018a). Dracunculiasis control measures that are leading to its ultimate eradication include the following:

Establishing a national programme office, conducting baseline surveys, and preparing and refining a national plan of action.
Educating the population in endemic areas that the source of Guinea worm comes from their drinking water.
Ensuring that people with blisters or emerging worms do not enter sources of drinking water through behaviour changes and by converting step-wells into draw-wells.
Promoting the boiling or filtering of water through a fine mesh cloth to remove copepods. Treating drinking water with chlorine or iodine will also kill the copepods and infective larvae.
Providing non-infected water through construction of wells or rainwater catchments.
In selected endemic villages, controlling copepod populations with temephos (Abate®) insecticide placed in reservoirs, tanks, ponds, and step-wells.
Implementing an intensified surveillance and aggressive case-containment strategy as programmes get close to achieving eradication.

The eradication of a disease requires a unique set of conditions, including the following: (1) a defined, accessible reservoir of the infectious agent; (2) affordable and effective control measures that can interrupt the chain of infection directed at the host, agent, or route of transmission; and (3) a surveillance mechanism adequate to monitor and ultimately certify the disappearance of the infectious agent.

It is likely that measles may be targeted for global eradication in the future. Some countries and geographic regions have already targeted measles for elimination—a term sometimes used to describe the eradication of a disease from a large geographic area. Other diseases that may potentially be targeted for eradication in the future include: mumps, rubella, hepatitis B, leprosy, and diphtheria.

Emerging infectious diseases

New, emerging, and re-emerging infectious diseases have become a focus for the attention of public health prevention and control programmes in both industrialized and developing countries. Such infectious diseases have

thwarted any expectation that infectious diseases will soon be eliminated as public health problems and resulted in a widening spectrum of diseases, many of which were once thought to be almost conquered. Krause has reflected on this as follows:

> Bacteria reproduce every 30 minutes. For them, a millennium is compressed into a fortnight. They are fleet afoot, and the pace of our research must keep up with them, or they will overtake us. (Krause 1998)

Many factors contribute to the emergence of new or re-emergence of those previously known (Lederberg et al. 1992; CDC 1994; Murphy 1994), including the following:

Human *demographic change* by which people begin to live in previously uninhabited remote areas of the world and are exposed to new environmental sources of infectious agents, insects, and animals.
Breakdowns of sanitary and other public health measures in overcrowded cities and in situations of civil unrest and war.
Economic development and changes in the use of land, including deforestation, reforestation, and urbanization.
Climate change, including global warming, which may extend the favourable habitats for vectors such as mosquitoes.
Other *human behaviours*, such as increased use of child-care facilities, sexual and drug use behaviours, and patterns of outdoor recreation.
International travel and commerce that quickly transport people and goods vast distances.
Changes in food processing and handling, including foods prepared from many different individual animals and transported great distances.
Evolution of pathogenic infectious agents by which they may infect new hosts, produce toxins, or adapt by responding to changes in the host immunity.
Development of resistance of infectious agents such as *Mycobacterium tuberculosis* and *Neisseria gonorrhoeae* to chemoprophylactic or chemotherapeutic medicines.
Resistance of the vectors of vector-borne infectious diseases to pesticides.
Immunosuppression of people due to medical treatments or new diseases that result in infectious diseases caused by agents not usually pathogenic in healthy hosts.
Deterioration in surveillance systems for infectious diseases, including laboratory support, to detect new or emerging disease problems at an early stage.

Examples of emerging infectious disease threats include the following:

Coronavirus disease 2019 (COVID-19) a severe acute viral respiratory illness caused by the the coronavirus SARS-CoV-2 was first identified in December 2019 in Wuhan, China. It clearly demonstrates the ability for a previously unknown virus with an animal origin (bat) to cross species and create a pandemic killing millions of susceptible people.

Toxic shock syndrome, due to the infectious toxin-producing strains of *Staphylococcus aureus*, illustrates how a new technology yielding a new product, super-absorbent tampons, can create the circumstances favouring the emergence of a new infectious disease threat.

Lyme disease, due to the infectious spirochete *Borrelia burgdorferi*, illustrates how changes in the ecology, including reforestation, increasing deer populations, and suburban migration of the population, can result in the emergence of a new microbial threat that has now become one of the most prevalent vector-borne disease in the United States.

Shigellosis, giardiasis, and hepatitis A are examples of emerging diseases that have become threats to staff and children in child-care centres as the use of such centres has increased due to changes in the work patterns of societies.

Opportunistic infections, such as *Pneumocystis* pneumonia caused by *Pneumocystis jirovecii* (formerly *carinii*), chronic cryptosporidiosis caused by *Cryptosporidium* species, and disseminated cytomegalovirus infections, illustrate emerging disease threats to the increasing number of people who are immunosuppressed because of cancer chemotherapy, organ transplantation, or HIV infection.

Food-borne infections such as diarrhoea caused by the enterohaemorrhagic strain O157:H7 of *Escherichia coli* and water-borne infections such as gastrointestinal disease due to *Cryptosporidium* species are examples of emerging disease threats that have arisen due to such factors as changes in diet, food processing, globalization of the food supply, and contamination of municipal water supplies.

Hantavirus pulmonary syndrome first detected in the United States in 1993 and caused by a previously unrecognized hantavirus illustrates how exposure to certain kinds of infected rodents can result in an emerging infectious disease.

Nipah virus disease first detected in Malaysia in 1999 and caused by a previously unrecognized Hendra-like virus demonstrates how close contact with fruit bats and pigs can result in an emerging infectious disease.

Severe acute respiratory syndrome (SARS), a viral respiratory illness caused by the SARS-associated coronavirus (SARS-CoV) that was first reported in Asia in February 2003 and spread to more than 20 countries in North America, South America, Europe, and Asia affecting over 8,000 people and killing over 700 before the SARS global outbreak was contained in 2004, shows the need for continued vigilance in surveillance and outbreak response capacity (WHO 2003).

The emergence of the novel influenza A virus in birds, highly pathogenic avian influenza (H5N1), that has occasionally infected humans is an example of an antigenic shift in an influenza virus that, should it develop into a form easily transmissible from human to human, could potentially lead to a pandemic.

Middle East respiratory syndrome (MERS), a viral respiratory disease caused by a novel coronavirus (MERS-CoV) was first identified in Saudi Arabia in 2012. There have been 1,888 laboratory-confirmed cases and 670 deaths since 2012. It is an example of a virus that had infected an animal reservoir (camel) for over 20 years before being noted in humans (Cotton et al. 2014). The majority of cases (>85%) and deaths have been reported in the Arabian Peninsula. The virus was exported and resulted in the largest outbreak outside the region of origin, in South Korea, in 2015.

Ebola virus disease (EVD), a viral haemorrhagic fever that was first discovered in 1976 in the Democratic Republic of Congo and Sudan, was responsible for an outbreak of unprecedented magnitude it was discovered for the first time in West Africa in 2014. It is an example of a major outbreak of a disease that had been previously confined to more geographically isolated areas. There have been over 11,000 fatalities and the outbreak led to individual cases in a number of other countries due to international travel (CDC 2016a).

Zika virus, a vector-borne disease, was first isolated in Uganda in 1947 (Dick et al. 1952), but became a global concern with the outbreak of Zika virus in the Americas beginning in 2015. Zika virus can be transmitted via vector (*Aëdes* mosquitoes), vertical transmission from mother to foetus, via sexual contact or possibly blood transfusion. It is an example of an infectious disease being introduced into large susceptible populations with dramatic numbers of people affected. While cases are typically mild or asymptomatic, the disease can result in birth defects such as microcephaly (CDC 2016b).

Antimicrobial drug resistance as a major factor in the emergence and re-emergence of infectious diseases deserves special attention. Although significant reductions in infectious disease mortality have occurred since the introduction of antimicrobials for general use in the 1940s, antimicrobial drug resistance has emerged because of their widespread use in humans.

Drugs that once seemed invincible are losing their effectiveness for a wide range of community-acquired infections, including tuberculosis, gonorrhoea, pneumococcal infections (a leading cause of otitis media, pneumonia, and meningitis), and for hospital-acquired enterococcal and staphylococcal infections. Resistance to antiviral (e.g. amantadine-resistant influenza virus and aciclovir-resistant herpes simplex), antifungal (e.g. azole-resistant *Candida* species), and antiprotozoal (e.g. metronidazole-resistant *Trichomonas vaginalis*) drugs is also emerging. Drug-resistant malaria has spread to nearly all areas of the world where malaria occurs. Concern has also arisen over strains of HIV resistant to antiviral drugs. Increased microbial resistance has resulted in prolonged hospitalizations and higher death rates from infections; has required much more expensive, and often more toxic, drugs or drug combinations (even for common infections); and has resulted in higher healthcare costs (CDC 1994).

Antimicrobial drug resistance has also emerged because of the use of antimicrobials in domesticated animals. For example, the use of fluoroquinolones in poultry has created a reservoir of quinolone-resistant *Campylobacter jejuni* that has now been isolated in humans. The Food and Drug Administration's Veterinary Feed Directive (VFD) (Final Rule effective 1 October 2015) was put into place to regulate drugs used in animal feeds, and stipulates that VFD drugs must be used only under supervision of a licensed veterinarian (FDA 2015).

An aggressive public health response to these new, emerging, and re-emerging infectious disease threats must be made to characterize them better and to mount an effective response for their control. For example, the 1999 outbreak of West Nile fever in New York City and surrounding areas that, within a 4-year period, spread throughout the United States demonstrates how a viral encephalitis, initially classified as Saint Louis encephalitis and later confirmed to be due to West Nile virus, can reach far beyond its normal setting.

The WHO has outlined the following high-priority areas (WHO 1995):

Strengthen global surveillance of infectious diseases.
Establish national and international infrastructures to recognize, report, and respond to new disease threats.
Further develop applied research on diagnosis, epidemiology, and control of emerging infectious diseases.
Strengthen the international capacity for infectious disease prevention and control.

(Emerging infectious diseases are addressed in detail in Chapter 8.17.)

Bioterrorism: the deliberate use of biological agents to cause harm

Another unfortunate source of an infectious disease threat is the spectre of biological warfare or bioterrorism, especially in an age where terrorist acts are frequent events (Christopher et al. 1997). The 2002 World Health Assembly resolution urges member states ‘to treat any deliberate use, including local, of biological and chemical agents and radionuclear attack to cause harm also as a global public health threat, and to respond to such a threat in other countries by sharing expertise, supplies and resources in order rapidly to contain the event and mitigate its effects’ (WHO 2002).

The WHO recommends the following (WHO 2004):

Public health authorities, in close cooperation with other government bodies, should draw up contingency plans for dealing with a deliberate release of biological or chemical agents intended to harm civilian populations. These plans should be consistent or integral with existing plans for outbreaks of disease, natural disasters, large-scale industrial or transportation accidents, and terrorist incidents.

Preparedness for deliberate releases of biological or chemical agents should be based on standard risk-analysis principles, starting with risk and threat assessment to determine the relative priority that should be accorded to such releases in comparison with other dangers to public health in the country concerned. Considerations for deliberate releases should be incorporated into existing public health infrastructures, rather than developing separate infrastructures.

Preparedness for deliberate releases of biological or chemical agents can be markedly increased in most countries by strengthening the public health infrastructure, and particularly public health surveillance and response, and measures should be taken to this end.

Managing the consequences of a deliberate release of biological or chemical agents may demand more resources than are available, and international assistance would then be essential.

Many countries are developing rapid response capability to deal with such contingencies, especially in the light of the 2001 bioterrorist attack using anthrax in the United States.

Conclusion

Only through worldwide concerted action will the effort to control infectious disease be effective. We are now in an era where, as Nobel Laureate Dr Joshua Lederberg has stated, ‘The microbe that felled one child in a distant continent yesterday can reach yours today and seed a global pandemic tomorrow’ (quoted in CDC 1994). As Hans Zinsser stated over 60 years ago:

> Infectious disease is one of the few genuine adventures left in the world. The dragons are all dead and the lance grows rusty in the chimney corner. . . About the only sporting proposition that remains unimpaired by the

relentless domestication of a once free-living human species is the war against those ferocious little fellow creatures, which lurk in the dark corners and stalk us in the bodies of rats, mice, and all kinds of domestic animals; which fly and crawl with the insects, and waylay us in our food and drink and even in our love (Hans Zinsser 1934, quoted in Murphy 1994).

ACKNOWLEDGEMENTS

Text extracts from Heymann, D.L. (ed.), *Control of Communicable Diseases Manual*, Nineteenth Edition, American Public Health Association, Washington, USA, Copyright © American Public Health Association 2008, with permission from American Public Health Association.

Text extracts from World Health Organization (WHO), *Public Health Response to Biological and Chemical Weapons: WHO Guidance*, Second edition *of Health Aspects of Chemical and Biological Weapons: Report of a WHO Group of Consultants*, WHO, Geneva, Switzerland, Copyright © 2004, reproduced with permission from the World Health Organization, available from http://whqlibdoc.who.int/publications/2004/9241546158.pdf.

REFERENCES

ACIP (Immunization Practices Advisory Committee) (2018). *Vaccine Recommendations and Guidelines of the ACIP*. [Online] Available at: https://www.cdc.gov/vaccines/hcp/acip-recs/general-recs/index.html.

Bennett J.E., Dolin, R., and Blaser, M.J. (eds.) (2020). *Mandell, Douglas and Bennett's Principles and Practice of Infectious Diseases* (9th ed.). Philadelphia, PA: Elsevier Saunders.

Beran, G.W. and Steele, J.H. (eds.) (1994). *Handbook Series of Zoonoses*. Boca Raton, FL: CRC Press.

Blumber, H., et al. (2003). American Thoracic Society/Centers for Disease Control and Prevention/Infectious Diseases Society of America: treatment of tuberculosis. *American Journal of Respiratory and Critical Care Medicine*, 167(4), 603–62.

Centers for Disease Control and Prevention (CDC) (1994). *Addressing Emerging Infectious Disease Threats: A Prevention Strategy for the United States*. Atlanta, GA: CDC.

Centers for Disease Control and Prevention (CDC) (1995). *Prevention Effectiveness: Making Prevention a Practical Reality*. Atlanta, GA: CDC.

Centers for Disease Control and Prevention (CDC) (2012a). *Estimates of Foodborne Illness in the United States*. Atlanta, GA: CDC. [Online] Available at: http://www.cdc.gov/foodborneburden/index.html

Centers for Disease Control and Prevention (CDC) (2012b). *Feature: Raw (Unpasteurized) Milk*. Atlanta, GA. [Online] Available at: http://www.cdc.gov/Features/RawMilk
Centers for Disease Control and Prevention (CDC) (2016). Summary of notifiable diseases, United States, 2014. *Morbidity and Mortality Weekly Report*, 63(54), 1–152.
Centers for Disease Control and Prevention (CDC) (2016a). Ebola (Ebola virus disease). [Online] Available at: https://www.cdc.gov/vhf/ebola/outbreaks/2014-west-africa/qa.html
Centers for Disease Control and Prevention (CDC) (2016b). Zika virus. [Online] Available at: https://www.cdc.gov/zika/about/overview.html
Cherry, J.D., et al (ed.) (2014). *Feigin and Cherry's Textbook of Pediatric Infectious Diseases* (7th ed). Philadelphia, PA: W.B. Saunders.
Christopher, G.W., et al. (1997). Biological warfare: a historical perspective. *Journal of the American Medical Association*, **278**, 412–17.
Committee on Infectious Diseases (2018). *American Academy of Pediatrics: Red Book: 2018 Report of the Committee on Infectious Diseases* (31st ed.). Elk Grove Village, IL: American Academy of Pediatrics.
Cotton, M., et al. (2014). Spread, circulation, and evolution of the Middle East respiratory syndrome coronavirus. *MBio*, 5(1), e01062–13.
Dick, G., et al. (1952). Zika virus. I. Isolations and serological specificity. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 46(5), 509–20.
Fenner, F., Henderson, D.A., Arita, I., Jecek, Z., and Ladnyi, I.D. (1988). *Smallpox and its Eradication*. Geneva: WHO.
Fine, P.E. (1993). Herd immunity: history, theory, practice. *Epidemiologic Reviews*, **15**, 265–302.
Food and Drug Administration (FDA) (2015). Veterinary Feed Directive; Final Rule, 80 FR 31707. 21 CFR Parts 514 and 558, 31707–35.
Gellert, G.A. (1993). International migration and control of communicable diseases. *Social Science & Medicine*, **37**, 1489–99.
Gorbach, S.L., Bartlett, J.G., and Blacklow, N.R. (eds.) (2004). *Infectious Diseases* (3rd ed.). Philadelphia, PA: Lippincott Williams and Wilkins.
Guerrant, R.L., Walker, D.H., and Weller, P.F. (eds.) (2011). *Tropical Infectious Diseases: Principles, Pathogens, and Practice* (3rd ed.). Philadelphia, PA: W.B. Saunders.
Heymann, D.L. (ed.) (2014). *Control of Communicable Diseases Manual* (20th ed.). Washington, DC: American Public Health Association.
Insel, R.A., Amstey, M., Woodin, K., and Pichichero, M. (1994). Maternal immunization to prevent infectious diseases in the neonate or infant. *International Journal of Technology Assessment in Health Care*, **10**, 143–53.
Kang, M., et al. (2010). Surgical treatment for multidrug-resistant and extensive drug-resistant tuberculosis. *The Annals of Thoracic Surgery*, 89(5), 1597–602.
Kempker, R.R., et al. (2012). Surgical treatment of drug-resistant tuberculosis. *The Lancet Infectious Diseases*, 12(2), 157–66.
Kim-Farley, R.J. (2017). The dangerous curve and the guardrail: disease and vaccination. *American Journal of Public Health*, 107(1), 15–17.
Kiple, K.F. (ed.) (1993). *The Cambridge World History of Human Disease*. Cambridge: Cambridge University Press.
Krause, R.M. (ed.) (1998). *Emerging Infections: Biomedical Research Reports*. New York: Academic Press.
Lederberg, J., Shope, R.E., and Oaks, S.C. Jr. (eds.) (1992). *Emerging Infections: Microbial Threats to Health in the United States*. Washington, DC: National Academy Press.
Lieberthal, A.S., et al (2013). The diagnosis and management of acute otitis media. *Pediatrics*, 131(3), e964–99.

Liu, C., et al. (2011). Clinical practice guidelines by the Infectious Diseases Society of America for the treatment of methicillin-resistant *staphylococcus aureus* infections in adults and children. *Clinical Infectious Diseases*, 52(3), e18–55.

MacNeily, A.E. and Afshar, K. (2011). Circumcision and non-HIV sexually transmitted infections. *Canadian Urological Association Journal*, 5(1), 58–9.

McKibben, L., et al. (2005). Guidance on public reporting of healthcare-associated infections: recommendations of the Hospital Infection Control Practices Advisory Committee. *American Journal of Infection Control*, **33**, 217–26.

McNeill, W.H. (1976). *Plagues and Peoples*. Garden City, NY: Anchor Press/Doubleday.

Mina, M.J., et al (2015). Long-term measles-induced immunomodulation increases overall childhood infectious disease mortality. *Science*, 348(6235), 694–9.

Murphy, F.A. (1994). New, emerging, and reemerging infectious diseases. *Advances in Virus Research*, **43**, 1–52.

Occhionorelli, S., et al (2014). Prophylactic appendectomy during laparoscopic surgery for other conditions. *Case Reports in Medicine*, **2014**, 292864.

Osterholm, M.T. and Hedberg C.W. (2005). Epidemiolologic principles. In: Mandell, G.L., Bennett, J.E., and Dolin, R. (eds.) *Mandell, Douglas and Bennett's Principles and Practice of Infectious Diseases* (6th ed.). New York: Elsevier Churchill Livingstone, pp. 158–68.

Patel, H.H., et al (2014). Indications for tonsillectomy: a 10 year retrospective review. *International Journal of Pediatric Otorhinolaryngology*, 78(12):2151–5.

Plotkin, S.A., Orenstein, W.A., and Offit P.A. (ed.) (2012). *Vaccines* (6th ed.). Philadelphia, PA: W.B Saunders.

Rogozov, V. and Bermel, N. (2009). Auto-appendectomy in the Antarctic: case report. *British Medical Journal*, 339, b4965.

Rosenthal, P.J., et al. (2017). Evidence-based policies on migration and global health are essential to maintain the health of those inside and outside the United States. *American Journal of Tropical Medicine and Hygiene*, 96(1), 5–6.

Scott, R.D., II (2009). *The Direct Medical Costs of Healthcare-Associated Infections in U.S. Hospitals*. Atlanta, GA: Division of Healthcare Quality Promotion, National Center for Preparedness, Detection, and Control of Infectious Diseases, CDC.

Siegel, J.D., Rhinehart, E., Jackson, M., Chiarello, L., and the Healthcare Infection Control Practices Advisory Committee (2007). *2007 Guideline for Isolation Precautions: Preventing Transmission of Infectious Agents in Healthcare Settings.* [Online] Available at: http://www.cdc.gov/niosh/docket/archive/pdfs/NIOSH-219/0219-010107-siegel.pdf

Sihoe, A., et al. (2009). The current role of thoracic surgery in tuberculosis management. *Respirology*, 14(7), 954–68.

Slaven, E.M., Stone, S.C., and Lopez, F.A. (ed.) (2006). *Infectious Diseases: Emergency Department Diagnosis & Management. Red and White Emergency Medicine Series*. New York: McGraw-Hill.

Smith, D.K., et al. (2010). Male circumcision in the United States for the prevention of HIV infection and other adverse health outcomes: report from a CDC consultation. *Public Health Reports*, 125(4), 517.

Solomon, S.L., Fraser, D.W., and Kaplan, S.L. (2009). Public health considerations. In: Feigin, R.D., Cherry, J.D., Demmler-Harrison, G.J., and Kaplan, S.L. (eds.) *Textbook of Pediatric Infectious Diseases* (5th ed.). Philadelphia, PA: W.B. Saunders, pp. 3221–44.

Wertheim, H.F.L., Horby, P., and Woodall, J. (ed.) (2012). *Atlas of Human Infectious Diseases*. Oxford: Wiley-Blackwell.

World Health Organization (WHO) (1995). *Communicable Disease Prevention and Control: New, Emerging, and Re-Emerging Infectious Diseases*. No. A48/15. Geneva: WHO.

World Health Organization (WHO) (2002). *Global Public Health Response to Natural Occurrence, Accidental Release or Deliberate Use of Biological and Chemical Agents or Radionuclear Material That Affect Health*. World Health Assembly Resolution, WHA55.16, Geneva: WHO.
World Health Organization (WHO) (2003). *Communicable Disease 2002: Global Defence Against the Infectious Disease Threat*. WHO/CDS/2003.15. Geneva: WHO.
World Health Organization (WHO) (2004). *Public Health Response to Biological and Chemical Weapons*. Geneva: WHO.
World Health Organization (WHO) (2005). *International Health Regulations (2005)*. Geneva: WHO.
World Health Organization (WHO) (2008). *The Global Burden of Disease: 2004 Update*. Geneva: WHO.
World Health Organization (WHO) (2008a). *Guidelines for the Programmatic Management of Drug-resistant Tuberculosis. Emergency Update 2008*. Geneva: WHO.
World Health Organization (WHO) (2012b). *International Travel and Health*. Geneva: WHO.
World Health Organization and United Nations Children's Fund (2015). *Progress on Drinking Water and Sanitation 2015 Update*. Geneva: WHO.
World Health Organization (WHO) (2017). *Male Circumcision for HIV Prevention*. Geneva: WHO. [Online] Available at: http://www.who.int/hiv/topics/malecircumcision/en/.
World Health Organization (WHO) (2018). *Poliomyelitis*. Fact sheet. Geneva: WHO. [Online] Available at: http://www.who.int/news-room/fact-sheets/detail/poliomyelitis.
World Health Organization (WHO) (2018a). *Dracunculiasis*. Fact sheet. Geneva: WHO. [Online] Available at: http://www.who.int/en/news-room/fact-sheets/detail/dracunculiasis-(guinea-worm-disease)
World Health Organization (WHO) (2020). The Top 10 Causes of Death. Geneva: WHO. [Online] Available at: http://www.who.int/newsroom/factsheets/detail/the-top-10-causes-of-death.

11.5

Medical screening

Theories, methods, and effectiveness

Tang Jin-ling and Li Li-ming

Introduction

Since the middle of last century, cancer, and cardiovascular disease have overtaken infectious disease as the top killers of human life. Cardiovascular disease can be successfully prevented through modification of risk factors, whereas against cancer we rely largely on early detection and early treatment. On one hand, checks for risk factors of cardiovascular disease has become almost a clinical routine with a hope to prevent severe complications. On the other hand, screening for cancer has rapidly become a widespread medical practice in the past decades. Genetic testing, one of the most exciting scientific breakthroughs in biomedical sciences in the twentieth century, further raised the enthusiasm for screening. Screening in various different disguises has become an important part of modern public health and clinical medicine.

Broadly speaking, screening exists in forms from firmly established and well-planned population screening programmes to ubiquitous use of diagnostic methods for early diagnosis to experiments with new diagnostic methods by individual clinicians in search of silent diseases. Well-organized screening programmes can indeed do more good than harm but the obsessive search of silent diseases may have often caused overdiagnosis and overtreatment. In this chapter, we will discuss the theories, methods, effects, and pitfalls of health screening in both the narrow and broad concepts with an aim to promote appropriate use of disease screening and efficient use of healthcare resources. Cancer will be used most of time as examples to illustrate when necessary.

What is medical screening?

After decades of tremendous efforts fighting against chronic diseases, we have yet to find preventive and therapeutic methods that are as powerful as hygiene, antibiotics, and vaccines used for combating infectious diseases.

Chronic diseases differ from infectious diseases in an important way that it takes years or even decades to become severe enough and cause symptoms, suffering, disabilities, or death. Moreover, rapid advances in biomedical technology, notably in testing and imaging, have made it possible to detect many chronic diseases when they are still asymptomatic. New diagnostic innovations are also applied to identify risk factors of chronic disease for primary prevention. It is further assumed that intervention on risk factors can prevent the occurrence of disease and treatment of silent disease can retard or stop disease progression, improve prognosis, and ultimately reduce suffering and mortality. Early detection and early intervention have formed the foundation for combating chronic disease in modern medicine. In public health, this strategy is known as medical screening, health screening or population screening (Wilson and Jungner 1968; Morrison 1992; NSC 1998; Raffle and Gray 2007).

Screening is defined as the systematic application of a simple test or inquiry to well persons to detect those having a specific preclinical disorder or increased risk of disease for the purpose of early treatment or prevention with an ultimate aim of improving the health of the screened population. In brief, screening is early detection and early intervention aimed to improve health. Figure 11.5.1 uses Pap smears for screening cervical cancer to show the participants flow and main components of a screening programme.

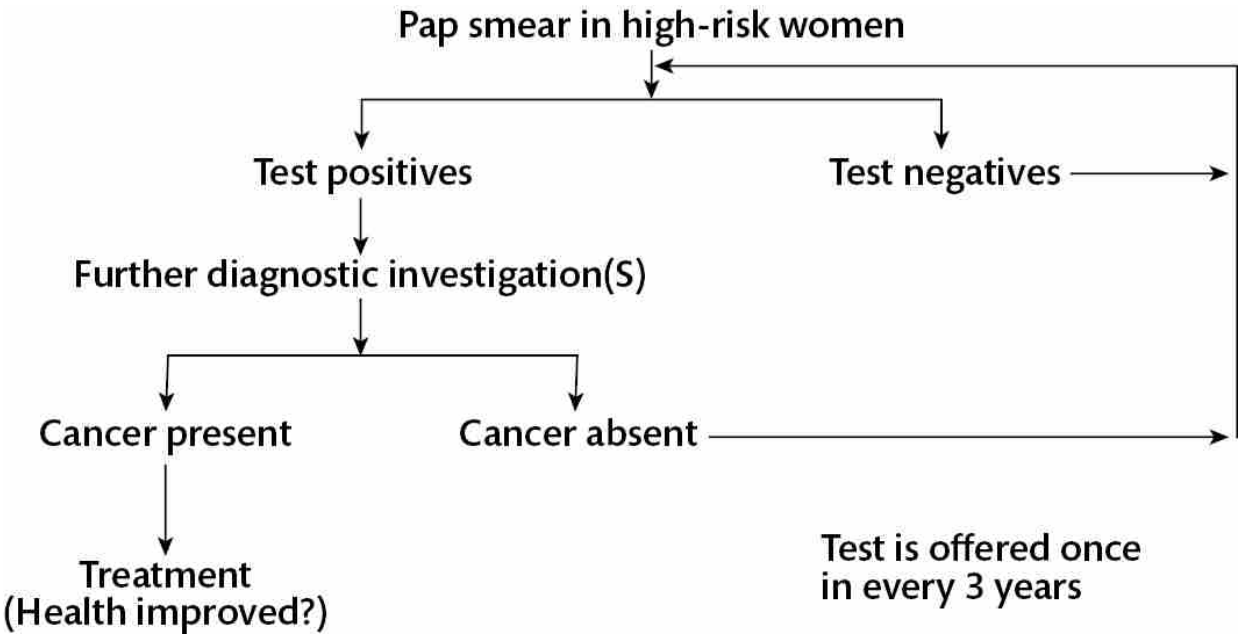


Fig. 11.5.1 Participants flow and components of a screening programme: Pap smear for screening cervical cancer.

Public health tends to emphasize that screening be a systematic approach. Screening is not simply the application of a new test; it is a carefully planned, systematically organized, and regularly repeated, collective effort that combines early detection and early treatment with an ultimate goal of improving those who are screened. 'Systematic' also means that the decisions on issues such as whether or not to screen, who to screen, and how to screen, how to further confirm diagnosis, and how to treat should be carefully determined based on scientific evidence before a screening programme is implemented so as to make sure that it is desired by and will benefit the population to be screened.

The 'test' is referred to a wide range of medical procedures that can be performed in order to detect a risk factor and a silent disease, such as laboratory tests and imaging techniques and endoscopic examinations. The 'inquiry' is referred to an interview or questionnaire used to obtain information about a person's occupational history, family history, disease history, lifestyle, and other information that can be used to estimate current or future risk of disease. Information from tests and inquiries can be used independently or more often in combination to predict the risk of disease, for example, by using scores or multiple regression models such as the Framingham cardiovascular risk equation.

Screening is offered to ‘well persons’ rather than patients. Well persons are not those who are totally free of any disorder. They may have disorders other than the one to be screened for and even the screened disorder but have not sought medical attention on account of symptoms of that disorder. Different from clinical practice, those who are screened are not those who have brought themselves to doctors for suspecting an illness but are those invited by healthcare professionals for checks and inquiries. This is a considerable deviation from the long-practised medical paradigm in which patients only see the doctor when they are sick and thus creates new ethical concerns.

For cancer, screening is mostly targeted at a disease at its preclinical stage. The goal is to prevent more severe complications or death from the disease. Pap smears, mammograms, and faecal occult bleed tests are examples in case. Screening can also be targeted at risk factors for prevention of the first occurrence of disease. For example, smoking, hypertension, hyperlipidaemia, and elevated blood glucose are often checked with a hope to prevent myocardial infarction and stroke through lifestyle and prophylactic interventions on these risk factors.

Although we normally talk about screening for a single disease, a screening programme can be multiphasic and targeted at more than one disease or all major risk factors of a disease at the same time. In this sense, periodical health checks can also be viewed as a kind of screening although with more complexities. Health checks can identify overt diseases. But health checks for silent diseases and risk factors should be assessed with similar scientific rigour and managed with similar standards as we do for screening.

Unlike drugs, a screening programme is a complex intervention with many components. Its potential benefits and harms are determined by multiple factors including the test, further diagnostic investigations, follow-up interventions, and the way they are organized. Screening will fail if any of the major components fail.

Last but not least, screening must be able to achieve the ultimate goal of improving the health of the screened population. Against this goal, many screening programmes and similar efforts may not be justified.

Some historical accounts

The aforementioned section depicts the picture of screening in the eyes of public health workers. However, screening can be a different concept to others including clinicians and the public. In fact, screening did not appear in this exact modern clothe at its onset. Screening is an evolving concept and new ideas are still being proposed and adopted.

Pre-modern screening before 1960

The idea of screening is not new. As early in 1861, Dr Horace Dobell, an eminent British physician recommended that physicians should perform periodic checks on everyone, irrespective of whether they were seeking medical help. In 1990, Dr George Gould, an American physician, argued that ranchers check their cattle, merchants check their stock, accountants check their books, generals check their armies, governments check their budgets, but why physicians do not regularly check the healthy people——and they should. However, their recommendation for regular and non-specific medical examinations of healthy people did not take firm hold immediately.

Historically, it is generally believed that the concept of modern screening originated in early last century from the practice of health checks organized by the army for selecting out those who were unsuited to the needs of the military machine, by insurance companies as a means of selecting premiums, and by large cooperations in selecting suitable employees (Raffle and Gray 2007). Early screening programmes also included testing for infectious disease, for example, tuberculosis, which was typically organized by the government. Most early screening activities, however, did not meet the stringent criteria for modern screening programmes. These criteria for screening are the results of scientific evaluation of health checks and early screening programmes and of better understanding of the natural history of chronic diseases in the past few decades.

In the 1950s, screening was already firmly in place in the United States, whereas it was just beginning in Britain. For example, in 1957, the Commission on Chronic Illness in the United States recommended a wide range of screening, including diabetes, glaucoma, and cancer of the mouth, skin, breast, cervix, and rectum. However, scientific evidence for the effectiveness of screening was lacking and doubts about screening started to mount.

Modern screening era since 1960

In particular, in 1968, two reports were published, one is by the Nuffield Provincial Hospital Trust and the other by the World Health Organization. Both questioned some of the general beliefs of screening. The Nuffield Trust evaluated 10 existing screening or proposed screening activities in the United Kingdom and concluded most of them were ineffective or even harmful. Wilson and Jungner's report for the World Health Organization discussed various aspects of screening and proposed, importantly, 10 criteria a screening programme should meet. They further argued that even if these principles were all met, research was still needed to provide evidence on the benefit, harm, and affordability of a screening programme.

Two randomized controlled trials were established to evaluate the effectiveness of screening, one at Kaiser Permanente in 1964 and the other in South-East London in 1967. Both trials failed to find a positive result on multiple outcomes including death that would support multiphasic screening. Decades later, these early reports and evaluation trials on screening have eventually reoriented the practice of screening from multiphasic health checks to specific disease screening: general health checks are not screening. Today, we have cumulated a great amount of evidence about modern screening and found often it was not effective or marginally effective. But they always cost and have arms.

Screening in the twenty-first century

In the past, screening was decided by experts, delivered as a utilitarian programme, and complied with by dutiful and grateful patients. But things have changed in the twenty-first century. There are pressures for change in screening. Transition away from paternal paradigm of medical care, heightened expectations from patients, right of patients for information, increasing new technologies among other factors have been changing the concepts, principles, and practice of screening.

The concept of screening that was firmly established in the late last century has never been universally accepted without challenges and variations. In fact, the early practice of screening has never truly disappeared; they continue to exist in different covers. In addition, new ideas of detecting silent diseases and checks for risk factors are being developed and adopted by millions of individual healthcare workers without

demonstrated evidence on their benefits and harms. In the name of early diagnosis of disease, many variations of screening truly exist, from experimental use by individual clinicians, to improvised or opportunistic testing in clinics and hospitals, to popular health checks, and to well-organized efficacy-proven screening programmes for specific diseases.

The rapid rise of general health checks with an increasingly large number of items included in it represents a worrying return to the early practice of screening before the 1960s. These new or so-called new ideas and atypical screening activities must also be brought into discussion as they constitute much of the inappropriate or even bad practice of screening and their effectiveness is uncertain. We should know what bad screening activities are like so we can prevent them from being introduced. It is particularly difficult to stop an ineffective intervention once they are already in practice. For example, prostate-specific antigen (PSA) was first introduced in the late 1980s in Western countries and is now still widely used although it has been proven ineffective (Hoffman 2011).

When would screening be potentially effective?

The concept of identifying future risks or the potential of disease of individuals is remarkably attractive and has permeated to nearly all aspects of medicine. The ever-increasing technological and biological wizardry of medicine raises expectations and further tempts clinicians and the public. However, so few medical interventions are like screening that is so appealing at the first glance but disappointing at the end. In addition, screening is so complex and difficult to be provided with high standard and quality so that they will eventually realize the benefits promised. As a result, only a few screening programmes can truly do more good than harm and are thus worthwhile (only programmes that can do more good than harm at a reasonable cost should be recommended). When implemented, they should be managed in such a way as to achieve a level of quality which can ensure that the balance of good and harm demonstrated in research is reproduced in the ordinary service setting.

What would make a screening programme potential beneficial? The potential benefits and harms of a screening programme are mostly determined by the following three factors: the disease, the screening test(s), and the treatment(s). We will explain why in the following sections.

The disease

About the disease, three important aspects are relevant and should be assessed: (1) the disease to be screened for must have a long detectable preclinical phase; (2) the severity of the disease must be considered as a large health burden to the population in terms of suffering, disability, destitution and/or death; and (3) the preclinical phase of the disease must be relatively prevalent.

The preclinical or asymptomatic phase of a disease is the time period from the onset of the disease to the time symptoms appear and a diagnosis is made routinely by a doctor (Fig. 11.5.2). A long preclinical phase is the prerequisite for the disease to be detected and treated early enough so that treatments would be more effective. For example, most infectious and acute diseases that have a very short asymptomatic phase are not suitable for screening, while most cancers may have a latent period of many years are often a suitable target for screening.

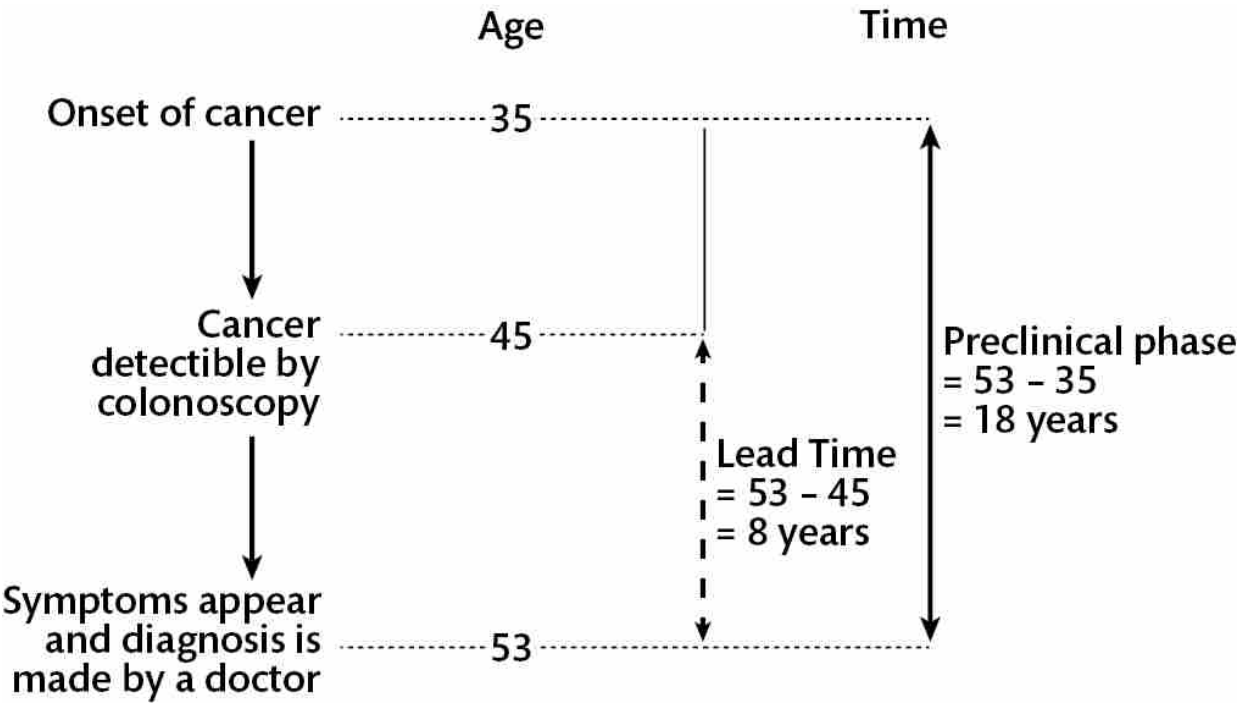


Fig. 11.5.2 Screening test and lead time in a hypothetical case of colon cancer.

Most cancers may have a long preclinical phase but the actual length of that period is normally unknown. First, it is almost impossible to know when a cancer started in a person. Second, most cancer patients currently diagnosed are asymptomatic and are almost always offered treatment and the natural progress of cancer is thus altered. This makes it impossible to know when symptoms would occur in these patients if there were no treatment. Lack of knowledge on the preclinical period of cancer makes it difficult to estimate accurately the effectiveness of screening and creates unsubstantiated belief and blind enthusiasm for screening.

The burden of disease is a relative concept. Diseases such as cancer that causes severe suffering, disability, and death are generally more suitable for screening than diseases such as minor skin problems. The view on the burden of a disease may vary greatly across populations and over time in the same population. For example, in some African countries where infectious disease is the major threat to life, cancer would be deemed much less important than in Western countries where cancer has become the major killer. Screening and management of hypertension, hyperlipidaemia, and diabetes mellitus for prevention of major cardiovascular complications were certainly less eminent a century ago than today in most populations.

The prevalence determines the total number of cases that can be detected by screening in the screened population. If the disease is rare and too few cases can be identified, no matter how effective the treatment would be, screening would not be able to have an important impact on the health of the population. When would a disease be deemed prevalent and justified for screening? There is no universal answer to the question either. Again the answer may vary across populations and over time dependent on resources available, opportunity costs, and values of people at the time the decision needs to be made.

The prevalence of a disease is positively related to the prevalence of the risk factors and the average duration of the preclinical phase. These relations suggest that screening be more likely effective in people with increased risk factors of the screened disease. Second, the prevalence of the undiagnosed but detectable preclinical disease also depends on whether or not the population has been screened recently. Recent screening in a population would reduce the number of undiagnosed asymptomatic patients. This suggests that in order for the screening programme to be cost-effective, two subsequent rounds of testing should not be kept too close to each other in

time. For example, breast mammogram or Pap smear should be performed again in the same person every 3–5 years rather than more frequently.

Risk factors and early-stage cancer can be considered similar in that both increase the future risk of severe complications, suffering, and death. Thus, cardiovascular risk factors that can increase the risk of myocardial infarction and stroke are often also screened and corrected with medications. These factors usually include hypertension, hyperlipidaemia, and type II diabetes, which are common and can effectively controlled through lifestyle modification and therapeutic drugs.

The screening test

The screening test refers to the test or a series of tests used in a screening programme to identify those who are likely to have the disease and warrant further confirmatory diagnostic investigations. Screening is feasible only if a good screening test is available. The test should be sufficiently accurate. Precisely, it should have relatively a high sensitivity and specificity. The ability of a test to correctly classify people who truly have the preclinical disease as test positive is called sensitivity and the ability to correctly classify those who are truly free of the disease as negative is called as specificity (Table 11.5.1).

Table 11.5.1 Classification of disease status by the gold standard and screening test

Test	Gold standard		
	Dis. (+)	Dis.(–)	Totals
Test (+)	a True positive	b False positive	a + b
Test (–)	c False negative	d True negative	c + d
Totals	a + c	b + d	N = a + b + c + d

If sensitivity and specificity are not perfect, which is always the case, errors will occur in the test results. Those who truly have the disease but are test negative, suggesting absence of disease are false negative results. False negative results will lead to missing the opportunity of early diagnosis and treatment. On the other hand, those who are truly free of the disease but test positive, suggesting presence of disease are false positive results. False positive results would result in unnecessary confirmatory diagnostic investigations and even further unnecessary treatments and harms. As screening can detect and treat conditions that will never cause symptoms and suffering in the rest of a person's life, unnecessary treatment is much commoner than usually expected.

Furthermore, the screening test should be able to give a sufficient lead time. Lead time is the period from the time an asymptomatic patient is detected by screening to the hypothetical time symptoms would appear and diagnosis be made by a doctor if the screening test had not been performed (Fig. 11.5.2). For a group of patients, this lead time can be approximated with some data and assumptions (Morrison 1992). For any individual patient, lead time is always a hypothetical concept and inestimable. For a screening programme to be effective in reducing morbidity and mortality in a population, there must be enough lead time in a sufficient number of cases detected. If lead time is insufficient, patients will not be treated in time to retard or stop the progression of the disease.

Finally, the test should also be simple, rapid, inexpensive, and safe so that it can be applied to a large number of people at an affordable cost and people and clinicians will accept it and make efforts to do it (Cochrane and Holland 1971; Sackett and Holland 1975; Morrison 1992) Otherwise, screening could become either too expensive to be affordable or too unattractive so people will not come for it. For example, conduct of the test should be easy to learn. Those that can be administered by non-medical personnel will cost less than those that require skills from years of medical training. Moreover, screening that requires hospitalization for a few days to go through complex and invasive procedures are unlikely to be popular to eligible participants.

The treatment

If screening can bring about any good, it will be from the treatment rather than the test and any other activities in it. Therefore, for a screening

programme to be potentially beneficial, there must be treatments available that are effective and acceptable in asymptomatic patients. Neither do treatments used for late-stage patients necessarily help early-stage patients nor would treatments acceptable to symptomatic patients equally acceptable to asymptomatic patients. In brief, the treatments for early stage of patients must be re-evaluated in randomized controlled trials if they are to be used in screening for treating preclinical patients.

Performance of the screening test

Screening is a combination of early detection and early treatment of preclinical diseases. Success of a screening programme is thus first determined by the test. The test is related to both the benefits and costs and harms of screening. This section will explain how they are related and how we can 'manipulate' the test so as to increase benefits and reduce costs and harms.

What would make a good test for screening? As already described, a good test for screening should be accurate, feasible, and acceptable. We will now analyse the role of accuracy for screening. Accuracy is referred to the ability of a test to correctly classify true disease status, either presence or absence. It is thus commonly expressed in two separate parameters: sensitivity and specificity.

Sensitivity and specificity

Sensitivity is the ability to detect the disease in those who truly have the disease, while specificity is the ability to rule out the possibility of disease in those truly free of the disease.

Sensitivity and specificity are estimated through scientific research. Such research normally starts with two groups of people, one truly with the disease of concern and the other truly free of the disease. Then the test is applied to all participants in both groups. Normally, a positive test result will suggest the presence of disease and a negative result absence of disease. According to true disease status and test results, the participants can be classified into four categories: (a) true positive; (b) false positive; (c) false negative; and (d) true negative (Table 11.5.1).

Sensitivity is expressed as the percentage of those having a positive result in those who truly have the disease and specificity is the percentage of those having a negative result in those truly free of the disease. Sensitivity can also be viewed as the probability of a true patient being detected by the test and specificity the probability of a person truly free of the disease being confirmed by the test. If sensitivity is not perfect, some true patients will turn out to be test negative. We call these negative results false negatives. Conversely, if specificity is not perfect, some of those truly free of the disease will turn out to be test positive. We call these positive results false positives (Table 11.5.1).

For example, assuming there are 100 participants truly having the disease and 200 truly free of the disease included in a diagnostic study, 95 out of the 100 participants truly having the disease turned out to be positive (a = true positives) and 5 negative (c = false negatives); 180 out of the 200 people truly free of the disease turned out to be test negative (d = true negatives) and 20 positive (c = false negatives). This gives

Sensitivity = a/(a + c) = 95/(95 + 5) = 95%
Specificity = d/(b + d) = 180/(20 + 180) = 90%

Predictive values

Information on the accuracy of a test is not directly useful for practice, where we have only the test results with which we wish to find out the true disease status. In practice, the questions concerned are: (1) how likely is a person to truly have the disease given a positive test result, and (2) how likely is a person to be truly free of the disease given a negative test result? The answer to the first question can be estimated by the positive predictive value and the second by the negative predictive value. The positive predictive value (PPV) is the probability of truly having the disease given a positive test result and the negative predictive value (NPV) the probability of being truly free of the disease given a negative test result.

However, predictive values cannot be estimated simply from sensitivity and specificity without knowing the prevalence of the disease in the population to which the test is applied. Put it differently, estimation of predictive values also requires knowledge on the prevalence of disease. Given sensitivity, specificity and prevalence, predictive values can be estimated by the following equations:

PPV = P.SN / {P.SN + (1–P)(1–SP)}

$$NPV = (1–P)SP / \{P(1–SN) + (1–P)SP\}$$

Where SN = sensitivity

SP = specificity

P = the prevalence of the disease in the population tested

From the aforementioned equations, we can see that predictive values are determined by characteristics of both the test and disease, precisely, the accuracy of the test and the prevalence of the disease in the population to which the test is applied. The prevalence of the disease is also called the pre-test probability of disease. The accuracy of the test is normally assumed to be a stable feature of a test and can be generalized from the research population to other populations, although this stability assumption may not be necessarily true in all circumstances. As a result, the same positive (or negative) test result may mean a different probability of having (or being free of) the disease in populations with a different prevalence of disease.

Assuming that a test has both sensitivity and specificity of 95%, a very good test for screening, the predictive values will differ hugely as the prevalence varies (Fig. 11.5.3). As the prevalence increases, the PPV will increase but the NPV will decrease. As screening is normally conducted in populations with a relatively low prevalence of disease, results of a screening test will generally have a low PPV and high negative predictive.

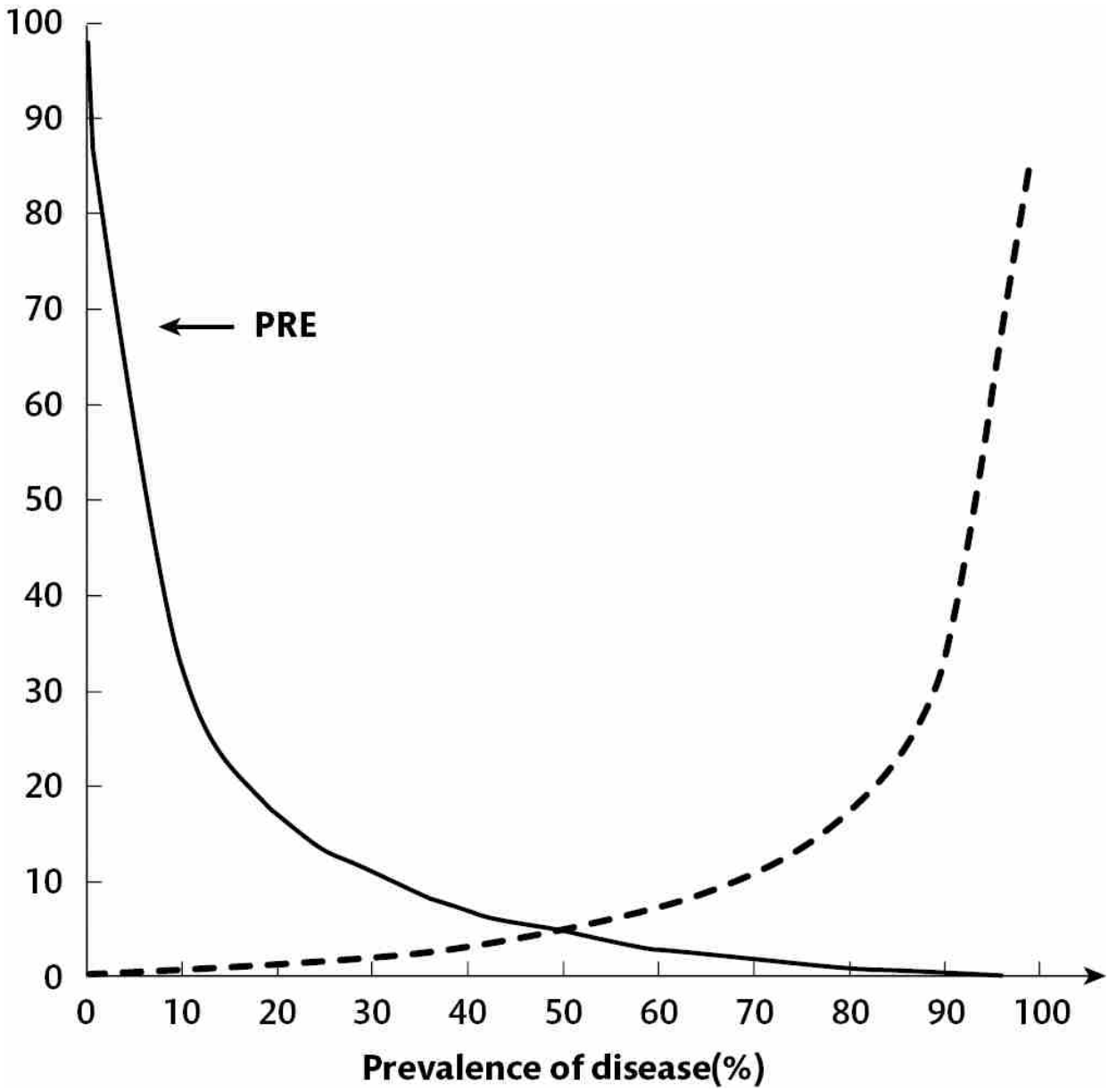


Fig. 11.5.3 Prevalence of disease and prediction error rate of test positive and negative results.

For screening, a high NPV is good as those with a negative test result can be reassured free of the disease. But a low PPV will be of a concern as it would mean many or most of test-positive results are not true positive and further expensive and time-consuming investigations would have to be conducted in them without a gain. The false positive rate in test-positive results is one minus the PPV and is often over 50% and can be as high as 90% or higher in some screening programmes. Thus, much of the efforts in

screening has to be made to reduce false positive results so as to minimize related costs and harms.

Two methods of reducing the rate of false positive results are obvious. One is to screen in people with a relatively high risk of the disease concerned. For example, the prevalence of lung cancer is higher in people of older age and smoking cigarettes, two major risk factors of lung cancer, than in the general adult population. Thus, screening for lung cancer can reduce cost in those of older age and smoking cigarettes. More complex methods can be used to identify high-risk people by incorporating more risk factors of the screened disease (Tammemägi et al. 2013). For the same reason, breast cancer screening should be focused on women aged 40 or above.

The other method is to ‘manipulate’ the sensitivity and specificity of the test so as to achieve a desired PPV for screening.

Balancing between sensitivity and specificity

The results of a medical test are in nature normally quantitative rather than qualitative and often arbitrarily divided into a dichotomous result by using a cut-off value to facilitate clinical interpretation. Those with a result equal to or above the cut-off value are considered test positive, while those with a result below the cut-off are deemed negative. Different cut-off values can be used and will give rise to different combinations of sensitivity and specificity. The sensitivity and specificity of a test is inversely related: specificity decreases as sensitivity increases and vice versa. This means importantly we cannot increase both sensitivity and specificity at the same time by choosing a particular cut-off value unless the test is improved technically.

Different combinations of sensitivity and specificity of a test may be used in different clinical settings. A high sensitivity test will help detect most patients, whereas a high specificity will help reduce false positive results and reduce the costs of further investigations. For example, in such settings as the emergency room, we may use a low cut-off value to have a high sensitivity so as not to miss even a small risk of a life-threatening problem. This will result in a high sensitivity but a large number of false positives will result due to the low specificity. Thus, those who are indicated free of a severe problem is certain but those who are indicated of having the problem will have to go through further investigations to confirm the diagnosis.

Conversely, in such settings as down-syndrome screening, a relatively high cut-off value is used to achieve a high specificity and high PPV so as to minimize the number of amniocentesis which may cause unnecessary abortion.

In screening, the prevalence is generally low and the percentage of those free of the disease which is complementary to the prevalence will be high. Consequently, a small error in the test results in those free of disease will generate a large number of false positives and considerably increase the costs and harms of further investigations in those who do not truly have the disease. Should we lower the cut-off so as to maximize the number of patients detected and potential benefits from treatments or should we increase the cut-off so as to minimize the costs and harms of further unnecessary investigations? They are the issues that need to be carefully debated in planning a screening programme. The choice of the cut-off value is central to balancing sensitivity and specificity and further to the consequent balance between benefits and costs of the entire screening programme.

How can we prove screening works?

A high risk of a severe disease, a good test, and an effective treatment are the prerequisites for a screening programme to be eventually beneficial. However, these criteria are not a guarantee that potential benefits of screening will be realized in practice. How can we find out whether a screening programme can truly bring about more good than harm?

Take cancer screening as an example, an intuitive method for evaluating the benefits of screening is to compare the survival between asymptomatic cases detected by screening and symptomatic cases diagnosed as a result of symptoms. Let’s call the former type of patients ‘screening-detected’ and the later ‘routinely diagnosed’. A higher survival rate in screening-detected patients than in routinely diagnosed patients is then considered as evidence that screening is beneficial in reducing fatality and increase survival.

Such comparisons used to be widely conducted. However, they are severely flawed and their results misleading. A few biases almost certainly exist to cause a false conclusion. The major ones include lead time bias, length time bias, and selection bias.

Lead time bias

Lead time is the duration of time from the time the disease is detected by screening to the time symptoms would appear in the same person if there were no screening (Fig. 11.5.2). Lead time always exists in screening; otherwise, screening will not detect patients any earlier and become equivalent to no screening. Purely due to lead time, screening-detected patients will surely survive longer than routinely diagnosed patients simply because the former was diagnosed and treated earlier and observed longer over time than the later. In other words, if early detection and early treatment is not effective and do not affect the natural progress of the disease, screening-detected patients will live longer by exactly the amount of lead time than routinely diagnosed patients. We call this false survival benefit lead time bias.

To illustrate, in scenario A of Figure 11.5.4, the patient is diagnosed and treated as a result of appearance of symptoms in the absence of screening, in scenario B the patient is detected and treated early as a result of screening but does not live any longer than that in scenario A as screening is not effective, and in scenario C the patient is also detected by screening but has lived longer than that in scenario A, showing a benefit of screening. As lead time always exists, the patient in scenario B will always look like surviving longer than that in scenario A. The extra survival in the patient in scenario B is however purely a result of lead time. Screening is beneficial only when the extra survival in screening-detected patients is longer than the lead time such as in scenario C. But we can distinguish scenario C from B if and only if we know the lead time.

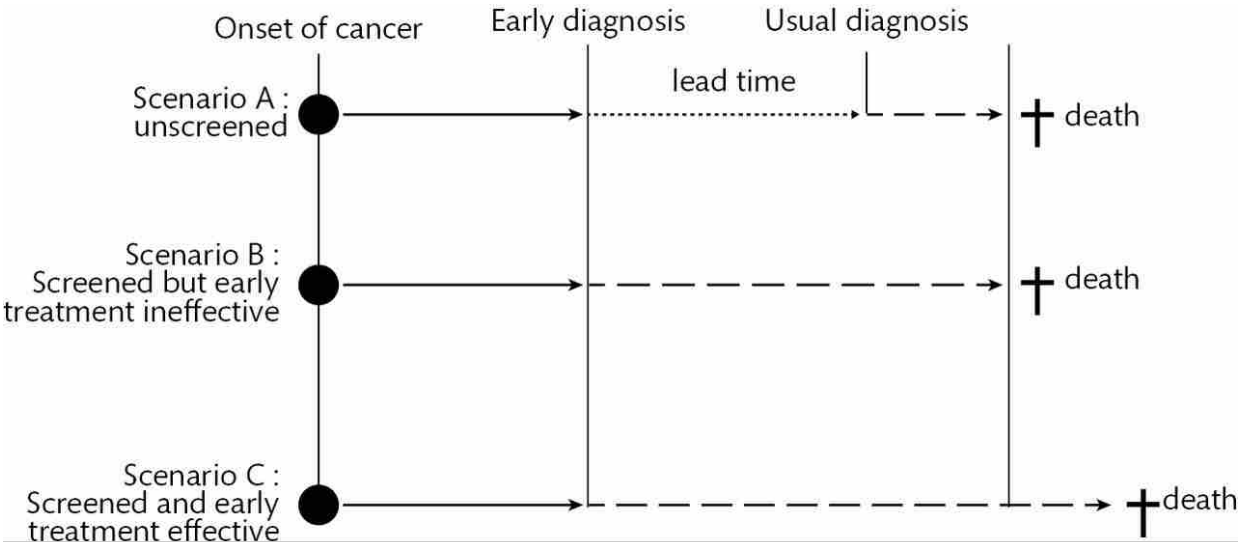


Fig. 11.5.4 How lead time affects survival time after screening?

But lead time cannot be estimated. It cannot be estimated by using the onset time of disease which is impossible to know for most chronic diseases. Nor can lead time to be estimated by the difference between the detection time by screening and the time symptoms appear as the two time points cannot be known in the same patient. In reality, the time symptoms appear cannot be known if screening is conducted and detected the patient. Conversely, the detection time by screening cannot be known if screening is not provided and diagnosis is made as a result of symptoms.

Lead time bias is like asking two persons who will live a longer life. If one is 25 years' old and the other 40 and they both will die at the age of 80, the younger person can live from now for another 55 years and the older only 40 years. The younger will live 15 years more than the older from now purely because of the lead time ——we see the younger 15 years earlier. Therefore, we cannot deduct from this observation that the younger will truly live a longer life as their life span is the same. In the context of screening, the younger person is like screening-detected patient and the older routinely diagnosed patient. Lead time is always a part, if not all, of the extra survival observed in screening-detected patients as compared with routinely diagnosed patients.

In conclusion, all comparisons of survival between screening-detected patients and routinely diagnosed patients are flawed and subject to lead time bias.

Length time bias

Length time bias is another inevitable bias in comparing the survival between screening-detected patients and routinely diagnosed patients. It occurs because the same cancer progresses at a different rate in different people.

Screening detects relatively benign cancer rather than the worst. To illustrate, Figure 11.5.5 shows three types of patients with the same cancer: slow-progress cancer (a, b, and c), cancer with a moderate progress rate (d, e, and f), and fast-progress cancer (g, h, and i). Assume cancer started all at the same time. By the time of screening, the fast-progress cancer would have progressed to a stage at which symptoms have appeared and diagnosis made at a routine clinical setting. On the other hand, slow-progress cancer is still too small to be detectable by the screening.

Screening tends to detect cancer that progresses more slowly and has inherently better prognosis than cancer that is routinely diagnosed. In addition, like slow-progress cancers, some screening-detected cancers may stop progress, regress, or even spontaneously resolve so that it will never cause a problem before people die from other causes. Thus, screening-detected patients are bound to have a better prognosis and survive longer than routinely diagnosed patients, even if screening is not effective. We call this false conclusion on the benefit of screening length time bias.

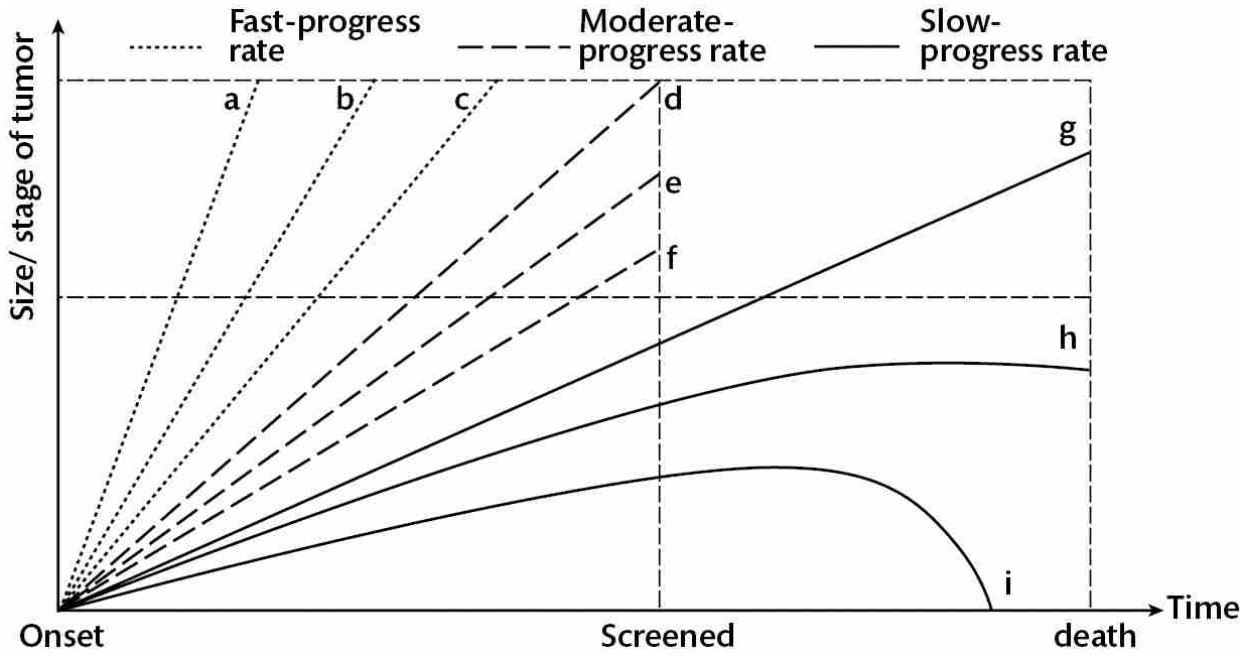


Fig. 11.5.5 Length time bias: fast-progress tumours come to medical attention before screening is performed, whereas tumours with a moderate progress rate allow more time for detection.

Confounding

Comparison of screening-detected patients with routinely diagnosed patients is inherently flawed for evaluation of the effectiveness of screening programmes. To avoid these biases, we can alternatively compare screened people with those who are not screened regardless of whether or not they are found to have the disease. The non-screened population can be those who are in the same community and offered screening but have not received screening or another population in which screening did not provide at all. Screening will be considered effective if the mortality of the screened disease in the screened population is found to be lower than that in the non-screened population. For example, the mortality of prostate cancer was reduced in screening region as compared with the rest of Australia, suggesting PSA screening be effective (Bartsch et al. 2008). However, such comparisons are like any other cohort studies and subject to the problem of confounding.

Confounding is a bias that arises in comparison due to incomparability in factors that may affect the prognosis of the screened disease between screened and non-screened populations. We would like to think that those who were screened had the same characteristics as those who were not

screened. However, there may exist many possible differences between screened and non-screened populations, which may cause confounding. For example, if people at a higher risk of the screened disease are more eager to be screened, e.g. women with a family history of breast cancer are more willing to get a mammography, a screening test will look worse than it truly is. This will cause underestimation of the true benefit of screening.

Confounding can also make screening look better than it truly is. For example, if one has to travel a long distance to get tested, more young and healthy people will be screened and fewer people in the screened population would die and screening would seem to make a positive difference. As screening is not compulsory and people mostly volunteer themselves to screening, this bias is also called volunteer bias or referral bias, a form of selection bias. Due to selection processes, the screened and non-screened will not be comparable and confounding will inevitably occur.

Like in clinical trials, the best way of tacking confounding, or selection bias some would like to call, is to conduct a randomized controlled trial. In randomized trials, people are randomly allocated to comparison groups so that any prognostic factors of the disease will be balanced among compared groups so that confounding will be eliminated.

Evaluation of screening by the randomized controlled trial

Randomized controlled trials in screening differ from usual clinical trials. In order to minimize contamination of screening service between compared groups, communities, or clusters of people rather than individuals are randomized. Screening service is offered to those in the intervention communities but not the control communities. Such trials are known as cluster randomized controlled trials. A lower mortality from the screened disease in the screened group than the control group will suggest screening is beneficial.

As illustrated in Figure 11.5.6, in a randomized controlled trial of screening for cancer, there are four categories of people in each of the compared groups: those free of cancer, fast-progress cancer, cancer with a moderate progress rate, and slow-progress cancer. Because of randomization, the proportion of each category will be comparable between screened and non-screened communities. Screening will affect the prognosis of only those with cancer at a moderate progress rate and mortality in the

other 3 categories are comparable and will cancel each other. Thus, any lowered mortality in the screened group can be attributed to screening. As only those with cancer at a moderate progress rate are compared, lead time bias and length time bias are also be avoided at the same time confounding is controlled by randomization.

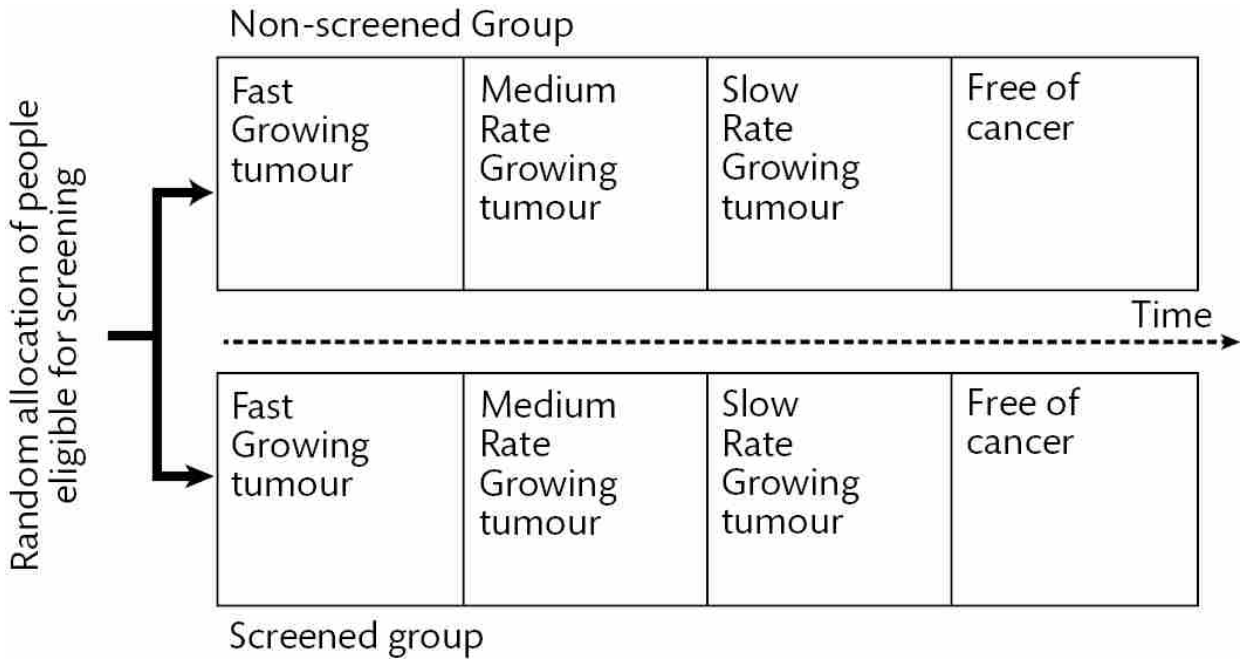


Fig. 11.5.6 Randomized controlled trial designed to evaluate a screening programme.

Now it is widely accepted that screening programmes for healthy people are justifiable only when randomized controlled trials have clearly demonstrated that benefits outweigh harms. The benefits should be estimated on important endpoint outcomes such as death rather than intermediate surrogate outcomes such as blood test and imaging results. For example, randomized controlled trials show that mammography can significantly reduce mortality from breast cancer in women over 50 years old (Wells 1998), whereas chest X-rays with or without sputum cytology do not affect mortality from lung cancer (Bech et al. 2003).

Cluster trials on screening are normally large and expensive and difficult to organize. They have also a low statistical power as most people in such a trial do not contribute to the comparison so the benefit if any is greatly diluted. Alternatively, trials are conducted in which only screening-detected

cancer patients are included and randomized to receive either a treatment or no treatment as control. It is generally considered unethical to give no treatment to cancer patients in the control group. Such trials are relatively rare. With a better understanding of the natural progress of cancers, many cancer patients are now willing to watchfully wait rather than to be treated immediately. This makes such trials more acceptable (Holmberg et al. 2002; Wilt et al. 2012). In theory, the control group can be given a treatment with proven benefit in screening-detected cancer. If there is such a treatment, it already forms evidence supporting screening.

However, trials of screening-detected patients, at best, can only be viewed as efficacy trials like those for drugs as they do not take into account the organizational aspects of a screening programme. If poorly organized, a screening programme may still fail even if the test is sufficiently accurate and treatment effective. In reality, compromises are often made in ways people are selected for testing, the test is conducted, and treatments delivered. Compromised management standard and quality may make a screening programme eventually fail to bring about any good to the screened population (Raffle and Gray 2007). Thus, the overall effectiveness of a screening programme can only be reliably evaluated in randomized controlled trials that randomly allocate eligible people to screening or usual care.

Balancing the benefits, costs, and harms of screening

Randomized controlled trial may show a screening programme is beneficial but the benefit comes at a cost. The costs include money, human resources and time, and inconvenience, worries and harms caused to the screened and incur at every stage of screening from the test to further investigations to the treatment.

Benefits

Screening will be futile if early treatment does not bring about extra benefit and can be effective only when treatments of early detected patients has greater benefits than usual care. The total population gain in the benefit from a screening programme will be determined by the total number of cases detected as well as the effectiveness of treatment. The former is in turn

determined by the prevalence of disease and sensitivity of the test. The higher the prevalence, the larger the number of patients can be potentially detected; the higher sensitivity is, the larger the proportion of patients can be eventually picked by screening. The total life years gained from screening is the product of the number of cases treated and the average extra survival gained from early treatments. Given that treatments are effective, the cost-effectiveness of screening can be increased by screening only persons at a high risk of the disease, by using a test of high sensitivity, by using a cut-off value that gives rise to a high sensitivity, and/or by testing the same persons more often.

Costs and harms

The benefits of screening come at a cost and with harm. Costs and harms can result from the screening test, further diagnostic investigations, and treatment. Harms may occur in the following forms:

harms and inconvenience caused by the initial test;
negative labelling effects: stress and anxiety caused by a false positive result;
harms and inconvenience caused by unnecessary investigations in false positive persons;
harms and inconvenience caused by unnecessary treatment of those who do not have the disease or have the disease that would never progress to the clinical stage and would have not been treated if there were no screening;
prolonged stress and anxiety if detected patients cannot be effectively treated;
a false sense of security caused by false negative results which could delay final diagnosis.

Increasing benefits and reducing costs and harms

The costs and harms of the initial test and of the final treatment are largely unavoidable. There is not much that can be done to reduce them. In contrast, confirmatory diagnostic investigations are where a large proportion of costs and labelling effects can be avoided. These further investigations are normally expensive and time-consuming, and are often invasive and can cause harm. Further investigations are unnecessary and the costs and harms

incurred can be reduced if false positive results can be reduced. All screening programmes should make efforts to do so.

As most people do not have the screened disease in the screened population, a relatively low specificity can result in a large number of false positive results which can outnumber many times the true patients detected. The lower the prevalence is, the higher the false PPV will be; the higher the specificity is, the lower the false PPV will be. The rate of false positive results can thus be considerably reduced by increasing the specificity and by screening in high people and at a longer interval between two tests. As sensitivity and specificity of a test are inversely related, specificity can only be increased with a compromise in sensitivity and missing diagnosis of a few more patients. This can be done by changing the cut-off value. A good screening programme must carefully choose the cut-off value so as to achieve a desirable combination of sensitivity and specificity, which will, in turn, help favourably balance the benefit against the cost and harm.

In brief, avoidable costs and harms of screening occur largely as a result of false positive test results and can be reduced by screening in high-risk people and by using a relatively high specificity. Having said that, time is an important factor to consider as well. As technology advances fast, a few years of waiting may substantively reduce the costs of the initial test and end up with an improved test with greater accuracy as well.

Illustration 1: screening for prostate cancer

In this section, we will use prostate cancer screening, in particular, data in the United States (US), as an example to demonstrate the rationale, principles, methods, and problems of screening.

Severity and prevalence of prostate cancer

Cancer is a terrifying disease. It relentlessly metastasizes throughout the body. It causes death if not treated, and often even it is treated. And the best way to treat it is to catch it early. These views about cancer seem strong enough to lend support to screening activities for cancer. Prostate cancer seems no exceptional.

Prostate cancer is common. It is a leading cause, and the second cause of death from cancer in men in the US. The lifetime risk of receiving a

diagnosis of prostate cancer was approximately 9% in 1985 and reached 16% in 2007 (Altekruse et al. 2010). Silent prostate cancer is also prevalent. One of the early studies showed that the prevalence of latent carcinoma of prostate at autopsy was 12% (Breslow et al. 1977). More recently, a systematic review of 29 autopsy studies of men who died of other causes showed the prevalence of incidental prostate cancer increases with age (Bell et al. 2015). It was found in some 5% in men aged 30 years or below and as common as 59% in men aged 80 years or above.

The median age at diagnosis is 67 years and the median age at death is 81 years (Altekruse et al. 2010). Some 60% of patients die from competing causes in 10 years after diagnosis of prostate cancer. These findings suggest prostate cancer be a slow progressing cancer and have a long preclinical phase. Studies estimated that asymptomatic prostate cancer can be detected with prostate-specific protein (PSA) by a lead time of 5–10 years (Draisma et al. 2009).

About 80% of prostate cancer patients are local and has not spread outside of the prostate. The 5-year survival rate for these patients is nearly 100%. In contrast, patients diagnosed with metastatic prostate cancer that has spread to distant lymph nodes, bones, or other organs, the 5-year survival rate is only about 29% (ACS 2018). The sharp contrast in survival of the local and distant cancers suggest that many patients can be detected early and their survival be greatly prolonged by early treatment.

The test—prostate-specific protein

Prostate-specific antigen (PSA) is the primary driving force for the rise of prostate cancer screening (AUA 2013; Hoffman 2018). PSA is a protein produced by normal as well as malignant cells of the prostate gland. The PSA test measures the level of PSA in a man's blood. The blood level of PSA is often elevated in men with prostate cancer. PSA testing was introduced in the late 1980s. In the US, the Food and Drug Administration (FDA) approved in 1994 the use of the PSA test in conjunction with a digital rectal exam to test asymptomatic men for prostate cancer.

Before the 1980s, solid cancers could be diagnosed only by symptoms, signs, and imaging and rarely by a simple blood test. PSA was considered an exciting breakthrough and caused great enthusiasm for screening for prostate cancer. By the early 2000s, approximately 75% of men aged 50 years or

above had undergone PSA testing and 90% of prostate cancer were detected by means of screening in the US.

PSA testing is safe, quick, simple, and inexpensive and thus generally acceptable to healthy men and feasible for testing a large number of people. The major concern is its accuracy. Using the cut-off value of 4.0 ng/ml as is normally recommended, specificity is 91% but sensitivity is only 21% which will miss 79% of cancers. Twelve per cent (12%) of the tested will turn out to be positive and only 30% of test-positives are truly positive. Using a lower cut-off value of 3.0 ng/ml, sensitivity will be increased to 32% which is still relatively low but will lower specificity to 85% (Wolf et al. 2010). Eighteen per cent (18%) of the tested will turn out to be positive and only 28% of the positive results are truly positive. The further investigation is biopsy of the prostate. Approximately, 70% of the test-positive men will not benefit from biopsy but may have harms including stress, bleeding, pain, clot formation, infection, and urinary difficulties.

The treatment

When prostate cancer is confirmed, the usual treatment is radical prostatectomy. PSA testing had become widespread before the benefits of aggressive treatment of asymptomatic prostate cancer was demonstrated in research. Till 2002, a Scandinavian randomized controlled trial of 694 men with asymptomatic prostate cancer and aged below 75 years showed radical prostatectomy as compared with watchful waiting could reduce the risk of death from 8.9% to 4.6%, a relative reduction of 48% during a median follow-up of 6.2 years (Holmberg et al. 2002). The mortality reduction persisted through 15 years of follow-up (Bill-Axelson et al. 2011). However, only about 5% of the cancer patients in this trial were detected by screening and over 75% were palpable. It is thus questionable whether this result is applicable to patients with prostate cancer mostly detected by screening and impalpable.

Unfortunately, the PIVOT trial (Prostate Cancer Intervention Versus Observation Trial) in the US failed to confirm this benefit of radical prostatectomy in men with prostate cancer mostly detected by screening and of earlier stage than those in the Scandinavian trial (PIVOT 2009). The trial randomized 731 men to either radical prostatectomy or watchful waiting.

After 12 years of follow-up, no statistically significant difference was found in overall and prostate-specific mortality (Phillips 2011).

It remains controversial whether current treatment methods for early stage of prostate cancer can prolong life. These trials are relatively small and did not take into account the standard and quality of screening, which may further compromise the benefit. Large randomized controlled trials were eventually mounted to directly compare the screened and non-screened populations. They are tougher and ultimate evaluations of PSA screening.

Randomized controlled trials of PSA screening

By 2010, six large randomized controlled trials on PSA screening have been completed and published with a total number of participants of 387,286 men studied (Djulbegovic et al. 2010). The median follow of the six trials ranged from 4 years to 15 years but was 9 years or longer in five trials. A systematic review of these trials showed that compared with usual care, PSA screening with or without digital rectal examination could on average detect 50% more, or 20 more in every 1000 screened, men with prostate cancer. The extra patients detected by screening are all stage I or stage II cancers. Harms are not reported in the majority of trials and found relatively rare if reported.

Importantly, five trials with 302 500 randomized participants showed a relative risk 0.88 on mortality from prostate cancer, a result which is statistically not significant. Four trials with 256 019 randomized participants showed a relative risk of 0.99 on all-cause mortality, again statistically insignificant. There is large heterogeneity in the proportion of extra prostate cancers detected but all the trials showed homogeneously no effect on reducing mortality. These results suggest that the increased detection of early-stage prostate cancer has not translated into a benefit in reducing overall mortality and mortality specific to prostate cancer.

In conclusion, some 30 years later since PSA was introduced for screening, the current best evidence does not seem to support that PSA screening is effective in reducing the outcomes that are important for patients. It may be argued that in some circumstances PSA may reduce mortality specific to prostate cancer. However, PSA screening may have increased the risk of death from other causes and thus led to no reduction in overall mortality, which constitutes strong evidence against PSA screening. Having made that conclusion, it still remains a possibility that the lack of

effect of PSA screening is partly a result from a likelihood that many silent prostate cancer patients may have also been detected and treated in the usual care group in these trials.

Illustration 2: screening for cardiovascular risk factors

Unlike cancers, cardiovascular diseases share many common risk factors. These factors are mostly well understood and can be effectively controlled with currently available interventions. The amenable cardiovascular risk factors include, notably, hypertension, hyperlipidaemia, smoking, body weight, lack of physical activities, and type II diabetes mellitus. In this section, we will use hypertension to illustrate how screening can be used for risk factors for the purpose of primary prevention, although these cardiovascular risk factors are usually checked together.

Hypertension pressure—the most frequently checked condition

Blood pressure has probably taken over the measurement of body temperature and become the most frequently checked clinical condition in today's medicine. It is checked often in routine clinical encounters, in general, health checks, and even at home with automated sphygmomanometers. Searching for hypertension is normally not taken as a part of screening programme and thus has not been debated and challenged for the evaluation we ask for cancer screening. This has further abetted the culture of measuring blood pressure.

Hypertension is common and its cardiovascular complications are severe and prevalent. Blood pressure can be easily measured. Blood pressure-lowering drugs can effectively reduce the risk of cardiovascular complications and are widely used. These all make hypertension suitable for screening. Indeed, worldwide hypertension has been widely screened and treated. However, treating hypertension is not merely for the sake of lowering blood pressure; it is for lowering the risk of severe cardiovascular complications and death.

Effectiveness of searching for hypertension

It is not clear whether screening for hypertension in the general population would help reduce future risk of cardiovascular complications and deaths. A systematic review of randomized controlled trials of general health checks, all or most of which included measuring blood pressure, found general health checks in general practice could effectively reduce cardiovascular risk factors such as blood pressure and blood cholesterol but had no effect on both cardiovascular mortality and overall mortality (Si et al. 2014). This finding is consistent with a previous larger systematic review which also included randomized trials in non-primary care settings (Krogsbøll et al. 2012).

That not all detected hypertensive patients are eventually treated may partly account for the lack of effect of routinely checking blood pressure and other cardiovascular risk factors on reducing mortality. In the context of screening, blood pressure checks should not be encouraged if the majority of detected patients are eventually treated. It is possible that blood pressure checks are already common in the general population so that the difference in the intervention and usual care groups is too small to be detectable. It is also possible that the benefit from checks for cardiovascular risk factors has been cancelled by harms of treatments and of checks for other diseases. More recently, a large randomized controlled trial provides evidence for checks for only cardiovascular risk. The trial randomized 59,616 participants from a general population and compared the effect of checks for hypertension and other major cardiovascular risk factors followed by repeated lifestyle interventions over 5 years. At the end of 10 years' follow-up, screening did not show any effect on ischaemic heart disease, stroke, and mortality (Jørgensen et al. 2014).

Screening for hypertension and other cardiovascular risk factors in the current medical context seems to have failed in the most important criteria for screening: screening should be able to reduce severe complications and deaths that the screened condition may cause. There is another important issue that needs to be addressed in searching for hypertension and in screening in general: what is the clinical condition we are screening for? If the condition screened is difficult to define, a screening programme is likely to be stretched and eventually become doubtful.

What constitutes hypertension?

Regular checks for hypertension reemphasized the ultimate criteria for screening: screening programmes should be recommended when and only when they can produce more good than harm at an affordable cost for a population. However, the good and harm ratio for hypertension screening changes as the diagnostic criteria for hypertension change. Hypertension is, in essence, an 'arbitrary' condition which is reflected in the choice of cut-off value for defining it. This has important implications for screening for hypertension and screening in general.

Epidemiological studies have shown beyond reasonable doubt that the risk of cardiovascular disease increases as blood pressure increases. The relation is largely linear for blood pressure between 120 mmHg and 180 mmHg systolic or between 70 mmHg and 110 mmHg diastolic, a range most adults' blood pressure falls into. There is no clear turning point below which cardiovascular risk does not increase with blood pressure (Law and Wald 2002; Lacey et al. 2018). Compared with those with the lowest blood pressure values, all others are hypertensive; conversely, compared with those with the highest blood pressure values, all others are normal tensive. There is no blood pressure value that can be used objectively and undebatably for defining hypertension. Any cut-off value will be a result of the trade-off in balancing among various factors. A higher cut-off value will result in fewer hypertensive patients for whom treatment will be more effective, whereas a lower cut-off value will result in more hypertensive patients for whom treatment will be less effective. Where should the cut-off value be set?

The fundamental requirement is that treatment of hypertension thus defined should be able to do more good than harm. Furthermore, resources available to, and the values of, the population concerned must also be taken into account (Jackson et al. 2005; MacMahon et al. 2005; Tang and Hu 2005). This was exactly how once the cholesterol cut-off for prescribing Statin was set in the United Kingdom (Pickin et al. 1999; Gaw and Packard 2000). Previously defined hypertension was widely accepted and treatments are effective. However, around the year 2000, the cut-off for hypertension was reduced to 90/140 mmHg. A systematic review of randomized controlled trials showed that antihypertensive drugs did not reduce risk of cardiovascular disease and related mortality in mild hypertension created by the new criteria (Lenzer 2012; Brunström and Carlberg 2018). This conclusion is very likely to be applicable to those who have an even lower blood pressure but will be diagnosed as hypertensive according to the new

US guidelines in which the cut-off is lowered to 130/80 mmHg (Whelton et al. 2017). The new definitions will make any screening programmes for early, mild hypertension questionable.

Importantly, changing of disease definitions has been happening to hyperlipidaemia, type II diabetes mellitus and many other risk factors and diseases including cancer (Welch et al. 2011; Doust et al. 2017; Yang et al. 2017). The implications of changing the definition of cancers for cancer screening will be discussed later in this chapter.

Health checks as multiphasic screening

Modern concept and practice of systematic screening originated in early attempts of general health checks. Today we emphasize systematic planning and rigorous evaluation of benefits and harms of screening programmes through randomized controlled trials. Since the 1960s, many randomized controlled trials on screening of individual diseases have been conducted and the results are often disappointing. In the same time, general health checks, the old practice of screening has not disappeared. In fact, in recent decades it has been on a continuous increase and become a large business in many countries including China (Ren 2016). Modern general health checks may include hundreds of tests and examinations in one package. Many of them are new tests including genetic testing and have not been rigorously evaluated and proven to be able to do more good than harm.

Health checks can be viewed as an application of many screening tests for many different diseases in one programme. On one hand, often only a fraction of people with identified problems are adequately followed up with further investigations and treatments and the benefit of the tests and examinations are not fully realized. On the other hand, further investigations and treatments of some detected problems may do harms without a benefit. Against the ultimate standard for screening that is a demonstration of benefits in reducing mortality and prolonging life in randomized controlled trials, are general health checks effective and worthwhile?

In 2012, a systematic review identified 14 randomized controlled trials with a total 182,880 participants included and summarized the evidence for the effects of general health checks in adults on reducing morbidity and mortality (Krogsbøll et al. 2012). In this review, general health checks were defined as screening for more than one disease or risk factor in more than

one organ system, regardless of whether they were performed once or repeatedly. Many trials included risk factors for cardiovascular disease and screening techniques for cancer.

Nine trials provided data on total mortality, 8 on cardiovascular mortality, 8 on cancer mortality. It showed no benefit in reducing mortality from health checks: the risk ratio is 0.99, 1.03, and 1.10 respectively for total mortality, cardiovascular mortality, and cancer mortality. None of these results are statistically significant. The review also failed to find any benefits on hospitalization, disability, worry, additional physician visits, and absence from work. Although the debate may continue, it is almost certain that benefits of general health checks for reducing mortality, the ultimate aim of screening, would be trivial, if any at all.

The authors of the review wrote in the Cochrane editorial as follows:

> In spite of the consistently negative results of screening and lifestyle counselling, general health checks remain a popular political tool for combating the burden of ischaemic heart disease and other chronic diseases. Lately, the British government has started a general health check for the population. Danish left wing parties went to the polls in 2011promising the same but changed their minds shortly after gaining power. In 1968 the World Health Organization proposed that several criteria should be fulfilled before systematic screening could be recommended. These criteria have often been interpreted as criteria for screening for early stages of diseases, but they also apply to systematic screening for high risk of diseases. Some of the main criteria are that the disease in question should be serious and that a valid screening instrument is available, both of which are fulfilled for ischaemic heart disease. However, the most important criterion is that screening should reduce morbidity and mortality in the target population.

The authors made a rather strong recommendation that doctors should not offer general health checks to their patients, and governments should abstain from introducing health check programmes (Gøtzsche 2014).

It was the early debates over the value of health checks that led to the modern concept of more organized screening programme focused on specific diseases (Raffle and Gray 2007). Given the widespread of different ideas and practices of screening that are often ineffective, it is time to revisit these debates. It is reasonable to argue that screening ideas such as general health checks should include only those tests and examinations that are clearly targeted at specific diseases and proven effective on important outcomes.

Early detection and overdiagnosis

Screening and overdiagnosis

With the availability of rapid-growth new technologies, the idea of early diagnosis and early treatment have been increasingly applied in clinical settings often before they are demonstrated effective. This has caused overdiagnoses. Overdiagnosis occurs when people without symptoms are diagnosed with a disease that ultimately will not cause them to experience symptoms or early deaths (Moynihan et al. 2012). Screening and overdiagnosis are related. Screening is about early detection and early treatment, whereas overdiagnosis is often in the form of early diagnosis that does not lead to a benefit.

Overdiagnosis can result from overdetection or overdefinition (Broderson et al. 2018). Overdetection is a major form of overdiagnosis in cancer and partly driven by advances in technologies (Orme et al. 2010). It can occur by identifying abnormalities that are never going to cause harm, which include those that do not progress, that progress too slowly to cause symptoms or harm during a person's remaining lifetime, or that resolve spontaneously.

Overdefinition is the major type of overdiagnosis for such conditions as hypertension, hyperlipidaemia, diabetes mellitus, and osteoporosis. It can occur by lowering the cut-off vale for disease diagnosis and by including more parameters in the diagnostic criteria without evidence that doing so helps people feel better or live longer.

Overdiagnosis in hypertension

As just discussed, hypertension is not black and white disease entity. Rather, it is an arbitrarily defined clinical condition by using a cut-off. With a hope to prevent early and in more people, the cut-off value for diagnosing hypertension has been reduced a few times so more and more people are now diagnosed, alerted, or treated.

In 1998, the cut-off value for hypertension was further lowered from 160/95 mmHg to 140/90 mmHg. In China, this change in the cut-off increased the prevalence of hypertension by 134% and created over 100 million new patients. A similar phenomenon was also observed for hyperlipidaemia and type II diabetes. It was estimated in 2002, the total drug costs for hypertension, hyperlipidaemia, and type II diabetes could amount to 56% of the total healthcare expenditure of the government if all the patients were treated with drugs (Hu et al. 2017).

In 2015, 23.2% or approximately 244.5 million of the Chinese adult population aged 18 years or above were hypertensive, and another 41.3% or approximately 435.3 million had pre-hypertension according to the current Chinese guidelines (Wang et al. 2018). In 2017, it was recommended in the US that the cut-off value for hypertension be further lowered to 120/80 mmHg (Whelton et al. 2017). Using these new criteria, the total number of hypertensive patients in China will be further doubled and increased to approximately 500 million or prevalence of 46.4% (Wang et al. 2018).

Given that there is insufficient evidence that drug treatment of mild hypertension can reduce morbidity and mortality (Krogsbøll et al. 2012; Jørgensen et al. 2014; Si et al. 2014) the new cut-off value for hypertension will create sweeping overdiagnosis worldwide.

Overdiagnosis in cancer

Cancer is usually considered a black and white disease entity. Once found, it kills. In fact, cancer is very much like hypertension which is a continuous condition and the risk of suffering and death increases as the severity or size of cancer increases. Early-stage, asymptomatic cancer is merely a ‘risk factor’ of future suffering and death rather than a current misfortune.

The relation of cancer to mortality can be illustrated in Figure 11.5.7. Cancer does not start with a large lump of cancerous cells. Instead, it begins with cancerous mutations in cells when they divide. These cancerous cells may multiply and form an *in situ* cancer. The *in situ* cancer may further grow in size and even become invasive spreading into surrounding tissues and the lymph nodes nearby. Local cancer may further progress into the metastatic stage and spread to another organ. Although most patients may die of a cancer at the metastatic stage, early-stage cancers are not necessarily lethal. In fact, many may regress or even resolve spontaneously, stop progress, or grow so slowly that they will not cause harms before the individual die of other causes (Black 1998). This means that some early-stage cancer may not harm in a person’s lifetime. Indeed, autopsy studies found that subclinical cancer exists in a large proportion of people who died of accidents or other causes rather than cancer (Welch et al. 2010).

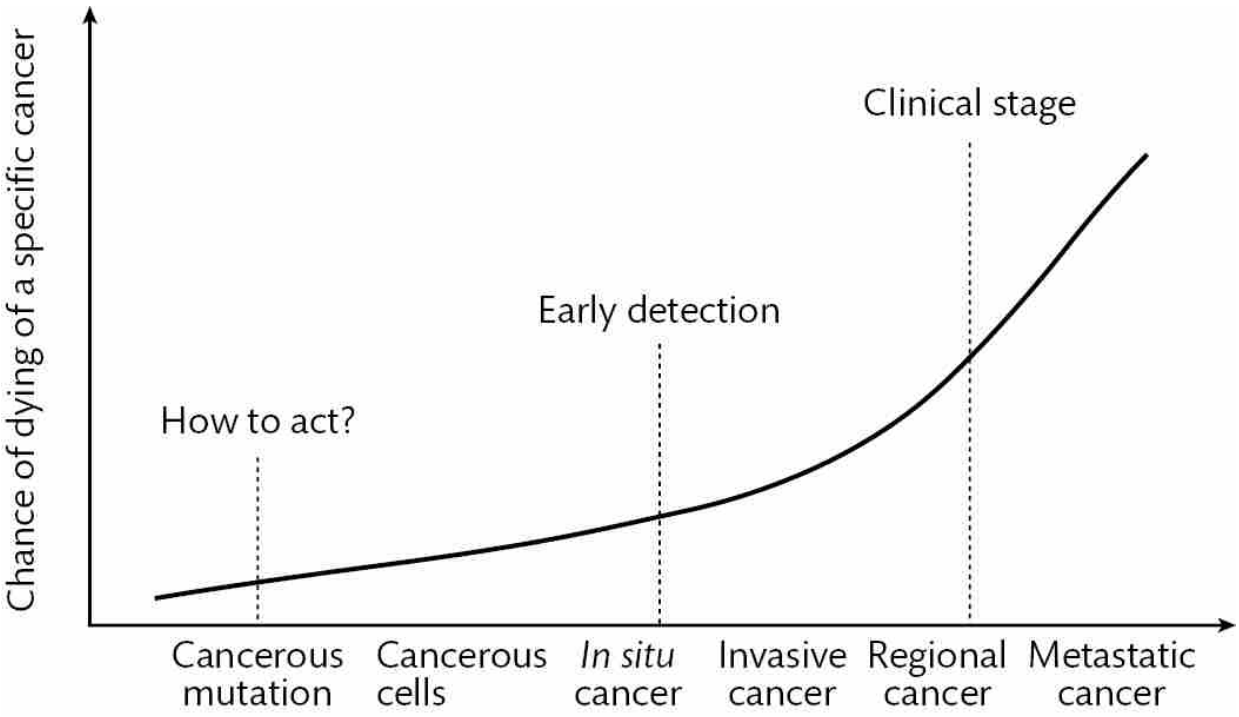


Fig. 11.5.7 Stage of cancer and the chance of dying from that cancer.

Now we can see that the relation between cancer and mortality is, essence, similar to that between hypertension and mortality. Again, where should we draw the cut-off for diagnosing cancer in such a circumstance? Unlike hypertension, we do not deliberately discuss and play with the cut-off value in diagnosis of cancer. But the truth is we keep pushing down the cut-off value in the size of the tumour: the cut-off is lowered every time more powerful diagnostic methods are made available. There is no doubt that cancers detected through current imaging technologies are much smaller those diagnosed through symptoms and physical examination. As a result, the definition of cancer has been changing as well but stealthily due to advances in diagnostic technology. Most cancer patients we diagnosed today are small and asymptomatic and would not have been diagnosed and treated in the days before the rise of modern technology. Have these advances in medicine truly helped reduce suffering and death? The evidence shows not necessarily.

Again, take prostate cancer as an example. A meta-analysis of 19 autopsy studies found that the prevalence of unsuspected prostate cancer in men aged 70–79 years was as high as 36% in Caucasians and 51% in African-Americans (Jahn et al. 2015). After PSA was introduced in the late 1980s in

the US, the incidence of prostate cancer quickly increased by over 100%. Between 1999 and 2006, the annual age-adjusted incidence of prostate cancer in men maintained above 150 per 100,000 men. About 80% of prostate cancers were, however, clinically confined to the prostate and only 4% had metastasized (Hoffman 2011). Given the great increase in the incidence, the mortality of prostate cancer, however, remained largely the same between 1975 and 2007. The combination of increasing incidence and stable mortality is a clear sign of overdiagnosis. Furthermore, evidence from six randomized controlled trials with a total of 387 286 participants do not support the routine use of PSA for screening prostate cancer (Djulbegovic et al. 2010).

Thyroid cancer is more striking an example of overdiagnosis. While the chances of tests to detect a thyroid 'abnormality' are high, the risk it will ever cause harm is low (Tang and Hu 2005; Welch et al. 2011). In the United Sates, an analysis in 2006 showed many of the newly diagnosed thyroid cancers were the smaller and less aggressive forms not requiring treatment (Davies and Welch 2006) but carried a risk of damaged nerves and long term medication if treated. In South Korea, the rate of thyroid cancer was increased by 14 times from 1993 to 2011. Most of the increase are believed to be overdiagnosis, attributable to the early detection of papillary thyroid cancer by ultrasound (Ahn et al. 2014; Park et al. 2016). Furthermore, a systematic review found that it is unclear that ultrasonography plus follow-up cytology for screening thyroid cancer can decrease mortality rates or improve important patient outcomes (Lin et al. 2017).

Other diseases

Overdiagnosis is not limited to hypertension and cancer and in fact is ubiquitous in medicine. Next are a few more examples.

Pulmonary embolism

Pulmonary embolism is usually considered a 'not to be missed' diagnosis. Historically, pulmonary embolism was nearly universally fatal, diagnosis made on clinical grounds and often only discovered after death. Anticoagulants were established management standard. In the late twentieth century, computed tomography pulmonary angiography allowed to detect pulmonary embolism reliably and non-invasively. Progressively smaller

pulmonary embolisms are now being found and the incidence of diagnosed pulmonary embolism has been doubled but the death rate remained largely unchanged, suggesting smaller subclinical pulmonary embolism is overdiagnosed (Prasad et al. 2012).

Abdominal aortic aneurysms

Abdominal aortic aneurysms is another typical example of detecting clinically less important abnormalities. A study of 201 men aged between 60 and 75 with hypertension or heart disease compared the number of abdominal aortic aneurysms that can be identified by ultrasound and that by physical examination. In five people with aneurysm of 5 cm or greater, physical examination can find 4. In 5 with aneurysm between 4 and 5 cm, physical examination detected only 1. In 8 with aneurysm smaller than 4 cm, physical examination failed to identify any. Ultrasound could detect all of them, creating three times more subclinical aneurysms. Most of these aneurysms are below the size for which surgery is recommended and thus represent typical overdiagnosis (Lederle et al. 1988).

Stroke

Silent stroke was found by MRI scans to exist in over 10% of people from the general population without clinically diagnosed stroke. The prevalence of silent stroke increases as age increases. It was 7%, 8%, 10%, and 15%, respectively, in those aged 30–49, 50–59, 60–69, and 70–79 years. It is not a certain treatment of silent stroke is beneficial (Das et al. 2008).

Chromic kidney disease

The new US guidelines recommend that an estimated glomerular filtration rate below 60 ml/min/1.73 m^2 and sustained for 3 months or longer is deemed abnormal. According to these criteria, more than 10% of adults in the US are now classified as having some form of chronic kidney disease (Levey et al. 2009).

Lessons from early detection

Advances in diagnostic methods have made diagnosis of disease at their clinical stage much easier, more quickly, and more reliably, which has

reduced unnecessary treatments and increased the effectiveness in those treated. However, at the same time, these advances are also being used widely to find diseases at their preclinical stage. Furthermore, diagnostic criteria for diseases are being expanded, making more and more people become patients. Modifying the definition of disease normally creates mild or early stage of diseases. However, it is often unclear whether the preclinical patients thus created and identified will benefit from treatment.

From the perspective of screening, many of the activities for detecting and treating early or mild diseases do not seem justifiable. Early detection is happening extensively at a large number and still on an increasing trend. The number of early patients thus created or discovered is likely many times more than those found by appropriately organized screening programmes. If there is no benefit, it is a considerable waste. We thus argue that new diagnostic methods for detecting asymptomatic patients should not be introduced into practice before they are rigorously evaluated in the way we do for screening. Guidelines should also be made for defining a new disease or modifying the diagnostic criteria of a current disease so that a benefit will be guaranteed to result.

Conclusions

Intuitively screening is highly appealing but in practice frequently ineffective.

All screening programmes cost and have harms. Decades of epidemiological research show that it is much less often than many would expect that screening can do more good than harm. This re-emphasizes urgently the ultimate criteria for screening: only those programmes that can do more good than harm at an affordable cost should be recommended. Policy makers should assess carefully as regards the disease, the test, and treatments to see whether they are suitable for screening. The ultimate evidence about the benefit and harm of a screening programme should be from randomized controlled trials. Screening programmes should also be organized in high standard and quality so that the benefit demonstrated in research can be reproduced in practice.

But not all efforts for early detection and early treatment occur in the form of systematic screening. In fact, more often they take place in routine clinical services and increasingly in general health checks. In some places, health

checks may account for most of early detection activities and should be planned and evaluated as carefully as we do for screening of specific diseases. Even if effective, a screening programme will benefit only a few at a large cost to the entire population. Even worse, the harm starts immediately but the benefit if any takes much longer to appear. Thus, the decision to introduce a new screening programme should be taken as seriously as we make important healthcare decisions like building a large new hospital.

Finally, is the benefit of a screening programme worth its costs and harms? The answer to this question inevitably varies across populations that have different resources available, different health needs to address, and different views on the value of the benefit and harm. Screening should be recommended to a population only when a benefit–harm ratio is deemed desirable to that population.

REFERENCES

Ahn, H.S., Kim, H.J., and Welch, H.G. (2014). Korea's thyroid-cancer 'epidemic'—screening and overdiagnosis. *New England Journal of Medicine*, **371**, 1765–7.

Altekruse, S.F., et al. (2010). *SEER Cancer Statistics Review 1975–2007*. Bethesda, MD: National Cancer Institute.

American Cancer Society (ACS) (2018). *Survival Rate of Prostate Cancer by Stage*. [Online] Available at: https://www.cancer.org/cancer/prostate-cancer/detection-diagnosis-staging/survival-rates.html

American Urological Association (AUA) (2013). *Early Detection of Prostate Cancer (updated 2018)* [Online] Available at: https://www.auanet.org/guidelines/prostate-cancer-early-detection-guideline

Bartsch G, et al. (2008). Tyrol Prostate Cancer Demonstration Project: early detection, treatment, outcome, incidence and mortality. *BJU International*, 101(7), 809–16.

Bech, P.B., Kelley, M.J., Tate, R.C., and McCrory, D.C. (2003). Screening for lung cancer: a review of the current literature. *Chest*, **123**, 72S–82S.

Bell, K.J.L., Del Mar, C., Wright, G., Dickinson, J., and Glasziou, P. (2015). Prevalence of incidental prostate cancer: a systematic review of autopsy studies. *International Journal of Cancer*, 137(7), 1749–57.

Bill-Axelson, A., et al. (2011). Radical prostatectomy versus watchful waiting in early prostate cancer. *New England Journal of Medicine*, **364**, 1708–17.

Black W. (1998). Advances in radiology and the real versus apparent effects of early diagnosis. *European Journal of Radiology*, **27**, 116–22.

Breslow, N., et al. (1977). Latent carcinoma of prostate at autopsy in seven areas. *International Journal of Cancer*, 20(5), 680–8.

Brodersen, J., et al. (2018). Overdiagnosis: what it is and what it isn't. *BMJ Evidence-Based Medicine*, **23**, 1–3.

Brunström, M. and Carlberg, B. (2018). Association of blood pressure lowering with mortality and cardiovascular disease across blood pressure levels: a systematic review and meta-analysis.

JAMA Internal Medicine, 178(1), 28–36.
Cochrane, A.L. and Holland, W.W. (1971). Validation of screening procedures. *British Medical Bulletin*, **27**, 3–8.
Das, R.R., et al. (2008). Prevalence and correlates of silent cerebral infarcts in the Framingham offspring study. *Stroke*, 39(11), 2929–35.
Davies, L. and Welch, H.G. (2006). Increasing incidence of thyroid cancer in the United States, 1973–2002. *Journal of the American Medical Association*, 295(18), 2164–7.
Djulbegovic, M., et al. (2010). Screening for prostate cancer: systematic review and meta-analysis of randomised controlled trials. *British Medical Journal*, 341, c4543.
Doust, J., et al. (2017). Guidance for modifying the definition of diseases: a checklist. *JAMA Internal Medicine*, 177(7), 1020–5.
Draisma, G., et al. (2009). Lead time and overdiagnosis in prostate-specific antigen screening: importance of methods and context. *Journal of the National Cancer Institute*, **101**, 374–83.
Gaw, A. and Packard, C.J. (2000). At what level of coronary heart disease risk should a statin be prescribed? *Current Opinions in Lipidology*, 11(4), 363–7.
Gøtzsche, P.C. (2014). General health checks don't work. *British Medical Journal*, 348, g3680.
Hoffman, R.M. (2011). Screening for prostate cancer. *New England Journal of Medicine*, **365**, 2013–19.
Hoffman, R. (2018). Screening for prostate cancer. [Online] Available at: https://www.uptodate.com/contents/screening-for-prostate-cancer
Holmberg, L., et al. (2002). A randomized trial comparing radical prostatectomy with watchful waiting in early prostate cancer. *New England Journal of Medicine*, **347**, 781–9.
Hu, X.F., Han, X.R., Yang, Z.Y., Hu, Y.H., and Tang, J.L. (2017). The impact of broadened diagnostic criteria on the prevalence of hypertension, hyperlipidemia and diabetes mellitus in China. *Chinese Journal of Preventive Medicine*, 51(5), 369–77.
Jackson, R., Lawes, C.M.M., Bennett, D.A., Milne, R.J., and Rodgers, A. (2005). Treatment with drugs to lower blood pressure and blood cholesterol based on an individual's absolute cardiovascular risk. *The Lancet*, 365(9457), 434–41.
Jahn, J.L., Giovannucci, E.L., and Stampfer, M.J. (2015). The high prevalence of undiagnosed prostate cancer at autopsy: implications for epidemiology and treatment of prostate cancer in the Prostate-specific Antigen-era. *International Journal of Cancer*, 137(12), 2795–802.
Jørgensen, T., Jacobsen, R.K., Toft, U., Aadahl, M., Glümer, C., and Pisinger, C. (2014). Effect of screening and lifestyle counselling on incidence of ischaemic heart disease in general population: Inter99 randomised trial. *British Medical Journal*, 348, g3617.
Krogsbøll, L.T., Jørgensen, K.J., Larsen, C.G., and Gøtzsche, P.C. (2012). General health checks in adults for reducing morbidity and mortality from disease: Cochrane systematic review and meta-analysis. *British Medical Journal*, 345, e7191.
Lacey, B., et al. (2018). Age-specific association between blood pressure and vascular and non-vascular chronic diseases in 0·5 million adults in China: a prospective cohort study. *Lancet Global Health*, 6(6), e641–9.
Law, M.R. and Wald, N.J. (2002). Risk factor thresholds: their existence under scrutiny. *British Medical Journal*, **324**, 1570.
Lederle, F.A., Walker, J.M., and Reinke, D.B. (1988). Selective screening for abdominal aortic aneurysms with physical examination and ultrasound. *Archives of Internal Medicine*, 148(8), 1753–6.
Lenzer, J. (2012). Cochrane review finds no proved benefit in drug treatment for patients with mild hypertension. *British Medical Journal*, 345, e5511.
Levey, A.S., et al. (2009). A new equation to estimate glomerular filtration rate. *Annals of Internal Medicine*, **150**, 604–12.
Lin, J.S., Bowles, E.J.A., Williams, S.B., and Morrison, C.C. (2017). Screening for thyroid cancer: updated evidence report and systematic review for the US Preventive Services Task Force.

Journal of the American Medical Association, 317(18), 1888–903.
MacMahon, S., Neal, B., and Rodgers, A. (2005). Hypertension—time to move on. *The Lancet*, 365(9464), 1108–9.
Morrison, A.S. (1992). *Screening in Chronic Disease* (2nd ed.). Oxford: Oxford University Press.
Moynihan, R., Doust, J., and Henry, H. (2012). Preventing overdiagnosis: how to stop harming the healthy. *British Medical Journal*, 344, e3502.
Orme, N.M., et al. (2010). Incidental findings in imaging research: evaluating incidence, benefit, and burden. *Archives of Internal Medicine*, **170**, 1525–32.
Park, S., et al. (2016). Association between screening and the thyroid cancer 'epidemic' in South Korea: evidence from a nationwide study. *British Medical Journal*, 355, i5745.
Phillips, C. (2011). Study questions benefit of surgery in some men with early-stage prostate cancer. *NCI Cancer Bulletin*, 8(11).
Pickin, D.M., et al. (1999). Cost effectiveness of HMG-CoA reductase inhibitor (statin) treatment related to the risk of coronary heart disease and cost of drug treatment. *Heart*, **82**, 325–32.
Prasad, V., Rho, J., and Cifu, A. (2012). The diagnosis and treatment of pulmonary embolism: a metaphor for medicine in the evidence-based medicine era. *Archives of Internal Medicine*, 172(12), 955–8.
Raffle, A. and Gray, M. (2007). *Screening: Evidence and Practice*. Oxford: Oxford University Press.
Ren, D. (2016). Prevention is better than cure: China's health-care market poised for gains amid shifting values: an emphasis on early screening and preventive medicine is emerging among China's newly-affluent consumers. *South China Morning Post*. Sunday 24 April. [Online] Available at: http://www.scmp.com/business/article/1938224/prevention-better-cure-chinas-health-care-market-poised-gains-amid-shifting
Sackett, D.L. and Holland, W.W. (1975). Controversy in the detection of disease. *The Lancet*, **2**, 357–9.
Si, S., Moss, J.R., Sullivan, T.R., Newton, S.S., and Stocks, N.P. (2014). Effectiveness of general practice-based health checks: a systematic review and meta-analysis. *British Journal of General Practice*, 64(618), e47–53.
Tammemägi, M.C., et al. (2013). Selection criteria for lung-cancer screening. *New England Journal of Medicine*, **368**, 728–36.
Tang, J.L. and Hu, Y.H. (2005). Drugs for preventing cardiovascular disease in China. *British Medical Journal*, **330**, 610.
The Prostate Cancer Intervention Versus Observation Trial (PIVOT) (2009). VA/NCI/AHRQ Cooperative Studies Program #407 (PIVOT) design and baseline results of a randomized controlled trial comparing radical prostatectomy to watchful waiting for men with clinically localized prostate cancer. *Contemporary Clinical Trials*, **30**, 81–7.
Wang, Z.W., et al. (2018). Status of hypertension in China: results from the China Hypertension Survey, 2012–2015. *Circulation*, **137**, 2344–56.
Welch, H.G. and Black, W.C. (2010). Overdiagnosis in cancer. *Journal of the National Cancer Institute*, **102**, 605–13.
Welch, H.G., Schwartz, L.M., and Woloshin, S. (2011). *Over-diagnosed: Making People Sick in the Pursuit of Health*. Boston, MA: Beacon.
Wells, J. (1998). Mammography and the politics of randomized controlled trials. *British Medical Journal*, **317**, 1224–30.
Whelton, P.K., et al. (2018). 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA guideline for the prevention, detection, evaluation, and management of high blood pressure in adults: a report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. *Hypertension*, 71, e13–115.

Wilson, J.M.G. and Jungner, F. (1968). *Principles and Practice of Screening for Disease (Public Health Paper 34)*. Geneva: WHO.

Wilt, T.J., et al. (2012). Radical prostatectomy versus observation for localized prostate cancer. *New England Journal of Medicine*, 367(3), 203–13.

Wolf, A.M., et al. (2010). American Cancer Society guideline for the early detection of prostate cancer: update 2010. *CA: A Cancer Journal for Clinicians*, 60(2), 70–98.

Yang, Z.Y., Fan, H., and Tang, J.L. (2017). Considerations concerning disease redefinition. *JAMA Internal Medicine*, 177(12), 1872–3.

11.6

The practice of environmental health in an era of sustainable development

Yasmin E.R. von Schirnding and Lynn R. Goldman

Introduction to environmental health practice

Roots of environmental health practice

The practice of environmental health goes back to earliest times. For example, cleanliness, water supply and sanitation have played an important role in disease prevention throughout recorded history (Rosen 1958). Many of the world's water and sanitation practices had their origins in the ancient civilizations of India, Egypt, and Mesopotamia—a testimony to their hydraulic engineering skills. It was the Romans who excelled in sanitary engineering however, with the Cloaca Maxima, the great sewer of Rome, still forming part of the drainage system of modern Rome. Personal hygiene and cleanliness were emphasized by early and prehistoric societies, and were regarded as a sign of spiritual purity (Rosen 1958).

For thousands of years the theory prevailed that disease was linked to the gods, and only gradually did the idea evolve (for example, in the writings of the Greek physician Hippocrates in *Airs, Waters and Places* (400 BCE) that disease was linked to natural causes such as climate and the environment (Rosen 1958). *Airs, Waters and Places* also provided practical advice on how to address environmental health issues (perhaps an early example of health impact assessment) by suggesting that before a place was colonized, it was necessary for the character of the soil to be assessed.

In the course of the subsequent demise of Greek-Roman civilization, much scientific knowledge was lost, and in the Middle Ages, unsanitary conditions prevailed, especially in cities, although regulations were implemented with penalties for pollution of water supplies and other

environmental nuisances. Quarantine was also a way in which environmental health was practised, to protect from the corruption of air (the theory of miasma—'bad air'), in addition to personal hygiene measures.

During the mid-nineteenth century and beyond, awareness of threats to public health from poor sanitation grew, and with this 'sanitary awakening', widespread sanitary reform measures were implemented. These were spearheaded by an enquiry by Edwin Chadwick, a Poor Law Commissioner, who, in 1842, published a major report on sanitary conditions in England (Chadwick 1842). Chadwick's investigation into the causes of poverty concluded that people often became poor as a result of ill-health from an adverse environment (that situation prevails today). In 1848, with the passage of the Public Health Act in the United Kingdom (Calman 1998), provision was made for the appointment of Inspectors of Nuisances—the predecessors of today's environmental health officers. A national board of health, and district health boards were also recommended.

It was John Snow who traced the origin of the cholera epidemics in London to contaminated water, and who thereby conducted one of the first comprehensive environmental epidemiology investigations—a key function of importance in the practice of environmental health. In the United States, the Shattuck report in 1850 (Winslow 1949) similarly drew attention to the links between poor living conditions and health, and recommended the establishment of state and local boards of health, as well as local sanitary associations and regular (environmental) health surveys.

From its roots in antiquity to the sanitation movement of the 1800s, environmental health practice was fairly narrowly focused on sanitation, environmental hygiene, inspection and control, and has since gradually broadened to address health and environment concerns in the context of sustainable development (von Schirnding 2002a), with emphasis on ecosystems and ecology (Rapport et al. 2009). In the early- to mid-twentieth century, air pollution disasters (e.g. Meuse Valley, Belgium in 1930, Donora, Pennsylvania in 1948, and London, United Kingdom in 1952); industrial disasters like Bhopal, India; as well as mounting evidence of the carcinogenic, reproductive and other effects of chemicals (highlighted in Rachel Carson's *Silent Spring*, 1962) led to increased awareness of toxic chemicals in the environment. This awareness was further heightened by

such incidents as the Love Canal toxic waste site disaster in the United States in the 1970s.

Environmental health has thus expanded beyond its horizons. Today environmental health practitioners have to concern themselves with a wide variety of issues, with virtually any aspect of the environment having the potential to affect human health. The environment can impact chronic diseases such as cancer, lung diseases, or birth defects, as well as infectious diseases like gastroenteritis, respiratory infections, or vector-borne diseases like malaria. Typical environmental health concerns include ambient and indoor air quality, food safety, water quality, sanitation, waste management, housing quality, noise pollution, electromagnetic radiation, occupational and industrial hygiene, vector control, and toxic chemical exposures.

The environment has much broader impacts too—both positive and negative. For example it may provide protection and resiliency to natural disasters or predispose people to harm (for example from building in earthquake-prone areas). It may promote health through the provision of nutritious and safe food, drinking water, provision of recreational opportunities, interaction with nature, and so on.

In particular, concern about health effects associated with natural ecosystems and global (climate) change has grown in recent years, necessitating the need to address health issues in the context of complex (eco) systems analysis (McMichael et al. 2006; Patz et al. 2008; McMichael 2013). With growing appreciation of the links between ecology and health, the integrity of natural ecosystems, and the transdisciplinary and intersectoral nature of environmental health, there have been calls for new epidemiological approaches to dealing with environmental health hazards (Soskolne et al. 2007). With the global spotlight on sustainable development, practitioners have begun to look at environmental health risks more holistically, from the viewpoint of various economic sectors like housing, transport, or energy, and settings or systems like the city at large.

This approach can be tracked through the various United Nations (UN) conferences held on the environment and sustainable development going back over the past 40 or more years. These are briefly highlighted next, as well as some of the key principles and goals articulated therein.

Global initiatives

Following the UN Conference on the Human Environment held in Stockholm, Sweden (UN 1972), in 1983 the World Commission on Environment and Development (UNCED) headed by former Norwegian Prime Minister Gro Harlem Brundtland (World Commission on Environment and Development 1987) linked the issue of environmental protection to economic growth and development. The Commission's report 'Our Common Future' (World Commission on Environment and Development 1987) launched the concept of sustainable development into the mainstream of world debate.

Sustainable development essentially means meeting people's development needs in such a way as not to exceed the earth's carrying capacity in the longer term. Although specific definitions for sustainable development vary, broadly speaking, the concept is concerned with balancing environmental, social, and economic objectives in order to maximize societal well-being, both now and in the future.

The UN Conference on Environment and Development (UNCED) (UN 1992)—a watershed event popularly referred to as the 'Earth Summit'—resulted in the adoption of Agenda 21—a global action plan with concrete goals, targets, and outputs for measures needed to achieve environmentally sustainable development into the twenty-first century. It also resulted in the 'Rio Declaration' which encompassed important agreements on 27 principles adopted by 178 nations, including the Precautionary Principle, the Polluter Pays Principle, the Common But Differentiated Responsibilities principle (CBDR), and other principles such as the principle of intergenerational equity, the principle of access to information, and the decision-making process, and the principle of integrated decision-making (these are discussed further in Section 3). Several key conventions were also adopted, such as those on climate change (including the Kyoto Protocol) and on biological diversity (under which was negotiated the Cartagena Protocol).

The Rio Declaration, in its preamble, alluded to the central role of health in sustainable development by stating that 'Human beings are at the centre of concerns for sustainable development. They are entitled to a healthy and productive life in harmony with nature'. This concept was further developed during the World Summit on Sustainable Development (WSSD) (UN 2002a) which took place in Johannesburg, to review progress 10 years after UNCED.

At WSSD, health was singled out as one of the five priority issues to be addressed, along with water, energy, agriculture, and biodiversity (Annan 2002; UN 2002b). In the Johannesburg Plan of Implementation (JPOI) which emerged from this conference, particular emphasis was placed on health issues in relation to environment and poverty concerns. Improving access to safe water, sanitation, clean air, improved waste management, and sound management of chemicals were among the key issues of importance for environmental health practitioners that received special attention in Johannesburg (von Schirnding 2005).

This momentum was somewhat lost in the subsequent run-up to the twentieth anniversary of UNCED. 'Rio+20' (UN 2012) aimed to review progress over the past 20 years and to renew political commitments to sustainable development. While many new green economy policies were endorsed in the final Rio+20 document 'The Future We Want', no new international commitments nor legally binding agreements were made, albeit voluntary agreements were made by countries to meet their own sustainable development targets. Coming out of WSSD where health made a strong showing, and out of the Millennium Development Goals (MDGs) process (UN 2000) where health had a central position, health did not feature strongly at Rio +20. This confirmed the pressing need for the health community, and especially environmental health practitioners, to remain actively engaged in the sustainable development agenda (Kleinert et al. 2012; Singh et al. 2012).

Of note, all three conferences (UNCED with its focus on environment, WSSD with its focus on the social dimension, especially—poverty, and Rio+20 with its focus on the green economy) failed to ensure the integration of the three pillars of sustainable development. The process of setting sustainable development goals in the follow-up process to Rio+20 (the 2030 Agenda for Sustainable Development), which was initiated at the Sustainable Development Summit in 2015 (Barbier 2012; UN 2015; Bhandary et al. 2015) aims to ensure that these three pillars of sustainable development are not considered separately from each other. Progress towards achievement of these goals is now reviewed under the newly constituted High-level Political Forum on Sustainable Development (UN 2018), one of the important outcomes of the Sustainable Development Summit. Health, and environmental health in particular, can be used as a tool by practitioners to integrate across the three dimensions of sustainable

development. Measurable targets (and indicators) are needed to ensure health remains a prerequisite for, and an outcome of, sustainable development—in all its dimensions. This is discussed further in Section 7.

Concept and scope of environmental health

In considering the practice of environmental health, it is apparent from the foregoing discussion that the nature and scope of the discipline has undergone considerable transformation, especially in light of the emerging concept of sustainable development and sustainability. In its broadest context, environmental health is concerned with the health effects associated with the environment. This includes both the physical (natural or 'built') as well as the social environment. It is less concerned with the health of the environment *per se*, as it is with its impacts on human health. In this respect, environmental health is an integral part of the discipline of public health, with strong links to the environmental sciences.

The World Health Organization (WHO) has defined environmental health in various ways to include 'Those aspects of human health and disease that are determined by factors in the environment. It also refers to the theory and practice of assessing and controlling factors in the environment that can potentially affect health' (WHO 1989). While *health* has been defined by the WHO as 'A state of complete physical, mental and social well-being and not merely the absence of disease or infirmity' (WHO 1946), environmental health has tended to focus mainly on the physical environment, with less emphasis on the social environment and on social and mental well-being. Nevertheless, both the physical as well as the social aspects of the environment are important aspects of the realm of environmental health.

An updated definition of environmental health by the WHO gives added emphasis to this social component and describes environmental health as comprising 'those aspects of human health, including quality of life, that are determined by physical, chemical, biological, social, and psychosocial factors in the environment. It also refers to the theory and practice of assessing, correcting, controlling, and preventing those factors in the environment that can potentially affect adversely the health of present and future generations' (WHO 2004). This definition incorporates, as well, a focus on sustainability, in so far as it alludes to the health of both present

and future generations. Environmental health is thus not only concerned with the identification, control, and remediation of factors in the environment which can affect health, but also with disease prevention and enhancement of intergenerational human health.

The intersectoral nature of environmental health is recognized in the WHO's Constitution (WHO 1948), which states that the Organization shall 'promote, in cooperation with other specialized agencies, where necessary, the improvement of nutrition, housing, sanitation, recreation, economic, or working conditions and other aspects of environmental hygiene'.

The concept has its roots in a number of international strategies and fora, including the 1977 'Health-For-All' strategy of WHO (and its renewal and subsequent updates: WHO 1977, 1981, 1998; Yach 1996), the Alma-Ata conference on Primary Health Care in 1978 (WHO 1978), the World Health Assembly (WHA) Technical Discussion on Intersectoral Action (WHO 1986a), the Ottawa Charter on Health Promotion (WHO 1986b), the Sundsvall meeting on Supportive Environments for Health (Haglund et al. 1996), the WHO Commission on Environment and Health (WHO 1992), and the international conference on 'Intersectoral Action for Health—A Cornerstone for Health-for-All in the 21st Century', held in Canada in 1997 (WHO 1997; Kreisel and von Schirnding 1998).

Addressing the underlying multifactorial determinants of health is essential in order to ensure health improvements which are sustainable over the long term. Primary prevention is a key aspect of the work of the environmental health practitioner, and involves implementation of interventions before ill-health occurs. It includes health promotion interventions geared towards general health and well-being, as well as interventions on specific known causes of disease-for example the phaseout of lead—a well-documented neurotoxicant—in gasoline. Pollution prevention is an example of a primary preventive measure, involving strategies such as source reduction, waste minimization, reuse, recycling, emissions control, waste disposal, and clean-up. Secondary preventive measures can be implemented which are targeted at an early stage of exposure before health impacts manifest (for example lead screening), while tertiary preventive measures involve the early identification and treatment of people with a disease (Goldman 2016).

By addressing health, environment, and development concerns in an intersectoral way, a win–win situation occurs: the health sector gains

because disease prevention can move upstream and take advantage of the intersectoral approach to development and environmental improvements. The environmental sector gains because environmental issues can become grounded in matters of local and direct concern to people.

In the following subsections, we look briefly at some of the major trends in health, environment, and development in recent years, before highlighting environmental health issues and the burden of environmentally related disease, which are of importance in directing the work of the environmental health practitioner.

Global trends

Of overriding significance is the fact that on a global scale there has been unprecedented wealth creation and economic development in the past decades, coupled with intensified global efforts targeting major causes of death and disease throughout the world. The sanitary movement of the 1800s mentioned earlier had a major impact on the reduction of infectious disease mortality and resulted in major improvements (increases) in life expectancy. More recently this has been accelerated in part due to such initiatives as the MDGs process which set in motion an ambitious programme of goals and targets to eradicate poverty by the year 2015 (UN 2000), and its successor, the 2030 Sustainable Development Goals (SDGs) process.

Life expectancies for men and women have increased, and a greater proportion of total deaths now occur among people over the age of 70 years, compared with 20 years ago. Infectious disease mortality rates have declined, and burdens of HIV, TB, and malaria have all fallen. Substantially fewer children under the age of five years are dying (Wang et al. 2013), with deaths in this age group having declined by almost 60% since 1970. Between 1990 and 2015, the under-5 mortality rate dropped from 90 to 43 deaths per 1,000 live births (UNDP 2015). Nonetheless, in 2015, 16,000 children continue to die every day, mostly from preventable and environmentally related diseases like pneumonia, diarrhoea, and malaria. This burden is greatest in sub-Saharan Africa (SSA).

Despite huge progress in human development and poverty eradication over the past decades (over one billion people were lifted out of poverty since 1990), today, nearly one billion people still live in poverty, including a

higher percentage of women than men, and big gaps in progress remain in some regions of the world, especially in sub-Saharan Africa and southern Asia. Moreover, the distribution of income at the country level has worsened in much of the world, despite narrowing in gaps in health and education (UNDP 2011). At the same time, global wealth has become increasingly concentrated in the wealthiest one percent of the population**,** who had 32% of global wealth in 2000, compared to 46% in 2010 (UNDP 2016).

Human population growth is increasing by 75–80 million people a year (Sachs 2012) and is projected to grow to 9.7 billion by 2050 (UNDP 2016). In 2014, over half of the world's population lived in urban areas (UNDP 2016), and, with the addition of another 2.5 billion people by 2050, this number may rise to two-thirds of the world's population. Over half the population will be under the age of 18 years. Urban growth will be greatest in Africa and Asia, followed by Latin America and Oceania. By 2030 it is expected that there will be 41 mega-cities with populations of over 10 million inhabitants. The fastest growing cities are likely to have the highest concentration of slums or informal settlements. Already around 880 million people live in slums, and it is predicted that 40% of the world's future expansion will occur in slums (UNDP 2016).

With rapid economic development and urbanization in developing countries, cities are becoming increasingly overcrowded and polluted. Air and water pollution levels have been increasing, as has the generation of municipal and hazardous waste. In the last 30 years, however, in many industrialized nations stronger environmental laws have resulted in cleaner and safer drinking and recreational water as well as in improved air quality. On a global basis nevertheless, pollution of air and water supplies is worsening (Goldman and Torres 2009). Drinking water is under pressure from such factors as pollution, ageing infrastructure, population growth, and resulting increased demand and consumption. While demand is increasing, supplies are becoming more erratic and uncertain, with water-related impacts on agriculture, health, and income levels (UNDP 2016). In general, growth in incomes has been associated with deterioration in key environmental indicators such as carbon dioxide emissions, soil and water quality, and forest cover (UNDP 2011).

Environmental health issues

The practice of environmental health is inextricably involved with disease prevention and control. Environmental health practitioners face a growing diversity of environmental hazards to deal with, in many different settings, in developed and developing countries alike. Traditional risks associated with poverty often coexist alongside risks associated with urbanization and industrialization, agriculture, and land-use practices. For example, water supplies may be contaminated by pesticides and faeces exposing consumers both to microorganisms and toxic substances, and air pollution may stem both from burning dirty household fuels such as coal, wood, and other sources of biomass fuel used for cooking and heating, as well as from the use of industrial fossil fuels. In poor countries, accidental chemical releases may occur more commonly, and there may be fewer means to protect people. Poor countries (and the poor in wealthy countries) may also be more vulnerable to the health effects associated with natural disasters and climate change.

Problems may have to be dealt with simultaneously at the global, national, and local levels. Some problems may be of special relevance at the local or micro level (for example water supply and sanitation, waste disposal, or toxic chemicals) while other problems may be more important at the regional or global level (for example global climate change, or long-range transport of air pollution).

Health impacts associated with global change and disturbances to natural ecosystems (disruption of bio-geophysical systems such as the climate system and associated rising temperatures, declining rainfall, extreme weather events, degradation of land and water, depletion of resources, and other effects) (McMichael 2013) have to be dealt with at all levels and, as indicated earlier, may impact severely on poor countries and on the poor in general. Potential impacts are numerous such as heat-related mortality, health impacts associated with extreme weather events, changes in vector-borne disease distribution, disruption of agricultural production, or threats to drinking water supply. While poor countries have contributed least to global climate change, they have experienced the greatest variability in weather (such as loss in rainfall in some regions and flooding and others) with implications for agricultural productivity and livelihoods (UNDP 2011). Without appropriate action, by 2030, over 100 million more people could be living in poverty (UNDP 2018).

Problems may differ in urban and rural environments, although increasingly the divide between urban and rural is becoming blurred. While cities provide numerous benefits associated with increased job opportunities and provision of essential services and facilities, environment and health problems may nevertheless be profound, especially in cities that are not economically homogeneous. In informal settlement areas in particular, infectious diseases associated with inadequate water and sanitation, and dirty household fuels may be prevalent. Rich and poor people live in very different epidemiological worlds, even within the same city, and this situation occurs in both high-income as well as low-income countries (WHO and UN-Habitat 2010). Infant mortality rates are often higher among poor segments of urban populations in developed countries.

Lead poisoning illustrates very well the unequal burden of risk borne by poor inner-city children who are not only more exposed to sources of lead in and around the home environment (from flaking lead-based paint, for example, or contaminated air and dust), but are also more vulnerable to the toxic effects of lead, being often nutritionally deprived and having inadequate access to healthcare facilities. This is true even of children living in such countries as the United States and parts of Europe (Tong et al. 2000). Within inner-city areas, studies from South Africa have shown substantial socioeconomic differentials in blood lead burden among inner-city children (von Schirnding et al. 1991).

Poorer countries also have higher levels of household deprivation as evidenced by such factors as indoor air pollution, and inadequate access to clean water and sanitation. The environmental burdens of poverty fall directly on poor populations. As income levels rise, however, and with increased affluence and associated lifestyles and consumption patterns, these burdens tend to lessen, concomitant with environmental burdens which are more diffuse, indirect, and delayed, and affect a more temporally and spatially displaced public (UNDP 2011). Risks associated with community urban air pollution are an example.

With further economic development, these more localized burdens tend eventually to lessen too (or be displaced elsewhere), while environmental risks associated with global effects—for example, greenhouse gas emissions—tend to rise with high levels of materials and energy consumption and waste generation (UNDP 2011). The most disadvantaged people carry a double (or triple) burden of deprivation—they are more

vulnerable to the global effects of environmental degradation, have often to cope with new and emerging risks associated with rapid industrialization and urbanization, but also have to deal with threats to their immediate environment like indoor air pollution, dirty water, and poor sanitation.

Burden of environmentally related disease

It has been estimated by WHO (WHO 2016) that in 2012, globally around 12.6 million deaths (23% of all deaths), were attributable to modifiable environmental risks, and 22% of the disease burden measured in DALYs (disability-adjusted life year—a measure of overall disease burden, expressed as the number of years lost due to ill-health, disability or early death) attributable to environmental factors. Children under 5, and older adults were most affected. Among children under 5, it was estimated that 26% of all deaths could be prevented if environmental risks were removed. Adults between the ages of 50 and 75 were found to be most affected by environmental factors. Diseases with the largest environmental fraction (in DALYs) included cardiovascular diseases, diarrheal diseases, and respiratory tract infections. These are related to such factors as ambient and household air pollution, and water, sanitation, and hygiene.

Recent findings indicate a predominance of non-communicable disease risks compared with 1990. With advances in economic development and other factors, as alluded to earlier, there has been a shift away from infectious, parasitic, and nutritional diseases towards non-communicable diseases (NCD)s. In contrast, deaths associated with ambient air, chemical and soil pollution associated with modern industrial development are increasing (Landrigan et al. 2018)

Nevertheless, still today, pneumonia and diarrhoea (associated with household air pollution or water and sanitation) are the leading killers of young children. Diarrhoea accounts for 20% of deaths among children under the age of 5 years, while lower respiratory infections account for account for around 18% of deaths in children under 5. Nearly all of these deaths occur in sub-Saharan Africa and South Asia, reflecting uneven progress in reducing child mortality (UNICEF 2012). The highest fraction of deaths and disease that could be addressed by improving environmental conditions is in low-and middle-income countries. In 2015, 844 million

people lacked a basic drinking water service, while 2.3 billion lacked a basic sanitation service (WHO and UNICEF 2017).

The global burden of disease (GBD) estimates provide a sense of where environmental health practice could make a big impact in future. Continued support to improve access to water and sanitation services and a transition to cleaner household fuels is a high priority in much of sub-Saharan Africa and Asia**.** Nearly 90% of the deaths due to diarrhoea worldwide have been attributed to unsafe water, inadequate sanitation, and poor hygiene. In addition, around three billion people worldwide use solid fuels as their main cooking fuel (UNICEF 2012).

Further information on the relevance of global burden of risk estimates for environmental health practice is provided in Box 11.6.1 and Table 11.6.1, which are based on the 2015 GBD study (Forouzanfar et al. 2016). It should be noted that while both the WHO study (WHO 2016) and the GBD study (Forouzanfar et al. 2016) found environmental factors to be a major cause of premature death, disease, and disability, the GBD estimates are slightly at variance with those reported earlier by WHO. WHO for example attributed 12.6 million deaths to unhealthy environments, and the GBD 2015 study 9 million premature deaths (16% of global mortality). This discrepancy is most likely the result of differing definitions of environment (Landrigan et al. 2018). In respect of specific risk factors like ambient air pollution, and household air pollution, resulting discrepancies are probably in part due to changes in computational methodology, and in regard to water and sanitation, differing definitions of access to safe water.

Box 11.6.1 Global burden of risk estimates main priorities for environmental health practice, e.g. in prevention, research, and policy development

The 2015 global comparative risk assessment of burden of disease and injury provides data on major environmental risks, discussed here (Forouzanfar et al. 2016). It should be noted however, that the 2017 comparative risk assessment of burden of disease and injury provides alternate estimates, based on, inter-alia, different modelling processes for air pollution (Stanaway et al. 2018).

Of the 79 risk factors and risk clusters examined, several related directly to environmental factors, including water and sanitation, air pollution (ambient particulate matter pollution, household air pollution from solid fuels and ambient ozone pollution), residential radon, and lead

exposure. It was estimated that 9 million premature deaths in 2015 were from environmentally related disease, contributing to 16% of total global mortality.

In 2015 4.2 million deaths were due to ambient particulate air pollution, and 2.9 million deaths were associated with household air pollution from solid fuels.1.8 million deaths resulted from diseases related to unsafe water, sanitation, and handwashing.

Environmental factors were responsible for 21% of all cardiovascular disease related deaths, 26% of deaths due to ischaemic heart disease, 23% due to stroke, 51% due to chronic obstructive pulmonary disease (COPD), and 43% due to lung cancer.

Between 1990 and 2015, global exposure to such factors as unsafe sanitation and household air pollution each decreased by more than 25%, and attributable DALYs declined significantly for unsafe sanitation and water, and household air pollution.

Environmental risks declined steadily in relation to the sociodemographic index (SDI).

In 1990, household air pollution ranked fourth in leading risk factors, while in 2015 it ranked eighth. Ambient particulate matter pollution ranked sixth in 1990 and 2015. In 1990 unsafe water ranked second in leading risk factors, whereas in 2015 unsafe water ranked as the fourteenth leading risk factor. Unsafe sanitation ranked seventh in 1990 and nineteenth in 2015.

From 2005 to 2015, global deaths which were attributable to unsafe water and no hand washing with soap, fell by more than 12%. DALYs decreased by more than 20%. Even greater declines occurred in deaths and DALYs for unsafe sanitation since 2005. Reductions in the attributable mortality and DALYs due to diarrheal diseases associated with unsafe water, sanitation, and hygiene where especially marked.

In regard to household air pollution, since 2005, attributable deaths decreased by 13% to 2.9 million deaths in 2015, and the disease burden decreased by 20.3% over this period. Both ambient particulate matter and household air pollution where among the 10 largest contributors to global DALYs in 2015.

Table 11.6.1 Leading environmental risk factors: ranking in 1990 and 2015

Risk factor	1990	2015
Unsafe water	2	14
Unsafe sanitation	7	19
Handwashing	9	18
Ambient particulate matter	6	6
Household air pollution from solid fuels	4	8

Sources: Forouzanfar et al. 2016; Landrigan et al. 2018.

In the following section we discuss the nature of environmental health effects and factors influencing exposure, and then go on to discuss how environmental health practitioners proceed to assess these.

Environmental health assessment

Nature of environmental health effects

Environmental health practitioners must concern themselves not only with addressing known exposures and disease outcomes associated with these, but also, importantly, diseases of unknown cause, with potential environmental links. In addition, still today, for many substances, it is not known whether or not there are safety thresholds for an adverse effect, and, if so, what those levels are.

Certain cancers and ‘subtle’ diseases and conditions related to environmental factors (e.g. lead in relation to intelligence, moderate hearing loss, or raised blood pressure) may not be recognized as such. Thus, the situation remains that much environmental disease goes unrecognized. With regard to chemicals, typically exposures to complex mixtures occur, which have not been well characterized.

Health outcomes

Environmental health effects result from a complex set of genetic, environmental, social, and other factors. Health effects may arise directly from a specific chemical, physical, or biological agent, or indirectly from factors associated with the broader physical and social environment (e.g. housing, land-use, transport, agriculture, or industry). While as indicated earlier, infectious agents resulting in such illnesses as diarrhoea and respiratory infections may be associated with inadequate water and sanitation and indoor air pollution, non-infectious agents are associated with non-communicable (chronic) illnesses. These include cancers, respiratory conditions like asthma, birth defects, and other illnesses.

Environmental health effects may include minor, temporary ill-health conditions, as well as acute illness and chronic disease. Effects are frequently irreversible. Acute ill-health effects may follow relatively shortly after an exposure (e.g. poisoning by pesticides, or following a chemical spill), whereas chronic exposure may occur as a result of an individual cumulative exposure over extended periods of time. A long time period may occur between the exposure and the health effect. An example would be asbestos exposure and mesothelioma (Selikoff et al. 1965). Dispersal of

the population at risk over time and the long incubation period make it difficult to reconstruct exposures. Some persistent pollutants like polychlorinated biphenyls (PCBs) and heavy metals like lead and mercury can accumulate in the environment for years, and may affect people over large distances, causing health effects long after their use has been discontinued. Acute health effects are usually easier to detect than chronic effects, which are difficult to relate to specific hazards or sources.

Multifactorial exposure

In addition, as indicated, many environmental diseases are multifactorial in so far as the causal factors are concerned, and it may be difficult to determine the effects of one exposure in the light of existence of effects of simultaneous exposure to other factors. Prevention has tended to focus on exposures associated with a single agent or medium (e.g. air or water); however, people are often exposed at the same time to a wide variety of physical, biological, and chemical agents, and may be exposed via several pathways (air, water, food) from one source. As indicated earlier, in developing countries people may be exposed both to hazards associated with poverty (unsafe drinking water and inadequate sanitation), pollution, unsafe agricultural and land-use practices, and so on. Relatively little is known, however, of the effects of cumulative exposures and how various risk factors might interact (Tickner 2005)—resulting in an additive or even synergistic (multiplicative) effect. The risk of health impacts associated with exposure to asbestos, for example, may be significantly increased (multiplied many times) in smokers (Selikoff et al. 1968).

Vulnerable groups

The poor, the elderly, women, and young children may be particularly affected. The elderly, for example, may be especially vulnerable to some toxic substances which may take longer to eliminate from the body. They may also be more vulnerable to respiratory illnesses associated with air pollution, for example. During pregnancy, women may have altered physiologies and metabolize toxic substances differently, while, during menopause, women may be especially vulnerable, for example, to the release of toxic substances such as lead from bones.

Children are particularly vulnerable to environmental insults, and are not merely ‘little adults’. They are more exposed than adults, as, due to their rapid metabolism, growth, and development, they drink, eat, and breathe in more of a contaminant relative to their bodyweight than do adults. Children also have enhanced hand-to-mouth activity and they play close to the ground, increasing their exposures still further. In addition, they have immature, and therefore vulnerable, physiologies. The unborn foetus may be especially vulnerable to toxic chemicals.

Thus, different groups have differing vulnerabilities and susceptibilities, including genetic dispositions, nutritional status, health, and physiological status which may influence their exposure (and response). Gene–environment interactions are known to be of importance and different subsets in the population may have different exposure–effect relationships. People may be more or less vulnerable at various stages of their lives.

To intervene on the causes of disease thus requires a good understanding of the health risks and their causes, as well as the susceptible individuals/population, so that surveillance and monitoring systems can be developed to identify in advance potential problems, thereby enabling effective interventions to be designed and implemented.

Role of epidemiology and toxicology

An important part of the work of the environmental health practitioner involves assessing the risks to which people may be exposed in the environment. While there is less uncertainty today about the extent of both direct and indirect risks to human health which are associated with factors in the environment, nevertheless there is still much to be learned about the way in which the environment influences human health.

In conducting health assessments, it is necessary to establish the nature of exposure and health outcome relationships, and to assess causality. Of the various fields that contribute to environmental health (e.g. environmental science, chemistry, physics, engineering, medicine, and many others) of key importance are epidemiology and toxicology, as well as exposure assessment.

Environmental epidemiology is concerned with the distribution and determinants of ill-health effects in human populations at risk from environmental exposures. As workplace exposures have decreased over

time, and environmental exposures have increased, an appreciation of the multiple causation of disease has grown. Various types of study design have been used in environmental epidemiology to assess the nature of the relationship between an exposure and health outcome(s) in a particular setting or population group. These include ecological studies, cross-sectional studies, case–control studies, cohort or follow-up studies, and intervention studies. They can be challenging to conduct due to the complexity of the health risks and exposures involved, and the need to take into account all the potential sources of confounding (extraneous factors associated with the outcome of interest) and bias.

No matter how well designed and executed, however, epidemiological studies suffer from several shortcomings. For a start, they can only disclose health risks after exposures have already occurred. Also, it is sometimes the case that there is too narrow a range of exposure for epidemiological studies to detect health effects in populations. For example, dioxin exposures are difficult to evaluate in the general population because people's dioxin levels occur within a relatively narrow range. In addition, it is difficult to measure exposure at the individual level (air pollution is an example) and exposure measurements made at an area or neighbourhood level may sometimes be (incorrectly) presumed to relate at an individual level, leading to what in epidemiological terms is referred to as the 'ecological fallacy'. This is where inferences about individuals are based on inferences about groups. Thus within an 'exposed' area, owing to micro-variations in exposure and other factors affecting people's exposure (e.g. behaviour patterns), people may not be homogeneously affected, even while assumed to be so.

Toxicological studies, which are experimental studies of the effect of pollutants on human beings or animals under controlled conditions, are also needed in order to identify in advance any adverse effects of substances under investigation. These have the benefit of allowing for a wider range of exposures to be assessed and can take account of the problem of confounding factors in epidemiological studies. However, to be cost-effective and humane, they are designed with as few animals as statistically possible, while dosing animals at high levels. Extrapolation from laboratory animals to humans, at high to low doses, is complex however, with many assumptions involved. It is not usually possible to reproduce all the contributing factors in the laboratory. National and international efforts are being made to substitute more humane methods of assessing toxicity for

animal testing. Such approaches involve employment of structure activity models, coupled with *in vitro* testing, to 'screen' substances for toxicity.

While there exist internationally harmonized guidelines for toxicity testing such as those established by the Organization for Economic Cooperation and Development (OECD), specific testing for toxicity varies by country, and depends on the type of substance and the statute under which it is covered. In the United States, for example, pesticides in food require tests of acute and chronic toxicity, neurotoxicity tests, cancer bioassays, tests of reproductive and developmental toxicity, and others (Goldman and Torres 2009).

Making causal inferences

The concept of risk is linked to that of causation. From the previous discussions it is evident that the relationship between many environmental exposures and health outcomes is complex. Even under optimal conditions and with the best study designs, it is difficult to make definitive assertions regarding causality. Bradford Hill (1965) listed the criteria shown in Box 11.6.2 as being key to assessing whether an observed association is likely to be causal or not.

Risk assessment

While epidemiology and toxicology are used to generate *new* evidence on health effects, risk assessment synthesizes *existing* information and weighs it up with the aim usually of addressing a particular regulatory or policy issue aimed at reducing or eliminating the risk (Bartel 2005). In this respect the National Research Council (NRC) has referred to it as a mixture of 'science and judgement' (NRC 1994).

Risk assessment is used in many fields, including environmental health. It is a science-based process concerned with estimating, usually in quantitative terms, the harm that will result from exposure to an environmental hazard (Fiorino 1995). Risk assessors might, for example, want to analyse the risks associated with being exposed to a factor like indoor air pollution, to a specific agent like lead in paint, or to a suspected contaminant in food or drinking water.

The conceptual framework for risk assessment was set out by the NRC in 1983 (NRC 1983) and divided the process into four key elements: hazard identification, dose–response assessment, exposure assessment, and risk characterization. Risk is a function of hazard and dose. *Hazard identification* concerns the identification of environmental agents and health effects for assessment, based on evidence for causality from epidemiological and toxicological studies.

Dose–response assessment describes the quantitative relationship between exposure and disease based on models that predict the level of response from any dose. Many exposures are considered to have a threshold for non-cancer effects, a point below which no effect is measured. While by default it is assumed that carcinogens have linear dose–response relationships at low doses, this is not true for all mechanisms for carcinogenicity. Risk assessors normally assume a monotonic dose–response relationship (increasing response with increasing dose), unless there is compelling evidence for a different dose–response relationship at low doses (for example between IQ and blood lead levels at low levels). *Exposure assessment* involves estimating or measuring the magnitude, duration, and timing of human exposure to the agent of concern. It necessitates defining the exposed population and the routes of exposure. Direct measures of exposure are rarely available to decision-makers, which consequently often rely on modelling and extrapolations for assessing exposure levels.

The final step is *risk characterization.* Information from the previous three steps is combined and the level of response for the identified health effect(s) at the particular level of exposure to the agent(s) in the defined population is estimated.

Box 11.6.2 Criteria for causality

The strength of the association refers to how large the effect or the size of the association is—a strong association (as measured by a relative risk or odds ratio) is more likely to have a causal component than is a modest association. Absence of a strong association does not, however, rule out a causal association. In environmental epidemiology frequently one is dealing with relative risks in the range of 1–2—since they often apply to the general population, however, the attributable risk or public health impact in terms of the absolute number of cases can be large.

Consistency refers to a relationship that is observed repeatedly in different settings using different methods—for example, in differing groups, places, circumstances, time—by others. Finding an

association consistently using different study designs, methods, and so on reduces the probability that an association is due to an error in the study design.

Specificity refers to an association that is specific for an outcome or group of individuals—that is, altering only the cause would alter the effect. An example would be mesothelioma in relation to asbestos exposure. Due to the multifactorial nature of many environmental diseases, this is not normally the case, however.

Temporality implies that the factor must precede the outcome it is assumed to affect. This is a critical factor in assessing causality.

Biological gradient refers to a dose–response relationship between the exposure and the disease. An increasing amount of exposure increases the risk.

Biological plausibility is concerned with the extent to which the association is consistent with what is already known about the response to the exposure and/or disease—that is, does it make sense?

Coherence of evidence implies that the causal association should not fundamentally contradict present substantive knowledge.

Experimental evidence stresses that causation is more likely if evidence is based on randomized experiments.

Reasoning by analogy implies that for analogous exposures and outcomes an effect has already been shown. For example, is the observed association supported by a similar association?

Source: data from Hill, A.B., The environment and disease: association or causation?, *Proceedings of the Royal Society of Medicine*, Volume 58, Issue 5, pp. 295–300, Copyright © 1965.

Tools to aid decision-making

While obtaining knowledge on various risks is fundamental to the practice of environmental health, it is equally necessary to put in place adequate health and environment surveillance and monitoring systems to assess environmental health conditions. In the following sections we deal with the need to develop indicators for health and environment, and with health impact assessment procedures, which have become critical tools to aid decision-making in environmental health.

Indicators

Monitoring systems

Surveillance and monitoring systems of potential hazards, exposures, and diseases are essential to the practice of environmental health. By tracking exposures and diseases, we can identify and respond to different kinds of environmental health problems, and assess the effectiveness of intervention measures implemented. Surveillance systems exist in various settings, and

have been used for differing purposes. Examples include air and water pollution monitoring, blood lead monitoring programmes, poison centre surveillance, and birth defect registries. The blood lead monitoring programme conducted by the US Centers for Disease Control and Prevention (CDC) played an important role in assessing how blood lead levels in the US population fell in association with reductions in the lead content of petrol during the 1980s, at a time in which there was much scepticism in many quarters about the role that lead in petrol played in influencing blood lead levels. This shows how essential it is that information systems are developed that are supportive of policy and decision-making.

It is important to appreciate that environmental health monitoring is not the same as environmental quality monitoring. From the perspective of environmental health, of key importance is that the exposure information collected is relevant to health concerns. Frequently it is the case that exposure information is based on data from high-exposure areas obtained for the purpose of monitoring the effectiveness of environmental control measures, or of assessing compliance with legislated exposure limits. Such compliance-oriented methods are not optimal for accurate estimation of likely exposure levels of the population.

For many pollutants there is a sharp decrease in concentration levels as one moves away from a source. Similarly, significant vertical variations may occur in the concentration level of a pollutant. Knowledge of spatial and temporal variability of concentrations of exposure may therefore be important from a health point of view, and it may also be necessary to take into account that the exposure may be mediated by more than one environmental pathway. In addition, health and environmental data from monitoring programmes are frequently available at differing levels of resolution, making it difficult to determine any associations. The complexity and intersectoral nature of environmental health surveillance can thus prove very challenging, and few routine monitoring systems are set up to assess exposure to potential contaminants comprehensively.

Role of indicators

Increasingly, communities are demanding access to relevant information on the state of the environment and its influence on health, sometimes as part of their ‘right to know’. The development of indicators is of importance in

helping to make scientifically sound and useful information accessible to decision-makers and the public alike. Environmental health practitioners have an important role to play in their design and use.

The Oxford Dictionary defines an indicator as ‘providing specific information on the state or condition of something’ (Oxford Dictionaries n.d.) They are a way of ‘seeing the big picture by looking at a small piece of it’ (Plan Canada 1999). Indicators can play a key role in turning data into useful information, thereby giving it added value, and also, importantly, in simplifying it. They can give a snapshot (synthesis) view of environment and health trends, and can be used to track progress, to help build the evidence base for effective policies, to identify issues and prioritize problems, provide data for the development of interventions, and for their evaluation.

Indicators have been used in many different fields, at the global, regional, national, local, and sectoral levels. The UN system has been actively involved in indicator development for many years, and there exist numerous data sets on indicators of health, environmental health, environment, sustainable development, poverty, housing, urban development, and others (von Schirnding 2002b). Indicators have also been applied and developed in various settings and sectors, and have been used for various purposes, including in ecosystem health, health promotion, healthy cities, and countless other ways (Cheadle et al. 1992; Takano and Nakamura 2001; von Schirnding 2002b).

Types of indicators

Descriptive indicators may be used to highlight what is happening to the environment or to human health, in order to assess trends, and to obtain baseline information on which to formulate subsequent policy options and plans. They may be of use at a local level (pollution concentration levels), at a national level (mortality rates, incidence of food poisoning), or at a global level (temperature increases in relation to climate change). A performance indicator on the other hand is one that is linked to a reference value or policy target, illustrating how far the indicator is from the desired level. Examples would be ‘the proportion of the population living in electrified areas’, ‘the number of children immunized against infectious diseases’ or ‘the proportion of food outlets inspected in an area’.

Composite indicators may be developed, which condense a wide range of information on different but related phenomena into a single measure or index. In the field of health, the DALY is an example of a composite measure of the burden of disease which combines the years of healthy life lost due to premature death, disability, or disease. Gross national product (GNP) is another example of a composite indicator, as is the Human Development Index (HDI) which combines information on life expectancy at birth, education level, and income level.

Indicators have an especially important role to play in the context of the planning process and setting measurable goals and targets. There are many examples of goal setting in the area of environmental health at all levels—global to local. One example at the global level would be the SDGs. Indicators should however be applicable to the users, rather than being merely technically relevant to data providers. In fact, indicators may be derived by community groups and used for community-based monitoring.

Indicator development

Criteria which have been developed to guide the development, selection, and evaluation of indicators include such factors as their general relevance, scientific soundness, and their applicability to users. Ideally, indicators should be credible, unbiased, and representative of the condition of interest, as well as be consistent, comparable in different settings in time and space, and readily understandable, based on information that is easily accessible. The definitions used for the indicators should be clear, and the data sources and methods used for their collection should be specified (Kjellstrom and Corvalan 1995).

In many countries, despite inevitable problems related to data quality and coverage, and the limitations of routine monitoring systems, it is nevertheless usually possible to find some basic data of use, even if rudimentary or incomplete. If data on health outcomes and exposures is truly unavailable/unusable/inaccessible, it may be prudent to focus instead on action indicators. It may also be possible to use proxy measures of actual exposure (e.g. in relation to air pollution) like traffic volume, distance from a road, or proximity to industry.

Indicator frameworks

Several indicator frameworks have been developed for presenting the various linkages between factors under consideration. One of the earliest of these, upon which many others have been subsequently based, is the OECD Pressure-State-Response (P-S-R) Framework (OECD 2001) which looked at the various pressures which impacted on the state of the environment and necessitated various policy responses.

Modifications to this model such as the D-P-S-I-R (driving forces, pressure, state, impact, response) model adopted by the European Environment Agency made provision for broader driving forces to be taken account of, as well as impacts (including on health) deriving from changes to the state of the environment. The WHO's 'DPSEEA' model (driving force, pressure, state, exposure, effect, action) divides the 'impact' component further into exposure and health effects separately (Kjellstrom and Corvalan 1995; von Schirnding 2002b).

Thus, for example, if applied in environmental health, one can think of 'driving forces' as including such indicators as the population growth rate, or the level of economic development, while the resulting 'pressures' which may be exerted on the environment may be in the form of indicators like pollutant emissions. This may in turn cause the 'state' (quality) of the environment to be degraded through the dispersal and accumulation of pollutants in various environmental media like air, soil, water, and food (indicators might include pollutant concentration levels in media). People may then become 'exposed' to these hazards in the environment (an indicator might be people exposed to air pollution levels above standards) and a variety of 'health effects' may subsequently occur (indicators like air pollution-related morbidity and mortality could be useful). Various 'actions' can then be implemented at different points in the framework (indicators might include clean air policies, technical control measures, treatment regimens, or education measures).

While this may be considered a somewhat linear and simplistic model, it can nevertheless be useful for presenting data in a way that shows the linkages between exposures, health effects, and policy actions in a straightforward and readily understandable, actionable way. Nevertheless, when it comes to dealing with complex ecosystems, for example, deciding what to measure, and how to measure it, and how to analyse and present the measurements, can be very challenging (Cole et al. 1998). In the realm of climate change, development of indicators of vulnerability and

preparedness to inform public health adaptation strategies and project the impact of climate change on human health, is an example (English et al. 2009). Various other indicator models and frameworks with complex webs and feedback loops, in the realm of health and sustainability, have been developed (Waheed et al. 2009; Reis et al. 2015).

Environmental health impact assessment

It is critical to conduct research on specific hazards, carry out epidemiological studies, conduct risk assessments, do environmental monitoring and surveillance of risks, and develop indicators. Most importantly, it is necessary to prevent risks to human health occurring from the outset. In many countries relatively little attention has been paid to the health impacts of planned developments. This, however, is a critical aspect of the practice of environmental health. Impact assessment provides a powerful (yet underutilized) tool to institutionalize a holistic, cross-sectoral approach to addressing health in public policy (Bhatia and Wernham 2008).

Key steps

In line with the principle of sustainability, development should minimize negative health impacts and maximize health benefits. Health impact assessment (HIA) is a decision-support tool aimed at adding value to the decision-making process by analysing potential effects of planned policies, programmes, or projects on health (Davenport et al. 2006). A practical way for collaborating with other sectors and promoting intersectoral collaboration (Kang et al. 2011), HIA has been defined as ‘A combination of procedures, methods and tools by which a policy, programme, or project may be judged as to its potential effects on the health of a population, and the distribution of those effects within the population’ **(**Gothenburg Consensus Paper, WHO 1999). It aims to identify and predict the impacts of proposed developments on environmental parameters that have a strong health significance. Using information from epidemiological and toxicological studies, researchers attempt to identify, predict, and evaluate possible changes in health which could occur as a result of the development under consideration. A risk assessment is then undertaken (see Chapter 7.4).

As a planning tool, it is meant to optimize not only health outcomes, but also environmental, social, and economic outcomes, being not only process-oriented (involving intersectoral collaboration and community participation) but also forward looking to future impacts in the longer term, and in this way bearing the hallmarks of sustainability and sustainable development. Indeed Strategic Environmental Assessment has an important role to play in advancing sustainable development, and, despite many barriers that still need to be overcome (White and Noble 2013), is nevertheless recognized by key bodies such as the World Bank as an important tool for promoting sustainable development (World Bank 2013). There have also been efforts to strengthen the health component of strategic environmental assessment (WHO 1999).

HIA is to an extent based on the environmental impact assessment (EIA) process, with a number of procedural steps accompanied by opportunities for stakeholder involvement. Steps to conduct an HIA include *screening* to identify projects or policies for which an HIA would be useful; *scoping* to identify which health impacts should be assessed and which populations are affected; *assessing* the magnitude, direction, and certainty of health impacts; *reporting* results to decision-makers; and *evaluating* the impact of HIA on the decision-making process. Some HIAs focus on biomedical outcomes (e.g. respiratory disease), while some define health holistically to include social, environmental, and economic conditions. Others focus on equity impacts (Dannenberg et al. 2006).

HIA approaches

Some approaches consider HIA independently of EIA while in other cases HIA is integrated with EIA. For example, in some European countries HIA has evolved independently of EIA and is applied in a wide range of public policy decisions not subject to EIA (Kemm et al. 2004). Despite much work done over the past years however, in a majority of countries, health impacts are all too infrequently incorporated into the impact assessment process. In comparison with EIA, HIA is still, in many ways, in a rudimentary phase. This is somewhat surprising given the substantial growth in HIA activity in recent years, as evidenced by the number of publications, training courses, and guidance materials developed (Davenport et al. 2006). Support has grown most rapidly in Canada, Europe, Australia, and New Zealand—countries where health promotion efforts have had a strong environmental

basis (Cole and Fielding 2007). HIA is also increasingly seen as an important tool for promoting Health in All Policies (Robert Wood Johnson Foundation 2011; Leppo et al. 2013).

Indeed it is interesting to note that even in the country where EIA originated (namely the United States with the promulgation of the National Environmental Policy Act of 1969 (NEPA 1969), and which established the foundation for environmental policy in the United States)—health impacts are still infrequently incorporated into the EIA process, although this is gradually changing. While the objectives and regulations of NEPA do indeed provide for the protection of human health and welfare, in practice where health impacts have been considered, these have been narrowly focused on chemicals and toxic exposures. A comprehensive and systematic approach to addressing human health impacts in impact assessment procedures has yet to evolve. A range of institutional, organizational, and disciplinary factors may account for this (Rattle and Kwiatkowski 2003).

It is also a consequence of the complexity of relationships between exposures and health outcomes as discussed earlier. Sometimes there may be inadequate evidence and methodological challenges involved in predicting health outcomes and it is necessary to make decisions at the pace required by political or economic priorities, which may be based on existing evidence, albeit flawed or inadequate, and professional opinion.

Some have contended that a voluntary approach to HIA is preferable to a regulatory approach, where there is more procedural rigidity and stricter rules of evidence required (Cole et al. 2004). That said, it is often the case that causal certainty and quantitative precision are unrealistic and unnecessary standards for impact assessment. What seems likely is that HIA linked to EIA will conform to the limits of EIA, rather than HIA expanding the scope of the EIA (Cole and Fielding 2007), although there is some evidence in Canada that there has been some success with expanding the traditional EIA approach.

Need for capacity building

Whatever the case may be, HIA methodology has been criticized for the lack of rigour in its use of evidence, and improved guidance is needed on the kinds of public health effects to be included in an impact assessment, as well as on the methods for analysing health effects. New analytical tools are needed to translate the impacts presented in an EIA into valid health impact

predictions. The scope of health issues that could be addressed is broad, and includes conditions as diverse as traffic injuries, social cohesion, traditional diets, domestic violence, psychological problems, and so on (Bhatia and Wernham 2008).

Impact assessment training for environmental health practitioners should be included in the curricula of schools of public health as well as in continuing education courses, and countries should advocate for formal guidance on incorporating health into EIA. Examples exist in Australia and Canada, where guidance for social impact assessment contributed to institutionalizing these issues as routine considerations within EIA.

A guide to reviewing available evidence for use in HIA has been developed by the WHO (Mindell et al. 2010). However, in order for HIA to be more effective, environmental health practitioners need to familiarize themselves with the EIA process and they need also to be more involved in community collaboration (Bhatia and Wernham 2008). In practice, there often exists a tension between the need to involve communities in the HIA process but at the same time to give priority to expert- and research-generated evidence (Wright et al. 2005). In the United States, efforts to promote the role of HIA in improving health have been intensifying and the PEW Charitable Trusts has, through their Health Impact Project in collaboration with the Robert Wood Johnson Foundation, developed a national initiative to promote the use of HIA in the United States, including an online searchable directory and interactive map of HIAs (Pew Charitable Trusts and Robert Wood Johnson Foundation, n.d.).

Intersectoral action and partnerships

The previous subsections on indicators for decision-making, and on HIA processes, have highlighted the possibilities for these tools to help institutionalize intersectoral approaches to policy and decision-making. In this section, we discuss the nature of the intersectoral planning process in addressing health and environment concerns at various levels, which is a key aspect of the practice of environmental health.

While governments share the main responsibility for ensuring healthy living environments, they can do so only by working in collaboration and partnership with other groups. Involving stakeholders in all stages of the

decision-making process is an important part of the practice of environmental health. Over recent years there has been an increasing trend towards decentralized government services and more emphasis on health and environmental actions by non-governmental organizations (NGOs) and the community itself. Intersectoral, participatory planning has become more central to the way governments do business in the field of environmental health.

Where partners have been involved in the definition of problems as well as in their solution, there has been a better chance of commitments being met in the long term. There have been a number of collaborative planning initiatives at various levels of government which involve partnerships with NGOs and local communities, whereby health authorities have been able to address problems more effectively. For example, there have been the national environment and health action plans at the national level (NEHAPS), and local health and environment planning initiatives such as local Agenda 21 initiatives, and initiatives like the WHO healthy cities movement, the UN Human Settlements programme, and United Nations Environment Programme's (UNEP) sustainable cities movement (von Schirnding 1997; UNEP n.d.), among others. These initiatives flourished in the context of the UNCED and Agenda 21—the Global Plan of Action on Sustainable Development.

These, and numerous others subsequently, have led to a better understanding of the importance of environment and development issues within health ministries, as well as a better understanding of health issues within environment ministries. In many cases successful intersectoral committees or commissions and fora have been set up, and in the process alliances and linkages formed with various sectors.

On a global basis, approaches to incorporate health and environment issues in national plans and programmes have varied from country to country depending upon priorities and planning mechanisms and the way in which responsibilities are divided. This has led to a wide variety of approaches being adopted. Sometimes health and environment plans have been included in national plans for sustainable development or for economic and social development. Sectoral plans might include health and environment issues or health considerations might be included in national environmental plans and so forth.

Healthy cities

The healthy cities project has been instrumental in helping to address health and environment issues at the local level in many countries and settings. What started as a small-scale initiative of the WHO arising from the Beyond Healthcare conference held in Toronto, Canada in 1984, became a major movement involving thousands of cities throughout the world.

The healthy cities concept and approach has been translated in different ways according to different priorities and emphases of different localities. Importantly, the healthy cities movement emphasizes process—in particular the building of political commitment and a common vision locally to push health higher up the urban policy agenda. Secondly, there is involvement of a wide range of stakeholders including local communities, and emphasizing community empowerment and participation rather than an expert-led health agenda. Thirdly, strategy development within local government, in the form of a city health plan based on intersectoral partnerships and the necessary stakeholder engagement, is important. This goes along with the development of health information systems, the integration of health into other urban policy and planning processes, and the use of tools to assess health effects (Rydin et al. 2012).

According to the commission by *The Lancet*, most cities evaluated are still struggling with the holistic approach, in which health is fully integrated into urban planning. This is no doubt in part due to lack of adequate understanding of the complexity of systems at work in these urban settings, prompting utilization of a systems approach to the analysis of healthy cities in the future (Rydin et al. 2012). Of key importance though, is that the health sector fit in with planning processes in other sectors rather than the other way around. As we have seen in other areas like HIA, the health sector is also frequently lacking the expertise and capacity needed to measure or estimate the health impacts associated with proposed projects or policies. Until it has developed that capacity its ability to meaningfully participate in intersectoral work may be limited.

What is needed is to be able to better characterize the health burden from the perspective of the sector, rather than merely the specific vehicle of exposure—for example, air or water. In all instances though, data challenges remain key. These include gaps in data availability and quality, difficulties in detecting trends (geographical and temporal), difficulty in disaggregating data, poor linkages of health and environmental data, and

sectoral data. Other problems include emphasis on symptoms rather than on causes of problems, little focus on analysis of management structures and capacity, and lack of clarity in priority-setting processes. Many of these issues have been discussed in earlier sections. Environmental health management issues are further elaborated on in the following section.

Environmental health policy and management

While there is more acceptance now of the links between health status and the determinants of health, especially those that arise from the activities of sectors other than health, there are still problems acting on this knowledge. Greater attention needs to be paid to developing institutional, managerial, administrative, and financial capacities to address problems. Definitions of environmental health aside, it has become increasingly complex to define roles and responsibilities in environmental health in light of its inherent intersectoral nature.

The Institute of Medicine (IOM) report *Investing in a Healthier Future* (2012) highlighted the need to develop a detailed description of a basic set of public health services that must be made available in all jurisdictions. They developed the concept of a minimum package of public health services which includes foundational capabilities and basic programmes which no health department can be without. These apply regardless of the specific area of control—for example, whether it be infectious disease control or environmental health. They are capabilities that are needed across all programmes.

Some of these foundational public health capabilities include the following:

Information systems and resources, including surveillance and epidemiology.
Health planning (including community health improvement planning).
Partnership development and community mobilization.
Policy development, analysis, and decision support.
Communication (including health literacy and cultural competence).
Public health research, evaluation, and quality improvement.

Environmental health policy

Environmental health policy is concerned with selecting actions to reduce or eliminate harmful effects of environmental exposures. Some measures may be short term and remedial, while others may be preventive and longer term. Actions can take many forms and involve such factors as regulations and standards, technological control measures, product design, personal or community protection measures, health and environment action plans, housing or transport policies, awareness raising, and education measures. Such policies can take the form of standards that directly regulate the allowable limit of a contaminant, for example, an air pollution or drinking water standard. Other environmental protection measures can benefit health indirectly, for example, via safeguarding biodiversity, or determining which fuels are used, and hence the impact on air quality and health.

We can maximize opportunities for health by better aligning the policies of sectors with the potential to affect health. Health professionals need constantly to motivate support for policies that are beneficial to health—for example, food and agricultural policies that promote integrated vector control over use of toxic pesticides or encouraging substitution of crops that harm health, and that promote consumption of diets balanced in fruits and vegetables. Similarly, energy policies need to support clean energy sources, cleaner and more energy efficient transport, and energy efficient buildings while transportation policies need to encourage use of bicycles and mass transit.

The overall cumulative impact of such policies can be substantial and can help avoid scenarios in which health is sacrificed for short-term sectoral or economic gain. That is not to imply that health and environmental goals always coincide—frequently they do not. DDT is one example where the immediate health gains were used to lobby for its continued use in sub-Saharan Africa despite its longer-term insidious negative impact on the environment.

Some laws and regulations may in turn contain policies embedded in them. For example, NEPA in the United States states that ‘Congress recognizes that . . . each person should enjoy a healthful environment and that each person has a responsibility to contribute to the preservation and enhancement of the environment’ (Johnson 2005).

Federal laws form much of the environmental framework in the United States and many of these have been emulated in many other countries. Examples of such laws include the Safe Drinking Water Act, and the Clean Air Act. A strong science base is normally required for setting standards, and a science-based risk assessment process undertaken to inform risk management policies and strategies, which of necessity also take into account social, economic, and political considerations.

Policies may also span the spectrum of government—from global to local. In general, policy approaches tend to become more specific and targeted as they go from national to state/provincial/regional to local. In the United States, for example, federal statutes and regulations may be issued and states and local governments will issue statutes or ordinances needed to comply with these, or they may develop their own regulations. Policies may be preventive and aim to reduce hazards directly in a prospective manner (e.g. air pollution regulations) or they may be retrospective and require, for example, clean-up of a toxic waste site. Other types of policies may be aimed at awareness raising and education of the public rather than regulatory control (Johnson 2005
).

International policy principles have been adopted as well. For example, as indicated earlier, the UNCED adopted the Precautionary Principle, which states ‘ … Where there are threats of serious or irreversible damage, lack of full scientific certainty shall not be used as a reason for postponing cost-effective measures to prevent environmental degradation’ (UNCED 1992). Another principle adopted was the Polluter Pays Principle—that is, those who cause pollution should pay the price to clean it up. Economic instruments such as pollutant trading systems are also in use which shift the societal cost of pollution to the polluter.

The principle of intergenerational equity states that ‘The right to development must be fulfilled so as to equitably meet developmental and environmental needs of present and future generations’. This principle has particular relevance in regard to chemicals with long-term adverse consequences, and to greenhouse gases and climate change, for example. Another UNCED principle, the principle of integrated decision-making states that ‘In order to achieve sustainable development, environmental protection shall constitute an integral part of the development process and cannot be considered in isolation from it’. Thus, environmental

considerations need to be incorporated into decision-making processes at all levels, global to local, and in all sectors, as alluded to earlier. Environment and HIA processes, and the 'health in all policies' movement, are example of this principle in action.

Impacts originating from a multitude of small sectors rather than from one large obvious source such as a major polluting factory, for example, are often more difficult to control. Smaller enterprises usually have fewer resources at their disposal and are more resistant to change, or may be less knowledgeable about standards which may apply, than large-scale industry. In developing countries, fewer controls are generally applied and control technology may be less well developed. In the absence of local standards (which demand relevant epidemiological and exposure data, which may often not be available), international standards are often applied but may not be appropriate (Goldman and Torres 2009) due to the differing prevailing lifestyles, cultures, climate, demographic attributes, institutional and policy framework, level of economic development, and so on, which differ from country to country.

Environmental standards may be risk based (i.e. based on assessments of environmental or health risks) or technology based (i.e. based on the best available technology) or a combination of both. Pollution prevention is key. Increasingly it has been recognized that controlling pollutants one medium at a time can result in transferring pollutants from one medium to another, for example, from air to water to land. Multimedia approaches are thus preferable. Reduction of pollution at source is optimal, but other strategies include waste minimization, reuse, recycling, emissions control, and, finally, clean-up. Environmental monitoring has an important role to play in evaluating the success or otherwise of the various control measures adopted, and the public may have an important role in advocacy, monitoring, and enforcement of environmental policies.

Financial incentives

As environmental health protection measures inevitably involve costs, cost–benefit analyses may be undertaken to weigh costs and benefits of an action, while cost-effectiveness analyses might be performed to look at alternative solutions to a problem. Attempts may also be made to assess economic inequities in impacts, taking into consideration issues of environmental justice. Economic and fiscal policies can influence the

potential for health gains and their distribution in society. Financial incentives can be established so that they can incentivize a healthier lifestyle and discourage the use of harmful products and substances. When combined with appropriate legislation and health education programmes, many of the negative health trends could be retarded or even reversed.

International regulations

When pollutants cross boundaries, or where there are global impacts and concerns, international agreements and standards come into play. In parallel with the role that international conferences have played in galvanizing action on health and environment issues, so too do international environmental treaties offer the prospect of further advances in public health. Due in part to the WHO's and others' increasing involvement in international conferences and treaty-making processes in recent years, health issues have received increasing recognition and attention. Mechanisms in international environmental law can be used to provide motivations for research aimed at further strengthening the evidence base related to health, and to achieve improved surveillance and monitoring systems concerned with ill-health. Increased efforts are needed to more fully realize the potential of international law to promote global environmental health (von Schirnding et al. 2002).

Many multilateral environmental agreements have health implications, for example, the Montreal Protocol on Protection of the Ozone Layer (1987), the Basel Convention on the Control of Transboundary Movements of Hazardous Wastes and their Disposal (1989), the Convention on Biological Diversity adopted in 1992 (including the Cartagena Protocol on Biosafety), the Kyoto Protocol of the UN Framework on Climate Change (1997), the Rotterdam Convention on the Prior Informed Consent Procedure for Certain Hazardous Chemicals and Pesticides in International Trade (1998), and the Stockholm Convention on Persistent Organic Pollutants (2001).

Environmental agreements in existence today, however, recognize the sovereign right and responsibility of nations to set their own standards. So, for example, the Basel Convention does not cover handling and disposal of wastes within a country, and the Rotterdam Convention does not require prior consent for chemicals manufactured and produced within a country (Goldman and Torres 2009).

Environmental health services

It is complicated to manage environmental health issues which are intersectoral, and whose determinants frequently lie outside the control of the health sector, or outside of jurisdictional boundaries, with different organizations and agencies having differing roles and responsibilities in respect of policy development and implementation, In addition, systems of organization in many government authorities worldwide may be based on concepts of bureaucracy, hierarchy, professional authority, disciplinary specialization, and sectoral analysis that were established for purposes other than the provision of environmental health services (von Schirnding 1997). To facilitate implementation of intersectoral policies and plans there may be a need for institutional restructuring and readjustment in order to overcome barriers and obstacles in coordinating action between sectors.

One approach is to make internal reforms within government departments. Issues that may need to be addressed might include creating mechanisms for interjurisdictional cooperation in implementing policies and plans, decentralizing of structures facilitate community involvement across sectors, and establishing better coordination between and within governance structures. Other types of jurisdictional reform include transferring powers and functions from one department to another or from one tier of government to another, changing geographical boundaries, consolidating multiple jurisdictions, and/or forming various coordinating mechanisms. In any case, increasing mechanisms for sharing power, exchanging information, and aligning financial resources and incentives are all important.

While there is no universal understanding of what precisely constitutes an environmental health service, it is essentially concerned with developing and implementing environment and health policies, largely through monitoring and control activities, but also increasingly through environmental health promotion efforts. Environmental health services worldwide have been established and developed within a complex framework of policies, institutions, and legislation. The concept of environmental health, the problems to be addressed, the way in which services are delivered, and the way in which governments are organized differ considerably throughout the world.

Environmental health activities are carried out in many different settings at various tiers of government, in the private sector, the workplace, healthcare facilities, and so on. In most countries a range of different agencies have responsibility for various aspects of environment and human health at different levels of government. Normally there are separate regulatory agencies dealing with the environment, health, food safety, agriculture, and so on, which have various environmental health responsibilities. Different tiers of government have different responsibilities. Thus, for example, regulatory frameworks and standards (e.g. for air, drinking water, and so on.) may be set at a national level, while implementation and enforcement may occur at a state, regional, or local level. Inspection of food preparation premises, food hygiene, sanitation, vector control, and so on, are normally done at the local level. This requires strong monitoring and enforcement abilities which are sometimes in short supply, especially in some developing countries.

Internationally a shift has occurred towards organizational diversification regarding environmental health services. In the context of the development of primary healthcare services and district health systems in many developing countries, and parallel development outside of the health sector of environmental management systems, the placement of environmental health services within government structures is a fundamental issue in need of addressing. Despite the fact that environmental health is recognized as a major and core function of a public health department, it is also the case that the public health sector has gradually lost organizational responsibility for many aspects of environmental health.

This includes, increasingly, such aspects as air quality, waste management, chemical safety, meat inspection, and many other areas. While it would be unmanageable for the full ambit of environmental health services to be located centrally in one department, in light of its intersectoral nature, what is needed is to develop organizational structures that facilitate the movement of professionals across different departments, sectors, and organizations. Box 11.6.3 gives a case study from the United States, which is typical of the situation in many countries around the world, who face similar challenges in regard to the placement of environmental health functions subsequent to the establishment of environmental departments and new environmental regulations.

Box 11.6.3 Organizational responsibility for environmental health in the United States

In the first part of the twentieth century in the United States, a typical scenario was a health department headed by a physician, with nurses responsible for immunization programmes and sanitarians/sanitary engineers responsible for sanitary inspection, water and sanitation, food safety, vector control, and so on. Sanitarians were generalists trained to respond to a broad array of environmental health issues.

With increasing industrialization and growing public awareness about the widespread contamination of the environment in the 1960s, a number of environmental laws dealing with water and air quality, toxic substances, and other issues were passed, and the US EPA was established. Separate state-level environmental agencies were also established, with responsibility for enforcement of these new environmental laws. At this time many of the traditional environmental health functions falling under health departments were transferred to environmental departments with positive (increased funding and visibility) but also negative (schism in environment and health actions) repercussions. Separate data systems, uncoordinated planning, and loss of oversight of community health and environment issues resulted. With the new environmental 'medium-specific' laws, and the emergence of environmental health specialists in particular areas like water or air quality, environmental health professionals became increasingly disconnected from their public health roots and counterparts in health departments at state or local level and the agencies grew apart too.

Subsequently in the 1970s and 1980s there was a growing public awareness and involvement in environmental health issues and a demand for more accountability and transparency in decision-making. With the advent of environmental impact assessment legislation and requirements for public involvement, watchdog groups, and advocacy organizations were formed with a strong focus on environmental justice, human rights, and equity concerns. Increasing initiatives to reconnect environmental and public health sprang up out of a sense of growing dissatisfaction over the inability of accountable agencies to address these in a coordinated and integrated way.

Source: data from Kotchian S., The practice of environmental health, pp. 895–925, in Frumkin H., (Ed), *Environmental Health: From Global to Local*, Jossey-Bass, San Francisco, USA, Copyright © 2005.

Environmental health practitioners

In general, environmental health practitioners and professionals have as their main responsibilities the identification, assessment, management, communication, and remediation of health risks resulting from environmental factors. They include people working in a wide range of professions and specialties and include environmental health officers (EHOs), sanitarians, health inspectors, clinicians, epidemiologists, toxicologists, environmental scientists, environmental engineers,

veterinarians, and many others. Examples of specialty areas of environmental health practice are given in Box 11.6.4.

Box 11.6.4 Specialty areas of environmental health practice

- Air quality management
- Consumer safety
- Emergency preparedness
- Environmental epidemiology
- Food safety
- Housing quality
- Noise control
- Outbreak investigation
- Radiation control
- Recreational health and safety
- Risk assessment
- Solid waste management
- Sanitation
- Toxic substances control
- Vector control
- Drinking and recreational water supply and treatment
- Water quality management

Environmental health officers

As stated at the beginning of this chapter, the roots of the environmental health profession today can be found in the sanitary and public health movement of the United Kingdom during the 1800s. The role of the health inspector has steadily changed over the years, and has evolved from a focus on inspection and control, to a broader role that also includes education, health promotion, community development, and indeed sustainable development (Guidotti 2015). Commensurate with the changing role of the health inspector has been a change in title—variously referred to over time and in different settings as sanitarian, inspector of nuisances, sanitary inspector, health inspector/technician/assistant, public health officer, environmental health officer, or environmental health practitioner/specialist/professional.

Notwithstanding changing roles and titles, the theory and conceptual basis of environmental health is firmly based in public health, and the practice of environmental health, as embodied in the day-to-day work of EHOs and others, is closely associated with inspection and environmental control, with EHOs being responsible for administering and enforcing legislation related to environmental health.

Normally an EHO would be a public employee in a health department in local government or at regional/provincial/state level who investigates environmental hazards and public health nuisances with the view to mitigating or eliminating these. EHOs are involved in inspection of housing, food establishments, schools, daycare centres, nursing homes, and many other premises, and are also involved in vector control, pollution control, and health promotion activities. They may also be employed by the military and by the private sector.

It is inevitable that, with the increasing involvement in environmental health of professionals and specialists from a wide variety of fields including environmental chemistry, physics, environmental science, geography, environmental engineering, public health, medicine, and others, the role of the EHO outside of its traditional ambit of inspection, monitoring, and control will continue to evolve. This also has a bearing on their role inside of, and outside of, the health service, and the challenges and constraints faced while working on the broader determinants of health, within the principles of sustainable development, but remaining firmly public health based.

Qualifications and training

Environmental health requires, in many countries, a university degree, with field training and professional certification needed for environmental health officers/practitioners. In the United Kingdom, a degree in environmental health is required, as well as certification and registration with the Chartered Institute of Environmental Health. Many universities in the United Kingdom offer accredited environmental health degrees, including undergraduate and postgraduate degrees. A postgraduate environmental health qualification is required for graduates from other scientific disciplines (Chartered Institute of Environmental Health n.d.). The Chartered Institute of Environmental Health offers training modules in

specialized areas of environmental health as does the Environmental Health Registration Board (n.d.).

In Canada, EHOs are required to hold a bachelor's degree in environmental health or a related field as well as a certificate in public health inspection. A number of universities in Canada offer degree programmes approved by the Canadian Institute of Public Health Inspectors (n.d.). A bachelor's degree and professional licence is also a requirement in parts of the United States and there exists the National Environmental Health Association (NEHA) offering environmental health credentials and registration/certification for practitioners. Interestingly, the premier NEHA credential is referred to as the Registered Environmental Health Specialist/Registered Sanitarian (REHS/RS) (NEHA n.d.).

There exists, as well the International Federation of Environmental Health (n.d.), the initiative to promote cooperation between countries where environmental health issues are cross-boundary. Regional groups exist in Africa, the Americas, Asia and the Pacific, Europe, and the Middle East. Member countries in turn have their own professional institutes and associations of environmental health. These include such bodies as the Environmental Health Workers Association of Uganda, the Public Health Inspectors Union of Sri Lanka, the Malta Association of Environmental Health Officers, the Kenya Association of Public Health Officers, and numerous others, reflecting the variety of names given to those practising environmental health around the world.

In some countries (South Africa for example), degree courses in environmental science or health are offered at universities, and diploma courses (e.g. the National Diploma in Environmental Health) at technical institutions. Degrees in environmental health technology are also offered. Environmental health practitioners are required to complete practical training and registration with the Health Professions Council of South Africa (n.d.). Training programmes have been updated to accommodate the changing role of the environmental health officer in South Africa from law enforcer to community facilitator, in line with changing political and economic realities which make legislation difficult to enforce, especially in the context of rapid urbanization, informal settlements, and a large informal sector.

Clinicians

Clinicians are becoming increasingly involved in the practice of environmental health. The diagnosis of illnesses that are associated with environmental exposures is an important clinical function that expanded from the field of occupational medicine, to a focus on the broader environment outside of the workplace (Laumbach and Kipen 2005). Treating environmentally related diseases is often the same as treating these illnesses from other causes, although in some cases (e.g. lead poisoning), specialized treatments may be required. Clinicians need to be alert to the role that environmental factors may play and adopt a high 'index of suspicion' in assessing underlying signs, symptoms, and diseases in patients —whether adults or children (Laumbach and Kipen 2005).

If an environmental cause is suspected, an environmental history should be taken in order to obtain information on potential sources and pathways of exposure (both current and past), and in the case of a paediatric evaluation the age and development stage of the child. In the case of infants, for example, breastfeeding could be a source of exposure to be considered, and the hand-to-mouth pathway could be of special significance.

Examples of environmental health problems dealt with by clinicians include lead poisoning, asthma associated with air pollution, poisoning by pesticides, hearing loss caused by noise, and cancer associated with excessive radiation. Apart from diagnosing and treating environmental diseases, clinicians also have an important role to play in prevention, including advocacy for protective standards, and patient education to avoid or reduce exposures.

Children's environmental health issues have become of increasing importance to paediatricians. In 1997, an Executive Order on Children's Environmental Health and Safety was signed in the United States and a cabinet-level oversight committee was formed between the US EPA and the Department of Health and Human Services. Since then, children's environmental health has become an increasingly important part of paediatric practice (Landrigan and Etzel 2013). Improved exposure and risk assessment and toxicity testing is needed to more adequately address children's environmental health issues, however (Galvez et al. 2005). As part of the WHO's contribution to the follow-up to the WSSD held in 2002, an international alliance on 'Healthy Environments for Children' (HECA)

was formed to help advocate for worldwide action on environmental health issues affecting children (WHO n.d. 2009).

Conclusion

This chapter has highlighted a number of key issues relating to the nature and practice of environmental health, which is a complex, dynamic, and evolving multidisciplinary field. It is increasingly evident that environmental health practitioners face numerous challenges in addressing the myriad factors influencing health and environment relationships in differing contexts and settings.

The multiplicity and interaction of chemical, physical, and biological risks in the environment, the various media and pathways of exposure, the differing vulnerabilities and susceptibilities of diverse groups, and the changing priorities and capacities for the management of environmental health risks in various parts of the world, all need to be addressed by practitioners of environmental health.

Despite dramatic advances in the recognition and amelioration of environmental health risks worldwide in recent decades, there nevertheless remain significant gaps in knowledge of new and emerging risks, and in approaches needed to more effectively address them. Improved indicators which are reflective of complex health and environment interactions, and health impact and risk assessment procedures which can better predict impacts associated with differing environment and development parameters and scenarios, are needed.

So too are systems approaches needed for addressing the complexity of issues associated with ecosystems, global change, urbanization, and sectors like energy, housing, and transport. Economic development, the environment, and human health are closely intertwined and interconnected, and need to be addressed as such.

As evidenced by the intersectoral health and sustainable development planning processes which have been underway in recent years at all levels —global to local—there will be more of a need for practitioners to work proactively with communities and other stakeholders and partners in addressing the complexity of health and environment issues. While increasing capacity is needed to assess and address intersectoral impacts,

the causes of which frequently lie outside the domain of the health sector, tools like HIA can help to institutionalize intersectoral processes in policymaking and decision-making.

With the continuing development of improved risk assessment procedures and environmental health management tools and policy instruments, and ongoing capacity development of environmental health practitioners and professionals in various settings and tiers of government worldwide, excellence in the practice of environmental health will culminate in increasingly healthier environments for current and future generations to come.

REFERENCES

Annan, K. (2002). *Towards a Sustainable Future. American Museum of Natural History's Annual Environmental Lecture*. New York: American Museum of Natural History.

Barbier, E.B. (2012). The green economy post Rio+20. *Science*, 338(6109), 887–8.

Bartel, S. (2005). Risk assessment. In: Frumkin, H. (ed.) *Environmental Health: From Global to Local*. San Francisco, CA: Jossey-Bass, pp. 940–60.

Bhandary, R., et al. (2015). *Taking Stock: A Brief Analysis of the UN Sustainable Development Summit*. Winnipeg: International Institute for Sustainable Development (IISD).

Bhatia, R. and Wernham, A. (2008). Integrating human health into environmental impact assessment: an unrealized opportunity for environmental health and justice. *Environmental Health Perspectives*, 116(8), 991–1000.

Calman, K. (1998). The 1848 Public Health Act and its relevance to improving health in England now. *British Medical Journal*, **317**, 596–8.

Canadian Institute of Public Health Inspectors (n.d.). Website. [Online] Available at: http://www.ciphi.on.ca/

Carson, R.L. (1962). *Silent Spring*. Cambridge: Riverside Press.

Chadwick, E. (1842). *Report of the Sanitary Condition of the Labouring Population and on the Means of its Improvement*. London: Printed by R. Clowes & Sons, for Her Majesty's Stationery Office.

Chartered Institute of Environmental Health (n.d.). Website. [Online] Available at: http://www.cieh.org

Cheadle, A., et al. (1992). Environmental indicators: a tool for evaluating community-based health promotion programs. *American Journal of Preventive Medicine*, **8**, 345–50.

Cole, B. and Fielding, J. (2007). Health impact assessment: a tool to help policy makers understand health beyond healthcare. *Annual Review of Public Health*, **28**, 393–412.

Cole, B., Wilhelm, M., Long, P., Fielding, J., Kominski, G., and Morgenstern, H. (2004). Prospects for health impact assessment in the United States: new and improved environmental impact assessment of something different? *Journal of Health Politics, Policy and Law*, **29**, 1153–86.

Cole, D.C., Eyles, J., and Gibson, B.L. (1998). Indicators of human health in ecosystems – what do we measure? *Science of the Total Environment*, 224(1–3), 201–13.

Dannenberg, A.L., et al. (2006). Growing the field of health impact assessment in the United States: an agenda for research and practice. *American Journal of Public Health*, 96(2), 262–70.

Davenport, C., Mathers, J., and Parry, J. (2006). Use of health impact assessment in incorporating health considerations in decision-making. *Journal of Epidemiology and Community Health*, 60(3), 196–201.

English, P.B., et al. (2009). Environmental health indicators of climate change for the United States: findings from the state environmental health indicator collaborative. *Environmental Health Perspectives*, 117(11), 1673–81.

Environmental Health Registration Board (n.d.). *Welcome.* [Online] Available at: http://www.ehrb.co.uk

Fiorino, D.J. (1995). *Making Environmental Policy*. Los Angeles, CA: University of California Press.

Forouzanfar, M.H., et al. (2016). Global, regional, and national comparative risk assessment of 79 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015. *The Lancet*, 388(10053), 1659–724.

Galvez, M., Forman, J., and Landrigan, P.J. (2005). Children. In: Frumkin, H. (ed.) *Environmental Health: From Global to Local*. San Francisco, CA: Jossey-Bass, pp. 805–45.

Goldman, L. and Torres, E.B. (2009). Environmental health practice. In: Detels, R., Beaglehole, R., Lansang, M.A., and Gulliford, M. (eds.) *Oxford Textbook of Public Health* (5th ed.). Oxford: Oxford University Press, pp. 1639–52.

Goldman, L. (2016). Environmental public health: from theory to practice. In: Frumkin, H. (ed.) *Environmental Health: From Global to Local*. San Francisco, CA: Jossey-Bass, pp. 725–46.

Guidotti, T. (2015). *Health and Sustainability: An Introduction*. Oxford: Oxford University Press.

Haglund, B., Pettersson, B., Finer, D., and Tillgren, P. (eds.) (1996). *Creating Supportive Environments for Health: Stories from the Third International Conference on Health Promotion, Sundsvall, Sweden*. Geneva: WHO.

Health Professions Council of South Africa (n.d.). *Website.* [Online] Available at: http://www.hpcsa.co.za

Hill, A.B. (1965). The environment and disease: association or causation? *Proceedings of the Royal Society of Medicine*, **58**, 295–300.

Hippocrates (400 BCE). *On Airs, Waters and Places*. (Translation by Francis Adams published 19 August 2015) *Hippocratic Corpus. Internet Classics Archives.* [Online] Available at: http://classics.mit.edu//Hippocrates/airwatpl.html

Institute of Medicine (2012). *For the Public's Health: Investing in a Healthier Future*. Washington, DC: The National Academies Press.

International Federation of Environmental Health (n.d.). Website. [Online] Available at: http://www.ifeh.org/

Johnson, B.L. (2005). Environmental health policy. In: Frumkin, H. (ed.) *Environmental Health: From Global to Local*. San Francisco, CA: Jossey-Bass, pp. 961–87.

Kang, E., Park, H., and Kim, J. (2011). Health impact assessment as a strategy for intersectoral collaboration. *Journal of Preventive Medicine and Public Health*, 44(5), 201–9.

Kemm, J., Parry, J., and Palmer, S. (2004). *Health Impact Assessment: Concepts, Techniques and Applications*. Oxford: Oxford University Press.

Kjellstrom, T. and Corvalan, C. (1995). Framework for the development of environmental health indicators. *World Health Statistics Quarterly*, 48(2), 144–54.

Kleinert, S., Bonk, M., and Horton, R. (2012). New voices in global health and sustainable development. *The Lancet*, 382, S1–2.

Kotchian, S. (2005). The practice of environmental health. In: Frumkin, H. (ed.) *Environmental Health: From Global to Local*. San Francisco, CA: Jossey-Bass, pp. 895–925.

Kreisel, W. and von Schirnding, Y.E.R. (1998). Intersectoral action for health: a cornerstone for health-for-all in the 21st century. *World Health Statistics Quarterly*, 51(1), 75–8

Landrigan et al. (2018). Lancet Commission on Pollution and Health. *The Lancet*, 391(10119), 462–512.
Landrigan, P. and Etzel, R. (eds.) (2013). *Textbook of Children's Environmental Health*. Oxford: Oxford University Press.
Laumbach, R. and Kipen, H.M. (2005). Health care services. In: Frumkin, H. (ed.) *Environmental Health: From Global to Local*. San Francisco, CA: Jossey-Bass, pp. 1010–31.
Leppo, K., et al. (2013). *Health in All Policies.* Finland: Ministry of Social Affairs and Health.
McMichael, A.J. (2013). Globalization, climate change, and human health. *The New England Journal of Medicine*, **368**, 1335–43.
McMichael, A.J., Woodruff, R.E., and Hales, S. (2006). Climate change and human health: present and future risks. *The Lancet*, 367(9513), 859–69.
Mindell, J., et al. (2010). Improving the use of evidence in health impact assessment. *Bulletin of the World Health Organization*, 88(7), 543–50.
National Environmental Health Association (n.d.). Registered Environmental Health Specialist/Registered Sanitarian (REHS/RS). [Online] Available at: http://www.neha.org/credential/REHS.html
National Environmental Policy Act (NEPA) (1969). *U.S.C. 4321-4347 Public Law* 91–190, 42.
National Research Council (1983). *Risk Assessment in the Federal Government: Managing the Process*. Washington, DC: National Academies Press.
National Research Council (1994). *Science and Judgment in Risk Assessment*. Washington, DC: National Academies Press.
Organisation for Economic Co-operation and Development (2001). *Environmental Indicators*. Paris: OECD.
Oxford Dictionaries (n.d.). *Indicator*. Oxford: Oxford University Press. [Online] Available at: http://www.oxforddictionaries.com/definition/english/indicator?q=indicator
Patz, J., Campbell-Lendrum, D., Gibbs, H., and Woodruff, R. (2008). Health impact assessment of global climate change: expanding on comparative risk assessment approaches for policy making. *Annual Review of Public Health*, **29**, 27–39.
PEW Charitable Trusts and Robert Wood Johnson Foundation (n.d.). *Health Impact Assessments in the United States*. [Online] Available at: http://www.healthimpactproject.org/en/projects/health-impact-project)
Plan Canada (1999). *Sustainable Communities Indicator Program (Vol. 39, No. 5)*. Toronto: Plan Canada.
Rapport, D.J., Gaudet, C.L., Constanza, R., Epstein, P.R., and Levin, R. (2009). *Ecosystem Health: Principles and Practice*. Chichester: John Wiley and Sons.
Rattle, R. and Kwiatkowski, R.E. (2003). Integrating health and social impact assessment. In: Becker, H.A. and Vanclay, F. (eds.) *The International Handbook of Social Impact Assessment*. Cheltenham: Edward Elgar, pp. 92–107.
Reis, S., et al. (2015). Integrating health and environmental impact analysis. *Public Health*, **129**, 1383–9.
Robert Wood Johnson Foundation (2011). *Health Impact Assessment: A Tool for Promoting Health in All Policies.* Issue Brief #11. Princeton, NJ: Robert Wood Johnson Foundation.
Rosen, G. (1958). *A History of Public Health*. Cambridge: Cambridge University Press.
Rydin, Y., et al. (2012). Shaping cities for health: complexity and the planning of urban environments in the 21st century. *The Lancet*, **379**, 2079–108.
Sachs, J.D. (2012). From Millennium Development Goals to sustainable development goals. *The Lancet*, **379**, 2206–11.
Selikoff, I.J., Churg, J., and Hammond, E.C. (1965). Relation between exposure to asbestos and mesothelioma. *Journal of Occupational Medicine*, 7(10), 542.

Selikoff, I.J., Hammond, E.C., and Churg, J. (1968). Asbestos exposure, smoking and neoplasia. *Journal of the American Medical Journal*, 204(2), 106–12.
Singh, S., Kalmus-Eliasz, M., and Yore, D. (2012). Health at the Rio+20 negotiations. *The Lancet*, 380(9846), e5–6.
Soskolne, C.L., Butler, C.D., Ijsselmuiden, C., London, L., and von Schirnding, Y.E.R. (2007). Toward a global agenda for research in environmental epidemiology. *Epidemiology*, 18(1), 162–6.
Stanaway, J.D et al. (2018). Global, regional, and national comparative risk assessment of 84 environmental and occupational, and metabolic risks or clusters of risks for 195 countries and territories, 1990–2017: A systematic analysis for the Global Burden of Disease Study 2017. *The Lancet*, 392(10159), 1923–94
Takano, T. and Nakamura, K. (2001). An analysis of health levels and various indicators of urban environments for healthy cities projects. *Journal of Epidemiology and Community Health*, **55**, 263–70.
Tickner, J.A. (2005). Prevention. In: Frumkin, H. (ed.) *Environmental Health: From Global to Local*. San Francisco, CA: Jossey-Bass, pp. 849–94.
Tong, S., von Schirnding, Y.E.R., and Prapamontol, T. (2000). Environmental lead exposure: a public health problem of global dimensions. *Bulletin of the World Health Organization*, 78(9), 1068–77.
United Nations (UN) (1972). *United Nations Conference on the Human Environment, Stockholm, Sweden, 5–16 June*.
United Nations (UN) (1992). *United Nations Conference on Environment and Development (UNCED), Rio de Janeiro, 3–14 June*.
United Nations (UN) (2000). *United Nations Millennium Assembly. 55th Session of the United Nations*. New York: UN.
United Nations (UN) (2002a). *World Summit on Sustainable Development (WSSD)*, Johannesburg, 26 August–4 September.
United Nations (UN) (2002b). *WEHAB Working Group. A Framework for Action on Health and the Environment*. New York: UN.
United Nations (UN) (2012). *The Future We Want: Rio +20 Outcome Document*, 20–22 June, Rio de Janeiro, Brazil. New York: UN.
United Nations (UN) (2015). *Sustainable Development Summit*, New York, 25–27 September. New York: UN.
United Nations (UN) (2018). *High-level Political Forum*, New York, 9–18 July. New York: UN.
United Nations Children's Fund (UNICEF) (2012). *Pneumonia and Diarrhoea: Tackling the Deadliest Diseases for the World's Poorest Children*. New York: UNICEF.
United Nations Development Programme (UNDP) (2011). *Human Development Report*. New York: UNDP.
United Nations Development Programme (UNDP) (2015). *The Millennium Development Goals Report 2015*. UN
United Nations Development Programme (UNDP) (2016). *Human Development Report 2016*. New York: UNDP.
United Nations Development Programme (UNDP) (2018). *Human Development Indices and Indicators: Statistical Update*. New York: UNDP
United Nations Environment Programme (UNEP) (n.d.). *International Register of Potentially Toxic Chemicals*. [Online] Available at: http://www.unep.ch/
Von Schirnding, Y.E.R. (1997). Addressing health and environment concerns in sustainable development with special reference to participatory planning initiatives such as healthy cities. *Ecosystem Health*, 3(4), 220–8.

Von Schirnding, Y.E.R. (2002a). Health and sustainable development: can we rise to the challenge? *The Lancet*, **360**, 632–7.
Von Schirnding, Y.E.R. (2002b). *Health in Sustainable Development Planning: The Role of Indicators*. Geneva: WHO.
Von Schirnding, Y.E.R. (2005). The World Summit on Sustainable Development: reaffirming the centrality of health. *Globalisation and Health*, **1**, 8.
Von Schirnding, Y.E.R., Bradshaw, D., Fuggle, R.F., and Stokol, M. (1991). Blood lead levels in South African inner-city children. Blood lead levels in inner city Cape Town children. *Environmental Health Perspectives*, **94**, 125–30.
Von Schirnding, Y.E.R., Onzivu, W., and Adede, A.O. (2002). International environmental law and global public health. *Bulletin of the World Health Organization*, 80(12), 970–4.
Waheed, B; Khan, F; and Veitch, B. (2009). Linkage-based frameworks for sustainability assessment: making a case for driving force-pressure-state-exposure-effect-action (DPSEEA) frameworks. *Sustainability*, **1**, 441–63.
Wang, H., et al. (2013). Age-specific and sex-specific mortality in 187 countries, 1970–2010: a systematic analysis for the global burden of disease study 2010. *The Lancet*, **380**, 2071–94.
White, L. and Noble, B. (2013). Strategic environmental assessment for sustainability: a review of a decade of academic research. *Environmental Impact Assessment Review*, **42**, 60–6.
Winslow, C. (1949). The message of Lemuel Shattuck for 1948. *American Journal of Public Health*, 39(2),156–162.
World Bank (2013). *Strategic Environmental Assessment*. [Online] Available at: http://www.worldbank.org/en/topic/environment/brief/strategic-environmnetal-assessment
World Commission on Environment and Development (1987). *Our Common Future*. Oxford: Oxford University Press.
World Health Organization (WHO) (1946). *Preamble to the Constitution of the World Health Organization as adopted by the International Health Conference, New York. June 19–22, 1946*. New York: WHO.
World Health Organization (WHO) (1948). *Constitution*. Geneva: WHO.
World Health Organization (WHO) (1977). *Health-for-All by the Year 2000*. Geneva: WHO.
World Health Organization (WHO) (1978). *Alma-Ata Primary Healthcare Conference*. Geneva: WHO.
World Health Organization (WHO) (1981). *Global Strategy for Health-for-All by the Year 2000*. Geneva: WHO.
World Health Organization (WHO) (1986a). *Intersectoral Action for Health. The Role of Intersectoral Cooperation in National Strategies of Health-for-All*. Geneva: WHO.
World Health Organization (WHO) (1986b). *Ottawa Charter for Health Promotion. First International Conference on Health Promotion, Ottawa, Canada, 17–21 November*. Geneva: WHO.
World Health Organization (WHO) (1989). *Environment and Health. The European Charter and Commentary. First European Conference on Environment and Health, Frankfurt, Germany, 7–8 December*. Geneva: WHO.
World Health Organization (WHO) (1992). *Commission on Environment and Health: Our Planet, Our Health*. Geneva: WHO.
World Health Organization (WHO) (1997). *Intersectoral Action for Health: A Cornerstone for Health-for-All in the 21st Century. Report of the International Conference. WHO/PPE/PAC/97.6*. Geneva: WHO
World Health Organization (WHO) (1998). *Health-for-All in the 21st Century*. A/51. Geneva: WHO.
World Health Organization (Regional Office for Europe) (EURO) (1999). *Health Impact Assessment: Main Concepts and Suggested Approach. Gothenburg Consensus Paper*. Brussels:

WHO.
World Health Organization (WHO) (2004). Public health, social and environmental determinants of health (PHE), air quality deteriorating in many of the world's cities. [Online] Available at: https://www.who.int/phe/health_topics/outdoorair/en/
World Health Organization (Regional Office for Europe) (EURO) (2009). *Health and Strategic Environmental Assessment*. WHO Consultation Meeting, Rome, Italy, 8–9 June.
World Health Organization (WHO) (2016). *Preventing Disease Through Healthy Environments*. Geneva: WHO.
World Health Organization (WHO) (n.d.). *Healthy Environments for Children Alliance*. [Online] Available at: http://www.who.int/heca/
World Health Organization and UN-Habitat (2010). *Hidden Cities: Unmasking and Overcoming Health Inequities in Urban Settings*. Geneva: The WHO Centre for Health Development, Kobe, and United Nations Human Settlements Programme.
World Health Organization and United Nations Children's Fund (2017). *Progress on Drinking Water, Sanitation and Hygiene*. New York: WHO/UNICEF.
Wright, J., Parry, J., and Mathers, J. (2005). Participation in health impact assessment: objectives, methods and core values. *Bulletin of the World Health Organization*, 83(1), 58–63.
Yach, D. (1996). Renewal of the Health-for-All strategy. *World Health Forum*, **17**, 321–31.

11.7

Strategies and structures for public health interventions

Sian Griffiths and Kevin A. Fenton

Introduction

The mantra that public health is everybody's business can be two edged. On the one hand, it implies wide support for interventions which would improve health. On the other, it implies such wide responsibility that no one makes public health interventions their priority. Thus, strategies for improving public health and structures for their implementation are complex and situation dependent. It is not always possible to translate a successful programme in city A to one in city B, country C to country D, or urban area E to rural area F (Rao et al. 2007). However, the knowledge base and evidence for effective interventions are growing and the appreciation of the critical role of cross-sectoral and multilevel approaches is deepening.

The use of evidence underscores the breadth of interventions which improve health. For example, the strongest evidence among the 43 most effective interventions identified for improving dietary habits, increasing physical activity, and reducing smoking were economic incentives to make healthier foods more affordable; improved safer roadsides and land-use design to encourage physical activity as well as smoking bans (American Heart Association 2012).

Thus, public health interventions require a wide range of strategies to translate policies into effective action plans (Kickbusch and Buckett 2010) and structures for their implementation if they are to be effective at producing population-level improvements in health.

Strategies may be global agreements, national plans, or local government or community agreements for action (Department of Health 2011a). To be implemented, they require structures which are both directly and indirectly concerned with health outcomes, evidence-based processes, and outcome

measures of their effectiveness. Strategies may directly involve the health services which make contributions both to improving health as well as to managing disease but, more often than not, strategies are multisectoral, address the broader determinants of health, and involve an ecological perspective which engages politicians, the corporate sector, civil society, local communities, and the public. This chapter will discuss the concept of strategic approaches for improving the public's health and give examples of health-promoting strategies at global, national, local, and individual levels.

Why a strategic approach?

The purpose of public health is to promote and protect health and prevent disease and injury. Improving the health and well-being of communities is a difficult process which takes time and involvement from a range of stakeholders including the public and communities. Having robust strategies to guide public health interventions is critical to guide the improvement of health of the public, to ensure effective use of resources and appropriate investments and to support successful implementation of plans in order that governments and organizations are best able to promote and protect health and prevent disease and injury.

The need for an increasing focus on effective strategies to guide public health interventions is welcome and has been a journey of continuous improvement over the past three decades. This has in part been driven by the understanding of the need to address the systemic causes of ill-health in a more coordinated, resource effective way, which delivers intended results and maintains trust among stakeholders and the community (WHO 2010).

Today, governments recognize that they have the responsibility to translate these calls for change into national health policies, strategies, and plans (WHO 2010). At the same time, they understand the complex systems in which we operate mean that plans must be fit for purpose, inclusive, and ideally integrated into existing assets and infrastructures to ensure success. It is now widely understood that national health policies, strategies, and plans have to go beyond healthcare delivery and address the broad public health agenda; and that they have to go beyond the boundaries of health systems, addressing the social determinants of health and the interaction between the health sector and other sectors in society.

In summary, the need for effective strategy to guide public health interventions is increasingly being driven by a range of other factors including:

Improvements in the understanding of the role of good governance, strong national capacities, and transparency to reap the full benefits of global, regional, and national cooperation and intelligence
A greater focus on accountability and transparency
A realistic assessment of current capacities and how bold vision for the future may be achieved through improved collaboration, coordination, and utilization of assets
The need to harmonize and align financial and technical health sector investments to reduce fragmentation and duplication and maximize health impact.

The need for strategy

Health is created by a complex variety of interacting influences, reflected in the World Health Organization's (WHO's) definition which not only highlights the physical and psychological aspects of good health, but also its socioeconomic and environmental dimensions (WHO 1948). More recently global leadership initiatives on Planetary Health (Whitmee et al. 2015), World Health Organization One Health (WHO 2017b), and the United National Sustainable Development Goals (United Nations n.d. c) provide examples of the global mobilization, engagement, and the strategic approach needed to tackle emerging and urgent public health threats.

Using England as an example, the government's 2010 public health strategy, *Healthy Lives, Healthy People* (HM Government 2010) underscored its commitment to protecting and improving health and reducing health inequalities. It recognized that although people are generally living longer, the overall burden of chronic physical and mental ill-health is increasing, and the poor are worse off than the rich. The strategy was evidence-based and rooted in an approach that engaged and integrated the efforts of national and local government, the National Health Service (NHS), charities, schools, higher education, voluntary groups, businesses, and employers. The strategy accepted a place for government leadership and, albeit when other approaches have failed, for regulation. However, it took

the view that the balance of responsibility and action should shift from central government to local communities, and that people should be ‘nudged’ towards taking on more responsibility for their health by living more healthily. A cabinet subcommittee was established to engage relevant government departments, and the strategy includes an encouraging list of related initiatives.

By 2013, this strategy was updated in the government’s Living Well for Longer: A call to action to reduce avoidable premature mortality strategy (DH 2013) recognizing that if we are to tackle the challenge of rising non-communicable diseases, improvements are needed across the three domains of prevention, early diagnosis, and treatment. Two-thirds of deaths under 75 are avoidable, and around 80% of those are preventable, largely through improved public health. This new strategic approach recognized changes within the health and care system in England where local government has become the major force for improvements in public health as it takes on responsibility for health improvement. This complements local responsibilities for tackling the wider determinants of health, such as air quality, road deaths, transport, noise, violence, housing, fuel poverty, and use of outdoor space. Bringing together lead responsibility for preventative action across this wide spectrum provides a unified approach to tackling the causes of ill-health. The updated strategy also stresses the importance of maximizing opportunities across primary care to reduce premature mortality, as local government roles span prevention, early diagnosis, and treatment. A strategic approach to strengthening primary care was seen as paramount for improving rates of early diagnosis, ensuring that key risk factors, diagnosing illness, and long-term conditions were done as early as possible, along with proper case management of people’s conditions and risks with appropriate and timely referral to secondary care for further treatment.

Strategic development is dynamic and linked to policy. For example, in China, among the first challenges faced by the new People’s Republic of China in 1949 was the need to improve the general health and hygiene of the population (Dai and Jiang 2013). Strategies for tackling inadequate sanitation, communicable diseases, and providing basic care at the community level for mothers and children led to the creation of the ‘barefoot doctor system’, a forerunner for many current models of community-based primary care. In the latter part of the twentieth century, China has become one of the largest forces in the world economy but during the ‘opening up’,

the lack of strategy for the health services meant that market forces dominated, resulting in greater inequity in health and decimation of the public health system. Health was not a priority and economic strategies held sway. As a consequence, the first decade of the twenty-first century brought the urgent need to bridge the gap between China's world-class economy and its poor health service provision in which many people were unable to afford basic care. The inequities between the affluent East Coast and urban areas have grown and China now faces not only the problems of an increasingly ageing population and the growing challenge from non-communicable diseases (NCDs), but the poorer, less-developed parts of the country still remain in greater need of infectious disease surveillance, improvements in maternal health, and implementation of basic primary care programmes, in addition to the challenge of a growing burden of NCDs. The massive movements of migrants from rural to urban communities, estimated to be around 200 million in 2012, add extra challenges for public health (Daar et al. 2007; Jin et al. 2013; Mohammed n.d.). At the national level, in addition to hospital service reform, new policies placing greater emphasis on prevention, a stronger public health system, and new strategies to improve primary care are in development to reduce the growing health burdens resulting from rapid urbanization, growing consumption of tobacco, and the increasing rates of lifestyle-related NCDs (Tan et al. 2018).

The experience of China demonstrates the need for national strategy to be interpreted at local levels according to the needs of the population and the resources available. The wide variations in health status, as well as access to care, underline that the delivery of strategy relies not only on national strategic intent, but on appropriate infrastructures and resources at the local level. This was demonstrated in 2003 by the severe acute respiratory syndrome (SARS) epidemic in Hong Kong.

CASE STUDY 11.7.1 Lessons from SARS in Hong Kong

Among the global lessons learned from the experience of SARS in 2003 was that better strategies and structures were needed across all sectors of the health and political systems to respond to the threat of new emerging infectious diseases and to the possibility of a pandemic. When SARS hit Hong Kong in 2003, its surveillance systems were focused on the possibility of an H5N1 (avian flu) pandemic and influenza-like illnesses, and the authorities were monitoring the hospital system. Local activity on the Hong Kong borders, such as shop shelves empty of white vinegar boiled by local residents to sanitize and protect their homes from a rumoured killer disease over the border, went unremarked. There was no formal exchange of accurate information about this atypical pneumonia outbreak in the Guangdong Province available to Hong Kong or to the

international community. Only when a group of healthcare workers and medical students in Hong Kong were admitted to one of the largest teaching hospitals with an unidentified respiratory illness within a few days in early March was an alarm triggered in the healthcare community (Hong Kong College of Community Medicine n.d.; Hong Kong Academy of Medicine n.d.).

While the story of the spread of the disease to the local community and to the rest of the world is now well known, the Independent Inquiry commissioned by the Hong Kong government (Hong Kong Special Administrative Region Government 2003) provided the opportunity to review what strategies had been in place and how well they had worked. The Inquiry found that while the healthcare staff could be commended for the care they gave the patients, the healthcare system itself had structural deficits and the strategy for action was inadequate. The ability to rapidly legislate to designate the new disease as notifiable was not in place, and interventions were hindered by outdated legislation. The organization of health protection functions, hospital management, and the relationships with central government all needed review. Better collaboration and communication outside Hong Kong, within Hong Kong, and within the healthcare sector was also needed, and it was recognized that to manage workforce capacity for future epidemics, training, and research were all needed. These findings had much in common with the Canadian experience (the country hardest hit by SARS outside of Asia), where it was found that amid the commendable response by public health and healthcare workers to contain SARS, there were still many systematic deficiencies to adequately address outbreak containment, surveillance, information management, and infection control (Naylor et al. 2004).

In response to the SARS epidemic, significant progress to strengthen structures and strategies to combat emerging infectious diseases have been put in place by both affected and unaffected jurisdictions. In Hong Kong, as in many other countries, surveillance information is now available through a real-time system of reporting of data not only across the Pearl River Delta but globally through the network coordinated by the WHO. Active networks not only of government officials responsible for resilience, but also of researchers, are in regular communication. Communication between the hospital and the community has been improved by streamlined systems and the creation of geographical coterminosity between catchment areas. Extensive plans are now in place for all sectors and staff to be trained through ‘drills’ which are intersectoral, and which have clear leadership and chains of command and control. The public and the media are better informed about communicable diseases and about preventive measures, such as hand washing. Efforts are continually being made to ensure that communication is open and transparent with press releases and letters to medical staff to keep them abreast of emergent diseases.

The new improved infrastructure was put to the test in 2008 with the threat of the H1N1 pandemic (Cowling et al. 2010). Unlike the SARS experience of 2003, the media and politicians worked together to ensure public health messages about H1N1 were clear and well communicated to the public, based on up-to-date information regularly provided by the Centre for Health Protection (CHP). The CHP, which was established post-SARS, was able to play its part in the global network of surveillance for communicable diseases acting as the central structure to coordinate local responses, and ensuing global collaboration and communication (Hong Kong Government n.d.).

The experiences of SARS and H1N1, and more recently H7N9 in Asia and the Middle East respiratory syndrome coronavirus (MERS-CoV), have demonstrated the need for cross-sectoral strategy, a clear plan for proposed implementation, as well as identified structures to support its implementation. The experience of H1N1 demonstrated that both strategy and structures had improved with the lessons learned from SARS and, in 2013, the flow of information about H7N9 to the press, various professions, and the public is significantly better, as is the support from the global scientific community.

CASE STUDY 11.7.2 Lessons from Ebola in Congo (WHO 2018)

On 8 May 2018, following the provisions of the International Health Regulations, the Ministry of Health of the Democratic Republic of Congo notified WHO of two confirmed cases of Ebola virus disease (EVD) in the Bikoro health district in Equateur province. The affected areas shared a border with the Republic of Congo. This is the ninth Ebola outbreak in Democratic Republic of the Congo, but the first in the province of Equateur.

The national and local authorities and their partners acted quickly to respond to this outbreak. A National Epidemic Response Coordination Committee was established to coordinate the interventions and partners' efforts following the official declaration of the epidemic as a public health emergency by the Ministry of Health (announced on 8 May 2018). A joint response plan for all partners to support the Ministry of Health was developed. This initial plan covered a 3-month period expiring at the end of July 2018. After this first stage, a consolidation and stabilization phase was necessary to develop resilience at the local and national levels.

By summer 2018, 54 cases were reported, including 38 confirmed cases and 16 probable cases. The last confirmed case of EVD in Equateur Province was reported on 6 June 2018. The last confirmed case admitted to the Ebola treatment centre was discharged on 12 June 2018 and monitoring of all registered contacts was discontinued on 27 June 2018.

A national strategic response plan was developed to guide the work of all stakeholders involved in the response around the Ministry of Health. The first phase of the EVD response was to protect the towns of Bikoro and Mbandaka against an increase in cases that could potentially threaten the country's major cities as well as neighbouring countries along the river. The strategy adopted by the Ministry of Health was to bring laboratory diagnosis of EVD closer to the populations affected by the epidemic. Infection prevention and control measures were strengthened in major hospitals and other health facilities.

Defining strategy

Strategy can be seen as a means of mobilizing an organization to get to its future. It is about giving an organization purpose and direction, motivating trustees, staff, and volunteers to engage meaningfully with the external environment—linking in with funders and beneficiaries (WHO 2010).

Strategy is about determining how an organization will match its diverse capabilities with the needs of stakeholders, now and in the future (in order to maximize impact). An organization's strategy will determine how it relates to the outside world—how it makes the most of opportunities, responds to change, and faces up to threats: opportunities to innovate, improvise, and improve.

A strategy makes it clear what the priorities are, and what the organization's resources should be concentrated on delivering. While most strategies focus on the medium to long term (3 to 5 years), they are often reliant on the short term too, since one cannot often predict when and how

things will change. Consequently, strategic plans need to be flexible and agile. Strategies are often developed ‘top-down’ but increasingly organizations seek to mobilize and motivate everyone in the organization and their strategic partners and key stakeholders by early involvement. Strategy will emerge from the grassroots of an organization—finding ways of creatively integrating top-down from bottom-up is vital.

Experience shows that the process for building comprehensive national health policies, strategies, and plans is as much a political process as a technical one. It requires attention for the inspirational side of defining vision and policy; it also requires attention for operational detail (WHO 2010). The balance between vision, policy, strategic planning, and attention for implementation varies considerably from country to country and between national, regional, and local levels. Given the overarching importance of context, blueprint approaches are unlikely to be of much assistance. Nevertheless, there are ways of going about policy dialogue that are more likely to produce robust policies, strategies, and plans.

Strategies are a necessary step in the translation of policies into plans and action. They create the bridge between policy and practice. Mintzberg (1987) has described five aspects of strategies:

Plan—with purposeful intended action.
Ploy—competitive manoeuvring.
Pattern—signalling a plan in a series of actions.
Position—relevant to the outside world.
Perspective—an organizational personality.

Translating the aspects of strategy into public health practice requires certain features:

Focus on populations rather than individuals.
Address the social determinants of health and their impact on poverty.
Base on values of social justice and equity.
Reflect a global perspective.
Be aware of the dynamic relationship between the environment and health.
Place priority on prevention rather than curative care.
Understand the importance of systems and structures.
Take a transdisciplinary and multisectoral approach.

Ensure decision-making is based on:

Vital statistics (population-based data).
Data, scientific evidence, and good research.
Surveillance and outbreak investigation.
Be committed to advocacy and in tune with politics.

Strategic plans may relate to the health status or needs of a defined group of people or of the population as a whole, and may take various forms, such as laws, rules and regulations, guidelines, recommendations, and mandates (Kawamoto 2004).

Health strategy and health service strategy are both important for delivering better public health interventions. Health services are responsible for dealing with disease, and they do not control the major forces which cause disease. They can, however, play a significant role in prevention both at institutional and individual levels through the services they provide and the institutional influences they exert. Health strategy is, however, conceptually broader and requires engagement at different population levels with a wide variety of forces which impact on health—social, economic, political—and with a wide range of multisectoral stakeholders.

CASE STUDY 11.7.3 The Sustainable Development Goals (WHO 2015)

The 2030 Sustainable Development Agenda is of unprecedented scope and ambition, applicable to all countries, and goes well beyond the previous Millennium Development Goals (MDGs), which came to the end of their term (United Nations n.d. a; United Nations n.d. b), and a post-2015 agenda, comprising 17 Sustainable Development Goals (SDGs), takes their place. While the MDGs were concerned with low and middle-income countries, SDGs are global in their aspiration. Poverty eradication, health, education, and food security, and nutrition remain priorities, but the SDGs importantly comprise a broad range of economic, social, and environmental objectives, and offer the prospect of more peaceful and inclusive societies (WHO n.d. a). Notable among the remaining challenges include acute epidemic diseases, disasters, and conflict situations, the burgeoning epidemic of NCDs and mental health disorders and large inequalities in all parts of the world.

The SDGs address many of these issues head on, setting a new health goal (‘Ensure healthy lives and promote well-being for all at all ages’) with a broad set of targets (WHO 2015). The SDGs also significantly broaden the scope of action with 17 goals and 169 targets, including one specific goal for health with 13 targets, of the new development agenda integrating the three dimensions of sustainable development around people, planet, prosperity, peace, and partnership. The health goal is broad: ‘Ensure healthy lives and promote well-being for all at all ages’. Health has a central place as a major contributor to and beneficiary of sustainable development policies. There are many linkages between the health goal and other goals and targets, reflecting the

integrated approach that is underpinning the SDGs. Universal health coverage (UHC), one of the 13 health goal targets, provides an overall framework for the implementation of a broad and ambitious health agenda in all countries (NHS n.d.).

Fundamental to achieving the SDGs will be the recognition that eradicating poverty and inequality, creating inclusive economic growth, preserving the planet, and improving population health are not just linked but interdependent. This has profound implications for development strategy, many of which will only be revealed as we move forward. But some seem fairly clear at the outset, including the fact that silo-based, vertical approaches to development will have to give way to broader, cross-cutting approaches coordinated around and aligned with countries' needs and priorities.

By taking a more integrated approach to development the SDGs present an opportunity to advance, seizing opportunities to collaborate and exploit synergies. However, there are clearly challenges too. For example, in order to take on cross-cutting issues, it will be necessary to achieve far greater intersectoral coherence and coordination of effort. To achieve that end, a new impetus will have to be given to global and regional partnerships and collaborations, with a focus on country action.

Public health strategy: from global to local

Global public health strategies

Strategies may address needs at multiple population levels and be influenced not only by local factors, but also by global policies, linking both vertically and horizontally. Vertical strategies may be concerned with single programmes, such as those addressing HIV/AIDS in sub-Saharan Africa, whereas horizontal strategies may focus on addressing a wider range of influences such as those creating or alleviating poverty.

Social determinants of health

Poverty and poor health are inextricably linked and strategies to reduce poverty are essential for improving health opportunity. Addressing the root causes of poverty while promoting greater equity are key strategic objectives underpinning the work of the WHO Commission on Social Determinants of Health (2008), an influential global think tank which has raised the profile of the debate on the need for strategic action across four main dimensions—political, economic, social, and cultural.

> Health is a universal aspiration and a basic human need. The development of society, rich or poor, can be judged by the quality of its population's health, how fairly health is distributed across the social spectrum, and the degree of protection provided from disadvantage due to ill-health. Health equity is central to this premise. (Marmot 2007)

Proposed strategies to reduce health inequity include:

Improving the conditions of daily life at home and at work, for women and men, in rural and urban settings, and for all ages.
Addressing the inequitable distribution of power and money, forces that shape the conditions of daily life.
Acting based on knowledge, a professional workforce, alliances with civil society, and evaluation.

The Commission did not stop with making these strategic statements, but has backed them up with a series of papers which provide examples of good practice to support their implementation.

Health promotion

The work of the Commission can be contextualized within a framework of strategic declarations from the WHO, particularly from the Declaration of Alma-Ata (WHO 1978). This declaration, and subsequent iterations, echo public health's historic roots by reflecting the importance of the socio-environmental conditions in which people live and the need for community-based action to improve health. Emphasis is placed on the importance of interactive roles of government, communities, and individuals alongside the healthcare sector to promote good health and prevent disease. While the health sector has a role in promoting health through the services it provides, its role beyond acute medical care is significant with potential for involvement with upstream activity to prevent ill-health in the entire population. This was proposed in the Declaration of Alma-Ata, which called for 'a comprehensive health strategy that not only provided health services but also addressed the underlying social economic and political causes of poor health'. De facto effective prevention would rely on working with, and across, other sectors and, in particular, promoting joined-up government at all its levels including greater community engagement through better primary care.

The conceptual framework of Alma-Ata was further developed by the Ottawa Charter (WHO 1986), which promoted the concept of healthy public policy rather than focusing only on health policy, extending consideration of

health to all policy environments, including the concept of health impact assessment.

The Ottawa Charter defined five key areas for strategic public health action:

Building healthy public policy.
Creating supportive environments.
Strengthening community action.
Developing personal skills.
Reorienting health services towards prevention and a holistic approach (WHO 1986).

The Ottawa Charter stated unequivocally that health is created in the context of everyday life where people live, love, work, and play. It expanded the concept of health determinants to include environmental challenges and people's empowerment. It included not only the basic well-described determinants but broadened the strategic thinking to the need to combine diverse but complementary approaches in order to promote lifestyles conducive to health. These now include, in particular, legislation, fiscal measures, taxation, and organizational change

Subsequently, in 2005, the Bangkok Charter (WHO 2005) further broadened the notion of 'Health in All Policies' (HiAP) by highlighting the need to work across the corporate sector and at a global level, stating that the promotion of health was:

Central to the global development agenda.
A core responsibility for all governments.
A key focus of communities and civil society.
A requirement for good corporate practice.

Underpinning concepts included those of social justice, stressing the importance of the values of gender and health equity, respect for diversity and human dignity, and the search for peace and security.

CASE STUDY 11.7.4 The Shanghai Declaration on Promoting Health (2016)

More recently the Shanghai Declaration on Promoting Health in the 2030 Agenda for Sustainable Development the outcome of the Ninth Global Conference on Health Promotion (Shanghai, 21 to

24 November 2016), which was jointly organized by the Government of China and WHO, under the auspices of a Scientific Advisory Group and Conference Organizing Committee. The concise Shanghai Declaration on Health Promotion was endorsed by the participants of the Conference. The declaration recognizes that that health and well-being are essential to achieving sustainable development and are essential to achieving the United Nations Development Agenda 2030 and its SDGs.

Key among the declaration was the reaffirmation that health remains a universal right, an essential resource for everyday living, a shared social goal, and a political priority for all countries. The United Nations (UN) SDGs establish a duty to invest in health, ensure UHC and reduce health inequities for people of all ages. The declaration committed to promoting health through action on all the SDGs and by engaging the whole of society in the health development process. This means acting decisively on all determinants of health, empowering people to increase control over their health and ensuring people-centred health systems.

The declaration also focused on tackling health inequalities, recognizing that this requires political action across many different sectors and regions. They also require global collective action. The importance of prioritizing good governance, local action through cities and communities, and people's empowerment by promoting health literacy was also stressed along with giving a high priority on innovation and development to support people's enjoyment of a healthy life and give precedence to the health of the most vulnerable (WHO n.d. e).

Finally, the declaration recognized that governments have a fundamental responsibility at national, local, and global level to address the damaging effects of unsustainable production and consumption. This includes offsetting economic policies that create unemployment and unsafe working conditions, and enable marketing, investment, and trade that compromise health. The declaration called on business leaders to demonstrate good corporate governance—profit must not stand above people's health. This is of particular concern in fighting the NCD epidemic.

Tackling non-communicable diseases

NCDs are a long-standing, progressive, and costly threat to healthy populations. Diseases such as diabetes are associated not only with the problems of diminished functional capacity, increased physical and psychosocial morbidity, comorbidities, and complications, but also lead to higher utilization of health services and higher medical costs, as well as opportunity costs that are associated with self-care and disability and shorter life expectancy than those without diabetes (Diabetes UK n.d.). Today, NCDs kill 40 million annually, more than any other cause of death. Of these deaths, 15 million occur prematurely among people aged 30–70 years, and 7 million in low- and low-middle-income countries. Low- and middle-income countries are particularly affected by NCDs with more than 80% of all deaths from NCDs occurring in these countries. Violence and injuries take an overwhelming toll on young people, particularly boys. In 2015, world leaders committed to reduce premature deaths from NCDs by one-third by

2030 as part of the SDGs. Recent WHO reports indicate that the world will struggle to meet that target.

The WHO continues to build on the Global Strategy for the Prevention and Control of NCDs (WHO 2013a), which was first adopted in 2000, and has been followed by strategies to address tobacco control, and lifestyle factors such as diet, exercise, and alcohol consumption. These have so far materialized into the Physical Activity and Health campaign (WHO 2004), the Reduction of Harmful Use of Alcohol campaign (WHO n.d. b), as well as an Action Plan for the Global Strategy for the Prevention and Control of NCDs for 2008–2013 (WHO 2009a). This set of actions also called for high-level political commitment and the concerted involvement of governments, communities, and healthcare providers for successful implementation. The 66th World Health Assembly (WHA) in 2013 concluded with Member States adopting an 'omnibus resolution' on NCDs. This resolution combined major decisions and recommendations in a single, comprehensive resolution, including:

Adopting the global monitoring framework (GMF), including nine targets and 25 indicators.
Endorsing the Global NCD Action Plan 2013–2020 (GAP).
Agreeing to establish a global coordination mechanism (GCM) for NCDs, and to develop terms of reference by November 2013.
Agreeing to establish a set of indicators to measure the implementation of the GAP (The NCD Alliance 2013a, 2013b).

The GAP lays out the vision of a world free of the avoidable burden of NCDs with the goal of reducing the preventable and avoidable burden of morbidity, mortality, and disability due to non-communicable diseases. The means of achievement will be through multisectoral collaboration and cooperation at national, regional, and global levels. The objective is for populations to reach the highest attainable standards of health and productivity at every age so that NCDs are no longer a barrier to well-being or socioeconomic development (WHO 2013b). The principles adopted should be:

Life-course approach
Empowerment of people and communities
Evidence-based strategies
Universal health coverage
Management of real, perceived, or potential conflicts of interest
Human rights approach
Equity-based approach
National action and international cooperation and solidarity
Multisectoral action (WHO 2013b)

The objectives of the plan are to raise the priority given to the prevention and control of NCDs at all levels, through strengthened international cooperation and advocacy; to strengthen national capacity, leadership, governance, multisectoral action, and partnerships to reduce modifiable risk factors and underlying social determinants through creation of health-promoting environments; and to strengthen and orient health systems to people-centred primary healthcare and UHC. This should be underpinned by promoting and supporting national capacity for high-quality research and development as well as monitoring the trends and determinants of NCDs to evaluate progress in their prevention and control.

Progress can be measured by adopting voluntary global targets (Box 11.7.1). By creating a framework and setting targets, emphasis is shifting from analysis of the problem to implementing changes that can make a difference, monitored through a set of measurable outcomes.

Box 11.7.1 Voluntary global targets

- A 25% relative reduction in risk of premature mortality from cardiovascular diseases, cancer, diabetes, or chronic respiratory diseases.
- At least 10% relative reduction in the harmful use of alcohol, as appropriate, within the national context.
- A 10% relative reduction in prevalence of insufficient physical activity.
- A 30% relative reduction in mean population intake of salt/sodium.
- A 30% relative reduction in prevalence of current tobacco use in persons aged 15+ years.
- A 25% relative reduction in the prevalence of raised blood pressure or contain the prevalence of raised blood pressure, according to national circumstances.
- Halt the rise in diabetes and obesity.
- At least 50% of eligible people receive drug therapy and counselling (including glycaemic control) to prevent heart attacks and strokes.
- An 80% availability of the affordable basic technologies and essential medicines, including generics, required to treat major NCD in both public and private facilities.

Reproduced with permission from World Health Organization (WHO), Noncommunicable diseases and mental health: about 9 voluntary global targets, Copyright © WHO 2014, available from: http://www.who.int/nmh/ncd-tools/definition-targets/en/

In 2017, Heads of State and Government and ministers from around the world re-committed to new and bold action to reduce suffering and death from NCDs, primarily heart and lung diseases, cancers and diabetes, the world's leading killers. Governments endorsed the *Montevideo Roadmap 2018–2030 on NCDs as a Sustainable Development Priority* (WHO 2017a) at the opening of the three-day Global Conference on Noncommunicable Diseases in Montevideo, hosted by WHO and the Presidency of Uruguay. The pledge followed agreement by world leaders to reduce 'premature' deaths from NCDs by one-third by 2030 as part of the United Nations' Agenda for Sustainable Development.

The *Montevideo Roadmap* (WHO 2017a) highlights the need for coordinated and coherent action from all sectors and the whole of society, as many of the main drivers of ill-health lie outside the control of health ministries, systems, and professionals. Non-state actors, including civil society and industry, have important roles to play. It also points out that the bulk of NCD deaths could have been prevented by action against tobacco, air pollution, unhealthy diets, physical inactivity, and harmful use of alcohol —as well as by improved disease detection and treatment. The roadmap will guide global preparations for the 2018 United Nations General Assembly Third High-level Meeting on NCDs which will assess countries' progress in meeting the target of reducing premature NCD deaths by 25% by 2025 and then by a third by 2030 (United Nations News Centre 2011). Finally, the plan identifies a range of challenges, including: uneven and insufficient

progress to reduce premature deaths from NCDs; influence of the private sector on governments to prioritize trade over public health goals; and lack of high-level political leadership to ensure that health promotion and NCD prevention and control are part of all areas of government policy.

Regional public health strategies

European region

The European Union (EU) health strategy creates a framework within which jurisdictions take forward their own strategies tailored to national needs. It raises not only the aspirations but also the need for effective implementation and for monitoring progress. Promoting good health is an integral part of Europe 2020, the EU's 10-year economic growth strategy which emphases smart, sustainable, and inclusive growth as way to overcome the structural weaknesses in Europe's economy, improve its competitiveness and productivity, and underpin a sustainable social market economy (Lefebvre and Flora 1988; Evans 2006; World Health Organization Europe Office n.d.). More specifically, health policy is important to Europe 2020's objectives for *smart* and *inclusive* growth because:

keeping people healthy and active for longer has a positive impact on productivity and competitiveness;
innovation can help make the healthcare sector more sustainable and find new cures for health conditions;
the healthcare sector has an important role to play in improving skills and creating jobs as it employs 1 in 10 of the most qualified workers in the EU;
with a projected 45% increase in the number of people aged 65 and over in the next 20 years, financing rising healthcare costs and access to a dignified and independent life for the ageing population will be central to the political debate.

Of the 7 EU2020 *flagship initiatives*, the four most relevant to public health are: Innovation Union; Digital agenda for Europe; Agenda for new skills and jobs; and the European platform against poverty (European Portal for Action on Health Inequalities n.d.).

CASE STUDY 11.7.5 Health in All Policies in the European Union (EU)

Health in All Policies has been adopted by the EU as part of its strategic approach seeking to complement national action on health. The EU health policies aim to: Protect people from health threats and disease; Promote healthy lifestyles; and Help national authorities in the EU cooperate on health issues (European Commission 2007; Leppo et al. 2013).

HiAP is one of the four principles adopted by the EU. The three further principles are:

A strategy based on shared health values: this principle places particular emphasis on reducing poverty and inequities and on the importance of indicators of comparable health data at all levels. Further work on how to reduce inequities in health and how to promote health literacy programmes for different age groups is part of the strategic approach.

Health is the greatest wealth: highlights that spending on health is not just a cost, it is an investment and that although health expenditure may be regarded as an economic burden, the real costs to society are the direct and indirect costs linked to ill-health. The report estimates that the annual economic burden of coronary heart disease can amount to 1% of the gross domestic product (GDP), and the costs of mental disorders to 3–4% of the GDP. It emphasizes that with 97% of spending going on acute care healthcare and treatment, there is an underinvestment in prevention and improving of the population's overall physical and mental health.

Strengthening the EU's voice in global health: the EU's contribution to global health requires interaction of policy areas, such as health, development cooperation, external action, research, and trade linking with the WHO and other relevant UN agencies, the World Bank, the International Labour Organization, the Organisation for Economic Cooperation (OECD), and the Council of Europe, as well as other strategic partners and countries (World Bank 2005, 2011, 2013).

The strategy has three objectives related to its policy priorities, which together culminate in an overarching framework for action at all levels across the population:

Fostering good health in an ageing Europe: promoting health and preventing disease requires efforts throughout the lifespan to tackle key issues including poor nutrition; physical activity; alcohol, drugs, and tobacco consumption; environmental risks; traffic accidents; and accidents in the home. Improving the health of children, adults of working age, and older people will help create a healthy, productive population and support healthy ageing now and in the future.

Protecting citizens from health threats: emphasizing the need for surveillance and mechanisms to respond to health threats, including emerging infections, as well as adaptation to climate change (The Lancet 2010).

Supporting dynamic health systems and new technologies: e-health, genomics, and biotechnologies can improve prevention of illness, delivery of treatment, and support a shift from hospital care to prevention and primary care. New technologies must be evaluated properly, including for cost-effectiveness and equity, and health professionals' training and capacity implications must be considered.

Africa region

Another example of a continent-wide strategy addressing different health needs comes from Africa where significant numbers of low-income countries face different challenges associated with poverty and infectious diseases. The strategy produced by the African Union Ministers of Health

was structured using the framework of Vision and Mission, Principles, Goals, and Objectives (NEPAD n.d.). Its goal is to contribute to Africa's socioeconomic development by improving the health of its people and by ensuring access to essential healthcare for all Africans, especially the poorest and most marginalized. The vision, expressed as working towards 'an integrated and prosperous Africa free of its heavy burden of disease, disability and premature death', was to build an effective, African-driven response to reduce the burden of disease and disability through strengthened health systems, scaled-up health interventions, intersectoral action, and empowered communities.

The strategic principles include health as a human right and a developmental concern requiring a multisectoral response, equity in healthcare as a foundation for all health systems, and effectiveness and efficiency as central to realizing the maximum benefits from available resources. Evidence is proposed as the basis for sound public health policy and practice and more of the same is not enough: 'new initiatives will endeavour to set standards that go beyond those set previously' (NEPAD n.d.). This has been followed more recently with the creation of the Africa Centres for Disease Control as a key driver for regional capacity building, collaboration, and response.

CASE STUDY 11.7.6 Strategic planning by the Africa Centres for Disease Control

More recently, the Africa Centres for Disease Control (CDC) was established and its statute approved by the 26th Assembly of the Heads of State and Government in January 2016 and officially launched on 31st January 2017 in Addis Ababa, Ethiopia (Africa Union and Africa CDC). The Africa CDC Governing Statute anchors the institution in the following guiding principles: Leadership, Credibility, Ownership, Delegated authority, Timely dissemination of information, Transparency, Accountability, and Value-addition. The Africa CDC endeavours to support Member States in improving the health of their people through building the capacity of public health institutes to focus on prevention of infection, surveillance and response to emergencies (including outbreaks, man-made and natural disasters, and public health events of regional and international concern), and building capacity to reduce disease burden on the continent. Building on these guiding principles and mission, the Africa CDC has developed a 5-year strategic plan (2017—2021). This strategic plan describes five strategic pillars, which will help the Africa CDC to achieve its mission and vision. These are: Surveillance & Disease Intelligence; Information Systems; Laboratory Systems and Networks; Emergency Preparedness and Response; and Public Health Research.

From these strategic pillars, several strategic objectives are derived with a defined set of activities and performance measures. Functionality and implementation of these pillars will require the support of cross-cutting enablers (workforce, partnership, innovation, finance and leadership and management) that provide the backbone for the operations of the Africa CDC.

National and subnational public health strategies

Canada

Changing Directions, Changing Lives is the first Canadian mental health strategy and was published in 2012 (Mental Health Commission of Canada 2012). The strategy was produced 'to help to ensure that people who experience mental health problems and illnesses—especially those with the most severe and complex ones—are treated with respect and dignity, and enjoy the same rights as all Canadians'. The strategy provides six strategic directions for change to improve mental healthcare (Box 11.7.2). Change will not be possible without a whole-of-government approach to mental health policy.

Box 11.7.2 Six strategic directions for changes to improve mental health and the mental healthcare system

1. To promote mental health across the lifespan in homes, schools, and workplaces, and prevent mental illness and suicide wherever possible.
2. To foster recovery and well-being for people of all ages living with mental health problems and illnesses, and uphold their rights.
3. To provide access to the right combination of services, treatments, and supports, when and where people need them. A full range of services, treatments, and supports includes primary healthcare, community-based, and specialized mental health services, peer support, and supported housing, education, and employment.
4. To reduce disparities in risk factors and access to mental health services, and strengthen the response to the needs of diverse communities and Northerners.
5. To work with First Nations, Inuit, and Métis to address their mental health needs, acknowledging their distinct circumstances, rights, and cultures.
6. To mobilize leadership, improve knowledge, and foster collaboration at all levels.

Reproduced with permission from Mental Health Commission of Canada, Changing Directions, Changing Lives: The Mental Health Strategy for Canada, Copyright © 2012, available from http://strategy.mentalhealthcommission.ca/pdf/strategy-summary-en.pdf.

England

In England, the mental health strategy published in 2011 followed many of the same themes. No Health Without Mental Health: A Cross-Government Mental Health Outcomes Strategy for People of All Ages sets shared objectives to improve people's mental health and well-being and plans to improve services for people with mental health problems. The strategy set out the policy for how the government, working with all sectors of the community and taking a life-course approach, would improve the mental health and well-being of the population and provide people with high-quality services that are equally accessible to all (Department of Health 2011b). The strategy incorporated the key elements of defining a policy, setting out aims and objectives, and proposing goals and how they will be monitored to ensure the highest quality of mental health for people of all ages, and all backgrounds and cultures (HM Government 2011).

More recently England's national strategic approach to mental health was updated in the *Five Year Forward View on Mental Health* report. This independent report of the Mental Health Taskforce (2016) set out the start of a ten year journey for transformation for mental health services, commissioned on behalf of the NHS. In preparing the report, particular attention was placed on ensuring that the lived experience of people with mental health problems were at the heart of the report and its recommendations. Over 20,000 people participated in sharing the changes they wanted to see in order to fulfil their life ambitions and take their places as equal citizens in society. Key priorities were prevention, access, integration, quality, and a positive experience of care.

The updated mental health strategy included a number of recommendations for the six NHS arm's length bodies to achieve the ambition of parity of esteem between mental and physical health for children, young people, adults, and older people. Recommendations were also included that cover areas where wider action is needed, underscoring that as well as access to good quality mental healthcare wherever they are seen in the NHS, another main ambition was to have a decent place to live, a job or good quality relationships in their local communities. Finally, the strategy placed a particular focus on tackling inequalities given that mental health problems disproportionately affect people living in poverty, minority ethnic groups, those who are unemployed, and who already face discrimination. To truly address these issues, the strategy recommended tackling inequalities at local and national level.

Global to national strategy and action

While at a global level the WHO produces strategies to provide direction for change, action to implement change is needed at a national population level. Examples of national strategies for mental health were given earlier in this chapter—each being country specific. Another approach is the proposal for a common framework for national action, an example of which is the Framework Convention on Tobacco Control (FCTC) (WHO n.d. c), which was the first international treaty negotiated under the auspices of the WHO, adopted by the WHO on 21 May 2003 and entered into force on 27 February 2005. It was developed in response to the threats posed by the globalization of the tobacco epidemic and is an evidence-based treaty that reaffirms the right of all people to the highest standard of health. By mid-2013, 176 jurisdictions had signed up to the Convention, including 90% of the world's population.

Support for national-level implementation has been provided by the WHO MPOWER package (WHO n.d. d), which gives strategic advice on six proven evidence-based policies to reduce tobacco-related harm (see also Chapter 9.1). This is a vertical strategic approach, focusing specifically on one risk factor, tobacco smoking. The recommended headline actions are:

Monitor tobacco use and prevention policies
Protect people from tobacco smoke
Offer help to quit tobacco use
Warn about the dangers of tobacco
Enforce bans on tobacco advertising, promotion, and sponsorship
Raise taxes on tobacco. (WHO n.d. d)

Importantly, MPOWER provides examples of successful initiatives which have reduced the burden of tobacco-related harm. For example, legislation is a key element to achieving the MPOWER measures and many countries have actively legislated to stop smoking in public places through the banning of tobacco smoking indoors in environments such as workplaces, shops, restaurants, bars, and residential care homes. In 2013, additional steps have been taken in Australia to introduce plain packaging of cigarettes and are now being studied by other countries who are deciding whether to follow suit (The Lancet 2013a, 2013b).

But while national legislation can provide environments in which tobacco harm is reduced, local strategies are also necessary particularly to engage the health sector.

Local-level action

Engagement of health services at a local level has been shown to be essential for the translation of the objectives of the FCTC. While health promotion in schools about the harms of tobacco in schools plays an important role in raising awareness, local services also need to provide support such as smoking cessation services in special clinics and within routine services in both public and private sectors. Non-government organizations (NGOs), for example, the Action on Smoking and Health (ASH) in the United Kingdom (ASH UK n.d.) and Australia (ASH Australia n.d.), and the Hong Kong Council on Smoking and Health (COSH) (n.d.) provide invaluable support in both leading and assisting the implementation of strategies. In Hong Kong, COSH advocates with government and also provides additional services and support for smokers who wish to quit, such as manning quitlines, organizing local campaigns, and providing smoking cessation information. Smoking cessation community clinics, such as the Hong Kong's Hospital Authorities' Smoking Cessation and Counselling Centres, are supported by other NGOs, as well as the government, and increase the effectiveness of smoking cessation services by providing nicotine replacement therapy and continuity of care. Among smokers who are aware of the dangers of tobacco, three out of four want to quit. Like people dependent on any addictive drug, it is difficult for most tobacco users to quit on their own, and they benefit from help and support to overcome their dependence from a variety of sources including local clinics, support groups, and advice from helplines.

While MPOWER seeks to provide global guidance that can be translated into local action, other strategies will be jurisdiction specific. Typically such strategies describe the epidemiology of the target population, its demography, the distribution of mortality and morbidity, its health needs, and local characteristics. Since the purpose of a strategy is to bring about change, plans for implementation will be included, as well as a framework for monitoring and evaluation, often promoting targets or objectives to be achieved. Many countries have produced strategies for tackling major health priorities, including strategies for addressing the epidemic of NCDs.

Typically, development of a strategic framework at a local level will analyse needs, priorities, and plans for implementation of a policy and engage stakeholders in its implementation. One example of this approach comes from Hong Kong where, in line with WHO and the Western Pacific Regional Action Plan (WPRO) NCD strategies, the Hong Kong Department of Health has developed a strategic framework for NCDs, revised in 2018 (Hong Kong Government 2013). This framework was produced following a series of processes engaging not only with the health sector but more widely, particularly with the NGO sector which plays a significant role in health promotion and service delivery. The first step was to review international literature as well as local epidemiology and health service information. The plan was then drawn up using a set of strategic principles (Box 11.7.3).

To flesh out the framework, the strategy group engaged other arms of government, professional groups, non-government agencies, representatives of the business sector, and the community to agree on a work programme based on promoting the need for prevention and early intervention, and to ensure that any action was to be evidence based, outcome focused, and addressing broader health determinants to reduce the burden of disease related to NCDs. Particular emphasis was placed on informing the population to promote responsibility for their own health, engagement of public and private sectors, engaged professionals, and a sustainable health system (Hong Kong Government 2013).

Implementing strategy

With frameworks and principles for strategy in place, action to address specific health behaviours and challenges needs to be planned and monitored. The development of a strategy can be a case for change, a new service, or a further development of an existing service and can involve major changes or relatively minor changes. The development of strategies within healthcare can be devised at a local or national level, and encompass a different size of area, or focus upon a specific topic or aspect of health. The strategy can be formulated through individual or cross-departmental /agency or cross-government programme. The majority of government strategies are currently developed following a period of consultation with a wide variety of partners.

Executing a strategy is perceived as more difficult than its development and represents a disciplined process or a logical set of connected activities that enables an organization to take a strategy and make it work (Hrebiniak 2005). Hunter has observed that, 'regardless of the precise nature of a policy or strategy, and the support that exists for it, if the means to implement it are either non-existent or inadequate in terms of capacity or capability, or both, then it will count for little' (Hunter 2007:20). There are key elements and activities that are known to affect the resources required for managing successful change efforts:

motivating change: creating readiness for change and overcoming resistance
creating a vision: mission, valued outcomes and conditions, midpoint goals
developing political support: assessing change agent power, identifying key stakeholders, influencing stakeholders
managing the transition: activity planning, commitment planning management structures
sustaining momentum: providing resources, building support system for change agents, developing new competencies and skills, reinforcing new behaviours

CASE STUDY 11.7.7 Implementing an alcohol harm reduction in strategy in Hong Kong

Having identified reducing alcohol harm as a priority for implementing the NCD strategy, the HK Working Group on Injuries and Alcohol Misuse proposed ten recommendations to address alcohol-related harm (Hong Kong Government 2011; Kim et al. 2013):

Strengthen surveillance of alcohol consumption through collecting data from all sectors.
Strengthen surveillance of alcohol-related harm by collecting data from the healthcare sector.
Promote research to understand drinking patterns, particularly among the young, and assessment of impact of measures taken.
Involve all health professionals through education and practice.
Encourage healthcare professionals to identify at-risk drinkers and explore the development of tools to enable this to happen.
Provide secondary, tertiary, and mental health services for those identified with needs.
Working together of government, NGOs, schools, and employers in partnership to maximize efforts.
Develop evidence-based and culture-relevant advice and information that can be disseminated in all sectors.
Strengthen community awareness and actions through publicity, media, and other means.
Review legislation and consider age restrictions for selling alcohol.

The recommendations were then used to establish process measures for monitoring implementation and various partners, such as NGOs and community groups, were involved in promoting the strategy. Healthcare providers were also included to provide health education and care for individuals and to engage with professional groups (Li et al. 2011).
Source: data from Hong Kong Government, Department of Health, Promoting Health in Hong Kong: A Strategic Framework for Prevention and Control of Non-communicable Diseases, available from http://www.change4health.gov.hk/filemanager/common/image/strategic_framework/promoting_health/promoting_health_e.pdf.

Local strategy implementation

The process of strategy development uses the whole range of public health tools and skills. Much of the focus of the work will be on the process of gaining commitment from partners, patients, and the public, and overcoming the barriers they perceive. Successful implementation relies on clarity of the health challenge, key aims, and objectives, as well as structures which support engaging the right people in the community to create change, and ensure that proposed changes are implemented and sustainable. Some key questions guiding local-level action are shown in Table 11.7.1.

Table 11.7.1 Questions guiding local-level action

How do we make the case?	By doing a local needs assessment, and taking into account local and national strategies
What are we aiming to do?	Expressed by clear aims and objectives with indicators and targets
How do we involve patients and the public?	Be clear about why patients and public need to be involved and how to engage them at all stages of the strategy identifying the barriers to change
How do we engage/communicate with partners?	Make sure all those who are affected by the strategy are involved including clinicians and partner organizations. Identifying who will support and who will oppose and how to deal with this; communicate with all involved as the work is taken forward
How can we make change happen?	Securing change is the core of the work and needs to be addressed in the planning stage. This stage should include a description of the actions that are required, and an assessment of the resource implications, securing adequate resources in order for implementation to be possible and appropriate
How do we know we have done what we wanted to do?	Evaluate the impact to demonstrate achievement against the standards and targets
How can we make change stick?	Change needs to be sustained to ensure that it becomes routine practice. This may require a change in work environment with a process of ongoing monitoring and audit to stop people reverting to old patterns of working

Source: data from Hill, A. et al., Public Health and Primary Care: Partners in Population Health, Volume 1, Oxford University Press, New York, USA, Copyright © 2007, by permission of Oxford University Press USA.

Numerous frameworks exist to guide how strategies are implemented. Once useful structure is the PEOPLE framework. The acronym PEOPLE (Hong Kong Government 2013, pp. xiii–xiv) can be used as a tool to review the structural support for strategy implementation:

Partnership
Environment
Outcome focused
Population-based interventions
Life-course approach
Empowerment

1. Partnership

Partnerships between different sectors of society and at different population levels are essential to maximize the strengths of different knowledge bases and skills. A strategic partnership or an alliance can provide a competitive advantage and an opportunity to access a broader range of resources and expertise. This means that the partnership can offer clients distinctive skill sets and product lines that are different from the competition.

Successful partnerships and alliances rely on the principle that the work involved in maintaining a partnership, and the benefits from the alliance are equally spread, rather than one partner carrying the load while the other reaps the benefits. The best results from a strategic partnership generally occur when each partner delivers excellence in service areas that are different but related to the other partner(s) and where the partners are not adversarial in the marketplace. That way, by partnering up, each partner can focus on its strengths, while having reliable people in other organizations to cover the areas outside its expertise.

CASE STUDY 11.7.8 Creation of Public Health England—emphasis on partnership

As part of the healthcare reform process in England, in 2013, the public health system was restructured by the creation of Public Health England (PHE) to reflect the need for strong central and local partnerships. The main drive for this reform was the recognition that public health is not the product of health service activity, and that addressing upstream broader social and economic determinants are key to promoting better health.

At a central level, the PHE strategy directorate is responsible for ‘a coherent, credible and internationally recognized strategy for the transformation of the public’s health’ (PHE n.d.) providing ‘the leadership for a widely-owned strategic direction for public health that is strongly supported by local authorities, the National Health Service (NHS) and other partners’. Key to its implementation was the repositioning of local directors of public health from the NHS to local government authorities that have duties and powers and promulgate local plans for improving health and well-being and reducing inequalities in their local communities. It is envisaged that there will be a continuous process of strategic assessment and planning involving all sectors to

‘develop local evidence-based priorities to improve the public’s health and reduce inequalities’, and helping determine what actions local authorities, the local NHS and other partners will need to take to meet health and social care needs, and to address the wider determinants that impact on health and well-being (Department of Health 2013a).

How well these structural changes work in delivering their objectives of reducing inequalities and addressing local health needs through joint strategies will need ongoing monitoring and evaluation.

When entering into a strategic partnership or alliance, there is a lot to consider, however the following strategies have been found to be especially helpful in creating positive and productive alliances:

Focus on outcomes. It is important to know what you can, and are prepared to deliver, and what you would like in return from your potential partner.
Commitment. The level of commitment required from each party needs to be considered. The nature of the partnership or alliance will determine how much each party needs to give in order to keep the partnership or alliance healthy.
Collaboration. The extent to which you will be happy to share resources with partners will need to be considered. Generally, the more collaborative the partnership or alliance, the more successful the endeavour is.
Communication. An emphasis on effective communication between partners is essential. Open and effective channels of communication between members of the partnership or alliance will ensure that there are no misaligned expectations between the parties.

2. Environment

Environmental influences play a key role in effective intervention. At a broad level, they may be economic, social, political, and physical. The healthy settings approach, which considers health promotion and disease prevention in the context of broader settings, provides an example of structuring environments to be health promoting.

The WHO defines a setting for health as the place or social context in which people engage in daily activities in which environmental, organization, and personal factors interact to affect health and well-being. Settings are characterized by physical boundaries, a range of people with defined roles, and an organizational structure (WHO n.d. f). Examples of settings include : Schools, Hospitals, Cities/villages, Workplaces. In healthy settings, people can create and solve problems related to health, and actively

use and shape the environment. Different settings provide diverse ways to implement health strategies.

Schools

Schools have an integral responsibility and role in encouraging health promotion as part of a larger social health model. The Healthy Schools movement, promoted by the WHO, has played a key role in developing and sharing experiences across many countries. For example, in India, the school fostering health programme has been emphasizing that a balanced diet, regular physical activity, and lifestyle management play a key role in the prevention and control of many diseases. One of its key programmes has been to work with schools to promote the 'Mid day Meal Program', which encourages healthy eating and hygiene (Mahal et al. 2009; Bangdiwala et al. 2011; Healthy India n.d.; Association of Schools & Programs of Public Health 2012, n.d.). The School Health Foundation of India has established a scheme for 'Healthy School' Certification based on the achievement of ten points of criteria (School Health Foundation of India n.d.). In Mississippi, United States, the state's Department of Education has established an Office of Healthy Schools committed to offering technical assistance and services to enable schools and communities to create programmes, based on CDC guidance, which encourage lifelong healthy behaviours (Etienne et al. 2013; Mississippi Office of Healthy Schools n.d.).

In England, the aims of the National Healthy Schools Programme (NHSP) are to: Support children and young people in developing healthy behaviours; help to raise pupil achievement; help to reduce health inequalities; and help promote social inclusion. Health-promoting schools have been characterized as having six domains for action: the formal curriculum, the school ethos, school policies and practices, school health services, school–home–community interactions, and organizational structures. National Healthy Schools also provide education and support of young peoples' personal and social health (including sex and relationship education), healthy eating, physical activity, and emotional health and well-being. The evidence regarding interventions to improve children's health indicates the value of intersectoral, comprehensive programmes, and these can be developed and implemented through the structure of health-promoting school programmes (Department of Health n.d.).

Hospitals

Hospitals are another example of a setting in which to promote health, as well as provide health services. The promotion of healthy hospitals can be described from two perspectives—through the care they give and through the lens of environmental sustainability. The initial declarations on health-promoting hospitals focused on the former with statements that hospitals should:

Promote human dignity, equity and solidarity, and professional ethics, acknowledging differences in the needs, values, and cultures of different population groups.
Be oriented towards quality improvement, the well-being of patients, relatives and staff, protection of the environment, and a realization of the potential to become learning organizations.
Focus on health with a holistic approach and not only on curative services.
Be centred on people providing health services in the best way possible to patients and their relatives to facilitate the healing process and contribute to the empowerment of patients.
Use resources efficiently and cost-effectively, and allocate resources on the basis of contribution to health improvement, forming as close links as possible with other levels of the healthcare system and the community (WHO Health Promoting Hospitals Network 1997).

The US CDC on Healthy Hospital Choices strategies suggests that hospitals could engage to promote better health where recommended—promoting healthy hospital food, physical activity, breastfeeding and lactation support, and tobacco-free choices (National Center for Chronic Disease Prevention and Health Promotion 2010; Rocco et al. 2011; CDC 2012; United States Global Health n.d.). However, over recent years, the responsibility of hospitals to contribute to sustainability has been a growing theme.

More recently the Global Green and Healthy Hospitals is an international network of hospitals, healthcare facilities, health systems, and health organizations dedicated to reducing their environmental footprint and promoting public and environmental health. The network has more than 1,000 members in 54 countries who represent the interests of over 32,100 hospitals and health centres. Members use innovation, ingenuity, and

investment to transform the health sector and foster a healthy, sustainable future.

The NGO ‘Healthier Hospital Initiative’ has created a network of hospitals committed to improving not only the health of patients and staff, but also to promoting the role hospitals can play in creating a sustainable environment through engaging leadership, healthier food, cleaner energy, less waste, safer chemicals, and smarter purchasing (Healthier Hospitals Initiative 2013).

In England, the NHS has created a Sustainable Development Unit which works across health, public health, and social care to help health and care organizations to fulfil their potential as leading environmentally sustainable and low-carbon organizations (NHS n.d.).

These roles of hospitals and healthcare facilities in civil society as providers of care, as providers of employment, and as influencers of the local environment and the local community are often forgotten.

Cities/villages

Cities/villages are also important healthy settings as they encompass the overarching areas where people live (Yang et al. 2018). The WHO healthy cities project (WHO n.d. g) is a global movement established to create health-supportive environments, which will enable the achievement of good quality life. They focus on ensuring basic sanitation and hygiene needs are met, as well as ensuring access to healthcare. In Europe, the primary aim is to put health high on the social, economic, and political agenda of city governments, and engage all sectors to provide policies and plans to address inequality and urban poverty, and address the needs of vulnerable groups and the social, economic, and environmental determinants of health (WHO Europe Office n.d.).

Workplaces are other important settings where health should be promoted. Just as the younger population spends most of their day at schools, adults usually spend more time at their workplaces than in any other setting. Workplaces can support programmes aimed at improving the health conditions and awareness among workers/employees for mental, as well as physical, well-being:

> Providing a psychologically safe workplace is no longer something that is simply nice to do, it is increasingly becoming a legal imperative. (Mental Health Commission of Canada 2013).

The WHO healthy workplace model suggests a comprehensive way of thinking and acting that addresses:

Work-related physical and psychosocial risks.
Promotion and support of healthy behaviours.
Broader social and environmental determinants (WHO n.d. h).

The programme for Health at Work seeks to ensure a positive approach to health and work which includes training managers to be aware and communicate about well-being and mental ill-health, as well as encouraging a culture of health (mental and physical) at work. This approach highlights the mutual benefits of health and productivity, linking traditional health and safety with well-being initiatives (World Health Organization Regional Office for the Western Pacific 1999; Healthy Workplaces 2011). This positive approach to health is demonstrated by the replacement of sick certificates with 'fit notes' in the United Kingdom. Introduced in April 2010, fit notes are issued to individuals by doctors to provide evidence that they 'may be fit for work' or 'not fit for work' (Department for Work and Pensions n.d. 2008).

3. Outcome focused

Too often, when doing strategic planning, public health practitioners and policymakers end up with global goals that only reflect a set of directions. The best-case scenario here is that they end up going in the 'right direction', but they never reach their community-specific destinations. Another typical result is that their successful areas or localities meet their success, but their needy or unsuccessful localities stagnate and remain the same.

Increasingly, global ambitions need to be met with more detailed locality specific strategies that identify their specific, desired destinations—that is, the objective and measurable outcomes that they want for local residents and communities. If 'outcomes-based' strategic planning is used in place of 'standards-based' strategic planning, localities, and implementers will be able to set their own strategic paths for achieving goals, cognizant of local assets, capacities, and capabilities. The clear result is that they will then have a much better chance of meeting their goals and attaining their outcomes.

A strategy that is outcome focused ensures an optimal investment of available resources with the greatest health gains through the monitoring of

health outcomes. Focusing on outcomes can help programmes measure the successes and failures of what has already been implemented, and allow room for evaluation, adjustment, and improvement.

CASE STUDY 11.7.9 Public health outcomes in England

In England, the Public Health Outcomes Framework *Healthy Lives, Healthy People: Improving Outcomes and Supporting Transparency* sets out a vision for public health, desired outcomes, and the indicators that will help us understand how well public health is being improved and protected. The Public Health Outcomes Framework was refreshed in May 2016, following a consultation in 2015.

The framework concentrates on two high-level outcomes to be achieved across the public health system, and groups further indicators into four 'domains' that cover the full spectrum of public health. The outcomes reflect a focus not only on how long people live, but on how well they live at all stages of life. A list of indicators updated, for the most recent and previous releases, can be found in the Public Health Outcomes Framework Collection within https://www.gov.uk. Data are published as part of a quarterly update cycle in August, November, February, and May each year. This data tool presents data for the indicators in the framework for the most recent period available and accompanying trend data where possible. Inequalities data are provided where these are available.

An example of outcome-focused strategies is the use of health inequality targets in England. These targets and headline indicators were initially promulgated in 2006 as part of a national strategy to promote health and identified 14 areas for measurement (Department of Health 2006). Two health inequalities targets on infant mortality and life expectancy in addition to 12 headline indicators for areas in which evidence suggested actions could lead to a greater emphasis on prevention were proposed. These have subsequently been refined and published as the Public Health Outcomes Framework for England 2013 to 2016 (Department of Health 2013b), which has the outcome targets of increased healthy life expectancy and reduced differences in life expectancy and healthy life expectancy between communities, as well as indicators in four domains: improving the wider determinants of health, health improvement, health protection and healthcare public health, and preventing premature mortality (Fig. 11.7.1).

This framework will guide public health action at the local level and regular monitoring will take place, allowing progress to be tracked and performance to be compared between different geographical areas based on local authority boundaries.

Public Health Outcomes Framework

OUTCOMES

Vision: To improve and protect the nation's health and well-being, and improve the health of the poorest fastest

Outcome 1: **Increased healthy life expectancy**
Taking account of the health quality as well as the length of life

(Note: This measure uses a self-reported health assessment, applied to life expectancy.)

Outcome 2: **Reduced differences in life expectancy and healthy life expectancy between communities**
Through greater improvements in more disadvantaged communities

(Note: These two measures would work as a package covering both morbidity and mortality, addressing within-area differences and between-area differences)

DOMAINS

DOMAIN 1:	DOMAIN 2:	DOMAIN 3:	DOMAIN 4:
Improving the wider determinants of health	Health improvement	Health protection	Healthcare public health and preventing premature mortality
Objective: Improvements against wider factors that affect health and well-being, and health inequalities	Objective: People are helped to live healthy lifestyles, make healthy choices and reduce health inequalities	Objective: The population's health is protected from major incidents and other threats, while reducing health inequalities	Objective: Reduced numbers of people living with preventable ill health and people dying prematurely. while reducing the gap between communities
indicators indicators indicators } *Across the life course*	*indicators indicators indicators* } *Across the life course*	*indicators indicators indicators* } *Across the life course*	*indicators indicators indicators* } *Across the life course*

Fig. 11.7.1 Example of strategic domains.
Reproduced from Department of Health, Public Health Outcomes Framework 2013 to 2016 and technical updates 2013, Crown Copyright 2013, licensed under the Open Government License v.2.0, available from https://www.gov.uk/government/publications/healthy-lives-healthy-people-improving-outcomes-and-supporting-transparency.

4. Population-based interventions

There are a broad range of prevention and early intervention strategies. Some prevention strategies target the *whole population*—for example, a mass media antismoking campaign. By their nature, these interventions are a *one size fits all* type, with no allowance for targeting specific populations. While these programmes are usually expensive, they are quite economic when calculated on a cost-per-person basis. Placing emphasis on the whole population produces collective health benefits and can change the patterns of behaviours and disease prevalence by having a small impact on many people. Population-based prevention strategies include legislation; media-based strategies (including mass media campaigns and reduction in the advertising and promotion of alcohol and tobacco), and drug education (including school drug education).

CASE STUDY 11.7.10 Less Salt, More Life campaign in Argentina

In 2011, bread makers in Argentina voluntarily reduced the salt in bread by 40% and restaurants in Buenos Aries removed salt from their tables as part of a campaign to reduce hypertension and reduce heart disease. Discussions also took place with the food industry on reducing the salt content of processed food. Prior to the campaign, the daily salt consumption was estimated to be two to three times that of recommended levels. Population-level salt reduction is considered to be one of the most cost-effective strategies to reduce hypertension and its consequences (He and MacGragor 2009). The government established a national committee with cross-government representation, working with the big food companies and unions to come to voluntary agreements. The coordinated monitoring plan will evaluate progress. It is estimated that with a 3 g salt reduction in each individual's diet, 6,000 cardiovascular deaths, 13,000 myocardial infarctions, and 13,000 strokes could be avoided overall (Laspiur 2012).

CASE STUDY 11.7.11 Sugar reduction and reformulation in England

Childhood obesity is a national crisis in England and a leading cause of ill-health and premature death where one in three children are overweight or obese by the time they leave primary school. Obese or overweight children are more likely to become obese or overweight adults, which increases their risk of developing cardiovascular disease, type 2 diabetes, and some cancers.

Public Health England's prioritization of childhood obesity recognizes that we live in an obesogenic environment that encourages weight gain and obesity, and excess sugar intake has

become the norm. Children and adults are consuming more than double the recommended amount of sugar and this is an underpinning factor of the obesity epidemic.

In 2016, PHE launched the first Sugar Reduction Programme in the world. This was supported by the government's commitment to tackling childhood obesity, and Childhood obesity: a plan for action (HM Government 2016), which included a commitment for PHE to lead a transparently monitored sugar reduction programme. The programme challenges the food industry—from retailers to manufacturers, restaurants, cafés, takeaways, pubs, entertainment chains, and delivery services—to reduce the sugar in the products that contribute most to children's intakes by 20% by 2020 (UK Government n.d.).

PHE determined where to focus efforts on sugar reduction by the food industry by identifying the main foods and drinks that contribute to children's consumption of sugar, and the biggest businesses that make them, using National Diet and Nutrition Survey (NDNS) data and market share analysis.

PHE has committed to engaging with food manufacturers and retailers who manufacture products that come under the high sugar categories to discuss sugar reduction in their products and targets for each category. PHE is also with the 'eating out of home' sector—those that provide food that we both buy and consume out of the home, including restaurants, pubs, takeaways, and fast food restaurants, sandwich, and coffee shops.

From 2017, PHE will work towards setting targets to reduce total calories in a wider range of product categories and across all sectors, including the eating out of home sector. Work on saturated fat will be further reviewed in 2018 onwards.

5. Life-course approach

A life-course approach addresses cumulative adverse effects on people's lives by fostering health from gestation through to old age. The risks of developing disease and disability accumulate with age and are influenced by social and physical factors at all stages of life (Fig. 11.7.2).

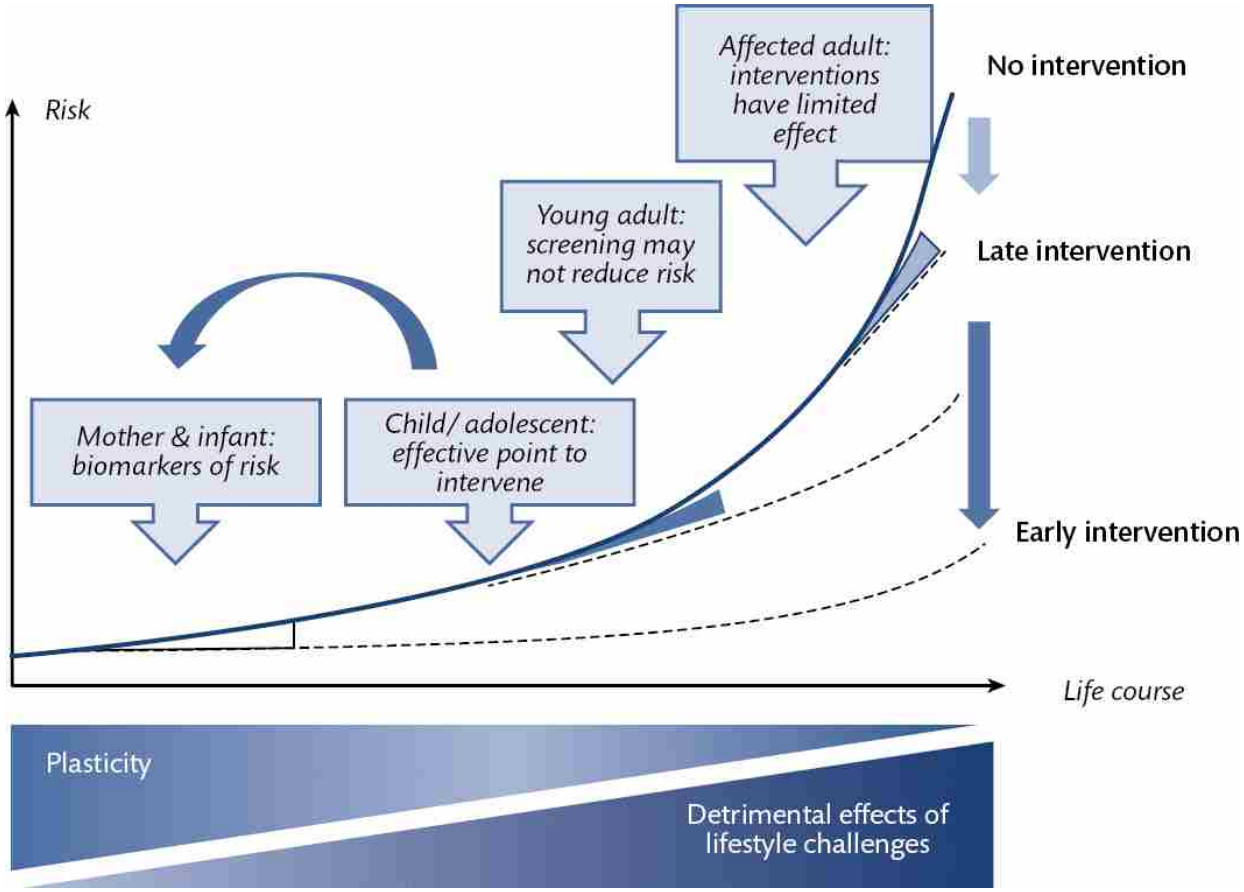


Fig. 11.7.2 Life-course view of NCD.

Reprinted from *Progress in Biophysics and Molecular Biology*, Volume 106, Issue 1, Hanson, M. et al., Developmental plasticity and developmental origins of non-communicable disease: Theoretical considerations and epigenetic mechanisms, pp. 272–80. Copyright © 2011, with permission from Elsevier.

There are critical periods, such as gestation, and early childhood and also adolescence, when environmental exposures do more harm to health and its potential, as well as sensitive developmental stages when social and cognitive factors are more easily acquired, than at other times. Thus, social behavioural and environmental factors impact on the potential for a healthy future and continue to do so throughout the life course to old age (Kuh Ben-Shlomo 1997; WHO n.d. i).

Understanding the importance of the life course has led to different initiatives to develop strategies to promote preconception care, promote optimum care in pregnancy, efforts to ensure safe delivery, a focus on vulnerable children and adolescents, and interventions tailored to promoting and maintaining health in middle and older age.

CASE STUDY 11.7.12 Everybody Active Every Day—promoting physical activity in England

Since the 1960s people in the United Kingdom have become less and less active. We are around 20% less active today, and estimates suggest that by 2030 we will be 35% less active. Half of women and a third of men in England today are not active enough to stay healthy. In October 2014, Public Health England (PHE) published Everybody Active, Every Day as the national physical activity framework for England. It drew on the international evidence base and was co-produced with over 1,000 different individuals and organizations. It set out the need for action across four domains at national and local levels (Active society, Moving Professionals, Active environments, and Moving at scale). The core document was supported by a 'What works—the evidence' document, a toolkit for MPs and a collation of local emerging and promising practice developed with ukactive and the National Centre for Sports and Exercise Medicine (NCSEM).

This vision of 'Everybody active every day' was supported by a practical 'what works' guide, which details promising local practice initiatives that already have strong evidence of success in increasing activity in local areas. In addition, the strategy aimed to promote physical activity through the life course focusing on five key opportunities for local areas:

- teach every child to enjoy, value and have the skills to be active every day
- create safe and attractive environments where everyone can walk or cycle, regardless of age or disability
- make 'every contact count' for professionals and volunteers to encourage active lives
- lead by example in every public sector workspace
- evaluate and share the findings so we are learning more about what works

A 2-year evaluation of the progress made in this strategy (Ahmad et al. 2018) highlighted that there has clearly been significant progress to tackle inactivity since the publication of Everybody Active, Every Day in 2014. The framework has informed local and national strategy and delivery plans and has made a visible difference to the approach across England. While the full impact of these first two years may take time to be reflected in the data, progress has clearly been made to embed the evidence into practice and build the foundations for a healthier future for the nation. International evidence demonstrates the need for continued leadership and coordination over several years to make sustainable change in population levels of physical activity.

6. Empowerment structures for local action—community-based initiatives

Community engagement, also called 'public participation', is the process of involving people in the decisions that affect their lives. It enables good governance and informed decision-making by promoting shared responsibilities for decisions. It supports an open approach to managing risk by providing a strong foundation for understanding decisions and building trust within our community about the decision-making process (Merson and Page 2009). Community engagement can involve a broad variety of activities, best considered as a continuum from informing (reflecting a low

level of engagement) through to active participation (reflecting a high level of engagement).

Informing takes place when a decision has already been made or action is required, and localities need to make sure that those affected are aware of the facts.
Consulting takes place when a project requires some input, feedback, or advice before part of the project or decision is progressed.
Active participation takes place when the locality collaborates with specific stakeholder groups or the community to work out what needs to be done and to develop solutions that are incorporated into decision-making.
All engagement processes need to inform; most will have some level of consultation and some will include active participation.

Engagement and empowerment give everyone in the community the opportunity to achieve their full potential by gaining control over decisions and actions that influence health. There are many levels of involvement when empowering people and their communities (Box 11.7.3).

Box 11.7.3 Principles for involving the public in public health strategy implementation

- The community must be seen as being part of the solution, not the problem.
- Community representatives need mechanisms to link to the wider community.
- Community input must be accountable to the local community. Any strategy has to take into account marginalized minority groups.
- To ensure the public can truly benefit, financial support is necessary.
- The community should raise issues that are of concern to them.
- The community needs to see concrete outcomes from their voluntary involvement, otherwise, they might lose interest.
- Meetings involving the community must avoid jargon.

Source: data from Crowley, P., Community development and primary care groups, pp. 41–2, in Communities Developing for Health, UK Health for All Network, Liverpool, UK, Copyright © 2000.

In addition to directly involving members of the public, the voluntary and community sector organizations also have a key role to play in helping to develop strategy. There is a considerable evidence base on the reasons why

partnership working is needed to address issues of exclusion, area-based disadvantage, and development opportunities, especially but not exclusively from the perspective of the community and voluntary sectors. For example:

It ensures the involvement of residents and excluded groups at different stages and on different issues in the strategy development process;
It enables community and voluntary organizations to highlight the contribution and assets they can bring to implementing solutions;
It can provide a framework for more joined-up approaches to addressing issues of equality, disadvantage, and exclusion;
It makes for better accounting for the impact of the strategy development processes, especially where they involved substantial public sector investment.

An example of the contributions of Diabetes UK to strategy development in England is highlighted in Box 11.7.4 next.

Box 11.7.4 Role of Diabetes UK

- Diabetes UK is the leading charity that cares for, connects with, and campaigns on behalf of every person affected by or at risk of diabetes. We help people manage their diabetes effectively by providing information, advice, and support.
- We campaign with people with diabetes and with healthcare professionals to improve the quality of care across the United Kingdom's health services.
- We fund pioneering research into care, cure, and prevention for all types of diabetes.
- We work to stem the rising tide of Type 2 diabetes—through risk assessment, early diagnosis, and by communicating how healthy lifestyle choices can help many people avoid or delay its onset.

Reproduced with permission from Diabetes UK, What we do, Copyright © Diabetes UK 2014, available from http://www.diabetes.org.uk/About_us/What-we-do/.

Delivering public health strategy

Political leadership

Undoubtedly, given the political environment in which public health policy development and implementation often sits, it is imperative that due

consideration is given to political context and the potential for this to be either a barrier or facilitator. In addition to having a positive effect on health, effective public health strategies can contribute to societal stability, through attaining health equity, national security, and economic dividends to those who invest in them. These benefits need to be articulated more effectively to Heads of State, government ministers, local politicians as well as civil servants, policy makers, and economists. Helping political leadership to understand that the health system and public health strategies can help to improve its functioning and outcomes, and are a critical source of employment, capital investment, contributes to economic growth via a healthy workforce that is more productive, and to national security.

CASE STUDY 11.7.13 Partners for Mental Health

The Canadian Mental Health strategy emphasizes the need to engage patients and their carers in determining what they want from health services:

The mobilization of the Canadian public to take action on mental health issues is essential in order to ensure that mental health remains a high priority (Mental Health Commission of Canada 2012).

Partners for Mental Health 'has one focused goal … to catalyse a social movement that will empower individuals and organizations to take action to improve mental health' and in particular to engage with the general public to increase engagement in and advocacy for mental health issues (Mental Health Commission of Canada 2012).

Effective strategy implementation also relies on community engagement and the conciliation of top-down directives with bottom-up initiatives. As such, it is important to engage the public and patient groups in implementation as their lack of engagement will result in a failing strategy. Ways of engagement may be through minimum participation, where people are merely passive receivers of information, for example, receiving health education materials or through active partnership, in which patient groups or the public are partners, empowered to play expanded roles and to take responsibility. For example, Diabetes UK plays a pivotal role in providing advocacy and advice for diabetic patients at national and local levels (Box 11.7.4).

Examples of mobilizing political leaders to support public health strategies are found both at global national and local level. Political leadership and engagement in WHO's strategic programmes on climate change, universal healthcare, tackling the HIV/AIDS are well documented and are key achieving consensus and accelerating progress. In England, strong political leadership from the Secretary of State for Health was essential to develop the National Tobacco Control Plan (DH 2017), National Sexual Health Strategy (DH 2013c), and more recently the Childhood Obesity Plans (DHSC 2018). At local levels, where public health strategies may become more challenging due to the resourcing, community concerns

and political mandates, local elected mayors, and councillors have been at the forefront of tackling issues that matter to local communities and promoting greater community engagement and co-production as strategies to promote inclusive policy development.

Politicians at global, national, and local level, have to balance a number of competing concerns and priorities, not just the evidence of effectiveness, before agreeing to champion, support, and fund implementation. Considerations may include alignment with political ideology and manifesto commitments, the need to balance competing interests, mitigating the risk and unintended consequences of public health policies on other sectors, how decisions will play out in the media and public domain, and the threat of loss or alienation of key constituencies including the public, donors, business etc. Recognizing these additional and important perspectives are critical to understand the window of opportunities for developing public health policies, the importance of developing coalitions, working with a diverse range of partners and sectors to demonstrate widespread buy-in and support, and clear demonstration of public support. If one is able to clearly align the policy with the values, priorities, and commitments of politicians then this will help mitigate barriers to implementation. Political leadership is therefore needed:

To establish the national or local public health strategy and business case action, providing support by necessary regulations.
To use it for demanding plans and budgets from public and private educators and employers of health workers.
To use innovative processes and technologies for more effective and efficient services delivered to protect, promote, and restore health at optimal levels.
To negotiate targeted technical and financial assistance from regional and global donors, and to require donor aid to be harmonized and targeted to meet priority health system needs as defined by countries.

One should also remember that this process of political engagement and support may happen quickly, for example spurred on by response to a national or local tragedy, or may happen over a prolonged period as time is taken to change public opinion, educate, engage, and enlist support of coalitions to demonstrate the need and value of the policy change. In England, national leadership for tobacco control in the United Kingdom unfolded over five decades and reflected the time taken to develop a robust

evidence base on the primary and secondary impacts and costs of smoking on individuals and society, strong leadership by the healthcare professions working in partnership with advocacy organizations, creating alliances with key political leadership, running effective antismoking campaigns, and successive development and implementation of national tobacco control strategies with clear commitments, funding, and accountability for progress (Raw and Joossens 2007; Joossens and Raw 2011).

Leadership and governance

The best developed public health strategies will fail if due regard is not given to leadership and governance. All over the world, public health managers and providers face the challenges of producing results in an increasingly complex healthcare environment. Health sector reform, changing donor priorities, shifting client needs, and new technologies require that managers at all levels take on responsibilities previously held by people at higher levels. Moreover, the drain on health systems workforce shortages, financial constraints, and dealing with new and emerging diseases has accelerated the need to learn new ways of leading and managing to achieve results. Through improved leadership, management, and teamwork, managers in public, private, and non-governmental organizations can face these challenges with more confidence and lead their teams to effect significant changes in health.

Appropriate leadership and governance in implementing public health strategy involves ensuring that strategic policy frameworks exist and are combined with effective oversight, coalition-building, regulation, attention to system design, and accountability. Leadership involves strategically developing and implementing plans and policies with accountability and transparency. Good leadership, which requires the ability to face challenges to achieve results in complex conditions, can and should be practised at all levels.

Governance of health systems comprises the actions adopted by a society to organize itself to promote the health of its population. Although governance is the least understood component of health systems, it impacts on all other health system functions. Good governance may result in strengthened performance of a health system, including effective delivery of public health strategies, health services, and improved health outcomes.

Governance has in recent years transitioned to the fore of the international development agenda, indicating a shift from attention to micro level, project specific objectives to macro level issues of policymaking. Governance in public health strategy is a cross-cutting theme, intimately connected with issues surrounding accountability. It is an integral part of the health system strengthening yet they remain inadequately monitored and evaluated.

In England, an Independent Advisory Group was established to oversee the new National HIV/AIDS Strategy (DH 2013c), bringing together advocates, scientists, clinicians, public health experts, and professionals to review progress and drive successful implementation. Within Public Health England the establishment of a delivery board reporting directly to the PHE Executive oversees implementation of major national public health strategies as well as PHE's own NCD prevention strategies. At the local level, oversight of strategy implementation often rests within existing governance structures including the local health and well-being board, providing oversight, challenge, and accountability for delivery.

Critical lessons are now emerging on some of the broad principles which must underpin good governance in the implementation of public health strategies. These include proactive participation and engagement of persons and communities who are the target of the policies to improve programme responsiveness, efficiency, effectiveness, and sustainability, and to strengthen service-user self-determination and satisfaction. There should be robust investment in data to support political momentum, decision-making of policymakers, evaluation, accountability, and equitable allocation of resources. There should be explicit promotion in policies of the importance of tackling inequalities and promoting inclusion, ensuring equitable and accessible services. And finally, linking back to the importance of political leadership, wherever possible there should be robust intersectoral coordination to ensure coherency across governmental departments (national and local) aligned with a commitment to promote programme sustainability by aligning implementation with pre-existing programmes and initiatives.

The public health workforce

Public health has been defined as 'the science and art of promoting and protecting health and well-being, preventing ill-health and prolonging life through the organized efforts of society'. While this means that a broad

range of people and organizations contribute to protecting and improving people's health, the World Health Organization has defined the three core functions of a public health system as:

the assessment and monitoring of the health of communities and populations at risk to identify health problems and priorities
the formulation of public policies designed to solve identified local and national health problems and priorities
to assure that all populations have access to appropriate and cost-effective care, including health promotion and disease prevention services

In the United Kingdom these functions are mainly delivered through a 'core' workforce (Box 11.7.5), including public health professionals such as directors of public health, specialists, academics and practitioners who are trained in all or some of the nine key areas of public health practice, working in a range of commissioning or delivery roles in a variety of organizations. The role of a broader group of people—the wider workforce—in the public health system, is increasingly being recognized. Improving public health requires the combined efforts of people from many professions and disciplinary backgrounds (Sim et al. 2007).

The formal definition of a public health professional is a person educated in public health or a related discipline who is employed to improve health through a population focus (Institute of Medicine 1988). A useful way of considering the workforce, adopted in the United Kingdom, is that there are three main groupings: (1) specialists; (2) those who practise public health as part of their role, such as primary care workers; and (3) those with an interest whose work impinges on the public's health.

However, it is now widely acknowledged that public health really is everyone's business and that our aim must be to influence and improve many aspects of an individual's life including transport, housing, education, and the environment if we are to begin to shift the pendulum towards a healthier, happier country. For example, voluntary NGOs frequently have greater insights into the health needs of disadvantaged groups within communities than do statutory organizations. As such, it is not enough to focus on any one sector or any one activity but to recognize the interactions and different levels at which these influences are working.

Box 11.7.5 The core public health workforce

Specialists—including directors of public health—working as system leaders at strategic or senior management level, with high levels of technical skills and knowledge—employed mainly in local authorities and PHE

Public health nurses, such as school nurses and health visitors, and those working in health protection, screening, and immunization, research, TB, sexual health, infection prevention and control, drugs, and alcohol; they may be employed by NHS trusts, PHE, the voluntary or private sectors

Public health midwives—improving and promoting the health and well-being of women, children, families, and communities throughout pregnancy, birth, and the postnatal period as well as specializing in quality assurance of antenatal screening programmes in PHE

Environmental health officers and practitioners—working to improve, monitor and enforce public health and environmental health standards in local authorities and in PHE health protection teams

Knowledge and intelligence staff—employed in data analytics, knowledge management, informatics and interpretation and presentation of information

Academics—developing the research and evidence base and training new entrants to public health

Scientists—leading the development and delivery of highly technical and specialized scientific functions in support of public health objectives, primarily within PHE, also working in NHS laboratories

Practitioners with a wide range of relevant skills delivering public health services and programmes in a range of settings including local authority public health teams and PHE

Health trainers, stop smoking advisors, weight management advisors, etc—providing advice, motivation, and practical support to individuals in their local communities, often employed in third sector organizations

In the United Kingdom, public health practice is recognized as having three overlapping domains of health protection, health improvement, and health services (HM Government 2010). This structural framework recognizes the interactions between the different aspects of practice and that practice needs to be underpinned by an understanding not only of epidemiology and biostatistics, but of ethical issues within a global context. The capabilities needed for public health practice have been described in the UK-wide skills and knowledge framework, which is currently being updated to reflect today's practice in both the core and the wider workforce. Four key functional areas have been identified:

Behaviours—principles, values, and ethical practice
Technical competence—traditional public health skill sets in which the core workforce will be more specialized
Context—strategic/policy; organizational/collaborative; contractual/commercial; political and democratic
Delivery—leadership, communication, programme management, prioritization, and management of resources

In addition to technical skills in public health, specialists, and particularly directors of public health in local authorities, need skills that enable them to work as system leaders at strategic or senior management level, or a senior level within a particular area of expertise.

Some overseas countries adopt a similar approach and standards including the Hong Kong College of Community Medicine (n.d.), which is part of the Hong Kong Academy of Medicine. The Academy is an overarching medical body representing the professional groups in medicine (Hong Kong Academy of Medicine n.d.). The Colleges within the Academy are professionally led and independent of government or licensing bodies. Their remit is to set professional medical standards and curricula from postgraduate to specialist levels. Licensing bodies, such as the UK General Medical Council (n.d.), Faculty of Health (n.d.), or the Hong Kong Medical Council (n.d.), act as professional regulators and issue the licences to practise.

Health service contribution: the role of primary care

How public health strategies and programmes are delivered is obviously key to their effectiveness not only in dealing with disease, but in promoting better public health. The role of health services and of public health within health services is important for delivering better health outcomes. For example, providing antenatal care, well-baby checks, postnatal support, and immunizations are all part of a public health service delivered in different places by different systems including within a primary care setting or within a government-run maternal or child health clinic. As such, public health services need to be considered within the framework of primary care provision.

That primary care should be the foundation of health services has been recognized by many countries who strive to develop or strengthen their public health and healthcare systems. The need to enhance primary care, based on the principles of Alma-Ata, has been reaffirmed in the World Health Report 2008, Primary Health Care: Now More Than Ever (WHO 2008) and further highlighted in the 2009 WHA's resolution on primary care policies (WHO 2009b).

In comparison with systems based on specialist care, primary care produces better population health outcomes, reduces the rate of avoidable mortality, improves continuity and access to healthcare, results in higher patient satisfaction, and reduces health-related disparities at a lower overall cost for healthcare (Starfield 1998). Primary care health services provide ongoing care at a community level and, as such, they are often closely interrelated with public health services. In many situations, primary care clinicians are indeed the main providers of public health services and increasing attention is being paid to effective and efficient models for integrating personal healthcare and public health, and organizing primary care on the principle of care for individuals in the context of an identified population over time. To quote Van Weel et al. (2008):

> The future of primary care, and health care in general, will depend on how effectively primary practices achieve this community-oriented primary care approach and contribute to equity and social cohesion. Current efforts to revitalize primary health care worldwide should go hand in hand with attention to the social determinants of health. Just as the social determinants' approach to improving health equity must involve health care so must programmes to control priority public health conditions include attention to the social determinants of health.

Now More Than Ever (WHO 2008) emphasizes yet again how implementing primary care strategies needs to be context specific, in particular with reference to income and resources. For example, the Chinese healthcare system is investing significantly in community health centres and community-based systems to provide universal access to primary care, to emphasize prevention, to enable choice between Western and traditional medicines, and to strengthen programmes within local communities.

The primary care strategy of the reforms in China aims to strengthen the establishment of the public health service system by:

Building sound public health networks of disease prevention, health education, maternal and child healthcare, mental health, first aid, blood collection and supply, health supervision, and family planning.
Improving the public health functions of the medical services system based on a basic medical services network.
Setting up an information-sharing and resources-sharing public health services system.
Enhancing the capacity of public health services.
Enhancing the capacity to respond to public health emergencies.
Decreasing the urban–rural inequity in utilization of public health services (Government of China 2009).

Another example of the promotion of the strategic focus for healthcare reform through developing primary care services to deliver public health comes from Chile. In Chile, the health policy has defined a detailed benefit package, which is well publicized among the population as an enforceable right. People are being informed about the kind of services, including access to specialized care, which they can claim from their primary care teams. In combination with sustained investment, such unambiguous entitlements create a powerful dynamic for the development of primary care (WHO 2008, chapter 6).

Research and evaluation

Research plays and important role in strategy development and implementation. It is essential that public health policies are developed on the most robust evidence available including a robust assessment of the need for change, evidence of effectiveness and cost-effectiveness of interventions, and where appropriate a robust evaluation, including the use of mathematical modelling, of potential options for action. Increasingly, we are seeing the importance of qualitative research as a tool to provide deeper insights into the feasibility, acceptability of potential policies from political, provider, and community perspectives.

Strategy design and implementation is a continuous process. Continued communication about key themes and the resulting behavioural changes should be identifiable and will need to be updated as evaluation strategies are embedded into the process. The two basic rules for successful evaluation:

The evaluation process must be thought through from the start, at the same time as developing the strategy's aims, objectives, and targets.
Adequate funding must be set aside for evaluation.

The key areas to evaluate must be agreed among the partners, including participants, to reflect different agendas. Evaluation will include: Measuring indicators of progress, including progress towards targets; Assessing how well various aspects of the strategy were perceived to work from the point of view of professionals from all sectors and by communities; and assessing whether the changes were a result of the intervention. The use of an impact screening tool can encourage a more realistic perception of the likely impacts of the strategic goals upon particular population groups, and can be used during all stages of planning and review, as highlighted to follow.

As discussed in earlier chapters, public health research and evaluation remain underfunded in most jurisdictions despite the importance of the evidence base to guide effective policies and professional practices. High-tech medical proposals with quick returns are often more attractive to funders.

Earlier in the chapter we have stressed the importance of evaluation and monitoring health interventions as key components of the strategic approach. Increasingly, strategies are developed which recognize the inter-relationship between health and health services. For example, in the United Kingdom when launching Investing in Research, Improving Health, the responsible minister wrote:

> We are committed to tackling disease prevention wherever possible and are excited by the emerging understanding of the biological links between inequalities and health. The benefits to health and healthcare are explicit but we are also alert to the potential for economic benefits. The life sciences have a key part to play and a tradition of excellence on which to build. (Chief Scientist Office, NHS Scotland 2009)

In England, the National Institute for Health Research has an arm which focuses on public health. The public health research (PHR) programme commissions research to provide new knowledge on the benefits, costs, acceptability, and wider impact of non-NHS interventions (NHS 2013) (Fig. 11.7.3). However, as in other countries such as Australia, the lion's share of available funding goes to basic sciences and clinical research, with allocations to public health and health services research in single-digit figures (Australian Government National Health and Medical Research

Council 2008; Pitcher et al. 2010; Australian Government Department of Health and Ageing 2012).

CASE STUDY 11.7.14 A strategic approach to public health research in England

Over the coming decades, the UK population will face a wide range of complex health challenges and opportunities, many of which can only be fully addressed through strategies to secure and improve the health of the public as a whole. The Academy's report, *Improving the Health of the Public by 2040* (AMS 2016) explored how to organize our research environment to generate and translate the evidence needed to underpin such strategies.

The report concluded that while PHR has provided fundamental insights into human health, there remains much we do not know about the complex array of interlinking factors that influence the health of the public, and about how to prevent and solve the many health challenges we face as a population. Solving these challenges will require shifting towards a 'health of the public' approach, involving disciplines that would not usually be considered to be within the public health field. This is turn requires six key developments: (1) Rebalancing and enhancing the coordination of research; (2) Harnessing new technologies and the digital revolution; (3) Developing transdisciplinary research capacity; (4) Aligning perspectives and approaches between clinical and public health practice; (5) Working with all sectors of society, including policymakers, practitioners, the commercial sector, and the public; and (6) Engaging globally.

The report emphasizes the need to significantly increase transdisciplinary research capability, in a way that harnesses the changing drivers of health. To contribute to this ambition, the Academy has launched a new grant scheme called Springboard—Health of the Public 2040 to help further the careers of newly independent researchers working in the health social sciences and medical humanities. The Academy's *Improving the Health of the Public by 2040* report also emphasizes that policy decisions should be guided by a robust understanding of how to most effectively support and improve the health of the public. To facilitate improved use of evidence in policymaking, the Academy is currently undertaking a joint project with the Royal Society in evidence synthesis.

Looking to the future, there needs to be closer alignment between PHR and policy. Ensuring programme relevant research is being undertaken to address contemporary and emerging public health problems, a stronger focus on implementation science, addressing the wider determinants of health, clarifying the costs and cost-effectiveness of interventions, and ensuring that efforts to reduce inequalities are addressed. Much of this change in focus will require a new focus on the research workforce and infrastructure as highlighted in the Academy of Medical Science in England's report on *Improving the Health of the Public by 2040* (AMS 2016).

Barriers to implementing strategy

We have already discussed the importance of the political leadership and support for the successful implementation of public health strategies, given the need for political support and advocacy, full engagement of a range of cross-sector partners, engaging the public, and releasee of resources. Failure to adequately consider this political reality (which may also include organizational or community level politics) can have a detrimental impact on implementation.

However, there may be other barriers to strategy implementation. While many organizations produce strategies, many do not achieve their goals. To some extent this is because they are aspirational and at times unrealistic. Management literature suggests that there are four main barriers to effective strategy implementation. These are:

The vision barrier: in the age of the knowledge-based economy, the workforce needs to understand the strategy.
The people barrier: personal incentives need to be tied to the long-term initiatives supporting the strategy, not to short-term (financial) results.
The resource barrier: strategies need to be linked to budgets.
The management barrier: managers need to discuss strategies but often focus on budgets (Kaplan and Norton 2001).

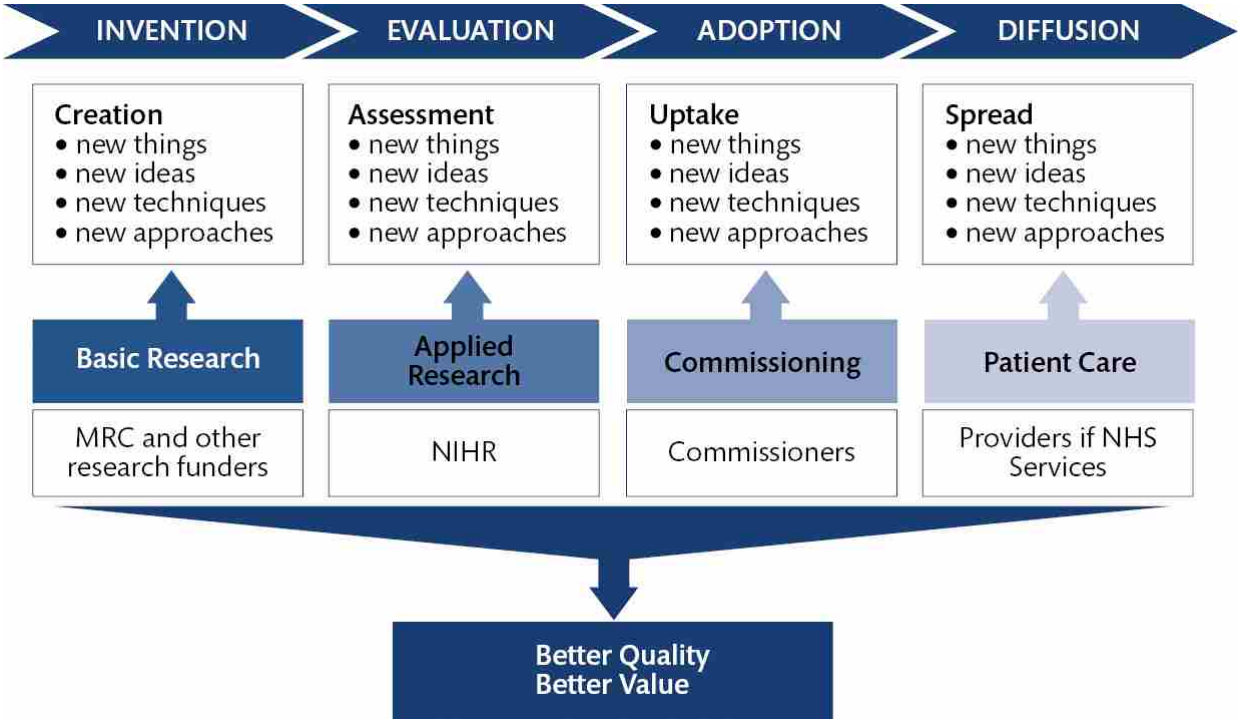


Fig. 11.7.3 The research pathway.

Reproduced with permission from The National Institute for Health Research (NIHR), NIHR Briefing Note 1.1, Version 9, Copyright © 2014, under the Open Government License.

Thus, there may be overattention being paid to the day-to-day duties rather than tackling the projects that have been identified as key to organizational sustainability and success. Important strategies may be perceived as overwhelming and too difficult to tackle. Lack of teamwork can inhibit cooperation across organizational divisions and different players have different priorities. Recommendations for successful implementation of strategic plans include:

Identify, understand, and plan for the political context within which the strategy is being developed and will be implemented.
Ensure the appropriate leadership and governance is in place to drive implementation and to ensure accountability for each step of the process including evaluation and dissemination of results.
Ensure clear discussion and agreement on the resources required for implementation and plans are in place to access those through the entire implementation phase.
Identify your coalition of diverse and inclusive partners for strategy development, implementation, advocacy, and strategic communications.
Focus: having too many priorities means no action and not everything can be or is a priority, so agree on what the top two to three goals are.
Communicate clearly and often, and understand each other's language and contextual background.
Emphasize and reward teamwork which may involve trade-offs, negotiation, and difficult conversations about resource allocation and expectations.
Regular review of progress: provide a framework for ensuring that all stakeholders reach their objectives.
Accountability needs to be clearly defined.
Emphasize implementation of the strategic plan is the most important aspect of planning.

Identifying the barriers to implementation and overcoming them is a key success factor for any organization. Teamwork, accountability, and resolution of conflicting priorities, or hidden agendas, will help ensure the implementation of the top two to three goals (Dane 2010).

Conclusions

This chapter has focused on strategies for promoting public health and structures for their implementation. As the scope of public health broadens and more consideration is given to the wider determinants of health and to the challenges of the future—the NCD epidemic, changing patterns of demography and disease, and climate change and its impact on health—more partners will need to be involved in signing up to strategies to tackle the threats they pose to public health. The example of SARS in 2003 demonstrated the problems that can occur when a strategy is missing. The

experiences of China during its economic transition demonstrate the need to constantly review the impact of broader determinants on equity in health and health services. Different examples of evidence-based and value strategies have been discussed with reference to examples and experiences in different countries and settings. The need for fit-for-purpose structures in the health systems, public health workforce, and research community has been highlighted and barriers to overcome in achieving the strategic goals described. All too often strategies remain merely fine words. The fundamental challenge is their implementation and continuous refinement in the light of delivery.

ACKNOWLEDGEMENTS

This chapter contains public sector information licensed under the Open Government Licence v2.0.

Text extracts from Save the Children, State of the World's Mothers 2012: Nutrition in the First 1,000 Days, Westport, Connecticut, USA, p.3, Copyright © 2012, reproduced with permission from Save the Children.

Text extracts reprinted from The Lancet, Volume 372, Issue 9642, van Weel, C., et al., Integration of Personal and community health care, pp. 871–2, Copyright © 2008, with permission from Elsevier, http://www.sciencedirect.com/science/journal/01406736.

Text extracts from World Health Organization, The Ottawa Charter for Health Promotion: First International Conference on Health Promotion, Ottawa, 21 November 1986, Copyright © 1986, reproduced with permission from the World Health Organization, available from http://www.who.int/healthpromotion/conferences/previous/ottawa/en/.

Text extracts from World Health Organization, The Bangkok Charter for Health Promotion in a Globalized World, Copyright © 2005, reproduced with permission from the World Health Organization, available from http://www.who.int/healthpromotion/conferences/6gchp/bangkok_charter/en.

Text extracts from World Health Organization, MPOWER: Tobacco Free Initiative, Copyright © 2014, reproduced with permission from the World Health Organization, available from http://www.who.int/tobacco/mpower/en/.

Text extracts from WHO Health Promoting Hospitals Network, The Vienna Recommendations on Health Promoting Hospitals, WHO Regional Office for Europe, Copenhagen, Sweden Copyright © 1997, reproduced with permission from World Health Organization.

REFERENCES

Action on Smoking and Health Australia (n.d.). The tobacco problem and our aims. [Online] Available at: http://www.ashaust.org.au/

Action on Smoking and Health UK (n.d.). About ASH. [Online] Available at: http://www.ash.org.uk/about-ash

Academy of Medical Sciences (AMS) (2016). *Improving the Health of the Public by 2040. Optimising the Research Environment for a Healthier, Fairer Future.* [Online] Available at: https://acmedsci.ac.uk/file-download/41399-5807581429f81.pdf

Ahmad, A., Rayment, M., Shah, P., Hannah, A., and Singh, K. (2018). *A Review of the Implementation of Everybody Active, Every Day at National and Local Level Final Report*. [Online] Available at: https://www.icf.com/-/media/.../everybody-active-every-day_public-health-england.pdf

Africa Union and Africa Centers for Disease Control. *2017–2021 Africa Centres for Disease Control and Prevention Strategic Plan*. [Online] Available at: http://www.ianphi.org/_includes/documents/Strategy%20Africa%20English%20CDC.PDF

American Heart Association (2012). Researchers identify evidence-based public health interventions for policy makers. [Online] Available at: http://newsroom.heart.org/news/researchers-identify-evidence-237575

Association of Schools & Programs of Public Health (n.d.). *MPH Core Competency Model*. [Online] Available at: http://www.aspph.org/educate/models/mph-competency-model/

Association of Schools and Programs of Public Health (2012). MPH expert panel. [Online] Available at: https://www.aspph.org/ftf-reports/the-mph/

Australian Government Department of Health and Ageing (2012). National mental health strategy. [Online] Available at: http://www.health.gov.au/internet/main/publishing.nsf/Content/mental-strat

Australian Government National Health and Medical Research Council (2008). *Report of the Review of Public Health Research Funding in Australia.* [Online] Available at: http://caphia.com.au/testsite/wp-content/uploads/2016/07/1327359298Review-of-PH-Research-Funding-in-Australia-2008.pdf

Bangdiwala, S.I., et al. (2011). Public health education in India and China: history, opportunities, and challenges. *Public Health Reviews*, 33(1), 204–24.

Centers for Disease Control and Prevention (CDC) (2012). *CDC Global Health Strategy 2012–2015*. [Online] Available at: http://www.cdc.gov/globalhealth/strategy/

Chief Scientist Office, NHS Scotland (2009). *Investing in Research, Improving Health*. [Online] Available at: https://www.cso.scot.nhs.uk/wp-content/uploads/2013/02/research.pdf

Commission on Social Determinants of Health (2008). *Closing the Gap in a Generation: Health Equity Through Action on the Social Determinants of Health*. Geneva: WHO. Available at: http://whqlibdoc.who.int/publications/2008/9789241563703_eng.pdf

Council on Smoking and Health (COSH) (n.d.). About COSH. [Online] Available at: http://www.smokefree.hk/en/content/web.do?page=ChairmanMessage

Cowling, B.J., et al. (2010). Community psychological and behavioral responses through the first wave of 2009 pandemic (H1N1) in Hong Kong. *Journal of Infectious Disease*, 202(6), 867–76.
Crowley, P. (2000). Community development and primary care groups. In: Hobbs, A. (ed.) *Communities Developing for Health*. Liverpool: UK Health for All Network, pp. 41–2.
Daar, A.S., et al. (2007). Grand challenges in chronic non-communicable diseases. *Nature*, **450**, 494–6.
Dai, C. and Jiang, M. (2013). Fake meat scandals add to Chinese food fears. *British Medical Journal*, 346, f3385.
Dane, M. (2010). Overcoming barriers to plan implementation. [Online] Available at: http://suite101.com/article/overcoming-barriers-to-plan-implementation-a213233
Department for Work and Pensions (2008). The role of the workplace in health and well-being. In: *Working for a Healthier Tomorrow: Dame Carol Black's Review of the Health of Britain's Working Age Population*. London: Department for Work and Pensions, pp. 51–60. Available at: http://www.dwp.gov.uk/docs/hwwb-working-for-a-healthier-tomorrow.pdf
Department for Work and Pensions (n.d.). Fit note. [Online] Available at: https://www.gov.uk/government/organisations/department-for-work-pensions/series/fit-note
Department of Health (DH) (2006). *Tackling Health Inequalities: Status Report on the Program for Action 2006. Update on Headline Indicators*. London: DOH.
Department of Health (DH) (2011a). *Public Health in Local Government.* [Online] Available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/216708/dh_131904.pdf
Department of Health (DH) (2011b). The mental health strategy for England. [Online] Available at: https://www.gov.uk/government/publications/the-mental-health-strategy-for-england
Department of Health (DH) (2013a). *Statutory Guidance on Joint Strategic Needs Assessments and Joint Health and Wellbeing Strategies 2013*. [Online] Available at: https://s3-eu-west-1.amazonaws.com/media.dh.gov.uk/network/18/files/2013/03/Statutory-Guidance-on-Joint-Strategic-Needs-Assessments-and-Joint-Health-and-Wellbeing-Strategies-March-20131.pdf
Department of Health (DH) (2013b). *Public Health Outcomes Framework 2013 to 2016*. [Online] Available at: https://www.gov.uk/government/publications/healthy-lives-healthy-people-improving-outcomes-and-supporting-transparency
Department of Health (DH) (2013c). *A Framework for Sexual Health Improvement in England*. [Online] Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/142592/9287-2900714-TSO-SexualHealthPolicyNW_ACCESSIBLE.pdf
Department of Health (DH) (2017). *Towards a Smokefree Generation A Tobacco Control Plan for England.* [Online] Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/630217/Towards_a_Smoke_free_Generation_-_A_Tobacco_Control_Plan_for_England_2017-2022__2_.pdf
Department of Health (DH) (n.d.). *National Healthy School Status: A Guide for Schools.* [Online] Available at: http://www.salisbury.anglican.org/resources-library/schools/schools-every-child-matters/be-healthy
Department of Health and Social Care (DHSC) (2018). Childhood obesity: a plan for action, Chapter 2. [Online] Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/718903/childhood-obesity-a-plan-for-action-chapter-2.pdf
Diabetes UK (n.d.). What we do. [Online] Available at: http://www.diabetes.org.uk/About_us/What-we-do/
Etienne, C., Teruel, J., Galvao, L.A., Garza, S.M., and Martinelli, R. (2013). *Health in All Policies: Experiences from the Americas*. Washington, DC: Pan American Health Organization.

European Commission (2007). *Together for Health: A Strategic Approach for the EU 2008–13. White Paper*. [Online] Available at: http://ec.europa.eu/health-eu/doc/whitepaper_en.pdf

European Portal for Action on Health Inequalities (n.d.). *Health in All Policies (HiAP)*. [Online] Available at: http://www.health-inequalities.eu/HEALTHEQUITY/EN/policies/health_in_all_policies/

Evans, W.D. (2006). How social marketing works in health care. *British Medical Journal*, **332**, 1207.

Faculty of Public Health (n.d.). About the faculty of Public Health. [Online] Available at: http://www.fph.org.uk/about-fph

General Medical Council (n.d.). About us. [Online] Available at: http://www.gmc-uk.org/

Global Green and Healthy Hospitals. [Online]. Available at: https://noharm-global.org/issues/global/global-green-and-healthy-hospitals

Government of China (2009). Opinions of the CPC Central Committee and the State Council on deepening the health care system reform. [Online] Available at: http://www.china.org.cn/government/scio-press-conferences/2009-04/09/content_17575378.htm

Hanson, M., Godfrey, K.M., Lillycrop, K.A., Burdge, G.C., and Cluckman, P.D. (2011). Developmental plasticity and developmental origins of non-communicable disease: theoretical considerations and epigenetic mechanisms. *Progress in Biophysics and Molecular Biology*, 106(1), 272–80.

He, F. and MacGragor, G. (2009). A comprehensive review on salt and health and current experience of worldwide salt reduction programmes. *Journal of Human Hypertension*, **23**, 363–84.

Healthier Hospitals Initiative (2013). About. 2013. [Online] Available at: http://healthierhospitals.org/about-hhi

Healthy India (n.d.). Mid-day Meals. [Online] Available at: http://healthy-india.org/healthy-spaces/schools/mid-day-meals.html

Healthy Workplaces (2011). *Healthy Workplace Guide: Ten Steps to Implementing A Workplace Health Program*. [Online] Available at: https://ses.library.usyd.edu.au/bitstream/handle/2123/8776/Healthy%20Workplace%20Guide_Ten%20steps.pdf?sequence=1&isAllowed=y

Hrebiniak, L.G. (2005). *Making Strategy Work*. New Jersey: Wharton School Publishing.

Hill, A., Griffiths, S., and Gillam, S. (2007). *Public Health and Primary Care: Partners in Population Health, Volume 1*. Oxford: Oxford University Press.

HM Government (2010). *Healthy Lives, Healthy People: Our Strategy for Public Health in England. London: The Stationery Office*. [Online] Available at: http://www.official-documents.gov.uk/document/cm79/7985/7985.pdf

HM Government (2011). *No Health Without Mental Health: A Cross-Government Mental Health Outcomes Strategy for People of All Ages*. [Online] Available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/135457/dh_124058.pdf.pdf

HM Government (2016). *Childhood Obesity: A Plan for Action*. [Online] Available at https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/546588/Childhood_obesity_2016__2__acc.pdf

Hong Kong Academy of Medicine (n.d.). Overview. [Online] Available at: http://www.hkam.org.hk/

Hong Kong College of Community Medicine (n.d.). Background. [Online] Available at: http://www.hkccm.org.hk/.

Hong Kong Government (2011). *Department of Health. Action Plan to Reduce Alcohol-Related Harm in Hong Kong*. [Online] Available at: http://www.dh.gov.hk/english/pub_rec/pub_rec_ar/pdf/ncd_ap2/action_plan_whole_document_e.pdf

Hong Kong Government (2013). *Department of Health. Promoting Health in Hong Kong: A Strategic Framework for Prevention and Control of Non-communicable Diseases.* [Online] Available at: http://www.change4health.gov.hk/filemanager/common/image/strategic_framework/promoting_health/promoting_health_e.pdf

Hong Kong Government (n.d.). *Department of Health. Centre for Health Protection*. [Online] Available at: http://www.chp.gov.hk/en/index.html

Hong Kong Special Administrative Region Government (2003). *SARS in Hong Kong: From Experience to Action. Report of the SARS Expert Committee*. [Online] Available at: http://www.sars-expertcom.gov.hk/english/reports/reports.html

Hunter, D. (2007) *Managing for Health*. London: Routledge

Institute of Medicine (1988). *The Future of Public Health*. Washington, DC: Committee for the Study of the Future of Public Health.

Jin, M., Griffiths, S.M., Fong, H., and Dawes, M.G. (2013). Health of China's rural–urban migrants and their families: a review of literature from 2000–2012. *British Medical Bulletin*, **106**, 19–43.

Joossens, L. and Raw, M. (2011). *The Tobacco Control Scale 2010 in Europe: A Report of the Association of European Cancer Leagues*. Brussels: Association of European Cancer Leagues.

Kaplan, R.S. and Norton, D.P. (2001). *The Strategy-Focused Organization*. Boston, MA: HBS Press.

Kawamoto, K. (2004). Learning about health care policy: part 1. *The Journal of Education, Community, and Values*, **2004**, 4–5.

Kickbusch, I. and Buckett, K. (eds.) (2010). *Implementing Health in All Policies, Adelaide 2010*. Adelaide: Government of South Australia, Department of Health. [Online] Available at: http://www.who.int/sdhconference/resources/implementinghiapadel-sahealth-100622.pdf

Kim, J.H., Wong, A.H., Goggins, W.B., Lau, J., and Griffiths, S.M. (2013). Drink driving in Hong Kong: the competing effects of random breath testing and alcohol tax reduction. *Addiction*, 108(7), 1217–28.

Kuh, D. and Ben-Shlomo, Y. (1997). *A Life-course Approach to Chronic Disease Epidemiology*. Oxford: Oxford University Press.

Laspiur, S. (2012). Argentine initiative: to reduce sodium consumption 'Less Salt, More Life'. [Online] Available at: http://new.paho.org/panamericanforum/wp-content/uploads/2012/08/less-salt-more-life_PAHO-consortium_ARG.pdf

Lefebvre, R.C. and Flora, J.A. (1988). Social marketing and public health intervention. *Health Education & Behavior*, 15(3), 299–315.

Leppo, K., Ollila, E., Peña, S., Wismar, W., and Cook, S. (eds.) (2013). *Health in All Policies, Seizing Opportunities, Implementing Policies. Finland: Ministry of Social Affairs and Health.* [Online] Available at: http://www.euro.who.int/__data/assets/pdf_file/0007/188809/Health-in-All-Policies-final.pdf

Li, L.M., Tang, J.L., Lv, J., Jiang, Y., and Griffiths, S. (2011). The need for integration in health sciences sets the future direction for public health education. *Public Health*, 125(1), 20–4.

Mahal, A., Karan, A., and Engelgau, M. (2009). *The Economic Implications of Non-Communicable Disease for India. The International Bank for Reconstruction and Development*. Washington, DC: World Bank.

Marmot, M., on behalf of the Commission on Social Determinants of Health (2007). Achieving health equity: from root causes to fair outcomes. *The Lancet*, **370**, 1153–63.

Mental Health Commission of Canada (2012). *Changing Directions, Changing Lives: The Mental Health Strategy for Canada*. [Online] Available at: https://www.mentalhealthcommission.ca/sites/default/files/MHStrategy_Strategy_ENG.pdf

Mental Health Commission of Canada (2013). National Standard of Canada. January 2013.

Mental Health Taskforce to the NHS in England. (2016). *The Five Year Forward View for Mental Health*. February 2016. [Online] Available at: https://www.england.nhs.uk/wp-content/uploads/2016/02/Mental- Health-Taskforce-FYFV-final.pdf

Merson, M.H. and Page, K.C. (2009). *The Dramatic Expansion of University Engagement in Global Health. Implications for US Policy: A Report of the CSIS Global Health Policy Center*. Washington, DC: Center for Strategic & International Studies. [Online] Available at: https://www.ghdonline.org/uploads/Univ_Engagement_in_GH.pdf

Mintzberg, H. (1987). The strategy concept i: five Ps for strategy. *California Management Review*, 30(1), 11–24.

Mississippi Office of Healthy Schools (n.d.). Health education services. [Online] Available at: http://healthyschoolsms.org/

Mohammed, A. (n.d.). The global burden of NCDs. [Online] Available at: http://ncdaction.org/page/the-global-burden-of-ncds

National Center for Chronic Disease Prevention and Health Promotion (2010). Healthy Hospital Choices. [Online] Available at: http://www.cdc.gov/nccdphp/dnpao/hwi/docs/HealthyHospBkWeb.pdf

National Health Service (NHS) (2013). *The National Institute for Health Research Report*. [Online] Available at: http://www.nihr.ac.uk/files/pdfs/Briefing%20documents/1.1%20The%20National%20Institute%20for%20Health%20Research.pdf

National Health Service (NHS) (n.d.). Sustainable development strategy consultation. [Online] Available at: https://www.sduhealth.org.uk/policy-strategy/engagement-resources.aspx

Naylor, C.D., Chantler, C., and Griffiths, S. (2004). Learning from SARS in Hong Kong and Toronto. *Journal of the American Medical Association*, 291(20), 2483–7.

NEPAD (n.d.). Africa Health Strategy 2005–2007. [Online] Available at: http://www.nepad.org/

Pitcher, S., Jordan, D., and Buckett, K. (2010). South Australia's strategic plan and health in all policies. In: Kickbusch, I. and Buckett, K. (eds.) *Implementing Health in All Policies Adelaide 2010*, pp. 87–92. Adelaide: Government of South Australia, Department of Health. [Online] Available at: http://www.who.int/sdhconference/resources/implementinghiapadel-sahealth-100622.pdf

Public Health England (PHE) (n.d.). What we do. [Online] Available at: https://www.gov.uk/government/organisations/public-health-england

Rao, M., Prasad, S., Adshead, F., and Tissera, H. (2007). The built environment and health. *The Lancet*, 370(9593), 1111–13.

Raw, M. and Joossens, L. (2007). *Progress in Tobacco Control Policies in 30 European Countries, 2005 to 2007*. [Online] Available at: https://www.slzt.cz/dokumenty/LJ_30_European_countries.pdf

Rocco, L., Kimie, T., Suhrcke, M., and Fumagalli, E. (2011). *Chronic Diseases and Labor Market Outcomes in Egypt*. *World Bank Policy Research Working Paper*, No. 5575. Washington, DC: World Bank.

Save the Children (2012). *State of the World's Mothers 2012: Nutrition in the First 1,000 Days*. [Online] Available at: https://resourcecentre.savethechildren.net/library/state-worlds-mothers-2012-nutrition-first-1000-days

School Health Foundation of India (n.d.). Healthy School Certification. [Online] Available at: http://schoolhealth.in/uncategorized/healthy-school-certification/

Shanghai Declaration on Promoting Health in the 2030 Agenda for Sustainable Development. [Online] Available at: http://www.who.int/healthpromotion/conferences/9gchp/shanghai-declaration.pdf?ua=1

Sim, F., Lock, K., and McKee, M. (2007). Maximizing the contribution of the public health workforce: the English experience. *Bulletin of the World Health Organization*, 85(12), 901–80.

Starfield, B. (1998). *Primary Care: Balancing Health Needs, Services and Technology*. New York: Oxford University Press.
Tan, X., Zhang, Y., and Shao, H. (2018). Healthy China 2030, a breakthrough for improving health. *Global Health Promotion*, 26(4), 96–9.
The Lancet (2010). *Climate Change Series*. [Online] Available at: http://www.thelancet.com/series/health-and-climate-change
The Lancet (2013a). *Women Deliver. Interview with Ana Langer, Director of the Women and Health Initiative at the Harvard School of Public Health*. [Online] Available at: http://www.lancet.com/themed/women-deliver-2013
The Lancet (2013b). Tobacco control—political will needed. *The Lancet*, 381(9877), 1511.
The Medical Council of Hong Kong (n.d.). *Homepage*. [Online] Available at: http://www.mchk.org.hk/
The NCD Alliance (2013a). *NCD Alliance Report 2012–2013: Putting Non-Communicable Diseases on the Global Agenda*. [Online] Available at: http://ncdalliance.org/sites/default/files/resource_files/NCD%20Alliance%20Report%202012-2013.pdf
The NCD Alliance (2013b). *66th World Health Assembly—May 2013*. [Online] Available at: http://www.ncdalliance.org/wha66
UK Government (n.d.). Childhood obesity: a plan for action, chapter 2. [Online] Available at: https://www.gov.uk/government/publications/childhood-obesity-a-plan-for-action-chapter-2
United Nations (n.d. a). *Millennium Development Goals*. [Online] Available at: http://www.un.org/millenniumgoals/
United Nations (n.d. b). *2010 Summit on the Millennium Development Goals*. [Online] Available at: http://www.un.org/millenniumgoals/bkgd.shtml
United Nations (n.d. c). *Sustainable Development Goals*. [Online] Available at: https://sustainabledevelopment.un.org/sdgs
United Nations News Centre (2011). Remarks to General Assembly meeting on the Prevention and Control of Non-Communicable Diseases. Secretary-General Ban Ki-moon General Assembly. [Online] Available at: http://www.un.org/apps/news/infocus/sgspeeches/search_full.asp?statID=1299
United Nations Population Fund (n.d.). *Empowering Women through Education*. [Online] Available at: http://www.unfpa.org/gender/empowerment2.htm
United States Global Health (n.d.). *Strategy Objectives of the HHS Global Health Strategy*. [Online] Available at: https://www.hhs.gov/sites/default/files/hhs-global-strategy.pdf
van Weel, C., de Maeseneer, J., and Roberts, R. (2008). Integration of personal and community health care. *The Lancet*, **372**, 871–2.
Whitmee, S., et al. (2015). Safeguarding human health in the Anthropocene epoch: report of The Rockefeller Foundation—Lancet Commission on planetary health. *The Lancet*, 386(10007), 1973–2028.
WHO Health Promoting Hospitals Network (1997). *The Vienna Recommendations on Health Promoting Hospitals*. Copenhagen: WHO Europe.
World Bank (2005). *Brail: Addressing the Challenge of Non-communicable Diseases in Brazil*. Washington, DC: World Bank.
World Bank (2011). Toward a healthy and harmonious life in China: stemming the rising tide of non-communicable diseases. [Online] Available at: http://www.worldbank.org/en/news/feature/2011/07/26/toward-health-harmonious-life-china-stemming-rising-tide-of-non-communicable-diseases
World Bank (2013). Improving nutrition through multisectoral approaches. [Online] Available at: http://www-wds.worldbank.org/external/default/WDSContentServer/WDSP/IB/2013/02/05/000356161_20130205130807/Rendered/PDF/751020WP0Impro00Box374299B00PUBLIC0.pdf

World Health Organization (WHO) (1948). *WHO Definition of Health*. [Online] Available at: http://www.who.int/about/definition/en/print.html

World Health Organization (WHO) (1978). Declaration of Alma-Ata. International Conference on Primary Health Care, Alma-Ata, USSR, 6–12 September 1978. [Online] Available at: http://www.who.int/publications/almaata_declaration_en.pdf

World Health Organization (WHO) (1986). The Ottawa Charter for Health Promotion. *First International Conference on Health Promotion, Ottawa*, 21 November 1986. [Online] Available at: http://www.who.int/healthpromotion/conferences/previous/ottawa/en/

World Health Organization (WHO) (2004). *Global Strategy on Diet, Physical Activity and Health. Strategy Documents. May 2004: Final Strategy Document and Resolution*. [Online] Available at: http://www.who.int/dietphysicalactivity/strategy/eb11344/en/

World Health Organization (WHO) (2005). *The Bangkok Charter for Health Promotion in a Globalized World*. [Online] Available at: http://www.who.int/healthpromotion/conferences/6gchp/bangkok_charter/en/

World Health Organization (WHO) (2008). *The World Health Report 2008—Primary Health Care: Now More Than Ever.* [Online] Available at: http://www.who.int/whr/2008/en/

World Health Organization (WHO) (2009a). *Noncommunicable Diseases and Mental Health. 2008–2013 Action Plan for the Global Strategy for the Prevention and Control of Noncommunicable Diseases.* [Online] Available at: http://www.who.int/nmh/publications/9789241597418/en/

World Health Organization (WHO) (2009b). *62nd World Health Assembly: Daily Notes on Proceedings.* [Online] Available at: http://www.who.int/mediacentre/events/2009/wha62/journal/en/index.html

World Health Organization (WHO) (2010). *A Framework for National Health Policies, Strategies and Plans June 2010*. [Online] Available at: http://www.who.int/nationalpolicies/FrameworkNHPSP_final_en.pdf

World Health Organization (WHO) (2013a). *Global Action Plan for the Prevention and Control of Noncommunicable Diseases 2013–2020. Revised Draft.* [Online] Available at: http://www.who.int/nmh/events/2013/revised_draft_ncd_action_plan.pdf

World Health Organization (WHO) (2013b). *Draft Action Plan for the Prevention and Control of Noncommunicable Diseases 2013–2020*. [Online] Available at: http://www.who.int/nmh/publications/ncd_action_plan2013.pdf

World Health Organization (WHO) (2014). *Noncommunicable Diseases and Mental Health: About 9 Voluntary Global Targets.* [Online] Available at: https://www.who.int/nmh/ncd-tools/definition-targets/en/

World Health Organization (WHO) (2015). *Health in 2015: from MDGs, Millennium Development Goals to SDGs, Sustainable Development Goals*. [Online] Available at: http://apps.who.int/iris/bitstream/handle/10665/200009/9789241565110_eng.pdf;jsessionid=FD9F8C6CDA1B640C2A95D889BEA62113?sequence=1

World Health Organization (WHO) (2017a). *Montevideo Roadmap 2018–2030 on NCDs as a Sustainable Development Priority*. [Online] Available at: http://www.who.int/conferences/global-ncd-conference/Roadmap.pdf

World Health Organization (WHO) (2017b). *One Health.* [Online] Available at: http://www.who.int/features/qa/one-health/en/

World Health Organization (WHO) (2018). *Strategic Plan for Consolidation and Stabilization Following the Epidemic of Ebola Virus Disease in the Democratic Republic of the Congo*. August–October 2018. [Online] Available at: http://www.who.int/emergencies/crises/cod/drc-consolidation-and-stabilization-plan-26July2018-en.pdf?ua=1

World Health Organization (WHO) (n.d. a). *Food Security.* [Online] Available at: http://www.who.int/trade/glossary/story028/en/

World Health Organization (WHO) (n.d. b). *Global Strategy to Reduce Harmful Use of Alcohol*. [Online] Available at: http://www.who.int/substance_abuse/activities/gsrhua/en/
World Health Organization (WHO) (n.d. c). *WHO Framework Convention on Tobacco Control*. [Online] Available at: http://www.who.int/fctc/en/
World Health Organization (WHO) (n.d. d). *MPOWER. Tobacco Free Initiative*. [Online] Available at: http://www.who.int/tobacco/mpower/en/
World Health Organization (WHO) (n.d. e). *Track 2: Health Literacy and Health Behaviour. 7th Global Conference on Health Promotion: Track Themes.* [Online] Available at: http://www.who.int/healthpromotion/conferences/7gchp/track2/en/index.html
World Health Organization (WHO) (n.d. f). *Introduction to Healthy Settings*. [Online] Available at: http://www.who.int/healthy_settings/about/en/
World Health Organization (WHO) (n.d. g). *Healthy Settings.* [Online] Available at: http://www.who.int/healthy_settings/en/
World Health Organization (WHO) (n.d. h). *Healthy Workplaces: A WHO Global Model for Action.* [Online] Available at: http://www.who.int/occupational_health/healthy_workplaces/en/
World Health Organization (WHO) (n.d. i). *Ageing and Life-Course. Care and Independence in Older Age*. [Online] Available at: http://www.who.int/ageing/en/
World Health Organization Europe Office (n.d.). *Environment and Health.* [Online] Available at: https://www.euro.who.int/en/health-topics/environment-and-health
World Health Organization Regional Office for the Western Pacific (1999). *Regional Guidelines for the Development of Healthy Workplaces.* [Online] Available at: http://www.who.int/occupational_health/regions/en/oehwproguidelines.pdf
Yang, J., et al. (2018). The Tsinghua-Lancet Commission on Healthy Cities in China: unlocking the power of cities for a healthy China. *The Lancet*, 391(10135), 2140–84.

11.8

Strategies for health services

Chien Earn Lee and Fran Baum

Introduction to strategies for health services

Two crucial questions this chapter needs to answer are: firstly, why public health professionals should be interested in health services, which leads to the second question: do health services actually make a meaningful contribution to population health?

Writers such as McKeown (1979) contended that the main drivers of improvements in health had been nutrition, environment, and behaviour rather than healthcare. He showed that three-quarters of the decline in mortality in England and Wales between 1841 and 1971 had been due to a reduction in deaths from infectious disease and that three-quarters of this reduction had preceded the widespread introduction of immunization or antibiotics.

Others such as Illich (1976) argued that instead of improving health, healthcare has a direct and adverse effect on health, for example, side effects of prescribed drugs, hospital-acquired infections, poorly performed surgery, and the consequences of following up spurious abnormalities from the increasing numbers of laboratory investigations.

However, as several commentators have noted, McKeown was describing a period in which healthcare still had relatively little to offer. It was only in the 1960s and 1970s that safe and effective drugs for many chronic diseases became widely available (Nolte and McKee 2004). Contrary to the situation before the 1960s, health services are now making an important contribution to overall improvements in life expectancies in industrialized countries. Many new treatments have been introduced that have been shown, in high-quality evaluative research, to prolong life. Examples include effective

treatment for hypertension and heart failure, secondary prevention following myocardial infarction, and chemotherapy for many childhood cancers.

International differences provide many natural experiments to examine this issue further (McKee and Bojan 1998). For example, prior to 1990 the death rates from causes amenable to medical care (i.e. deaths from causes that are potentially preventable with timely and effective healthcare) were higher in former communist countries of Central and Eastern Europe than in the West (Bojan et al. 1991; Boys et al. 1991). Excluding early neonatal deaths, for which only incomplete data are available, Velkova et al. (1997) estimated that amenable causes accounted for 24% of the East–West gap in male life expectancy between birth and the age of 75 years in 1988. The corresponding figure for females was 39%.

There is now considerable evidence that, unlike the situation prior to the 1970s, health services, although only one of the determinants of health, others being genetic predisposition, individual behaviour and lifestyle, social, economic and commercial determinants (see Section 10) and environmental circumstances, can make a substantial difference to population health (McKee and Figueras 2009).

The Ottawa Charter for Health Promotion (WHO 1986) recommended the reorientation of health services and their resources, beyond its responsibility for providing clinical and curative services, towards the promotion of health. The Charter also emphasized the need to work towards enhancing life skills to increase the options available for people to exercise more control over their own health and over their environments and to make choices conducive to health. More recently, the World Health Organization (WHO) has added to the work of the Charter and argued that government should adopt a Health in All Policies approach (Leppo et al. 2013) which involves all sectors in examining and then seeking to maximize their health input so that they govern for health (Baum 2019).

Public health professionals have a key advocacy role to ensure that resources in hospitals (which invariably receive the lion’s share of the healthcare budget) are used effectively, efficiently, and equitably to optimize outcomes, and that community-based services which are less urgent but no less important than acute services are not neglected. The capabilities of public health professionals in developing systems-based solutions and mobilizing intersectoral action will need to grow to improve population health

The remainder of this chapter explores the changing context of healthcare services, the emerging themes in healthcare services, and strategies for delivering and improving healthcare services

The context for healthcare services

Healthcare services do not exist in a vacuum. While this statement seems self-evident, it has often not been adequately taken into account in developing and evolving healthcare strategies. As a result, strategies have either not taken root (especially those related to 'transplanted' best practices from other systems) or like the proverbial frog being slowly boiled alive, failed to respond in a timely manner to meet new challenges. Contextual factors that need to be considered include societal and professional norms and values as they influence the central principles of the system as well as the desire for reform. Economic factors such as the level of economic development would affect the level of disposable resources available for health services (Maxwell 1981), while the upward pressure on healthcare expenditures has been the key driver for healthcare reforms around the world. Healthcare systems must also be able to respond proactively to factors affecting population health needs such as the changing demographics and patterns of disease as well as the impact of technological change. The role of education in producing sufficient professional and managerial capability will also influence the health system's capacity for serious and sustainable action. These contextual factors are discussed here under the headings of defining norms and values, the reality of macroeconomics, demographic pressures, changing disease patterns, technological changes, the quest for quality and reducing waste, and rising patient expectations.

Norms and beliefs

Healthcare systems act as mirrors that reflect deeply rooted social and cultural expectations of the population that they serve (Contandriopoulos et al. 1998). Although these norms and beliefs are generated outside the formal structure of the healthcare system, they play a major role in defining the system's overall characteristics.

Societies have dominant belief systems where the daily tensions between differing values and beliefs in a society are held in relative stability (Benson

1975). A useful approach to understanding belief systems sees these tensions as grouped around four poles (Habermas 1987): values, understanding of phenomena, definition of jurisdictions and allocation of resources, and logic of regulation.

In this framework, values include the tensions and trade-offs between equity, individual autonomy, and efficiency (Clark 1988). Understanding of phenomena relates to how concepts such as life, death, sickness, health, and pain are interpreted and thus viewed as relating to the objectives of a healthcare system (Gillett 1995). Definition of jurisdictions and allocation of resources comprise the perceptions of the role and functions of those working in the healthcare sector, as well as the allocation of resources between prevention and cure and between healthcare and broader determinants of health. The logic of regulation relates to how society chooses to regulate the delivery of healthcare. This may be technocratic with trained experts guiding the system on the basis of their knowledge and position within the hierarchy; professional self-regulation which has the physician (as the best agent of the patient) at the centre of the system; the market-based model in which regulations reflect supply and demand in a competitive market; or the democratic model, in which the population, either directly or, more often, through elected or appointed representatives, is responsible for setting out the framework for delivery of healthcare.

An understanding of the dominant belief system in a society is important for public health professionals as it contributes to knowledge of why systems are as they are, how they have changed, and the objectives that individuals within the healthcare system are pursuing. It will also influence the choice of strategies that should be adopted to bring about meaningful and sustained change (McKee and Figueras 2009). For example, some societies view healthcare as predominantly a social and collective good. Such societies would more readily accept policies designed to create intentional cross-subsidies, from young to old, rich to poor, and from healthy to sick, than societies that view healthcare as a commodity that can be bought and sold in the market.

Norms and beliefs can and do evolve over time. For example, years of spectacular growth have seen more people pulled from abject poverty in modern Asia than at any other time in history. But as they become more affluent, the region’s citizens want more from their governments. Across the continent, pressure is growing for public pensions, national health insurance,

unemployment benefits, and other hallmarks of social protection. As a result, the world's most vibrant economies are shifting gear, away from simply building wealth towards building states with better and more comprehensive welfare provisions (*The Economist* 2012). In addition, the United Nations have adopted the goal of Universal Health Coverage (UHC) as part of the Sustainable Development Goals designed to be achieved by 2030 (WHO 2020). The current Director General of the WHO Dr Tedros Adhanom Ghebreyesus said that achieving 'Universal health coverage is a political choice' (WHO 2019). While such changes increase the demands on healthcare systems, they also provide opportunities to introduce changes and strategies that may not have been acceptable based on previous norms and beliefs.

Macroeconomic factors

The historically high rates of economic growth seen during the 1960s and 1970s have not been sustained in the 1990s. This has been compounded by economic volatility with shorter boom and bust cycles with globalization. With economies across the world becoming more and more connected to one another, adverse economic conditions in one country can rapidly spread to other countries and escalate to global proportions, as was seen in the global financial crisis in the late 2000s.

At the same time, the expansion of healthcare coverage both in terms of scope of services and population covered is adding significantly to the upward pressures on healthcare expenditure. There are already concerns that the current level of expenditure on healthcare is not sustainable and is adversely affecting other sectors of the economy. For example, the United States spent more than 17% of its gross domestic product on health, raising concerns about the heavy burden placed on companies doing business in the United States which can put them at a substantial competitive disadvantage in the international marketplace.

The macroeconomic environment influences the political decision-making process, in particular, the allocation of resources to competing sectors. The growth in health spending slowed or fell in real terms in 2010 in almost all Organisation for Economic Cooperation and Development (OECD) countries, reversing a long-term trend of rapid increases. Overall, health spending grew by nearly 5% per year in real terms in OECD countries over

the period 2000–2009, but this was followed by zero growth in 2010. Preliminary figures for a limited number of countries suggest little or no growth in 2011. Countries like the United Kingdom that invested heavily in healthcare in 1990s are having to scale back in the face of the tightening economic situation. The UK government aims to deliver £20 billion (4%) efficiency savings in the National Health Service (NHS) by the end of 2014–2015. The challenge at present is how to balance increased demand from patients with the requirement for productivity and efficiency, while maintaining excellent, quality care.

Demographic pressures

Dr Gro Harlem Brundtland, former Director General of the WHO, called ageing one of humanity's great achievements and also one of its greatest challenges. In rich countries, the number of people over 65 who will require medical care, is expected to increase by between 50% and 120% from 1995 to 2025 and the elderly incur disproportionately more healthcare cost than the young. For example, in the United States, the elderly account for less than 15% of the total population but incur more than one-third of healthcare spending (and this is before taking into account social or long-term care costs).

The greatest challenge lies in the developing countries. While developed countries grew affluent before they became old, developing countries are growing old before they get affluent. The demographic shift towards an older population that took place gradually in Europe over a century, is taking place in the developing countries at a pace which far outstrips these countries' socioeconomic development. The impact of ageing on demand for healthcare is often compounded by falling birth rates, leading to an increasing dependency ratio.

It is, however, necessary to enter three caveats. Firstly a study of American Medicare data showed that while estimated lifetime cost increases with age, the cost associated with an additional year of life and the average annual cost over a person's lifetime both decreased as the age at death increased (McGinnis and Foege 1993; Lubitz et al. 1995). This could be due to less intensive treatment the older the patients are in their respective last year of life (Scitovsky 1988). This suggests that healthcare cost will be more

substantially affected by the expected increase in the absolute number of elderly people than increased longevity beyond the age of 65.

Secondly, the health and social profile of the elderly of the future may be different from current elderly. They are likely to be better educated and would have benefited from a lifetime of better nutrition. On the other hand, they are also likely to have adopted a more sedentary lifestyle and have higher expectations of healthcare. For these reasons, we should not make simplistic extrapolations of demand based on current data of the elderly to a future population.

The final caveat is that pessimistic predictions of the costs of ageing are based on the assumption that health services will continue to expand in their current form which is highly hospital centric. The impact of ageing on disease burden and costs can be moderated by greater emphasis on prevention of disease and disability and transforming models of care for the elderly.

Changing healthcare workforce

Specialization in the pursuit of greater knowledge and expertise has characterized scientific progress in medicine for nearly two centuries (Cassel and Reuben 2011). A more recent development is the creation of subspecialties, recognizing that there are some specific populations of patients who would benefit from the highly focused knowledge and skills obtained by additional training and certification beyond that of a ‘general’ specialist. As these tertiary subspecialists require high patient volume to maintain proficiency in their skills, they are usually based at referral centres.

Many new and emerging subspecialties are cross-disciplinary. In the United States, sleep medicine crosses six different specialties, and palliative medicine is a subspecialty option for ten different primary specialties. The proliferation of specialties has led to concerns that medical care would become increasingly fragmented with the concomitant loss or undervaluing of the generalist physician with decreased attractiveness as a career choice for young doctors.

As trade barriers come down, the forces of globalization have increased the movement of all types of professionals throughout the world, and healthcare is no exception. Richer countries are increasingly relying on immigration as a means of coping with domestic shortages of healthcare

professionals. This trend has led to concerns that the outflow of healthcare professionals is adversely affecting the healthcare system in many of these source countries. Insufficient human resources for health in developing countries have been identified as one of the main constraints limiting progress on such initiatives as 3by51 and the Millennium Development Goals (MDGs) (Institute of Medicine (IOM) 2004). Recipient countries on the other hand need to develop the means to effectively assimilate the influx of healthcare professionals from diverse backgrounds and cultures.

Countries are also exploring expanded scope of practice for healthcare professionals as a means of addressing access-to-care problems. Nurse practitioners, physical and occupational therapists, and other allied healthcare professionals are growing in independence from physicians' oversight. For example, several states in the United States have passed legislation allowing pharmacists limited ability to evaluate and manage drug regimens as well as dispense drugs.

Changing burden of disease

The term 'epidemiological transition' refers to the complex changes in patterns of health and disease as societies develop. It describes the interactions between these patterns and their demographic, economic, and sociological determinants and consequences.

In the classical (Western) model, Omran (1971) described the transition from the Age of Pestilence and Famine with its high mortality and low average life expectancy (20–40 years), through the Age of Receding Pandemics, to the Age of Degenerative and Man-Made Diseases with its relatively low levels of mortality (average life expectancy of more than 50 years) and morbidity overshadowing mortality as an index of health.

Although the classical model assumes linear unidirectional change, counter transitions have also been observed where age-specific mortality rates rise rather than fall over time due to causes such as the HIV/AIDS pandemic in sub-Saharan Africa (Gaylin and Kates 1997). There is also an emerging awareness of old problems that never went away resulting in an overlap of eras, often with large disparities emerging between and within countries. On the other hand, although chronic diseases such as diabetes have usually been considered as diseases of affluence, four out of five people with diabetes now live in low- and middle-income countries

(International Diabetes Federation 2012). In short, health systems in these countries must continue to grapple with the unfinished agenda of communicable diseases while addressing the new agenda of non-communicable ones in an ageing population.

As the burden of disease is dynamic (Murray and Lopez 1996), the configuration of health services must evolve accordingly to meet the changing health needs of the population. Health services cannot remain trapped in structures designed for a time when disease patterns were different. For example, the increasing incidence and prevalence of chronic diseases requires health systems to organize health services with greater emphasis on prevention and community-based services instead of the current hospital-centric model which is more suited for acute episodic care.

Current care delivery models are, however, often overly complex and uncoordinated, requiring many steps and patient 'hand-offs' that slow down care delivery and decrease safety. Healthcare organizations, hospitals, and physician groups also typically operate as separate 'silos', acting without the benefit of complete information about the patient's condition, medical history, services provided in other settings, or medications provided by other clinicians (IOM 2001).

Public health professionals have a central role in tracking and explaining trends in diseases (including predicting emerging trends), designing mechanisms to address them, supporting the changes in service delivery that are required, and monitoring responses to these changes. With appropriate interventions, Olshansky and Ault (1986) have postulated the 'Age of Delayed Degenerative Diseases' as the fourth stage of the epidemiological transition.

Technological changes

The pace of change in healthcare is steadily accelerating. Hip replacements have been joined by knees, shoulders, and finger joint replacements. Transplant surgeons have added heart, liver, and pancreas transplants to their initial successes with kidneys. AIDS has been transformed from a rapidly progressive fatal disease to one in which increasing numbers of patients are keeping the disease under control with complex cocktails of antiviral therapy. The list is almost endless.

The growth of healthcare technology is widely held to have contributed substantially to the upward pressure on healthcare expenditure. For example, the rapid growth in cardiology services throughout the developed world is driven more by angioplasty and other technological advances in care than by changes in the prevalence of heart disease (Dash et al. 2009). Although its precise contribution is controversial (Mossialos and Le Grand 1999), a study by the Health Care Financing Administration estimated that approximately half of the growth in real per capita healthcare costs in the United States is attributable to the introduction and diffusion of new medical technology for the 1940–1990 period, within an estimated probable range of 38–62% of growth (Smith et al. 2000).

There are various reasons for this increase. New technologies are often more expensive than the ones they replaced. Even where the actual technology is less expensive, it may lead to increased costs as other aspects of the service are reorganized to reflect changing patterns of treatment. For example, the introduction of new treatments may lead to an expansion in the number of individuals with indications for treatment, either because a previously untreatable condition becomes treatable or, as side effects or contraindications are reduced, the threshold for treatment falls. Finally, the diffusion of technology from tertiary centres, in some cases into primary care, can markedly reduce barriers to access and thus increase uptake.

Since the 1980s, there has also been an unparalleled revolution in the pace and volume of communication. The development of new ways of storing and delivering knowledge has transformed the way healthcare services are delivered and used, bringing benefits but also risks. It enables the separation of assessment from service provision allowing healthcare professionals to seek and provide advice without the need for face-to-face consultation. Telemedicine has led to the expansion of roles and the need for new competencies in healthcare workers. It also provides opportunities to optimize the value of highly trained subspecialists by making their expertise accessible to a wider patient population through physician-to-physician telemedicine approaches.

Quest for quality and reduction of waste

In tandem with the rapid growth in medical science and technology, healthcare is becoming more complex with more to know, more to do, more

to manage, more to watch, and more people involved than ever before (IOM 2001). Healthcare systems have, however, struggled to translate knowledge into practice. For example, it takes an average of 17 years for research evidence to reach clinical practice (Green et al. 2009). Small-area variation studies by Wennberg (1984) showed that decisions by clinicians in different areas regarding patients presenting with identical complaints could vary by as much as 40%.

Geographical variation of expenditure has also been identified. For example, in 2004, per capita spending on health ranged from roughly $4,000 in Utah to $6,700 in Massachusetts in the United States (Congressional Budget Office 2008). Higher spending does not necessarily mean better outcomes. The primary reason for such variations has been previously cited as clinical uncertainty about the most appropriate treatment in any given circumstance (Anderson and Mooney 1990). This position is, however, becoming increasingly untenable as medical knowledge advances and evidence-based practice guidelines are developed for many conditions.

Healthcare systems are also grappling with the consequences of unsafe care. A study in *The New England Journal of Medicine* (Landrigan et al. 2010) reported that 25% of hospitalized patients are harmed by medical errors. IOM's landmark study 'To Err is Human' (1989) estimated that there were 98,000 annual deaths from medical errors in the United States. The IOM's follow-up publication 'Crossing the Quality Chasm' has identified six dimensions where healthcare needs to improve:

Safe—reducing harm to patients.
Effective—providing appropriate care based on scientific knowledge to avoid underuse or overuse of services.
Patient-centred—providing care that is based on what matters to patients taking into account their preferences, needs, and values.
Timely—reducing unnecessary delays that may compromise care outcomes.
Efficient—avoiding waste not just of material resources but also ideas and energy.
Equitable—providing consistent quality of care regardless of personal characteristics such as gender, ethnicity, geographic location, and socioeconomic status.

Inappropriate care and errors waste precious resources. Waste occurs not just at the individual clinician–patient interface but as a result of poor

coordination between providers and across subsectors, inadequate management of capital resources, and operational budgets. The IOM Committee that produced the report 'Best Care at Lower Cost: The Path to Continuously Learning Health Care in America' estimated that there was $750 billion in unnecessary health spending in 2009 in the United States alone. The IOM Committee recommended a more nimble healthcare system that is consistently reliable and that constantly, systematically, and seamlessly improves. Healthcare systems must learn by avoiding past mistakes and adopt newfound successes.

The International Cochrane Collaboration has played a major role in providing evidence of what healthcare interventions can (and cannot) achieve by undertaking systematic reviews of available evidence in the literature using standardized methods, and disseminating the information (Chalmers and Altman 1995). Unfortunately such information is not always translated into daily practice as it challenges traditional models based on clinical autonomy.

Increasing patient expectations

The development of new and more expensive technologies coupled with the ready access to information, particularly with the growth of the internet, has led to higher and growing public expectations which may be difficult to meet.

Increasingly people are also taking over responsibility for their own care aided by online advice and other sources of information. The public increasingly expects healthcare systems to be responsive at all times (24 hours a day, every day) and that access to care should be provided over multiple channels such as the internet, by telephone, and by other means in addition to face-to-face visits. Patients expect healthcare services to make information available to patients and their families that allow them to make informed decisions when selecting a health plan, hospital, or clinical practice, or choosing among alternative treatments.

Through the internet, patients draw on the perspectives not only of physicians but also other sufferers from their condition. However, it comes with risks too. Those searching the web must be aware of its unregulated nature, with posts of views ranging from Nobel laureates to quacks (Eysenbach et al. 1998; Sandvik 1999). Internet prescribing may offer price

savings for some, but it may also circumvent regulatory safeguards designed to reduce risks to patients.

Emerging themes in health services

This section explores some of the themes that are emerging in response to the changing context of health services. The evolving macroeconomic environment coupled with the demand for health services brought about by changing demographics, epidemiology, and patient expectations have led to a reassessment of the relationship between the state and the market as well as the way health services are organized and financed. Developments in technology underpinned by changes in norms and beliefs are challenging traditional hierarchies and stimulating new ways of delivering health services with greater involvement of patients in their own care. These themes will be considered in turn.

State and market

During the 1980s and 1990s, many countries began to re-examine the role of the state in the regulation, funding, and delivery of health services. Countries where the state has played a central role are exploring new pluralistic solutions with involvement of a range of actors at varying distances from the state. Some functions are being devolved to local government or to statutory agencies at arm's length from government or transferred altogether to private ownership. Conversely, countries in which the state has traditionally played a less central role, acting only as a source of regulation between competing private interests, are now reassessing the need to expand their role and intervene more actively (Saltman and Figueras 1997).

It would be quite wrong, however, to see these developments as a process of convergence to some ideal model of a health system. Instead, the situation in each country represents the complex interplay of the economic, social, demographic, epidemiological, managerial, and technological forces described earlier and deep-seated cultural beliefs. The debate is often characterized as a dichotomy between the state and the market but this is a gross oversimplification (Saltman and von Otter 1995). In some instances elements of both have been combined albeit in differing proportions to allow

an increased use of market-style incentives with continued ownership and operation of facilities by the public sector in quasi- or internal markets.

The introduction of market mechanisms into the health sector creates a series of conceptual and practical challenges. Conceptually, healthcare is considered in many societies as a social good, in which the provision of services to each individual is also valuable to the community as a whole, whether because of ideas of self-interest, as in the treatment of dangerous mental illness or contagious disease, or altruism. Such principles have led to the creation of systems to ensure universal coverage and solidarity. In contrast, market incentives are intrinsically based on the assumption that services provided are a commodity, with demand driven by individual preference (McKee and Figueras 2009).

Another difficulty is that the neoclassical concept of a market is based on the relationship between a buyer and a seller. However, healthcare systems typically comprise a more complex set of relationships involving the patient, various health professionals, a provider institution, and a payer, with health professionals acting as both suppliers of services and agents for the patient (Young and Saltman 1985). Finally, the assumption in neoclassical economic theory that cross-subsidies between different patients are inefficient is incompatible with the social function of the healthcare system where there are substantial cross-subsidies between young and old, rich and poor, and healthy and sick.

The key practical challenge in introducing market mechanisms into healthcare is the development of real prices. This is a tedious and expensive process which requires a robust costing system to compute the real cost (including overheads) at the service as well as patient level. Some attempts to implement market mechanisms have floundered as the transaction costs and the costly duplication of facilities to secure a competitive advantage (Saltman and von Otter 1995) outweighed the improvements in operational efficiency. Another reason for the failure of market mechanisms in healthcare is that competition operates at the wrong level. Porter and Teisberg (2006) argue that currently competition is both too broad and too narrow. Competition is too broad because much of the competition now takes place at the level of health plans, networks, hospital groups, physician groups, and clinics. Instead, it should occur in addressing particular medical conditions. Competition is too narrow because it now takes place at the level of discrete interventions or services. Instead it should take place at the

patient level, addressing medical conditions over the full cycle of care, including monitoring and prevention, diagnosis, treatment, and the ongoing management of the condition.

Some other attempts have, however, been more successful. In these systems, the introduction of market principles does not mean that the government intervenes less but it intervenes differently. Even as we are familiar with market failures in healthcare, there can be government failure too—which can occur in a state- or market-based system. The manner of government intervention matters more than the fact of intervention. The purported failure of market mechanisms may at times be due more to poor implementation than the mechanism itself. Oftentimes markets fail in healthcare, because they are allowed to fail. For markets to work, government intervention such as setting the ground rules for competition and providing clear standards for market participants with tight monitoring and evaluation as well as making prices transparent is needed.

A related issue is the changing approach to regulation. Governments need to find the appropriate balance between regulatory and competitive measures —which will change with time—depending on the context in which they are implemented. Regulation is becoming more sophisticated with greater recognition that traditional ‘command and control’ measures need to be complemented by a more extensive regulatory toolkit (which includes ‘softer’ tools such as education and community engagement). Governments also need to be cognizant of the risk of perverse incentives with poorly drafted regulations.

Many countries are beginning to question whether self-regulation of healthcare professionals can retain credibility as a system that serves the best interests of the public. Self-interest of the profession plays out in various ways, at times negatively. For example, one professional group may use the regulatory structure to exercise control and influence over another group, or to keep hold of their monopoly on very profitable services. The key question for policymakers, then, is one of governance: to what extent are healthcare professional regulators sufficiently independent of the groups they regulate —and to what extent do they act in the best interests of patients and the public? Several countries have moved to ‘shared regulation’ whereby the government and other credible experts from outside the regulated profession sit with members of the profession around the table. The proportion of

places around the table is a matter for each country to determine, based on local context (Dzau et al. 2012).

Governments that have successfully incorporated market principles rely on competition and market forces to improve service delivery and efficiency, while retaining the prerogative of the state to intervene directly when the market failed to keep healthcare costs down (Atwal 2010). Markets are necessary to help government work better, and good government is necessary to help markets work better (Menon 2010). In fact some researchers have concluded that governments have to be more competent to supervise contracting and other market-style arrangements than to run services directly (Kettl 1993).

Reorganizing the system: decentralization, recentralization, and privatization

Decentralization has been defined as the transfer of authority from a high, or national, level to a lower, or subnational, level (Rondinelli 1981). Decentralization has multiple meanings, interpretations, and implementation in different countries and different contexts. The common essence of decentralization lies in strengthening local authorities through transfer of power and resources from the central government.

Decentralization has become attractive because of disillusionment with the perceived failings of large centralized, bureaucratic institutions. These include inefficiency, lack of responsiveness, and failure to innovate. Decentralized institutions are seen to have several advantages. Those working in them are better equipped to identify changing circumstances and needs, and thus can respond more flexibly. They can be more innovative than centralized institutions and they can generate higher morale, more commitment, and greater productivity. Decentralization can also facilitate enhanced partnerships with the local population, thus increasing the democratic input into policymaking.

The experience of many countries in recent years demonstrates that there are certain areas where decision-making power should not be decentralized. Four such areas include:

Basic framework for health policy, strategic decisions on the development of health resources such as capital, personnel, expensive equipment.
Research and development as they shape the future of the entire system.
Regulations concerning public safety.
Monitoring, assessment, and analysis of the health of the population and healthcare provision (as these are tools for influencing the behaviour of decentralized units).

Borgenhammer (1993) has also identified certain social and cultural conditions that are necessary for successful decentralization. These include adequate local administrative and managerial capacity, ideological certainty in the implementation of tasks, and readiness to accept several interpretations of a particular problem. Organizations undertaking decentralization must engage in a process of articulation of their mission, create internal cultures around core values, and establish mechanisms to monitor results (Osborne and Gaebler 1993).

With respect to health services, decentralization enables better access with people benefiting from the convenience of having core health services close to where they live. However, decentralization can have negative effects (Collins and Green 1994). These include fragmentation of services, loss of critical mass in services, and unacceptable variations in practice (Borgenhammer 1993). For health services there is substantial clinical evidence that centralizing certain complex conditions and procedures, specialization with co-location of specialist staff and services in facilities to generate a critical volume of cases has led to better clinical outcomes and higher productivity. The aim then is to ‘decentralize where possible, centralize where necessary’ (NHS 2007). This entails the creation of a primary and community care infrastructure that is able to deliver integrated services close to patients’ homes with consolidation of centres that perform only a small number of procedures (Dash et al. 2009).

Privatization requires separate consideration as it is fundamentally different, in several important aspects, from other types of decentralization. Its advocates argue that it brings benefits from the introduction of economic incentives for greater efficiency and higher quality in the management of healthcare (Cabinet Office 1999). It is also attractive to some governments as a means of attracting private capital into the health sector, thus reducing demands on scarce public funds.

These benefits, however, come with considerable risks. Private management and capital investment require financial returns that are consistent with those obtainable in other private markets. The pressures involved in achieving these returns can result in adverse selection (e.g. practices that intentionally discriminate against the poor and other vulnerable groups). Activities that do not generate immediate returns such as primary and secondary prevention activities are likely to be relegated to lower priority (Himmelstein et al. 1999). If privatization is not to undermine wider health policies, privatization must be accompanied by strengthened central regulatory controls. This increases administrative cost for both regulators and providers.

Two broad findings emerge from the debate on decentralization. The first is whether the advantages of decentralization outweigh the disadvantages, such as fragmented and duplicated services and high transaction costs, an issue that must be resolved on a case-by-case basis (Saltman and Figueras 1997). Secondly, a policy of decentralization requires a series of measures to create an appropriate framework, including development of a shared vision and a means of monitoring outcomes.

Ecosystem approach to healthcare

Health status is largely determined by the interaction of four linked factors: genetic susceptibility, behaviour and lifestyle, socioeconomic status, and environmental conditions (Lalonde 1974). Addressing these factors requires the concerted effort of partners outside of the traditional health sector such as in housing, agriculture, and education, as well as improved coordination and continuity of care between hospital and community-based services (including primary healthcare).

Healthcare systems are moving, albeit at different speeds, from an illness to a wellness/health model, shifting the allocation of funds from curative to preventive care. This is, however, a delicate balancing exercise due to the need to maintain current services while preparing for the future. The Institute for Healthcare Improvement in America has developed a framework that seeks to optimize health system performance by simultaneously pursuing three dimensions, the ‘Triple Aim’ of improving the patient experience of care (including quality and satisfaction), improving the health of

populations, and reducing the per capita cost of healthcare (Berwick et al. 2008).

The components of the Triple Aim are not independent of each other. The pursuit of any one goal can result in changes that affect the other two, sometimes negatively and sometimes positively. For example, improving care for individuals can raise costs if the improvements are associated with new, effective, but costly technologies or drugs. Conversely, eliminating overuse or misuse of therapies or diagnostic tests can lead to both reduced costs and improved outcomes.

The balanced pursuit of the Triple Aim is unfortunately not congruent with the current business models of any but a small number of healthcare organizations. Payment dynamics often lead hospitals to focus only on care within their walls, viewing readmissions for chronic diseases such as congestive heart failure as the result of defects outside the hospital and not as their responsibility to avert (Bisognano and Kenney 2012). For hospitals paid based on number of patients treated, averting readmission incurs costs from additional effort such as patient education and post-discharge follow-up (which may not be reimbursed as these occur outside the hospital) and reduces revenue by decreasing the number of patients being admitted. In such instances, rational common interests and rational individual interests are in conflict. Lack of system knowledge among clinicians and organizations leads them to optimize the components of the system with which they are most familiar, at the expense of the whole.

Movement between the various care levels (primary, secondary, tertiary, and community care) typically involves passage across an interface that is governed by rules of varying degrees of formality. Examples include referral to hospital by a primary care physician or discharge from an acute hospital to a long-stay facility. However, these movements are often not seamless, thus hampering efforts to enable patients to be treated in the most cost-effective settings.

The pursuit of the Triple Aim requires health services to move from the current silo-based approach to a more systematic and integrated model of care. Integrated care brings together inputs, delivery, management, and organization of services related to diagnosis, treatment, care, rehabilitation, and health promotion as a means to improve the services in relation to access, quality, user satisfaction, and efficiency (Gröne and Garcia-Barbero 2001).

Measurement drives behaviour. As such, the move towards better integration of care has been accompanied by efforts to measure health status rather than just healthcare. The proposed framework shown in Figure 11.8.1 (Arah et al. 2006) illustrates a measurement system that takes into account non-healthcare determinants of health, health system performance across the various care sectors, and embeds healthcare within the broader economic, social, and political context of a country. A framework that is unduly narrow and clinical in its focus will miss this big-picture perspective.

Patient engagement and empowerment

Very few healthcare delivery models currently view the patient as a resource who can help improve healthcare delivery for themselves and each other. Patients are either viewed as liabilities who consume scarce resources, or as passive vessels into which treatments are poured (Dzau et al. 2012). At times, healthcare services not only fail to take advantage of the patient's capabilities, but they also diminish them: studies have shown that during hospitalization, older people actually lose the ability to function independently (Creditor 1993; Sager et al. 1996).

Patients with chronic diseases can profoundly change the development and progression of their conditions through effective self-care, goal-setting, and adherence to prescribed medication, exercise, and nutrition regimes. A survey by the Commonwealth Fund (Schoen et al. 2011), however, suggests that this opportunity is being lost in many healthcare systems, owing to the professionals' failure to prepare patients for this role by discussing treatment goals/priorities, making treatment plans that patients can carry out in daily life, and giving clear instructions on symptoms and when to seek care.

On the other hand, we have the positive example in Jönköping, Sweden, where patients with kidney failure have been trained to use hospital dialysis machines themselves, thereby improving the efficiency, productivity, convenience, and safety of care delivery. These patients have also been empowered to provide advice to one another on ways of managing their illness (Dzau et al. 2012).

Several important issues need to be addressed in increasing patient participation in healthcare systems (Saltman and Figueras 1997). One is whether choice is equally important in all healthcare decisions. Another is, given the special characteristics of healthcare, do users have access to

enough information and are they able to understand it sufficiently to make an informed choice? Several countries have sought to inform the public by means of publication of rankings of outcomes by hospital or physician. However, these are subject to substantial methodological limitations and also provide scope for opportunistic behaviour by providers, either to exclude patients with higher risks or to manipulate the data that are presented. Finally, enhancing choice risks increasing health inequalities as those with most resources (time, money, and energy) are likely to have the greatest opportunity to make choices.

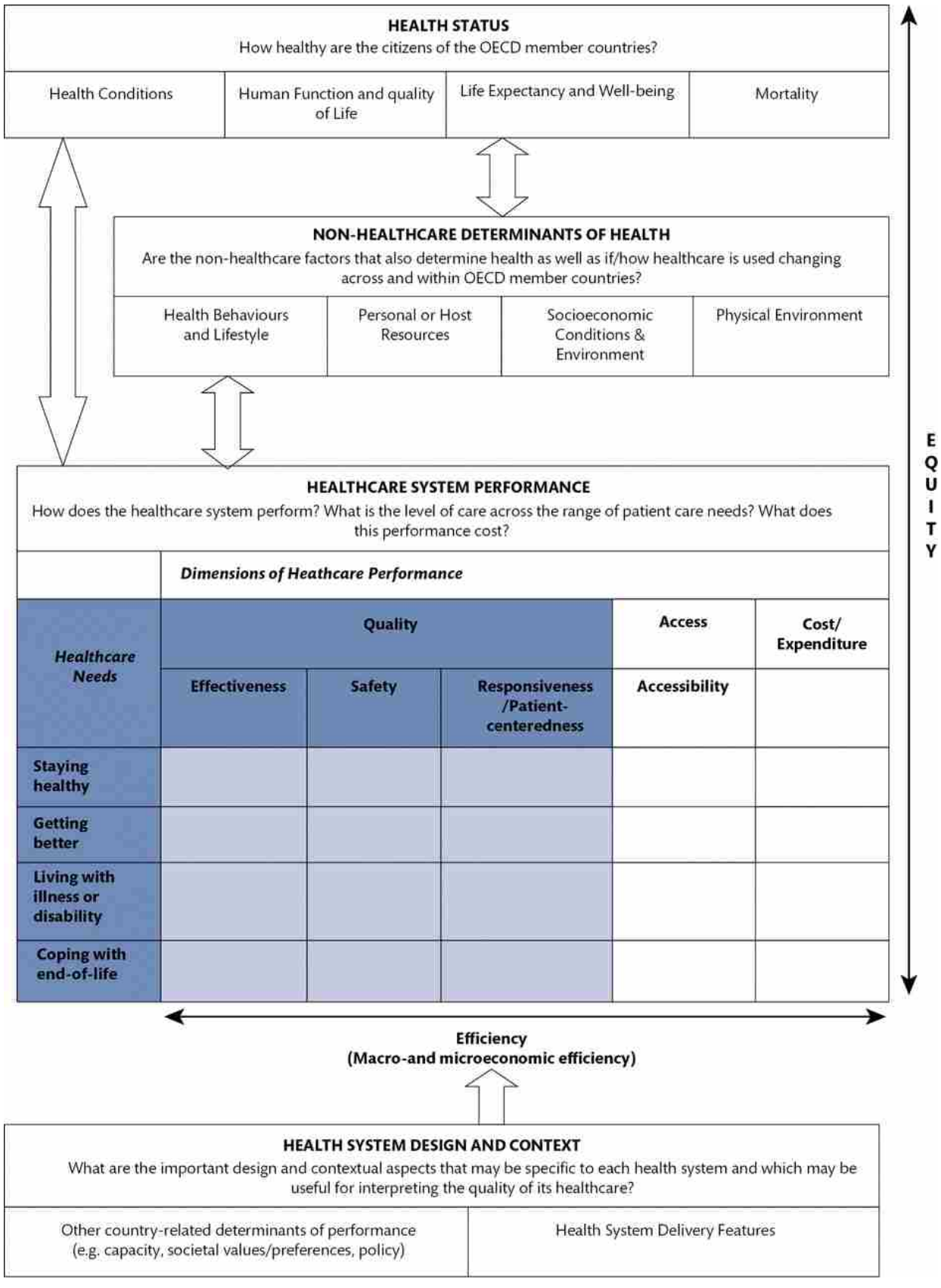

HEALTH STATUS
How healthy are the citizens of the OECD member countries?
Health Conditions
Human Function and quality of Life
Life Expectancy and Well-being
Mortality
NON-HEALTHCARE DETERMINANTS OF HEALTH
Are the non-healthcare factors that also determine health as well as if/how healthcare is used changing across and within OECD member countries?
Health Behaviours and Lifestyle
Personal or Host Resources
Socioeconomic Conditions & Environment
Physical Environment
HEALTHCARE SYSTEM PERFORMANCE
How does the healthcare system perform? What is the level of care across the range of patient care needs? What does this performance cost?
Dimensions of Heathcare Performance
Healthcare Needs
Quality
Access
Cost/ Expenditure
Effectiveness
Safety
Responsiveness /Patient-centeredness
Accessibility
Staying healthy
Getting better
Living with illness or disability
Coping with end-of-life
EQUITY
Efficiency
(Macro-and microeconomic efficiency)
HEALTH SYSTEM DESIGN AND CONTEXT
What are the important design and contextual aspects that may be specific to each health system and which may be useful for interpreting the quality of its healthcare?
Other country-related determinants of performance (e.g. capacity, societal values/preferences, policy)
Health System Delivery Features

Fig. 11.8.1 Conceptual framework for OECD Health Care Quality Indicator (HCQI) Project. Reproduced from Arah, O.A. et al., A conceptual framework for the OECD Health Care Quality Indicators Project, *International Journal for Quality in Health Care*, Volume 18, Issue Supplement 1, pp. 5–13, Copyright © 2006 International Society for Quality in Health Care and Oxford University Press, by permission of Oxford University Press.

Cost sharing

Increasingly, the debate about healthcare costs is no longer about 'who pays?' but about the optimal formula for cost sharing, bearing in mind the need to balance efficiency goals against equity concerns (Lim 2004). Proponents of this approach argue that there is 'no free lunch'. Health insurance premiums have to be paid by the insured, employee medical benefits form part of wage costs, and public expenditures on healthcare are ultimately paid for by taxpayers. The key question is how best to structure the healthcare system, and the means of financing it, in order to keep it efficient, while making it accessible to all at rates that are affordable to them (Ministry of Health 1993).

People with the most control over what drugs get prescribed and what procedures get done have little incentive to lower these costs (indeed, to the extent that they get paid for the procedure, their incentives are often quite the opposite). Furthermore in a fee for service system, it is not in the economic interest of doctors and hospitals to put themselves out of business by promoting disease prevention. Likewise, patients often feel little need to control the costs of their own medical care if it is fully covered by insurance or the state. These approaches fuel overconsumption of healthcare services, hence contributing to rising healthcare costs. Some healthcare systems have thus tried to structure their financing system to give the individual maximum incentive to stay healthy, save for his or her medical expenses, and avoid using more medical services than is absolutely required.

Cost sharing for health services can help limit overconsumption of healthcare and could instil greater personal responsibility for health. On the other hand, cost sharing (or patient co-payment) can and has led to access and equity problems in some countries. Hence, cost sharing or co-payment should be one component of a multifaceted, multilayered healthcare financing system (Lim 2004). It is important to achieve an appropriate balance between personal responsibility with sufficient incentives for people to be prudent and to look after their own health, and risk pooling to provide a comprehensive safety net should they fall sick.

Cost sharing also allows governments to target their social spending more carefully. For example, greater subsidies could be directed at financially needier patients through financial means-testing. Social provision should be about protecting the poor more than subsidizing the rich. At the same time, in fast-ageing societies, especially, increased expenditures on the elderly must not squeeze out healthcare investments in the young (*The Economist* 2012).

These emerging themes bring new opportunities for public health practitioners to reorient health services towards the maximization of health gains. The next section will explore strategies for health services to translate these opportunities into reality.

Strategies for health services

In devising strategies for health services, we need to take into account the complex nature of healthcare where reductionist, mechanical, and linear methods of problem-solving tend to flounder against the laws of unforeseen consequences and incomplete information. It is also important to differentiate the means from the ends. For example, while healthcare systems aim for universal coverage to ensure that no one is denied healthcare that they need because of inability to pay, the means they have employed to achieve this can markedly differ.

Another consideration in developing strategies for health services is the mismatch between the lifecycles of different elements in the healthcare system such as major infrastructure (e.g. hospitals), technological advances (including information technology), service models, and government/regulatory policies. This can pose a major problem for planning and coordination, in particular to ensure alignment and timeliness in response to a constantly changing environment.

It is impossible in the space available to provide a comprehensive overview of every strategy that might be employed to enhance health services. Instead, this section focuses on key steps in the strategic planning process for health services where public health professionals have a potentially vital role to play. It starts by looking at the needs assessment and priority-setting process to ensure that the health services provided reflect the health needs of the population that they serve. Following that, it reviews the design and organization of health services to enhance the efficiency with

which services are delivered. It concludes by discussing the enablers that are necessary to implement, sustain, and enhance these services.

Needs assessment and priority setting

‘Need’ may best be defined as the ability to benefit from ‘healthcare’, which depends both on morbidity and on the effectiveness of care. Need should not be confused with supply or demand. The relationship between need and demand is mediated by differing illness behaviours. Demand is what people would be willing to pay for in a market or might wish to use in a system of free healthcare, and supply is what is actually provided. It is misleading to measure existing service provision as though it were an indicator for need. Similarly, the measurement of waiting lists, reflecting as they do demand mediated by doctor as agent, is also misleading (Stevens and Gabbay 1991).

Healthcare needs are those that can benefit from healthcare (health education, disease prevention, diagnosis, treatment, rehabilitation, terminal care). Health needs incorporate the wider social and environmental determinants of health, such as deprivation, housing, diet, education, and employment. This wider definition allows us to look beyond the confines of the medical model based on health services, to the wider influences on health. Health needs of a population will be constantly changing, and many will not be amenable to medical intervention (Wright et al. 1998). Figure 11.8.2 illustrates how need, demand, and supply might overlap or differ. It shows seven fields of services divided into (E) those for which there is a need but no demand and supply, (F) those for which there is a demand but no need or supply, (G) those for which there is a supply but no need or demand, and then (A–D) the various degrees of overlap (NHS Confederation 2008).

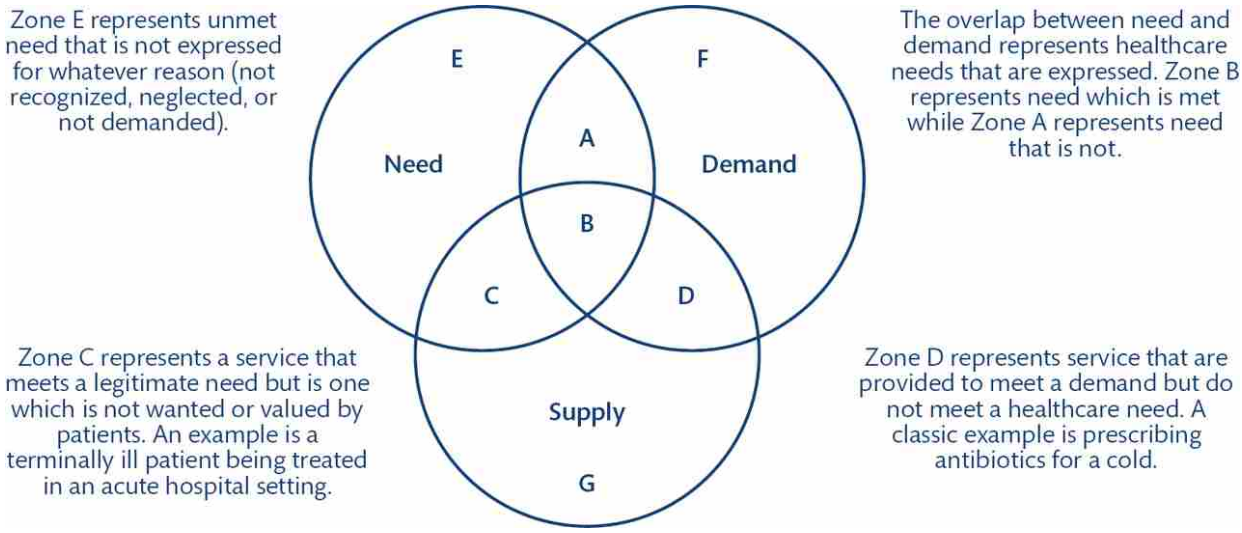


Fig. 11.8.2 Need, demand, and supply of healthcare.

Reproduced with permission from NHS Confederation, *Priority Setting—Strategic Planning*, The NHS Confederation, London, UK, Copyright © The NHS Confederation 2008.

The key challenge is to make the three circles of need, demand, and supply more congruent. This includes demand management and changing supply as well as better definition of need. Demand management may involve curtailing it where it is inappropriate (areas F and D in Fig. 11.8.2) or stimulating it (areas A and B). Varied mechanisms through which demand can be influenced include: the accessibility and organization of services, the provision of health information, education, and financial incentives. Demand can also be induced by supply: geographical variation in hospital admission rates is explained more by the supply of hospital beds than by indicators of mortality (Feldstein 1964); referral rates of general practitioners owe more to the characteristics of individual doctors than to the health of their populations (Wilkin 1992).

Strategies that act on supply of health services include reducing the number of healthcare professionals or facilities, setting global budgets for providers (Schwartz et al. 1996), giving professionals incentives to reduce the amount of care provided, and reducing access to care (Abel-Smith et al. 1995). Even need can be shifted somewhat, with new findings on disease pathology and new treatments that are effective.

Distinguishing between individual needs and the wider needs of the community is important in the planning and provision of local health services. Without a proper need assessment process, the pattern of health services will frequently reflect only what a few people perceive to be the

needs of the population rather than what they actually are and often those whose needs are greatest will receive the least (Hart 1971).

Assessment of health needs is a systematic method of identifying unmet health and healthcare needs of a population and making changes to meet these unmet needs (Wright et al. 1998). The principal activities involved in healthcare needs assessment are the assessment of incidence and prevalence (how many people need the service/intervention), the effectiveness and cost-effectiveness of services (do they confer any benefit, and if so at what cost, i.e. what is the relative benefit?) and baseline services. Changing the provision of healthcare services for the better necessitates knowledge of existing services, both to know which services ought to change and to identify opportunities for the release of resources to enable the change to happen.

Health needs assessment should not just measure ill health, as this assumes that something can be done to tackle it. It should incorporate the concept of a capacity to benefit which requires an assessment of the effectiveness of health interventions and attempts to make explicit what benefits are being pursued, at the level of individuals and the population.

Health needs assessment should include two other objectives. Firstly to improve the spatial allocation of resources so that broadly equal populations receive broadly equal resources. Secondly the accurate targeting of resources so that those who get a service need it, and those who need a service get it.

As the capacity to benefit is always going to be greater than available resources, health needs assessment should also incorporate questions of priority setting (Donaldson and Mooney 1991). Some commentators have argued that rationing should not be necessary if either sufficient funding was made available, typically by redirecting it from other areas of public expenditure or raising taxes, or by ensuring that available resources are used more efficiently.

In reality, some form of rationing has always taken place in all healthcare systems although, in most cases, this was implicit, inextricably linked to clinical judgement. Priority setting or rationing in healthcare continues to be a politically charged topic, but recently its necessity has gained wider recognition as it is impossible to provide everyone with every effective intervention they might need or want (Sabik and Lie 2008). Explicitly addressing priority setting is necessary to develop fairer methods to allocate

scarce healthcare resources (Fleck 1994) through a transparent process that balances clinical, ethical, and economic considerations (Black 1994).

Priority-setting approaches can be broadly grouped into two categories: outlining principles and defining practices. Countries such as Norway, the Netherlands, Sweden, and Denmark decided to develop principles that would guide prioritization efforts. In the Netherlands, the Dunning Committee (Committee on Choices in Health Care 1992) proposed four criteria that any intervention to be funded from social insurance should meet: necessity, effectiveness, efficiency, and individual responsibility. They described these principles as forming a sieve for sifting out services that should not be publicly funded ('Dunning's funnel'). The idea was that one should apply these successively beginning with the principle of necessity, and then limiting the number of services provided. The Committee excluded services that could easily be paid for by the individuals themselves. One example the Committee gave was routine adult dental care. Although such dental care is necessary, effective, and efficient, adults can easily pay for such services out of pocket, and it is therefore best left to an individual's responsibility. The drawback of this approach is that because of their abstractness, government commissions that outlined principles have had little direct impact on their countries' policies.

Other countries, such as the United Kingdom, Israel, New Zealand, and the state of Oregon in the United States established bodies that would actually recommend what services should be provided within the system. Decisions may lead to blanket exclusions of interventions or of condition–intervention combinations, as in the approach taken in Oregon, or production of guidelines, as has been done in New Zealand (The National Advisory Committee on Health and Disability) and by the National Institute of Health and Care Excellence in the United Kingdom. The experience of Oregon has been especially interesting because of the technical and ethical issues it has raised. It was designed to create a list of condition–intervention pairs, ranked on the basis of cost–benefit, with a cut-off point based on available resources below which combinations would not be funded. This initial method was, however, abandoned because of a public outcry over the resulting ranking of services, for example: appendectomy was rated lower than cosmetic dentistry. In response to this public reaction as well as criticism of the methodology, the Commission set out to rank healthcare services based on more broadly defined criteria, where expert knowledge and intuitive

judgements by the Commissioners about appropriateness played a much larger role. The process also raised important ethical questions as the exercise was confined to poor women and children in the Medicaid programme and services denied to them would remain available to others (Rosenbaum 1992).

In the United Kingdom, the National Institute for Clinical Excellence (now the National Institute for Health and Care Excellence (NICE)) was established in 1999 to: appraise new health technologies, develop clinical guidelines, and assess interventional procedures (Rawlins 1999). In conducting these activities, NICE addresses questions including what constitutes necessary and appropriate care, how to incorporate cost-effectiveness considerations, and what interventions should be publicly funded.

The challenge of healthcare prioritization is to identify an appropriate balance between an expert-led process and a process that emphasizes public involvement in decision-making to ensure that decisions reached are legitimate. Ultimately, while public health professionals have an important role in providing the evidence on which any debate on priority setting must be based, priority setting in a publicly funded system is the responsibility of politicians. Decision-making inevitably involves trade-offs between objectives as a balance is sought among universal coverage, comprehensiveness of services, equity, efficiency, cost containment, and broader social values (Sabik and Lie 2008).

Design and organization of health services to ensure efficient service delivery

Conventional service delivery is often based on institutional hospital care and a supply of trained health professionals—all relatively expensive with inpatient care typically consuming 45–75% of resources dedicated to healthcare (Saltman and Figueras 1997). Moreover, different professionals often work in organizational silos, and fail to integrate services to achieve best value and best care. New technologies in fields such as genetics, imaging, miniaturization, pharmaceuticals, and information are changing the mix of preventive, diagnostic, and treatment practices, enabling services to be delivered less expensively to wider populations in their own homes or communities instead of hospitals.

Health systems have adopted several strategies to meet these challenges. They include organizing the network of hospital services to achieve a more efficient and effective configuration through decentralization and centralization initiatives (see 'Reorganizing the system: decentralization, recentralization, and privatization'). Three other strategies are described in the following subsections: improved targeting of interventions, substitution, and integration of services in the hospital and community.

Improved targeting of interventions

Bohmer (2011) identified specification and planning as a common habit of high-value healthcare organization. These organizations specify decisions and activities in advance. Whenever possible, both operational decisions, such as those related to patient flow (admission, discharge, and transfer criteria), and core clinical decisions, such as diagnosis, tests, or treatment selection, are based on explicit criteria. Criteria-based decision-making is manifested in the use of clinical decision support systems and treatment algorithms, severity and risk scores, criteria for initiating a call to a rapid-response team, or triggering the commitment of a future resource (e.g. a discharge planner, pre-procedure checklists, and standardized patient assessments). Specification also applies to separating heterogeneous patient populations into clinically meaningful subgroups—by disease subtype, severity, or risk of complications—each with its own distinct pathway. For example, genomic testing has allowed oncology units to divide patients into separate groups according to their probable response to specific therapies (for instance, Kirsten RNA-associated rat sarcoma 2 virus (*KRAS*) gene testing for cetuximab therapy). And at Intermountain Healthcare in Utah and Idaho, the needs of psychiatric patients are divided into mild (routine care by a primary care physician), moderate (team care), and severe (specialist referral), using a scoring system based on published guidelines (Rathod et al. 2010).

To promote such standardization, we need to accelerate the healthcare evidence base and support the rapid sharing of clinical protocols, best practices, and outcomes data. There is concurrently a need to make effective use of information technologies to automate clinical information and make it readily accessible to patients and all members of the care team (IOM 2001). Excellent information infrastructure and well-thought out and implemented

modes of ongoing communication can reduce the need to craft laborious, case-by-case strategies for coordinating patient care.

A second common characteristic of high-value organizations is infrastructure design. They deliberately design microsystems (Nelson et al. 2011)—including staff, information, and clinical technology, physical space, business processes, and policies and procedures that support patient care—to match their defined subpopulations and pathways. Thus, different conditions or patient groups have different microsystem designs. The various tasks of care are allocated to different members of a clinical team (including the patient), with the skills and training of each staff member matched to the work.

Substitution

Substitution refers to the process of regrouping resources (staff, equipment, information, facilities, etc.) across and within care settings to exploit the best and least costly solutions to achieve better clinical, financial, and patient-related outcomes. A useful typology differentiates substitution into three categories: moving the location at which care is delivered, introducing new technologies, and changing the mix of staff and skills. There may be other substitution types according to different combinations of these three (Warner 1996). For example, moving care from the hospital to community-based care may involve:

Location only such as for physiotherapy.
Location and technology such as for point-of-care laboratory testing and telemedicine that allows remote consultations.
Location and staff and/or skills such as nurse-led home care.
All substitution types such as home dialysis services.

In addition to substitution, the weightage of services across the spectrum of care is also changing with greater emphasis being placed on prevention and self-management to reduce the demand for care. Prevention programmes have historically focused on younger or working-age adults, but there is increasing recognition of the value of prevention programmes aimed at older people with services focused on increasing functionality. These health promotion and primary prevention strategies should ultimately reduce the total demand for healthcare services and healthcare costs. However, the

difficulties in measuring health promotion in terms of cost-effectiveness so that it can be compared with other health services interventions, has meant that health promotion's role has tended to be undervalued. Public health professionals have to develop more and better ways to evaluate health promotion that satisfy the needs of policymakers, managers, and clinicians so the full potential of health promotion can be realized (McKee and Bojan 1998).

Health services should be reoriented from a hospital-centric model to a model with primary healthcare at the core of the healthcare system, and secondary and tertiary care serving as supporting, referral levels. Healthcare systems configured around primary care produce healthier populations at lower cost. In the United States, for example, overall health is better in regions with more primary care physicians and where primary care functions well. By contrast, overall health is worse where specialists predominate (Starfield et al. 2005a, 2005b). An effective primary care system has the following virtues (Oldham et al. 2012):

It can coordinate the care of the many people with multiple, complex health needs, in a way that individual specialists cannot.
It can deliver care closer to people, increasing convenience, especially in areas remote from hospitals.
As a first point of contact for patients it facilitates the early detection of illness, and thereby improves outcomes.

In addition to substitution within healthcare, substitution should also occur between healthcare and other sectors, recognizing the importance of lifestyle and social determinants in improving health outcomes. The importance of broader comprehensive primary healthcare approach has recently been argued as being vital to achieving UHC (Sanders et al. 2019). The upstream cause of ill health has been shown to be social and economic determinants and that these factors also shape the health behaviours people adopt (Baum 2019). The health and social sectors in some countries are thus pooling resources particularly in the provision of services to the elderly to tackle social isolation, engage communities more widely, adapt or develop homes, and make cities more age-friendly places.

Integration of care

The involvement of increasingly more parties in the delivery of healthcare accentuates the need to adopt a systems perspective to foster evolution of care delivery as a whole rather than focusing on single elements. Care processes need to be redesigned to enable coordinated, seamless care across patient conditions, disciplines/specialties, settings, and clinicians, and over time. The key challenge is to match evidence-based, appropriate, and responsive care to patients according to their needs (such as various chronic diseases) across the various stages of their life. Capital investments and operating expenses should encourage thinking across boundaries. As an illustration, questions to be asked might include, 'Might two additional home outreach nurses be better for heart failure patients than another cardiologist?' (Berwick et al. 2008).

One way of achieving this is by designing distinct integrated care platforms (Leape et al. 2009) for conditions that share common work and support requirements, such as chronic disease care, complex acute care, palliative, and end-of-life care. Every care platform must have the following characteristics:

Patient centredness: personnel, facilities, and services are organized to meet all patients' needs efficiently and responsively; to be available when and where needed, and to include the patient and family as partners in care.
Work assignment: work is assigned to the individuals who are responsible for its completion. Assignments strive to maximize the performance capability of each individual while ensuring that work is done by the least expensive qualified caregiver or multidisciplinary team at the location most accessible to the patient. The physician participates when his/her special expertise is required and when patient expectations permit no alternative.
Support: the support framework—people, systems, and tools (e.g. technologies, IT, telecommunications)—is defined by the work and patient participation.
Community linkage: linkages to community advocacy, support, and education groups (especially health literacy) are incorporated into the design as appropriate (e.g. for patients with chronic conditions).
Variation management: ensuring quality and efficiency requires determining whether variations in process are appropriate. Process and outcome measures should thus be incorporated into daily work. Use of such measures makes it possible to understand the degree to which performance is consistent with best practices, and the extent to which patients are being helped. Exception analysis assesses whether variations result from: (1) adaptations to a specific patient requirement, (2) evolution of new evidence (good), (3) lack of training in appropriate care, or (4) poorly defined care pathways (bad).
Transparency: the results of the delivery process and outcomes of care should be shared with all concerned, including patients. This would enable patients, caregivers, organizations, and managers to know the 'state of the system' with respect to its reliability, adherence to evidence, cost, and progress in improvement.

For the integrated platform to work we need an 'integrator' which accepts responsibility for the Triple Aim of improving the patient experience of care, improving the health of populations, and reducing the per capita cost of healthcare for a specified population. The simplest such form, such as Kaiser Permanente, has fully integrated financing and either full ownership of or exclusive relationships with delivery structures which enables it to use those structures to good advantage. A unified financing or a single delivery system is not, however, necessarily a prerequisite for an integrator. That role might be within the reach of a powerful, visionary insurer; a large primary care

group in partnership with payers; or a regional health system that will link healthcare organizations (as well as public health and social service organizations) whose missions overlap across the spectrum of delivery (Berwick et al. 2008). At the patient level, the integrator ensures that members of the population, especially those with chronic illnesses, have someone who can work with them to establish a plan for their ongoing care, guide them through the jungle of health services, and be their advocate.

The integrator has to segment the population according to need such as health status, level of support from family or others, and socioeconomic status, to facilitate efficient and equitable resource allocation (Lynn et al. 2007). In most cases, the integrator would not be a direct provider of all necessary services. Instead, it would need to commission some services from suppliers through business relationships consciously designed to facilitate pursuit of the desired outcomes.

Enablers

Three key enablers are needed to translate strategies into reality: intelligent purchasing, a workforce with the appropriate capacity and capability, and a robust information and knowledge management system.

Intelligent purchasing

All health systems exercise some form of purchasing but the prevailing approach to paying for healthcare, based predominantly on individual services and products, encourages wasteful and ineffective care (IOM 2012). Payment systems that are based on workload tend to disincentivize providers from coordinating care for patients across settings and over time as the benefits of investments (and cost) made by one group of providers are realized by a different group of providers. For example, investments in primary care preventive efforts tend to reduce demand for hospital services rather than primary care services. Hospitals may not be keen on these systems-level savings since their own reimbursement may decline with the decreased workload.

Health systems need to move from passive forms of purchasing—the mere reimbursement of providers—to more proactive and strategic forms of purchasing that consider which interventions should be purchased, how they should be purchased and from whom (Figueras et al. 2005). In an increasing

number of countries intelligent purchasing is seen as an alternative to a traditional command and control approach to implement health policy objectives, including as a means of reorienting the focus of health services, to ensure that they reflect health needs and provide cost-effective care. Its essential characteristic is that it separates purchasers from providers but binds each party by means of contracts to explicit commitments, with creation of the economic motivation to fulfil these commitments (McKee and Figueras 2009).

Payers/purchasers should adopt outcome- and value-oriented payment models where payment systems align financial incentives with the implementation of care processes based on best practices and the achievement of better patient outcomes. Payment systems should also reduce fragmentation of care by supporting high-quality, team-based care focused on patients' needs. Contracts can be used to develop intersectoral responses to health problems and to reorient healthcare providers so that they integrate prevention with curative care.

Contracts should also provide an opportunity for providers to share in the benefits of quality improvement. Substantial improvements in quality are most likely to be obtained when providers are highly motivated and rewarded for carefully designing and fine-tuning care processes to achieve increasingly higher levels of quality and safety. Rewards should thus be located close to the level at which the reengineering and process redesign needed to improve quality are likely to take place (IOM 2001).

Purchasers represent the interests of their populations, in particular the vulnerable and disadvantaged groups as well as underserved communities, by allocating resources and purchasing services in accordance with their needs. There is, however, a need to guard against the risk of providers underemphasizing or phasing out services that are less profitable. Contracts should thus provide fair payment for good clinical management of the types of patients seen. Clinicians should be adequately compensated for taking good care of all types of patients, neither gaining nor losing financially for caring for sicker patients or those with more complicated conditions. The risk of random incidence of disease in the population should reside with a larger risk pool, whether that is large groups of providers, health plans, or insurance companies (IOM 2001).

Purchasing also offers a means for enhancing participation by the population in the organization of healthcare by providing the opportunity for

consumers and purchasers to recognize quality differences in healthcare and direct their decisions accordingly. To do this, consumers need to have good information on quality and the ability to use that information as they see fit to meet their needs.

Implementation of an intelligent purchasing system is a complex process requiring a high level of skills and well-developed information systems. At a minimum, information is required on patient flow, cost, and utilization across specialties or diagnostic groups and demographic and risk groups. Purchasing decisions also need to take into account the reality that medical care is extremely complex. Diagnostic labels are often imprecise and clouded by a degree of legitimate uncertainty. Decisions on clinical management also incorporate values and beliefs relating to factors such as attitude to risk and the utility placed on different health states. It is thus important that expectations of what can be achieved are realistic (McKee and Clarke 1995). Contracts must incorporate sufficient flexibility and reflect the views of all those concerned if they are to retain any credibility. It should also be recognized that purchasing may not be the best option for some services especially where clear specifications in terms of the desired product/service cannot be defined clearly.

While the focus of purchasing should ideally be on outcomes, outcomes are typically more difficult to measure and may only become apparent long after the intervention took place. Some outcomes may also be rare and the sample size required to detect a deviation from what is expected may be very large. For example, Mant and Hicks (1995) have shown, on the basis of knowledge of effective treatment for myocardial infarction, that, in a comparison of two typical hospitals, it would take 73 years of data to detect a significant 3% reduction in mortality whereas the corresponding difference in process measures, relating to uptake of treatments, would emerge after only 4 months. Where process or structure measures are used there should be evidence that they correlate with a good outcome. Measures based on structure can be of some value, based on the assumption that high-quality care cannot be provided in the absence of basic prerequisites, such as adequately trained staff, but this is a necessary rather than a sufficient measure and should normally be supplemented by measures of either process or outcome.

All these activities (e.g. needs assessment, performance analysis, negotiating, and monitoring) incur transaction costs. A substantial increase

in quality and efficiency is required to justify these additional costs. Increasingly, purchasers and providers are entering into long-term contractual relationships that reduce transaction costs (Rosen and McKee 1995). If transaction costs can be minimized without compromising the pursuit of the objectives of equity and health gain, intelligent purchasing can provide a formidable instrument to promote population health.

Workforce

A major challenge in transitioning to the desired models of care described previously is preparing the workforce to acquire new skills and adopt new ways of relating to patients and each other. At least three approaches can be taken to support the workforce in this transition:

Redesign the way health professionals are trained. Healthcare professionals should be taught evidence-based practice and equipped with the skills and tools to access and apply the expanding knowledge base. They should also have sufficient communication and behavioural skills to work effectively with patients, not least in helping to promote treatment adherence and health-enhancing lifestyle changes. In addition they need to acquire the skills to improve the processes and systems of care through proven methods and actions. Furthermore as effective health services will increasingly be delivered by collaborative teams of clinicians, with each member playing a vital role, the training should adopt a multidisciplinary approach and continually advance the effectiveness of teams instead of training health professionals purely by discipline. As almost every country wants more patients to be treated in the community rather than in hospitals, we need to ensure that healthcare professionals spend considerable time during training in the community, not just in hospitals. We also need to review the necessity for the expensive additional years of education for an increasing number of doctors and nurses to be trained in very narrow subspecialties (Emanuel and Fuchs 2012) and its alignment with the wider interests and needs of the population.

Modify the ways in which health professionals are regulated. Regulation of healthcare professionals is designed to protect the public and ensure that patients are kept safe, but providing such protection needs to be balanced against the regulatory burden and the need to retain sufficient freedom for innovators to succeed. Regulation should be focused on the patient and should be independent of professional interests. Scope-of-practice laws and other workforce regulations need to allow for innovation in the use of all types of clinicians to meet patient needs in the most effective and efficient way possible. Statutory regulation should be used to protect the public in areas of high risk, but other processes should be used when the risk is low. Ongoing licensure and certification requirements should also reflect the need for lifelong learning and evaluation of competencies to ensure that healthcare professionals have the specific skills for the time and contexts in which they work (Dzau et al. 2012).

Examine how the *liability system* can constructively support changes in care delivery while remaining part of an overall approach to accountability for healthcare professionals and organizations.

In addition to the professional workforce, it is also important to work on securing the supply of informal care. Care provided by families is the backbone of long-term care. But it is difficult to maintain or increase it in the face of falling support ratios, urbanization, and migration, and changing work patterns. One option is to direct cash and services to support carers. Another is to use technological innovations to provide information and support. Governments can also innovatively incentivize or reinforce traditional family caring responsibilities through the tax system and other regulatory measures (Hope et al. 2012).

Information and knowledge management

The knowledge of how to configure structures and processes to attain the best possible clinical outcomes will become one of the health service's most important assets. Healthcare organizations must thus create, maintain, and disseminate the institutional knowledge required to realize good outcomes.

As healthcare is complex and constantly changing, we need to build healthcare systems that continually learn, in real-time and with new tools, how to better manage problems. Data generated in healthcare delivery—whether clinical, delivery process, or financial—should be collected in digital formats, compiled, and protected as resources for managing care, capturing results, improving processes, strengthening public health, and generating new insights. Decision support tools and knowledge management systems can be included routinely in healthcare delivery to ensure that decisions are informed by the best evidence. The characteristics of a continuous learning healthcare system are summarized in Box 11.8.1 (IOM 2012).

Box 11.8.1 Characteristics of a continuously learning healthcare system

Science and informatics *Real-time access to knowledge*—a learning healthcare system continuously and reliably captures, curates, and delivers the best available evidence to guide, support, tailor, and improve clinical decision-making and care safety and quality. *Digital capture of the care experience*—a learning healthcare system captures the care experience on digital platforms for real-time generation and application of knowledge for care improvement.

Patient–clinician partnerships *Engaged, empowered patients*—a learning healthcare system is anchored on patient needs and perspectives and promotes the inclusion of patients, families, and other caregivers as vital members of the continuously learning care team.

Incentives *Incentives aligned for value*—in a learning healthcare system, incentives are actively aligned to encourage continuous improvement, identify, and reduce waste, and reward high-value care. *Full transparency*—a learning healthcare system systematically monitors the safety, quality, processes, prices, costs, and outcomes of care, and makes information available for care improvement and informed choices and decision-making by clinicians, patients, and their families.

Culture *Leadership-instilled culture of learning*—a learning healthcare system is stewarded by leadership committed to a culture of teamwork, collaboration, and adaptability in support of continuous learning as a core aim. *Supportive system competencies*—in a learning healthcare system, complex care operations and processes are constantly refined through ongoing team training and skill building, systems analysis and information development, and creation of the feedback loops for continuous learning and system improvement.

Source: data from National Academy of Sciences, *Best Care at Lower Cost: The Path to Continuously Learning Healthcare in America*, Copyright © 2013.

Regulators have a key role in facilitating this learning process by clarifying and improving rules governing the collection and use of clinical data to safeguard patient privacy while promoting the seamless use of such data for better care coordination and management, improved care, and enhanced knowledge. Legal protections should be put in place to shield individuals from improper use of de-anonymized health data and educational campaigns should be conducted, aimed at providers and public alike, to raise awareness of the rights, responsibilities, risks, and benefits related to health data.

Conclusion

Health services have an important role to play in improving the health status of populations. This chapter has identified mechanisms through which public health can play a major role in maximizing the health gain obtained from health services, in particular in the design of systems that ensure that patients are managed at the level of the healthcare system that is most appropriate, the creation of mechanisms that identify and promote high-quality clinical care, and the reorientation of curative services towards prevention.

Public health professionals, with their emphasis on improving population health, have a key role in ensuring that the pursuit of health gain becomes a central objective of healthcare services, whatever other objectives are being pursued by others (McKee and Figueras 2009). To do so, they must promote the equitable use of interventions that are effective and appropriate for the population in question, reduce interventions that are ineffective or harmful,

and maximize the health gains obtained with the available funding. Much will depend on the ability of the public health profession to adapt and bring its portfolio of tools and skills to bear on rapidly changing health services.

REFERENCES

Abel-Smith, B., Figueras, J., Holland, W., McKee, M., and Mossialos, E. (1995). *Choices in Health Policy: An Agenda for the European Union*. Aldershot: Dartmouth Press/Office for Official Publications of the European Communities.

Anderson, T.F. and Mooney, G. (eds.) (1990). *The Challenges of Medical Care Variations*. Basingstoke: Macmillan.

Arah, O.A., Westert, G.P., Hurst, J., and Klazinga, N.S. (2006). A conceptual framework for the OECD Health Care Quality Indicators Project. *International Journal for Quality in Health Care*, **2006**, 5–13.

Atwal, P. (2010). How do you solve a problem like health reform: has Singapore got it right? [Online] Available at: http://healthaffairs.org/blog/2010/11/12/how-do-you-solve-a-problem-like-health-reform-has-singapore-got-it-right/

Baum, F (2019) *Governing for Health: Advancing Health and Equity through Policy and Advocacy*. New York: Oxford University Press.

Benson, J.K. (1975). The interorganisational network as a political economy. *Administrative Science Quarterly*, **20**, 229–49.

Berwick, D.M., Nolan, T.W., and Whittington, J. (2008). The triple aim: care, health, and cost. *Health Affairs*, **27**, 759–69.

Bisognano, M.A. and Kenney, C. (2012). *Pursuing the Triple Aim: Seven Innovators Show the Way to Better Care, Better Health, and Lower Costs*. San Francisco, CA: Jossey-Bass.

Black, D. (1994). *A Doctor Looks at Health Economics. Office of Health Economics Annual Lecture*. London: Office of Health Economics.

Bohmer, R.M.J. (2011). The four habits of high-value health care organizations. *The New England Journal of Medicine*, **365**, 22.

Bojan, F., Hajdu, P., and Belicza, E. (1991). Avoidable mortality. Is it an indicator of quality of medical care in Eastern European countries? *Quality Assurance of Health Care*, **3**, 191–203.

Borgenhammer, E. (1993). *At vårda liv: orgnisation, etik, kvalitet. [Looking after life: organisation, ethics, quality.]* Stockholm: SNS Förlag.

Boys, R.J., Forster, D.P., and Jozan, P. (1991). Mortality from causes amenable and non-amenable to medical care: the experience of Eastern Europe. *British Medical Journal*, **303**, 879–83.

Cabinet Office (1999). *Modernising Government*. London: HMSO.

Cassel, C.K. and Reuben, D.B. (2011). Specialization, subspecialization, and subsubspecialization in internal medicine. *The New England Journal of Medicine*, **364**, 1169–73.

Chalmers, I. and Altman, D.G. (1995). *Systematic Reviews*. London: BMJ Publications.

Clark, D.G. (1988). Autonomy, personal empowerment and quality of life in long-term care. *Journal of Applied Gerontology*, **7**, 279–97.

Collins, C.D. and Green, A.T. (1994). Decentralisation and primary health care: some negative implications in developing countries. *International Journal of Health Services*, **24**, 459–75.

Committee on Choices in Health Care (1992). *Choices in Health Care*. Rijswijk: Ministry of Welfare HCA.

Congressional Budget Office (2008). *Geographic Variation in Health Care Spending*. Washington, DC: Congressional Budget Office.

Contandriopoulos, A.P., Lauristin, M., and Leibovich, E. (1998). Values, norms, and the reform of health care systems. In: Saltman, R., Figueras, J., and Sakallarides, C. (eds.) *Critical Challenges for Health Care Reform in Europe*. Buckingham: Open University Press, pp. 339–62.

Creditor, M. (1993). Hazards of hospitalization in the elderly. *Annals of Internal Medicine*, **118**, 219–23.

Dash, P., Llewellyn, C., and Richardson, B. (2009). Developing a regional health system strategy. *Health International*, **8**, 26–35.

Donaldson, C. and Mooney, G. (1991). Needs assessment, priority setting, and contracts for health care: an economic view. *British Medical Journal*, **303**, 1529–30.

Dzau, V., et al. (2012). *A Neglected Resource: Transforming Healthcare through Human Capital. Report of the Innovative Delivery Models Working Group*. London: Imperial College.

Emanuel, E. and Fuchs, V. (2012), Shortening medical training by 30%. *Journal of the American Medical Association*, 307(11), 1143–4.

Eysenbach, G., et al. (1998). Towards quality management of medical information on the Internet: evaluation, labelling, and filtering of information. *British Medical Journal*, **317**, 1496–502.

Feldstein, M.S. (1964). Effects of differences in hospital bed scarcity on type of use. *British Medical Journal*, ii, 562–5.

Figueras, J., Jakubowski, E., and Robinson, R. (eds.) (2005). *Purchasing to Improve Health Systems Performance*. Maidenhead: Open University Press.

Fleck, L.M. (1994). Just caring: health reform and health care rationing. *Journal of Medicine and Philosophy*, **19**, 435–43.

Gaylin, D.S. and Kates, J. (1997). Refocusing the lens: epidemiologic transition theory, mortality differentials, and the AIDS pandemic. *Social Science & Medicine*, 44(5), 609–21.

Gillett, G. (1995). Virtue and truth in clinical science. *Journal of Medical Philosophy*, **20**, 285–98.

Green, L., Ottoson, J., García, C., and Hiatt, R. (2009). Diffusion theory and knowledge dissemination, utilization, and integration in public health. *Annual Review of Public Health*, **30**, 151–74.

Gröne, O. and Garcia-Barbero, M. (2001). Integrated care: a position paper of the WHO European Office for Integrated Health Services. *International Journal of Integrated Care*, 1, e21.

Habermas, J. (1987). *Théorie de L'agir Communicationnel*. Paris: Fayard.

Hart, J.T. (1971). The inverse care law. *The Lancet*, i, 405–12.

Himmelstein, D.U., et al. (1999). Quality of care in investor-owned vs not for profit HMOs. *Journal of the American Medical Association*, **282**, 159–63.

Hope, P., et al. (2012). *Creating Sustainable Health and Care Systems in Ageing Societies. Ageing Societies Working Group 2012*. London: Imperial College.

Illich, I. (1976). *Limits to Medicine*. London: Marion Boyars.

Institute of Medicine (IOM) (1989). *To Err is Human: Building A Safer Health System*. Washington, DC: National Academy Press.

Institute of Medicine (IOM) (2001). *Crossing the Quality Chasm: A New Health System for the 21st century*. Washington, DC: National Academy Press.

Institute of Medicine (IOM) (2004). *The Migration of Health Care Workers: Creative Solutions to Manage Health Workforce Migration*. Washington, DC: National Academy Press.

Institute of Medicine (IOM) (2012). *Best Care at Lower Cost: The Path to Continuously Learning Health Care in America*. Washington, DC: National Academy Press.

International Diabetes Federation (2012). *IDF Diabetes Atlas* (5th ed.). Brussels: International Diabetes Federation.

Kettl, D.F. (1993). *Sharing Power: Public Governance and Private Markets*. Washington, DC: Brookings Institution.

Lalonde, M. (1974). *A New Perspective on Health of Canadians: A Working Document*. Ottawa: Information Canada.

Landrigan, C.P., Parry, G.J., Bones, C.B., Hackbarth, A.D., Goldmann, D.A., and Sharek, P.J. (2010). Temporal trends in rates of patient harm resulting from medical care. *The New England Journal of Medicine*, 363(22), 2124–34.

Leape, L., et al. (2009). Transforming healthcare: a safety imperative. *Quality and Safety in Health Care*, **18**, 424–8.

Leppo, K., et al. (eds.) (2013). *Health in All Policies—Seizing Opportunities Implementing Policies*. Helsinki:, Ministry of Social Affairs and Health.

Lim, M.K. (2004). Shifting the burden of health care finance: a case study of public–private partnership in Singapore. *Health Policy*, **69**, 83–92.

Lubitz, J., Beebe, J., and Baker, C. (1995). Longevity and medical care expenditures. *The New England Journal of Medicine*, **332**, 999–1003.

Lynn, J., Straube, B.M., Bell, K.M., Jencks, S.F., and Kambic, R.T. (2007). Using population segmentation to provide better health care for all: the 'bridges to health' model. *Milbank Quarterly*, **85**, 185–208.

Mant, J. and Hicks, N. (1995). Detecting differences in quality of care: the sensitivity of measures of process and outcome in treating acute myocardial infarction. *British Medical Journal*, **311**, 793–6.

Maxwell, R.J. (1981). *Health and Wealth: An International Study of Health Spending*. Cambridge, MA: Lexington Books.

McGinnis, J.M. and Foege, W.H. (1993). Actual causes of death in the United States. *Journal of the American Medical Association*, **270**, 2207–12.

McKee, M. and Bojan, F. (1998). Reforming public health services. In: Figueras, J., Saltman, R., and Sakallarides, C. (eds.) *Critical Challenges for Health Care Reform*. Buckingham: Open University Press, pp. 135–54.

McKee, M. and Clarke, A. (1995). Guidelines, enthusiasms, uncertainty and the limits to purchasing. *British Medical Journal*, **310**, 101–4.

McKee, M. and Figueras, J. (2009). Strategies for health services. In: Detels, R., Beaglehole, R., Lansang, M.A., and Gulliford, M. (eds.) *Oxford Textbook of Public Health* (5th ed.). Oxford: Oxford University Press, pp. 1668–81.

McKeown, T. (1979). *The Role of Medicine: Dream, Mirage or Nemesis?* Oxford: Blackwell Scientific.

Menon, R. (2010). *Markets and Government: Striking a Balance in Singapore*. Singapore: Economic Policy Forum.

Ministry of Health (1993). *Affordable Healthcare: A White Paper*. Singapore: Ministry of Health.

Mossialos, E. and Le Grand, J. (1999). Cost containment in the EU: an overview. In: Mossialos, E. and Le Grand, J. (eds.) *Health Care and Cost Containment in the European Union*. Ashgate: Aldershot, pp. 1–154.

Murray, C.J.L. and Lopez, A. (1996). *The Global Burden of Disease*. Cambridge, MA: Harvard University Press.

National Health Service (NHS) (2007). *Healthcare for London: A Framework for Action*. London: NHS.

Nelson, E.C., Batalden, P.B., Godfrey, M.M., and Lazar, J.S. (2011). *Value by Design: Developing Clinical Microsystems to Achieve Organizational Excellence*. San Francisco, CA: Jossey-Bass.

NHS Confederation (2008). *Priority Setting—Strategic Planning*. [Online] Available at: https://www.nhsconfed.org/~/media/Confederation/Files/Publications/Documents/Priority%20setting%20strategic%20planning

Nolte, E. and McKee, M. (2004). *Does Health Care Save Lives? Avoidable Mortality Revisited*. London: The Nuffield Trust.

Oldham, J., Richardson, B., Dorling, G., and Howitt, P. (2012). *Primary Care—The Central Function and Main Focus Report of The Primary Care Working Group 2012*. London: Imperial College.

Olshansky, S.J. and Ault, A.B. (1986). The fourth stage of the epidemio-logic transition: the age of delayed degenerative diseases. *The Milbank Quarterly*, **64**, 355–91.
Omran, A.R. (1971). The epidemiologic transition: a theory of the epidemiology of population change. *The Milbank Memorial Fund Quarterly*, **49**, 509–38.
Osborne, D. and Gaebler, T. (1993). *Reinventing Government*. Reading, MA: Addison-Wesley.
Porter, M. and Teisberg, E.O. (2006). *Redefining Healthcare: Creating Value-Based Competition on Results*. Boston, MA: Harvard Business School Press.
Rathod, R.H., et al. (2010). A novel approach to gathering and acting on relevant clinical information: SCAMPs. *Congenital Heart Disease*, **5**, 343–53.
Rawlins, M. (1999). In pursuit of quality: the National Institute for Clinical Excellence. *The Lancet*, **353**, 1079–83.
Rondinelli, D. (1981). Government decentralisation in comparative theory and practice in developing countries. *International Review of Administrative Sciences*, **47**, 133–45.
Rosen, R. and McKee, M. (1995). Short termism in the NHS. *British Medical Journal*, **311**, 703–4.
Rosenbaum, S. (1992). Poor women, poor children, poor policy: the Oregon Medicaid experiment. In: Strosberg, M., Wiener, J., Baker, R., and Fein, I.A. (eds.) *Rationing America's Medical Care: The Oregon Plan and Beyond*. Washington, DC: Brookings Institution, pp. 103–4.
Sabik, L.M. and Lie, R.K. (2008). Priority setting in health care: lessons from the experiences of eight countries. *International Journal for Equity in Health*, **7**, 4.
Sager, M.A., et al. (1996). Functional outcomes of acute medical illness and hospitalization in older persons. *Archives of Internal Medicine*, **156**, 645–52.
Saltman, R.B. and Figueras, J. (1997). *European Health Care Reform: Analysis of Current Strategies*. Copenhagen: WHO.
Saltman, R.B. and von Otter, C. (eds.) (1995). *Implementing Planned Markets in Health Care*. Buckingham: Open University Press.
Sanders, D., Nandi, S., Labonté, R., Vance, C., Van Damme, W. (2019). From primary health care to universal health coverage—one step forward and two steps back. *The Lancet*, 394(10199), 619–21.
Sandvik, H. (1999). Health information and interaction on the internet: a survey of female urinary incontinence. *British Medical Journal*, **319**, 29–32.
Schoen, C., et al. (2011). New 2011 survey of patients with complex care needs in eleven countries finds that care is often poorly coordinated. *Health Affairs*, **30**, 2437–48.
Schwartz, F.W., Glennerster, H., and Saltman, R.B. (eds.). (1996). *Fixing Health Budgets—Experience from Europe and North America*. Chichester: Wiley.
Scitovsky, A.A. (1988). Medical care in the last twelve months of life: the relation between age, functional status, and medical care expenditures. *The Milbank Memorial Fund Quarterly: Health and Society*, **66**, 640–60.
Smith, S.D., Heffler, S.K., and Freeland, M.S. (2000). *The Impact of Technological Change on Healthcare Cost Spending: An Evaluation of the Literature. Health Care Financing Administration Working Paper*. [Online] Available at: http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/downloads/tech_2000_0810.pdf
Starfield, B., et al. (2005a). The effects of specialist supply on populations' health: assessing the evidence. *Health Affairs*, Suppl. Web Exclusives: W5–97–W5–107.
Starfield, B., Shi, L., and Macinko, J. (2005b). Contribution of primary care to health systems and health. *The Milbank Quarterly*, **83**, 457–502.
Stevens, A. and Gabbay, J. (1991). Needs assessment needs assessment. *Health Trends*, **23**, 20–2.
The Economist (2012). Rethinking the welfare state: Asia's next revolution. *The Economist*, 8 September.
Velkova, A., Wolleswinkel-van den Bosch, J.H., and Mackenbach, J.P. (1997). The east–west life expectancy gap: differences in mortality from conditions amenable to medical intervention.

International Journal of Epidemiology, **26**, 75–84.
Warner, M. (1996). *Implementing Healthcare Reforms Through Substitution*. Cardiff: Welsh Institute for Health and Social Care.
Wennberg, J.E. (1984). Dealing with medical practice variations: a proposal for action. *Health Affairs*, 3(2), 6–31.
Wilkin, D. (1992). Patterns of referral: explaining variation. In: Roland, M. and Coulter, A. (eds.) *Hospital Referrals*. Oxford: Oxford University Press, pp. 76–91.
World Health Organization (WHO) (1986). *Ottawa Charter for Health Promotion: First International Conference on Health Promotion*. Geneva: World Health Organization.
World Health Organization (WHO) (2019). WHO welcomes landmark UN declaration on universal health coverage. [Online] Available at: https://www.who.int/news-room/detail/23-09-2019-who-welcomes-landmark-un-declaration-on-universal-health-coverage
World Health Organization (WHO) (2020). Universal Health Coverage. [Online] Available at: https://www.who.int/health-topics/universal-health-coverage#tab=tab_1
Wright, J., Williams, R., and Wilkinson, J.R. (1998). Development and importance of health needs assessment. *British Medical Journal*, **316**, 1310–13.
Young, D.W. and Saltman, R.B. (1985). *The Hospital Power Equilibrium—Physician Behaviour and Cost Control*. Baltimore, MD: Johns Hopkins University Press.

11.9

Training of public health professionals in developing countries

San Hone and Roger Detels

Introduction to training of public health professionals in developing countries

Public health and medical systems and training are essential for the health of a country (Heller et al. 2007; Bangdiwala et al. 2010). This is particularly true for developing countries, in which great disparities in wealth exist and there are limited funds to invest in the health of their citizens. Developing countries have being exerting tremendous efforts towards achieving universal health coverage (UHC) through primary health-led approaches for which public health services and medical resources must be made available at national, provincial, subprovincial, and local district levels. The skills and requirements of public health and medical professionals at each of these levels differ and are unique within each level. The mandate of public health is broad, and therefore requires some level of specialization within the profession. Thus, various public health professionals are responsible for assuring access to primary healthcare, as well as maternal and child health services, protection of the environment, including provision of adequate and safe levels of water and safe disposal of human and other waste, implementation of community health programmes such as immunization coverage, prevention, and control of endemic and epidemic health threats from microbial agents, addressing increases in chronic illnesses, and administration/management of the broad and diverse public health workforce at each level. Thus, public health training requires providing training not only in medical sciences, but also in technical fields such as microbiology, chemistry, the physical sciences, management sciences, and behavioural, social, and political sciences, statistics, epidemiology, and research methodologies.

The World Economic Forum launched a 4-year US$100 million fund to support two non-governmental organizations, Last Mile Health and Living Goods, to train 50,000 health workers in six countries, bringing global attention to the need to improve healthcare at the lowest levels (Winters et al. 2018). The focus is to demonstrate how technology can reinvent community healthcare at scale to ensure quality healthcare to marginalized populations in lower-/middle-income countries, using utilitarian but cutting edge strategies to train health workers; e.g. digital education in the implementation of 'M Health' discussed next.

The public health workforce draws individuals from a variety of backgrounds relevant to their intended public health roles, including medicine, nursing, and the behavioural, management, and biological/physical sciences (Beaglehole and Dal Poz 2003). Generally, health training programmes provide training at the certificate level mainly for health professionals and others serving at the local level, which may include health volunteers (who are the core of many health systems and provide service at the village level), specialized training (e.g. midwife), as well as training at the master's level (usually requiring 1–3 years of graduate study) and the doctoral level (requiring 4–6 years of graduate study) for those who become leaders of public health programmes and research. Training at the graduate level also includes specialized training in specific fields for those who focus on environmental, social, behavioural, and healthcare issues of public health essential for a broad mandate of public health. Although most public health training programmes provide basic initial training in public health, there is also an important need to provide postgraduate and refresher training to ensure that the public health workforce continues to utilize the most modern strategies and resources.

In the opinion of some public health professionals, including the authors, there is too often a mismatch between the roles and responsibilities of public health professionals at each level and the training that they receive. Too often, public health curricula in developing countries were developed by the European colonialists based on the European system, and do not adequately address current health problems of the country. This mismatch reflects both the content of the training programmes and the strategies/content of the country's public health system itself. Public health planners, managers, and administrators are confronted by a range of barriers to achieving optimal public health, including funding constraints, politics, and the scarcity of

qualified personnel and public health leaders. It is the responsibility of public health trainers to develop strategies and curricula to overcome these barriers and to ensure that the training of public health professionals is appropriate and has the flexibility to provide trainees with the ability to devise new strategies to address constantly changing public health issues and problems. In some instances, this requires reconsideration of the current public health structure to ensure maximum effectiveness. Public health programmes need to ensure that health professionals in training see a clear career path for themselves, that there is an optimal distribution of health professionals, particularly to rural and remote areas, and that the system provides both monetary and non-monetary incentives to motivate them. In this chapter, we address each of these issues in greater depth and make suggestions for improvement and enhancement of health training in developing countries.

Typical public health structures in developing countries—health system structure in developing countries

Typically, government health systems in developing countries have a tiered system for providing health, comprising the national, provincial, subprovincial, and local levels. In addition, in most developing countries there are private healthcare sectors that provide services primarily to the wealthy, and national and international non-governmental organizations that also provide public health services through community volunteers and clinical services. Sometimes these operate outside the purview of the government public health system which can create problems (see chapter 11.12). In this chapter we will focus on training relevant for the government-sponsored public health system.

The greatest number of health personnel work at the lowest level, with decreasing numbers at the subprovincial, provincial, and national levels. The highest levels of training and responsibility focus on the national level. The duration and intensity of training for health workers increase with greater levels of responsibility for programme leadership.

The public health systems in developing countries are usually divided into three levels: (1) the central administrative level that develops health policies, recommends strategies for the entire country and administers national

programmes, national training institutions, and national hospitals that provide complicated tertiary care; (2) the middle level comprised of provincial and subprovincial health departments that are responsible for implementing national policies and programmes, and provincial hospitals that often see referrals from the local healthcare facilities and conduct regional training programmes for local health professionals and volunteer health workers; and (3) the local or district level that implements national and provincial programmes such as primary healthcare, maternal and child health programmes, and immunization campaigns (Fig. 11.9.1).

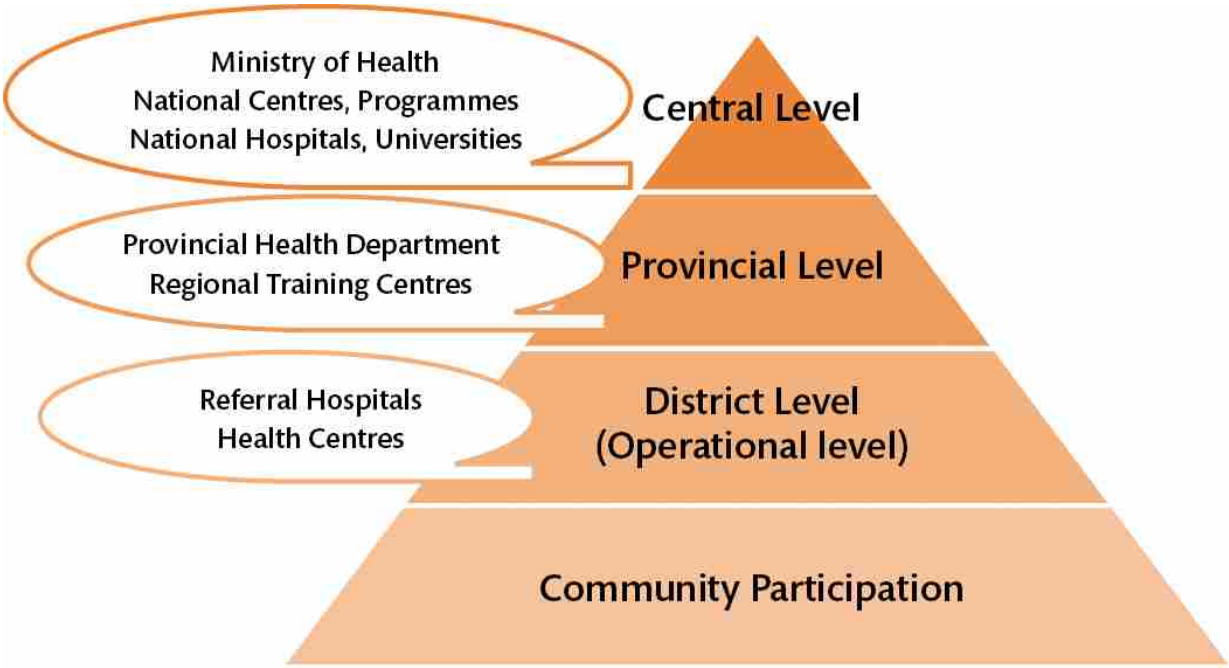


Fig. 11.9.1 The health system.
Reproduced by kind permission of Dr Lo Veasnakiry.

The district level (health centres or health posts) usually provides a minimum package of primary healthcare and public health activities to a population of approximately 5,000–30,000 each and a referral hospital for low-level uncomplicated health problems. The health force at the village health centre includes primary/community healthcare workers, nurses and/or nurse midwives who receive training for one or more years and are responsible for primary healthcare services, including maternal and child, family planning, childhood immunizations, and delivery of treatment for

non-communicable diseases, HIV and tuberculosis which required daily observed treatment. With the recent trend of moving towards the global elimination of certain communicable diseases such as malaria, vaccine-preventable childhood diseases (polio, measles), and elimination of mother to child transmission of HIV and syphilis, etc., the role of the primary health care providers has expanded.

In some countries, these primary care providers are assisted by public health staff who receive about six or more months of training focusing on health education for the endemic and epidemic diseases and environmental sanitation activities. At the village or commune level, there are often volunteer health workers who are selected by the local population and provide minimal first aid and referral to the district health centre or hospital, and who assure directly observed therapy for tuberculosis and other diseases. They also help organize immunization sessions, give oral vaccinations (e.g. polio) and high-dose vitamin A to children, but are not allowed to give injections. In addition, they promote health education activities, including promoting environmental sanitation. The position carries with it a level of prestige that is its main reward, although some level of remuneration is usually advisable. Volunteer health workers usually receive training for several weeks. The characteristics of an ideal volunteer health worker have been outlined by Taylor et al. (2018), and include knowledge of the local population and culture, good interpersonal and communication skills, the ability to see the world from others' perspectives, a non-judgemental attitude, 'imbeddedness' within the community, shared demographic characteristics, and a desire to assist people.

Responsibilities/roles of public health professionals at each level in developing countries

Levels of training and duration vary, depending on the role that the graduate is expected to play in the health system. At the highest level, usually for bachelor's or master's programmes, the trainees receive instruction in administration of health programmes (although care is usually taught through the medical education system), content and strategies for health education, and behavioural modification to achieve a healthy lifestyle, strategies for assuring a safe water supply, appropriate disposal of human

and other waste products, oral health, systems of health financing, control strategies for the major infectious diseases affecting the country (which usually include tuberculosis, malaria, parasite infections, and HIV, as well as any locally endemic diseases), and administration and strategies for immunization programmes. Some developing countries require a medical degree or other health science or technical degree as a prerequisite for entering into a master of public health (MPH) programme. Graduates of these programmes are expected to lead public health programmes. Programmes at this level usually require the trainee to complete an internship in the field at the local or rural level to ensure their understanding of the system and the culture of the populace at the point of delivery. Research methodologies, if available, are usually taught in the relatively few PhD/doctoral programmes in medical and public health institutions. However, many developing countries are only recently investing in schools of public health, and in most, research is not conducted. Recently there has been a movement to use the internet to provide distance learning for public health workers at the MPH level and for postgraduate/refresher courses that target developing countries, which are offered through various universities and government agencies (Mokwena et al. 2007). Although these programmes are probably quite useful for teaching methodology and public health strategies, they may have difficulties in training relevant to practising at the local level in different countries, as they cannot offer hands-on field experience. One strategy to overcome this difficulty is to have discussion sessions on specific topics through the web, which include practitioners from the local level. Because most developing countries do not have the resources to invest in research programmes, most public health programmes are not evidence-based.

At the village level, many countries employ volunteer health workers who have had only a few weeks of training and perform very routine functions, such as tending to minor wounds, routine first aid, and encouraging villagers to take their medications as recommended by physicians and allied health workers at the district or higher levels. The role of the volunteer health worker may include home visits, first aid, treatment of simple and common ailments, health education, nutrition and surveillance, family planning, referrals, collection of data on vital events (births, deaths), and assisting more highly trained health professionals such as the nurse midwives. They usually have the best knowledge of the village and the villagers in terms of

health status and needs, and are often selected by the communities themselves. To be successful they need to be an integral part of their community and have the support of that community. It is important that they receive some reward from the villages, as they are not usually paid by the government.

A typical health structure at the district/township level is presented in Figure 11.9.2.

Health professionals at the village health post include nurse midwives and allied health personnel and health workers who usually have one to three years of training. Nurse midwives are responsible for encouraging and implementing enrolment in prenatal programmes, staffing antenatal clinics, delivering babies, and providing postnatal guidance. They need to be able to recognize difficult pregnancies that will require more highly trained medical personnel to ensure safe delivery. These problem pregnancies are usually referred to a higher-level facility. Transportation to the hospital, particularly for emergency and complicated deliveries, can be a serious problem. Allied health personnel with nurse midwives usually perform the other medical, sanitation, and health education roles, and direct and facilitate health campaigns such as polio immunizations days. In some countries, nurse midwives perform both roles, although they may be assisted by assistant nurse midwives and/or public health staff who have had less but some training. Nurse midwives may be responsible for as many as 10–15 villages. The rural and urban health centres may have facilities for care of patients, including injuries requiring more than simple first aid. Patients with complicated illnesses, however, are referred to higher-level health facilities, and are diagnosed and treated by physicians.

With the trend towards UHC, many countries are striving to identify feasible and cost-effective strategies for providing health services. Major strategies include decentralization of some services that were previously not available at the village health level; for example, performing HIV screening tests and introducing a package of disease reduction services for primary healthcare workers. Under these schemes, primary healthcare providers, nurse midwives, and lower-level health workers are also responsible for conducting packages of activities to target communicable diseases for which the government has pledged to achieve elimination by a specific time. These usually include active case surveillance, implementing appropriate screening

tests, implementing immunization campaigns, introducing standardized treatment, and implementing timely reporting.

The district hospitals may have a laboratory to perform routine procedures, but are usually not able to do more sophisticated biochemical and microbiological assays, which are usually done at the provincial level. One problem with the laboratories at each of the different levels is that quality control procedures are often not implemented and standardized. Few of these laboratories participate in quality control programmes, although some wealthier developing countries (e.g. China) have implemented quality control programmes at the county and provincial levels. Thus, there is a problem with achieving valid test results and safety in many of these laboratories (Callaway 2012). A second problem is staffing these laboratories with well-trained, competent technicians who need to periodically update their laboratory skills to ensure valid results and safe procedures.

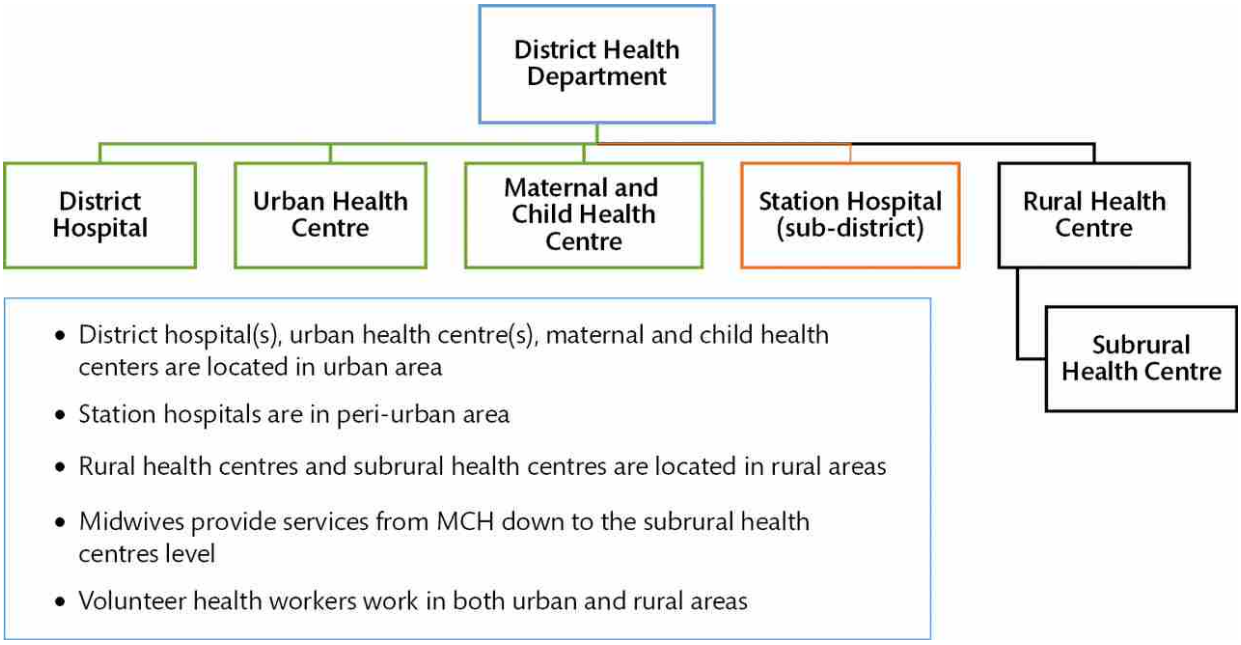


Fig. 11.9.2 Health structure/system at the district level.

Complicated surgical procedures and medical management of serious health problems are usually performed only at the provincial and national/tertiary levels. A serious problem in developing countries with

limited resources and populations residing in remote areas is diagnosis and timely transport of patients to tertiary care facilities to ensure successful treatment.

Training of health workers may occur at each of the levels of the public health system, although training of higher levels of health professionals is usually limited to universities, medical schools, and schools of public health, dentistry, health sciences, and nursing.

Education for health professionals in developing countries

In all developing countries, completion of medical school is required for healthcare professionals working in the various healthcare facilities at the national, provincial and district level hospitals. Many developing countries require graduating physicians to work a minimum number of years at the district level as a strategy for providing healthcare at the peripheral levels of the health system. A medical degree or a master's degree in public health is often also required at the national and provincial levels for staff responsible for administering national programmes. Specialized and advanced medical training is usually only required at the national level healthcare facilities, and sometimes at provincial-level health facilities. Specialized training is usually required for health professionals conducting dental care and pharmaceutical services.

Advanced training is required for health professionals who will design and direct health information systems such as disease surveillance, and who will design and conduct health research to inform policy decisions and evaluate health programmes. These activities are usually directed at the national level, although they may involve staff at the provincial, district, and local levels.

Nurses usually require bachelor's degree level training. However, midwives who work at the local level may receive an associate midwife degree, which may require only 1–2 years of certificate-level training, depending on the country. Some countries also have midwife assistants who receive shorter periods of training. Staff providing lower-level nursing, medical imaging, sanitary services, low-level laboratory services, and physical therapy usually have limited training.

Problems with providing public health training in developing countries

Health professionals at the district level, who often have had only 1–3 years of training, are expected to assist in planning and management of public health programmes, recognition of the need for and initial conduct of outbreak investigations; implementation and monitoring of immunization programmes; prevention, treatment, and monitoring of tuberculosis and HIV programmes (e.g. directly observed therapy); implementing specific communicable disease elimination strategies (e.g. active case investigation for malaria elimination); and basic health education in strategies to prevent disease and promote healthy lifestyles. Their training may include the various skills and knowledge required to carry out these responsibilities, including the need for collection of accurate data that can be used to plan and support effective health systems (International Institute for Viral Registration and Statistics 1989). Thus, there is a need for congruency in the training of public health professionals and the responsibilities they will assume (Sudhinaraset et al. 2013). A major barrier to achieving proper training at all levels of public health training is the scarcity of qualified faculties in these programmes. Refresher training for health workers is important to ensure that they keep up with current and changing strategies and policies. The training of health professionals at the local level is often not integrated, so staff members know only their own responsibilities but do not have a holistic understanding of where their efforts fit into the broad public health system.

Funding for public health programmes and training is often a low priority for political leaders and decision-makers. The majority of health programmes in developing countries are subsidized by international donors who have their own health agendas, which may differ from those of the country (see chapter 11.12).

Clear public health career tracks are often not apparent for qualified individuals who wish to choose a career in public health. Complicating recruitment of public health professionals at every level is the low level of remuneration. Often physicians who are willing to work in public health must also have private practices to generate enough income to support their families (Chomitz et al. 1998; Ferrinho et al. 2004). Thus, they are not able to dedicate their full efforts to their public health responsibilities. Physicians

are often reluctant to work in remote rural areas (WHO 2010a). Public health workers at the peripheral level are usually paid low wages or are expected to volunteer their services. In these situations, other reward structures must be created to motivate volunteers and low-level workers (Henderson and Tulloch 2008; Mangham and Hanson 2008).

Strategies to overcome barriers to optimal public health training

The fundamental principles for designing public health training involve both academic and practical components. For training programmes to be responsive to needs, the needs must be clearly delineated. It is important that the tasks and responsibilities of health professionals at every level of the public health system be designed to meet the needs of the constituency and the level of responsibility expected at that level. A first step is to conduct needs assessments at each level, from the local and provincial levels to the national level, and from primary health care providers to researchers and policy-makers (Bangdiwala et al. 2010). The assessment should involve members of the community or constituency at that level. Formulation of the responsibilities and tasks at each level can then be clearly defined, and documentation of formal job descriptions provided so that an appropriate training curriculum may be developed for health workers expected to perform the required tasks at that level of the system. Training resources should then be identified and/or developed for the required training at each level of the public health system. At each level, the curriculum should be designed and evaluated for responsiveness to the job descriptions of the positions that graduates are intended to fill. The faculty of the training programmes at each level should include individuals who have had experience in delivering services at that level. The training should include providing an understanding of where the responsibilities of the health professional being trained fit into the public health system of the country.

Incorporating new technologies

Development and widespread use of hand-held computers/phones, global positioning systems, and increasingly sophisticated communications

systems, called ‘M Health’, provide opportunities for monitoring, mentoring, and enhancing health activities, particularly at the local level. Implementation of ‘M Health’ strategies can lower costs, while enhancing the quality of the health service delivered, and the completeness and ease of real-time events and disease reporting and transparency and accountability for both funders and donors. For example ‘M Health’ strategies can be used for training and updating of skills via videos sent to mobile devices for: connecting the primary health care providers with supervisory staff for assistance in health and logistic problems that they encounter in the field; accessing higher-level health personnel for advice about care; linking individuals with serious health problems to a higher level of care; reporting health activities such as immunizations and community outbreaks in real time; facilitating rapid public health responses; and increasing the confidence of local health workers because their mobile devices give them access to a higher level of support. Mobile phones used in conjunction with global positioning systems can be used to monitor primary health care providers performance by assuring that they are where they report being and having them check in and out for each home visit, etc. These low-cost strategies are particularly relevant for developing countries with limited budgets and many disease problems. It is important to deploy mobile devices (technology) to facilitate the global goal of UHC, which is dependent on the performance of community-based public health professionals. Updated mobile devices can serve as a platform for practise-based education, training and mentoring through which knowledge and skills for clinical service delivery can be transferred from super-tertiary centres to the community level, and thus ultimately ensure not only the quality of services but also the sustainability and feasibility of UHC (Project-ECHO 2017). In addition, the strategies of using mobile devices can be configured according to specific needs, such as a referencing job aids providing standardized health messages and guiding stepwise procedures for service delivery, a trigger point for initiating case investigation, and to advance the response for effective disease surveillance, etc. The use of these emerging strategies and technology to enhance the performance of health workers and meet the needs of villagers at low costs need to become part of health systems and part of the curriculum for training at every level (Rotheram-Boris et al. 2012; Tomlinson et al. 2013; Taylor et al. 2018).

It is important for those designing public health systems to delineate clear career paths for health professionals at each level, which will inform and, if done well, attract potential students to pursue those careers. An important component of defining career paths is to include incentives for individuals in these career paths (Calcutta Declaration 1999). A problem in many developing countries is the high attrition rate of trained health volunteers, in part due to a lack of monetary incentive. In some developing countries, training, especially for nurses, may better prepare trainees for work in developed countries. Thus, migration of health professionals whose training was funded by the developing countries, particularly doctors and nurses, to better working conditions and pay in developed countries reduces the potential workforce (Stilwell et al. 2004; Naicker et al. 2009; WHO 2010b). In these situations, developing countries may be considered to be subsidizing healthcare in developed countries!

Incentives for health workers should include monetary rewards, but it is important that they also include recognition of the importance of the health professionals at every level in the entire system for advancing the health of their constituencies. This recognition is particularly important at the lowest and least financially rewarding levels. Prestige and recognition of worth are powerful incentives!

Assuring that the public health system remains responsive to changing needs

Finally, it is important to monitor and periodically evaluate the performance of the health system at each level. Over time, specific needs change due to a range of factors and new technologies are developed; therefore, training programmes at every level must adapt their curricula to meet current needs and rapidly emerging technologies. Previously trained health workers need to update their skills. Thus, it is important to also include continuing training at all levels for health workers that will assure that they can continue to meet the changing demands of the health system.

REFERENCES

Bangdiwala, S.I., et al. (2010). Workforce resources for health in developing countries. *Public Health Reviews*, **32**, 296–318.
Beaglehole, R. and Dal Poz, M.R. (2003). Public health workforce: challenges and policy issues. *Human Resources for Health*, **1**, 4.
Calcutta Declaration on Public Health (1999). Regional Conference on Public Health in South East Asia in the 21st Century, Calcutta, 22–24 November 1999. *Journal of Epidemiology & Community Health*, **54**, 749.
Callaway, E. (2012). Biosafety concerns for laboratories in the developing world. *Nature*, **485**, 425.
Chomitz, K., et al. (1998). What do doctors want? Developing incentives for doctors to serve in indonesia's rural and remote areas. *Unpublished World Bank Policy Research Working Paper 1888.*
Ferrinho, P., et al. (2004). Dual practice in the health sector: a review of the evidence. *Human Resources for Health*, **2**, 14.
Heller, R.F., et al. (2007). Capacity-building for public health. *Bulletin of the World Health Organization*, **85**, 901–80.
Henderson, L.N. and Tulloch, J. (2008). Incentives for retaining and motivating health workers in Pacific and Asian countries. *Human Resources for Health*, **6**, 18.
Institute for Viral Registration and Statistics (1989). *Health Data Issues for Primary Health Care Delivery Systems in Developing Countries*. Bethesda, MD: International Institute for Viral Registration and Statistics.
Mangham, I.J. and Hanson, K. (2008). Employment preferences of public sector nurses in Malawi: results from a discrete choice experiment. *Tropical Medicine & International Health*, **13**, 1–9.
Mokwena, K., et al. (2007). Training of public health workforce at the National School of Public Health: meeting Africa's needs. *Bulletin of the World Health Organization*, **85**, 949–54.
Naicker, S., et al. (2009). Shortage of health care workers in developing countries—Africa. *Ethnicity & Disease*, 19(Suppl. 1), S1–60–4.
Project-ECHO (2017). Website. [Online] Available at: https://hsc.unm.edu/echo/
Rotheram-Boris, M.J., et al. (2012). Standardized functions for smartphone applications: examples from maternal and child health. *International Journal of Telemedicine and Applications*, **2012**, 1–16
Stilwell, B., et al. (2004). Migration of health-care workers from developing countries: strategic approaches to its management. *Bulletin of the World Health Organization*, **82**, 595–600.
Taylor, B., Mathers, J., and Parry, J. (2018). Who are community health workers and what do they do? Development of an empirically derived reporting taxonomy. *Journal of Public Health*, 40(1), 199–209.
Tomlinson, M., et al. (2013). Value of mobile information system to improve quality of care by community health workers. *South African Infection Management,* 15(1), 1–17.
Sudhinaraset, M., et al. (2013). What is the role of informal healthcare providers in developing countries? *PLoS One*, 8, e54978.
Winters, N., O'Donovan, J., and Geniets, A. (2018). A new era for community health workers in countries of low and middle income? *Science Direct*, **6**, 489–90.
World Health Organization (WHO) (2010a). *Increasing Access to Health Workers in Remote and Rural Areas Through Improved Retention. Global Policy Recommendations; Executive Summary*. Geneva: WHO.
World Health Organization (WHO) (2010b). *Migration of Health Workers. Fact Sheet No. 301*. Geneva: WHO.

11.10

Transformative learning for health professional in the twenty-first century for the future health workforce

Wanicha Chuenkongkaew and Suwit Wibulpolprasert

Why transforming health professionals learning?

Since the 1910 Flexnor report, a set of studies of health professional education system has been proposed for a continuous reform. The empirical experiential apprenticeship was replaced by a science-based curricula at non-commercial and university-based medical schools. The reforms equipped health professionals with the knowledge and skills that doubled life expectancy during the twentieth century (Frenk et al. 2010).

At the start of the twenty-first century, new complex challenges emerged that affect worldwide healthcare services systems. Inequities in term of healthcare access and quality underscore failure to share health advances both domestically and internationally. An unprecedented rate of technology development, rapid demographic transition towards aged and urban populations with increasing chronic noncommunicable diseases (NCD) and disability, rapid globalization favours easy spread of emerging such as COVID19- coronavirus pandemic and re-emerging infectious diseases, have moved the world into a situation of VUCA (Volatility, Uncertainty, Complexity, and Ambiguity). All these emerging challenges, together with the top political commitments to achieve Sustainable Development Goals (SDGs) by 2030, require a totally new type of health professionals.

The current health professional education systems have not been well adapted to address the global challenges. This is mainly caused by outdated, static and fragmented curricula, institutes, and faculty members. The graduates are thus equipped with insufficient knowledge, attitude, and competencies to match the present and future healthcare systems (Wasi 2000). There are many challenges including:

Poor teamwork and inadequate collaboration among health professionals
Gender inequity of professional status
Narrow attention on technique without understanding the broader context
Episodic encounters with patient illnesses rather than holistic continuous care
Emphasizing hospital-based care rather than community and home-based, primary care and disease prevention
Inadequate understanding of social determinants of health
Quantitative, qualitative imbalances, inequitable distribution, and inappropriate skill-mix in the professional labour market
Poor leadership in health system development

Thus, the need to transform the health professional education systems is imminent. The movements must ensure that the new graduates fit well with the changing health challenges and health systems development.

Health challenges and future healthcare delivery system

Global emerging challenges including rapid globalization, technologies development, demographic, and epidemiological transition have been affecting the healthcare delivery system. We can envisage more chronic diseases, both Communicable (CD) and Non-Communicable Diseases (NCD), and Public Health Emergencies, linked to social determinants of health (SDH). We see the rapid increases in health expense and equity. Socio-economic development has also resulted in changing demand and provider-patient relationships.

The World Bank and the World Health Organization (WHO) reported in 2017 that half of the world's population lacked access to essential healthcare. Moreover, almost 100 million people were pushed into extreme poverty, forcing them to survive on only $1.90 or less a day (Tracking Universal Health Coverage 2017). To end people's extreme poverty, which is the greatest challenge and ambitious goal of SDGs, inclusive and sustainable economic growth is needed. This can be achieved through policy coherence of economic, social, and environmental sectors.

Universal Health Coverage (UHC) is an effective social measure to tackle health disparity and medical impoverishment at the same time. The United Nation SDGs 3.8 aims to achieve UHC in 2030.

UHC is defined as ‘all people can access the health services they need without incurring financial hardship. It includes the access to essential comprehensive health promotion, disease prevention, treatment, rehabilitation, long-term, and palliative care services and technologies, without financial barriers’.
To achieve UHC, five criteria must be met.

First, the *availability* of essential health services. This includes an adequate number of skilled Health Work Force (HWF) employed with decent working conditions.
Second, *affordability* to ensure no financial hardship through minimizing out-of-pocket (OOP) payments at point of services.
Third, *accessibility* with equitable geographical distribution and near home primary and community-based healthcare.
Fourth, *acceptability* with culturally appropriate, gender sensitive, non-discriminatory and responsive to special needs of specific populations,
Fifth *quality and accountability*, including quality of care, patient safety, and responsiveness to patients’ demand, as well as people’s engagement.

Public health (PH) is defined as ‘… the science and art of preventing disease, prolonging life and promoting health and efficiency through organized community effort’ (Grant 2012).

Essential PH functions include strengthening primary and community-based healthcare; increasing capability, accessibility, and quality of individual and people-based healthcare services; reinforcing evidence informed policy decision processes through Health Systems Researches; strengthening qualified and equitable HWF; protecting health and safety both in normal and crisis situations and creating social innovation for PH problems. With emerging threats from health in crisis, including man-made, natural disasters, and pandemics, PH capacity is even more essential to ‘prevent, detect, and respond’ to protect health and humanity.

Thus, UHC and PH are closely linked and should be developed in parallel with good integration.

Investing on HWF is the key to achieve health-related SDGs

To achieve SDG targets in 67 low- and middle-income countries (LMIC), more investment on health is needed. A prospective investment of $41-$57 per person, per year on health by 2030 could provide a 9:1 return of global economic growth. The International Labour Organization (ILO) determines a threshold of 41.1 HWF per 10,000 population to achieve UHC. The strategic investment should focus more in the frontline HWF (potentially for women) in rural and underserved communities.

Currently, the world's urbanized population is projected to double from 30% in 1950 to 66% in 2050. Despite the rapid urbanization projected for the next two decades, African and Asian rural populations are estimated to be almost 90% of the global rural population, mostly in India, China, and Nigeria (United Nations 2014), (Fig. 11.10.1). Only 22% of the global urban population has limited healthcare access, whereas 56% of the rural population lacks access to essential healthcare services (International Labour Office 2015) (Fig. 11.10.2). The rural dwellers in Africa are the most deprived group in healthcare access. The challenge to achieve health-related SDGs is to focus to a greater degree in the LMIC where small city growth becomes urbanized rapidly.

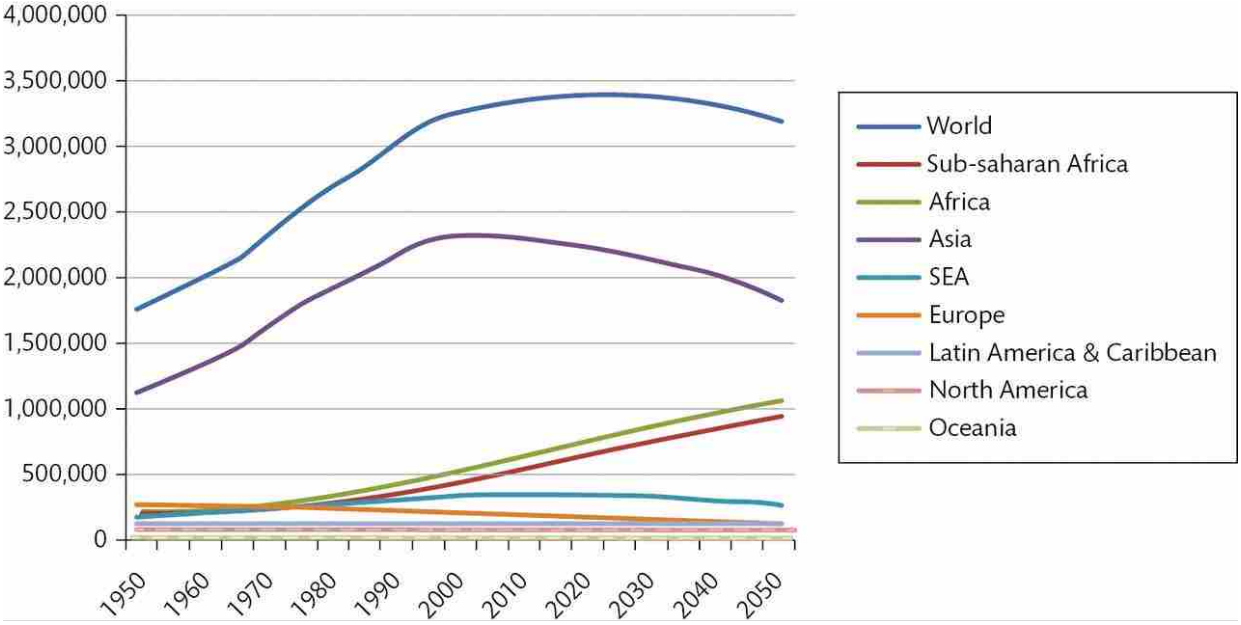


Fig. 11.10.1 Annual rural population at mid-year (thousands).

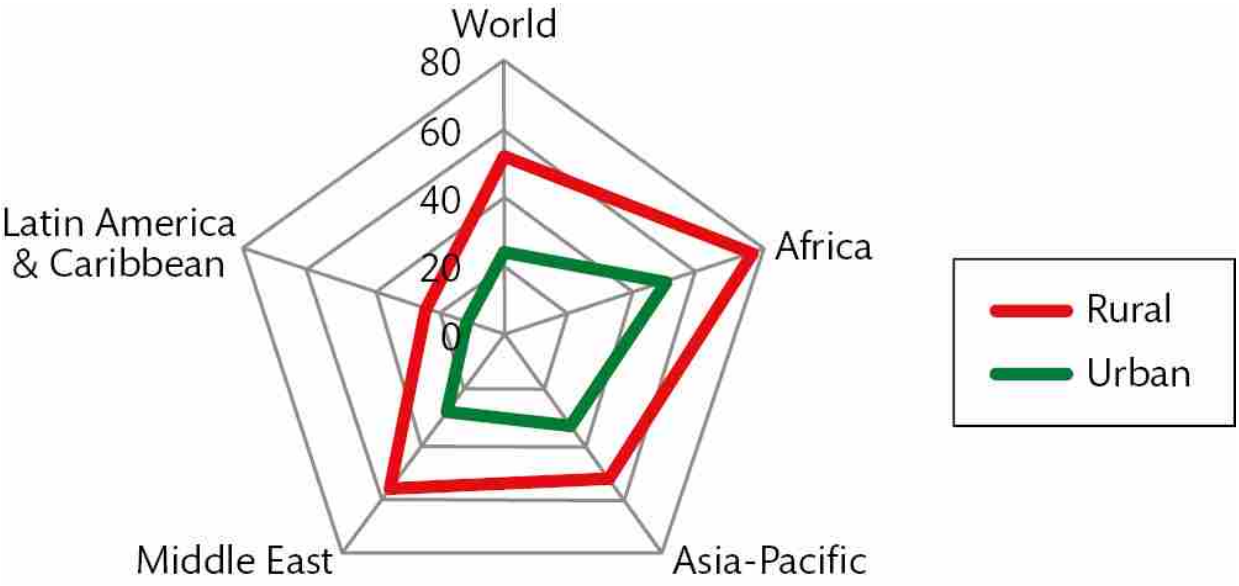


Fig. 11.10.2 Rural/urban population with lack of health coverage due to HWF shortages, 2015 (ILO HWF lack of access to healthcare services, percentages) The 'urban' population is defined as the definition using in each country. The definitions are applied by national statistical offices in completion of the latest available census.

Accessibility, availability, financial protection of efficient healthcare services, OOP expenditure and maternal mortality ratios (MMR) are five indicators for measurement of key dimensions of healthcare coverage and access (Fig. 11.10.3).

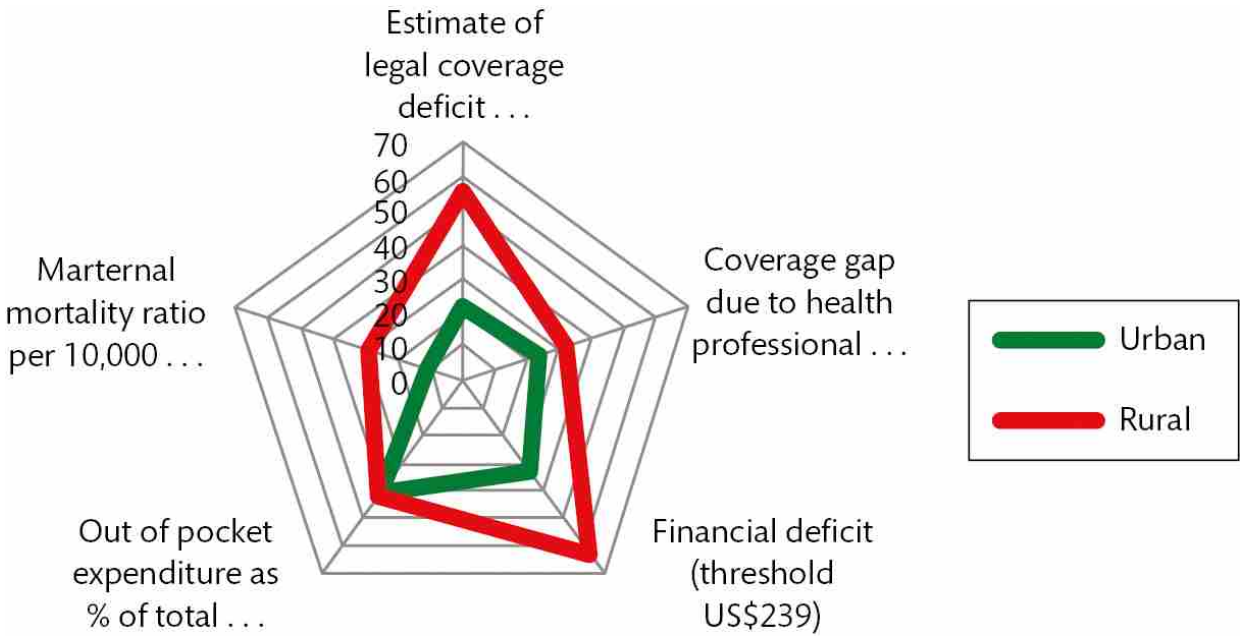


Fig. 11.10.3 The global inadequacy in access to quality healthcare service: rural/urban population 2015 (percentage) Regional averages weighted by rural and urban dwellers except for MMR, by skilled birth attendants.

Source: ILO estimates, 2015.

Although we are facing rapid urbanization, more than 50% of the world population is in the rural underserved areas (Fig. 11.10.2), where severe shortage of HWF is a major challenge (Table 11.10.1). Strengthening primary and community-based healthcare and mobilizing community partnerships are essential to address the urban/rural divide. It requires strong basic primary healthcare (PHC)-based health systems manned by qualified committed health volunteers and professionals. Its success depends on active engagement of families, local government, community leaders, and intersectoral collaboration. Empowering HWF through transforming learning and practice in the community is a key contribution to address the disparity in UHC.

Table 11.10.1 A mismatch in the global and national health labour market

	Stock			Needs-based shortage			Labour market demand	
	2013	2030	Change (%)	2013	2030	Change (%)	2013	2030
Income								
High	17.6	24.3	38	0.1	0.1	-7	17.7	23.8
Upper-middle	14.7	22.6	54	3.7	1.8	-50	19	33.3
Lower-middle	10	18.3	82	9.1	6.6	-28	10.9	21.7
Low	1.1	2.1	86	4.6	6.1	33	0.6	1.4
Region								
Africa	1.9	3.1	63	4.2	6.1	45	1.1	2.4
Americas	9.4	14	50	0.8	0.6	-17	8.8	15.3
Eastern Mediterranean	3.1	5.3	72	1.7	1.7	-1	3.1	6.2
Europe	12.7	16.8	32	0.1	0.1	-33	14.2	18.2
South-East Asia	6.2	10.9	75	6.9	4.7	-32	6	12.2
Western Pacific	10.3	17.3	68	3.7	1.4	-64	15.1	25.9
World	43.5	67.3	55	17.4	14.5	-17	48.3	80.2

All health worker figures in millions, rounded to nearest 100,000

Stock and shortage figures refer to 193 Member States

(Source: WHO Global Strategy on HRH http://who.int/hrh/resources/globstrathrh-2030/en/)

Labour market demand figures refer to 165 Member States with sufficient data (source: World Bank)

A priority approach is community-engaged health professional learning, a mechanism to envision learners how to work in and with communities, in other words, learning how to involve the people that live in community with local redevelopment. It is a key component to produce health professionals with skills and spirit to the community for the reduction of geographical inequities. An engagement of learning develops when the community and academic partnership sustainably collaborate on the design, delivery, and

programme evaluation with equal value to empower the community by training qualified and motivated health professionals to provide community-based service for health equity.

Transformative learning for effective community engagement needs socially accountable values including honesty, respect, authenticity, openness and transparency, and trustworthiness.

Global movement on transformative learning for health professional in the twenty-first century

In 2010, the Global Commission on Health Professionals Education for the twenty-first century was launched, on the 100th anniversary of the Flexner Report. The Global Commission report strongly proposed a system-based approach to replace scientific based curricula in health professional learning. The report triggered global movements to promote transformation, for instance:

The Southeast Asia Regional Committee resolution, SEA/RC/R6 on ‘Strengthening health workforce education and training in the region’ (WHO 2012).

World Health Assembly resolution, WHA 66.23 on Transforming health workforce education in support of UHC (WHA 2013).

The WHO recommendations 2013 for transforming and scaling up health professional education and training (WHO 2013). Eleven recommendations include faculty development in term of continuous professional development (CPD); accreditation; and recruitment process for faculty and teaching staff, curriculum development, simulation method, direct entry of graduates, admission procedures, streamlined educational pathways and ladder programmes, interprofessional education, accreditation, and CPD. The global recommendations provide a guide for transformative learning which include policy and technical guidelines for establishing undergraduate and graduate programmes It gives guidance on how best to produce graduates in response to health needs of the populations.

In the Prince Mahidol Award Conference (2014) ‘Transformative learning for health equity’, several new movements to tackle the current inadequacies of health professional learning have been gathering pace across a range of different constituencies (WHO 2006).

In 2015, as replacement to ‘unfinished’ health-related MDGs, a SDG 3.C on substantially increasing health financing and the recruitment, development, training and retention of the HWF in developing countries is a primary aim for the attainment of UHC (SDG 2015). To achieve UHC, all people must have access to an essential national package of healthcare services and essential, safe, effective, and good-quality medicines without financial barriers.

In 2016, the Global Strategy on Human Resources for Health: Workforce 2030 (GSHRH 2030), reiterates and revisits through a contemporary lens the agenda to adopt transformative strategies in the scale-up of HWF learning. The strategy outlined a set of milestone and key indicators on HWF learning. These include the establishment of an accreditation mechanism for health training institutions by 2020, progress towards halving current levels of disparity in HWF distribution (disaggregated by place of employment, for instance subnational and/or urban/rural/remote) and reduction of the prequalification attrition rates in medical, nursing and allied health professional schools to 20% or less by 2030.

Education strategies should focus on investment in trainers who have a high potential of social return as a priority. It should prioritize regulating curricula to balance on producing graduates in response to local needs and to international markets. A coordinated approach is essential to create close linkage between health professional planning, learning, and motivating interprofessional education (IPE) and interprofessional practice (IPP). National policies should include setting and monitoring the education standards and sufficient financial support. The collaboration between higher education and health sector should be implemented by the integration between health systems and competency-based learning.

Health professionals should be equipped with knowledge, skills, values, and attitudes in term of SDH and public health systems to build the most collaborative and effective interprofessional team. Building PH capabilities on epidemic preparedness and the global health security agenda to prevent, detect, and respond to emerging threats is a priority in current education strategies. The high quality of health training institutes should be regulated through accreditation mechanisms.

The United Nation Secretary General appointed a High-Level Commission on Health Employment and Economic Growth (HEEG 2016) which reported 10 concrete recommendations. The commission brought in the active involvement, at the top level relevant Intergovernmental Organizations including the ILO, the Organization for Economic Co-operation and Development (OECD), and the WHO. The commission recommends stimulation of the health and social sector jobs to drive inclusive economic growth for greater economic stability and security. The achievement can also make gains across SDG 1 (poverty elimination), SDG 3 (good health and well-being), SDG 4 (quality education), SDG 5 (gender equality) and SDG 8 (decent work and economic growth). The report also shows that between 2000 and 2014, in OECD countries, there was a 48% increase in health and social sector employment while the industrial and agriculture sector declined.

The commission recommended an increase of at more than 40 million new jobs in health and social sectors to reduce projected shortage of 18 million HWF, especially in LMIC and LIC by 2030. More health demand will create more jobs. The global economic return on health investment is 9:1 and one extra year of life raises gross domestic product (GDP) by 4%.

More than 70% of the global health and social service workforce are women, compared with less than 40% of those across all employment sectors.

The commission also recommended 10 specific actions. Six actions are on health system and HWF reform, including job creation; gender and woman's right; education, training, and skills; health service delivery and organization; technology; crisis management and humanitarian setting. The remaining four recommendations on enabling actions include financing and fiscal space; partnership and cooperation; international migration; data and information and accountability.

Transforming and scaling up high-quality learning should shift away from narrow specializations to general practice. It should highlight more on lifelong learning to build highly competent HWF. It has to generate motivated HWF in response to local and global health needs. Furthermore, reduction of unnecessary entry barriers is needed. These include addressing geographical inequities, strengthening youth education for better chance of employment, and reinforcing the effectiveness of community-based health workers.

New skills for the future health professionals and leaders

Addressing global emerging challenges, health professionals need to move towards better health promotion, health protection, disease prevention, people-centred care, and community-based and primary healthcare based health systems with extensive geographical coverage.

In the twenty-first century, competency-based curricula and team-based learning will replace traditional time-based learning. As defined by OECD, 'competency' is not only 'knowledge and skills' but the ability to meet complex needs by making use of psychosocial resources in a particular context. Effective health promotion needs to integrate health literacy, communication skill, and personal attitude. A competency-based approach is tailored to identify learner's outcome or performance to align learning and training with health needs of populations in different settings (Gruppen et al. 2012).

Increasing social and economic development allow increasing and timely access to healthcare and knowledge. The patient–doctor relationship has gradually shifted from vertical patron–client, to contractual relationship with reducing trust and increasing cost. To bring back trust-based relationship, the

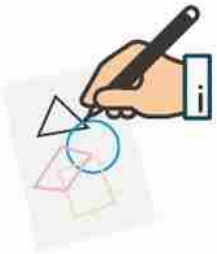
communitarian relationship based of community family doctors would be the new direction.

In the twenty-first century, basic skills should include learning, literacy, and life skills (Fig. 11.10.4) in the training process. In addition, health professionals are also required to understand the high intrinsic value of social justice; ethics; morality; human rights; equity in health; altruism and social accountability. These will facilitate their communication with patients and work with other professionals, families, and community reflecting a mutual respect. They need the skills to enquire, search, interpret, and use evidence for appropriate decision-making. They also need to comprehensively understand the influence of social, political, and economic effects on health status including a concept of Health in All Policies and the SDH.

21st CENTURY SKILLS

learning Skills

CRITICAL THINKING



CREATIVITY

COLLABORATION

COMMUNICATION

Literacy Skills

INFORMATION

MEDIA

TECHNOLOGY

Life Skills

FLEXIBILITY

LEADERSHIP

INITIATIVE

PRODUCTIVITY

SOCIAL

Fig. 11.10.4 Twenty-first century skills: learning, literacy and life skills.

New health professional leaders need five key soft skills for change as illustrated in Table 11.10.2 (Piirto 2010).

Table 11.10.2 Five ‘soft skills’ for changes

Intelligent	A good understanding or a high mental capacity; comprehend quickly
Imagination	A capacity to create a solution that is both novel and appropriate, divergent thinking, being a good critical thinker and problem solver
Inspiration	A sudden arrival of a novel idea, a ‘leap’
Insight	An ability to clearly understand the inner nature of things, particularly by intuition An insight manifests itself suddenly, requires preparation of hard work, believe in reconceptualization, connects old and new information, and applies to ill-structured problems
Intuition	An immediate and effortless knowing from the collective unconscious (a dark or wild animal side and a social mask to get along with other people) of prototypes of the deep human experience Many people do not trust intuition despite their presence but creative people trust and prefer using it. An intuitive person is more likely to operate the tasks by using trial and error instead of using an instruction

Transformative learning for health professional

Transformative learning for health professional is the sustainable expansion and reform of health professionals’ learning to achieve required new skills. It aims to reduce health inequity and improve quality of care as well as addressing issues of health professionals’ competencies for the country health systems. Achieving this will ensure improvement of population health and health equity.

Transformative learning theory was outlined by Jack Mezirow in 1978 (Jack Mezirow 2014). It comprises three dimensions:

psychological (changes in understanding of the self)
convictional (revision of belief systems)
behavioural (changes in lifestyle)

The learning of health professionals is shifting from ‘informative’ to ‘formative’ and on to ‘transformative’.

The ‘informative’ learning is the fundamental level of fact memorization and skill building to become *experts*.

The ‘formative’ learning focuses at the higher level of critical reasoning and systematic thinking to produce *professionals*.

The ‘transformative’ learning aims at the highest level of analysis, synthesis, and enlightenment to be *leader for changes*.

Transformative learning requires reform of the health professional educational system. Reform includes both institutional and instructional reforms. The institutional reform focuses at reforming the structural and functional education system. The instructional reform specifies curricular design, learning process development, and outcome.

Institutional reform (structure and function)

See Box 11.10.1.

Box 11.10.1 Transformative health professional learning

Institutional reform (structure and function)

- Professional reform
- Leadership development and management reform
- Collaboration among stakeholders
- Legal and regulatory reform
- Expenditure and financing reform

Instructional reform (process and outcome)

- Recruitment reform
- Curricula reform
- Learning process reform, including Interprofessional education (IPE), application of IT and authentic learning

Professional reform that focus on faculty development

Faculty development includes academic and career development, teaching improvement, research capacity, and leadership development to ensure teaching, research, and service congruence (Steinert 2012). Effective faculty development is a crucial component in promoting individual and organizational change for its 'grounded in growth' perspective.

Parker Palmer addressed that 'growth of any craft depends on shared practice and honest dialogue among the people who do it' (Palmer 1998). One of the key components of faculty development is to share 'community of practice' dialogue. This would allow collective reflection on human issues of the profession rather than only technical ones. Among others, faculty development help to create values or merits as 'role models' or 'inspirational teachers', rewards system, and attempts for excellence.

Leadership development and management reform

Leadership is critically significant to guide, direct, and influence others to bring about fundamental changes to the health systems. The institutional mechanism is empirically applied to: (1) emphasize fostering leadership development through transformative learning for health professional education and research at all levels; (2) connecting health professional development to evolving local health needs, national priorities and the SDGs; (3) building health equity and eliminating health disparity (WHO 2016) through collaborative institutional and instructional reforms; (4) generating and interpreting data with appropriate stakeholder input to inform evidence-based policy and evaluate innovative practice for education and training; and (5) developing a national HWF plan that harnesses market forces towards development of a workforce with adequate skill-mix for delivery of comprehensive PHC.

Management reform of the education system is collectively pointing to decentralizing the educational administration to shift from central to local control in terms of authority and decision-making power based on good governance and social accountability.

Collaboration among stakeholders

Synergy and coherence among multisectoral key stakeholders from professional educational institutions and councils and the systems that are collaboratively responsible for healthcare services is a necessary component to bring about transformative change and leadership through collective action in developing future health professionals who have a commitment to work in underserved areas where health inequities and inequalities remain. Such collaboration includes public-private partnerships for an inclusive and sustainable healthcare system.

Legal and regulatory reform

In terms of legal and regulatory reform, quality assurance, accreditation and re-accreditation of training institute\ions and curriculum are key instruments to ensure quality and competency of graduates while national licensing examination, re-licensing process, and CPD are instruments are necessary to ensure continuity of health professionals' competency. Accreditation of health professional learning should be introduced and strengthened. It can be achieved through implementation of accreditation procedures. The procedures should emphasize continuous improvement of quality educational processes and outcome, with inclusion of social accountability. Mutual learning with gradual movement towards mutual recognition among countries should be supported, for example the Mutual Recognition Arrangements on the medical, nursing, dental, and pharmacy professions among the South-East Asia Nations.

Finally, with the increasing number of private health professional education institutes, the accreditation and licensing systems should be applied equally in both public and private institutions.

Reform of financing systems

Health professional education requires big investment. Producing a medical doctor in Thailand, for example, costs around US$60,000. Most public institutes are highly subsidized. Students in the private schools may take publicly subsidized loans. The challenge is equity in access and efficiency in using the public resources. Students from low socio-economic families,

especially in the rural settings, find it difficult to compete with those from urban better off families.

‘Rural recruitment, local training and hometown placement’ is an important measure to ensure better equity both in the access to education and the distribution of the graduates.

In some instances, local communities recruit and support their own students with the condition that students must return to work in their communities.

An education financing reform should aim to restructure and reallocate resources to cover local communities through public-private partnerships.

Demand-side financing schemes, where funds are channelled directly to the local communities or students, will empower them and create greater community responsiveness among the supply side.

Instructional reform

See Box 11.10.1.

Recruitment reform

Student enrolment and admission policies should focus more on increasing the socio-economic, demographic, psycho-physiologic, ethnic, and geographical diversity of students. It would reflect local population diversity and ensure their social return. As such, special systems should be in place to increase inclusion of students from disadvantaged, underserved populations, and marginalized groups. These systems should also align with the requisite academic standards and other desirable characteristics. In addition, special training programmes should be implemented to support underserved population groups to achieve admissions requirements and to reduce dropouts before completion of their training.

Curricula reform

The existing conventional, and fragmented curricula should be transformed. They should focus on population health needs, new models of service delivery and health team skill-mix to support UHC for health equity. The

‘silo’ competencies across different relevant curricula have to be reviewed to ensure better alignment across different professionals.

Essential new issues should include:

Broader and deeper competency-based learning with emphasis on social and contextual determinants of health in situ
Broader understanding of overall integrated people-centred health systems and roles of different health professionals
Continuous leadership development to be active ‘change agents’
Home-based comprehensive primary care, health promotion and disease prevention
Community engagement, inter- or transprofessional health teams, and immersion education experiences
Inclusion of community leaders, health service managers, and users from heath facilities on curriculum planning committees

Interprofessional education (IPE)

WHO defines IPE as ‘Learning “With”, “From” and “About” different professions, for collaborative practice to improve healthcare quality’ (WHO 2010). It is one of the 11 WHO’s key recommendations. It supports and enhances the performance of community health systems. It deepens mutual understanding among different professions to bring about an effective collaboration and improve health outcomes in the community.

In order to scale-up IPE, all health professions should be incrementally focusing on IPP in the real-life services (de Villiers and Naidoo 2011; IOM 2013).

To optimize health professional learning, the transformation should also focus on team-based education in addition to the competency-based learning.

The key challenges are:

Personal barriers, including failure to appreciate other professionals
Misunderstanding and miscommunication
Fear of change
Differences in income
Different power in the systems
Language or gender
Institutional barriers such as practice setting, structure, and priority

Information and communication technology (ICT) for transformative learning

Digitalized Information Technology (IT) and social media skills are key elements of twenty-first century teaching and learning. It can provide a common platform for knowledge exchange and sharing. However, this platform does require substantial infrastructure support to connect all the basic facilities, services, and information together.

ICT for Transformative learning has five objectives:

Fostering and deepening the learner-centred learning system and learning environment
Diminishing costs by mobilizing and sharing resources
Increasing the application of ICT in learning settings and sharing knowledge on informatics
Strengthening teachers' and learners' confidence, competency, and capacity in ICT utilization
Developing an online course from the existing course

The four strategies for applying ICT in transformative learning include:

Fostering ICT across all disciplines
Increasing the use of technology platform s in courses
Promoting learner utilization of mobile devices
Implementing a comprehensive learning management system to achieve learning outcome improvement

The key successful indicators are:

Percentage of teaching and technical support staff with ICT literacy
The number of electronic resources on learning, educational research studies, and innovations that can be applied to improve the learning process and achieve better learning outcomes
The number of ICT skilful faculty members

In the technology-rich digitalized innovative environment, shortfalls in soft skills development can occur. To overcome such challenges, a few measures should be addressed:

Social interaction by means of communicating through dialogue
Encouraging
Explaining
Hugging and thanking

We should always emphasize that 'The digitalized technologies are helpful but artificial, the real health services depend on the *ten digits* of the health professionals that touch on the patients and people'.

'Appropriate Technology with High Human Touch' is the most favoured service, not the 'High Technology with Low Touch' (Gade and Chari 2013).

Authentic learning

Authentic learning or experiential based learning will engage learners to actually participate in solving real health challenges. The learners' skills and attitude gained from experience and practice will keep them motivated to enhance their lifelong learning. This is to ensure sustainable quality of care and ultimately achieve community satisfaction. The key challenge is to create learning experiences to effectively provoke deep reflective thinking.

Authentic learning should be supported by the effective and appropriate application of ICT.

Networks on transformative learning for health professional

A number of global and regional transformative learning networks to actively enrich its momentum have been established, including:

The Joint Learning Initiative (2004)
The Global Health Workforce Alliance (GHWA)
The Asia Pacific Action Alliance on HRH (AAAH)
THEnet
The President's Emergency Plan for AIDS Relief (PEPFAR) supported Medical and Nursing Education Partnership Initiative (MEPI and NEPI) or AFREhealth
The Asia-Pacific Network on Health Professional Education Reform (ANHER)
The Pacific Open Learning Health Net (POLHN)

The Joint Learning Initiative (2004)

The Joint Learning Initiative (2004) is a network of more than 100 great healthcare leaders who are determined to confront the global health system crisis by mobilizing and strengthening HWF primarily in low-income countries. Reflecting their initiative, the HWF shortage has become a common concern and the collective action to tackle the HWF crisis is needed.

The Global Health Workforce Alliance (GHWA)

When the Rockefeller supported Joint Learning Initiative ended in 2005, one of its recommendations was to establish global and regional networks on HWF.

In 2006, the GHWA was launched and hosted by WHO in response to the HWF crisis. It is a global network of various agencies including civil society, financing sectors, workers, national governments, international, academic, and professional organizations. In 2016, GHWA transitioned to the Global Health Workforce Network (GHWN), operated by WHO to support the national actions on the Global Strategies on HRH 2016–2030 and the recommendations of the UN High-Level Commission on Health Employment and Economic Growth.

The Asia-Pacific Action Alliance on Human Resources for Health (AAAH)

At more or less the same time as the GHWA, the HWF leaders in Asia and Pacific established the AAAH. It is a regional partnership mechanism to strengthen country health workforce planning and development actions. AAAH's prominent roles are to advocate HRH policy; monitor information system; strengthen capacity; generate, manage, and share knowledge; and coordinate technical issues. AAAH has convened annual conference, since 2006 and biennial since 2014. Network activities of 18 countries, in WHO SEAR and Western Pacific Regions (WPR), include training workshops and study visits. Intersession research workshops to drive national agenda by country partners through evidence generation and capacity building activities.

Training for Health Equity Network (THEnet)

In 2008, Training for Health Equity Network (THEnet) aims to increase the impact of academic institutions on community health and develop more equitable health systems. The identification of successful strategies to indicate health workforce shortages is primary goal, THEnet is located in outreach areas worldwide and steps forward to convince schools and policymakers that socially accountable education can increase quantity and quality of HWF practising in rural and underserved areas.

The Medical Education Partnership Initiative (MEPI) and the Nursing Education Partnership Initiative (NEPI)

In 2010, the United States (US) PEPFAR, supported the MEPI administered by the National Institutes of Health (NIH) and the NEPI administered by the Health Resources and Services Administration (HRSA). These two networks have together strengthened innovative evidence-based strategies for transformative health professional learning, training, and research in support of HIV epidemic control and achievement of UHC in Africa. Both MEPI and NEPI were a response to the critical shortage of HWF in sub-Saharan

Africa by increasing the quantity, quality, and capacity of physicians, nurses, and midwives. From these two multicounty initiatives, a new African designed and led African Forum for Health Research and Education in Health (AFREhealth) has emerged which embraces interprofessional collaboration since 2016.

Asian Network on Health Professional Education Reform (ANHER)

To envision the 2010 Global Commission's report on Transformative Learning, Bangladesh, China, India, Thailand, and Vietnam, committed to implement and conduct distinct and suitable health professionals' education reform. The strategies will be flexible with respect to the nation's socio-economic and cultural situation and trends, as well as its domestic health service system. In 2011 at a meeting in Hanoi, they initiated ANHER. The network has agreed on joint actions including joint surveys on health professional learning, information sharing, and learning in medical nurse and public health professional reform, plus other internationally collaborative activities. The situation analysis survey focused on the introduction of innovative institutional and instructional reforms. The reforms aimed to improve quality, enhance teamwork among students, schools, and networks, improved curricula, and practicum to prepare health workers for the new challenges. The innovations are introduced into the learning systems such as e-learning methods and introduction of new skills in the curricula such as public health and management training. Articles on cross-country analysis and country experiences were published for experience learning and sharing and drive the reform (Chuenkongkaew et al. 2016; Pudpong et al. 2017; Zodpey et al. 2018).

The Pacific Open Learning Health Net (POLHN)

The POLHN is a key programme of the WHO in collaboration with the Pacific Ministries of Health to increase health professional standards and quality by offering free, online, and blended courses. The network cooperates through 47 learning centres and a team of dedicated focal points in 15 countries. POLHN offers different specialized online accredited

programmes from Fiji National University, Pacific Paramedical Training Centre, and Penn Foster Dental School. Since Internet access in the Pacific is getting better and an increasing number of online courses are available globally, online continuing education will be increasingly important in the Pacific. In this regard, POLHN needs to reinforce its role in overall CPD, including better regional coordination with more countries.

Conclusion

The successes of health-related SDGs depend on the successes of the health system reform towards primary care-based, patient-centred health service. These reforms, in turn, depend on the successes of the Health Professional Education reform.

The reforms will ensure adequate numbers and equitable distribution of competent, inspired, and motivated HWF with leadership capacity.

Transformative health professional learning is a set of essential reforms, from hospital-based to health systems and community-based and authentic learning. The reforms focus on building skills and attitude on rationalizing appropriate technology with a high human touch. It moves from individual silo-based practice to integrated team-based practice. It also ensures the effective and constructive collaboration between the producers and users of HWF in the health systems.

REFERENCES

Chuenkongkaew, W.L., Negandhi, H., Lumbiganon, P., Wang, W., Mahmud, K., and Cuong, P.V. (2016). Attitude towards working in rural area and self-assessment of competencies in last year medical students: a survey of five countries in Asia. *BMC Medical Education*, **16**, 238.

de Villiers, M.R. and Naidoo, C. (2011). The consultation—a different approach to the patient: the patient-centered clinical method. In: Marsh, B. (ed.) *Handbook of Family Medicine* (3rd ed.). Cape Town: Oxford University Press.

Frenk, J., et al. (2010). Health professionals for a new century: transforming education to strengthen health systems in an interdependent world. *The Lancet*, 376(9756), 1923–58.

Gade, S. and Chari, S. (2013). Case-based learning in endocrine physiology: an approach toward self-directed learning and the development of soft skills in medical students. *Advances in Physiology Education*, 37(4), 356–60.

Global Strategy on Human Resources for Health: Workforce (GSHRH) (2030). [Online] Available at: http://www.who.int/hrh/resources/glob-strat-hrh_workforce2030.pdf

Grant, R. (2012). A bridge between public health and primary care. *American Journal of Public Health*, 102(Suppl 3), S304.

Gruppen, L.D., Mangrulkar, R.S., and Kolars, J.C. (2012). Competency-based education in the health professions: implications for improving global health. *Human Resources for Health*, **10**, 43.

High-Level Commission on Health Employment and Economic Growth (HEEG) (2016). *Working for Health and Growth*. [Online] Available at: http://apps.who.int/iris/bitstream/handle/10665/250047/9789241511308-eng.pdf?sequence=1

Institute of Medicine (IOM) (2013). *Interprofessional Education for Collaboration: Learning How to Improve Health from Interprofessional Models Across the Continuum of Education to Practice: Workshop Summary*. Washington, DC: The National Academies Press.

International Labour Office (2015). *Global Evidence on Inequities in Rural Health Protection: New Data on Rural Deficits in Health Coverage for 174 Countries.* Scheil-Adlung, X. (ed.). Geneva: ILO.

Jack Mezirow (2014). Jack Mezirow, who transformed the field of adult learning, dies at 91. Teachers College—Columbia University. [Online] Available at: https://www.tc.columbia.edu/articles/2014/october/jack-mezirow-who-transformed-the-field-of-adult-learning-d/

Palmer, P.J. (1998). *The Courage to Teach: Exploring the Inner Landscape of a Teacher's Life*. San Francisco, CA: Blackstone, p. 144

Piirto, J. (2010). The five core attitudes and seven i's for enhancing creativity in the classroom. In: Kaufman, J. and Beghetto, R. (eds.). *Creativity in the Classroom*. New York: Cambridge University Press, pp. 142–71.

Prince Mahidol Award Conference (2014). Transformative learning for health equity. Conference Proceedings, 27–31 January 2014. [Online] Available at http://www.pmaconference.mahidol.ac.th/

Pudpong, N., Suphanchaimat, R., Batra, B., Hou, J., Vu, L.T.H., and Dipika, P. (2017). A final-year nursing student survey: rural, attitudes, perceived competencies and intention, to work across five Asian countries. *BMC Nursing*, **16**, 13.

Steinert, Y. (2012). Perspectives on faculty development: aiming for 6/6 by 2020. *Perspectives on Medical Education*, 1(1), 31–42.

Sustainable Development Goals (SDG) (2015). *Sustainable Development Goals to Transform Our World*. [Online] Available at: https://www.un.org/sustainabledevelopment/sustainable-development-goals/

Tracking Universal Health Coverage (2017). *Global Monitoring Report*. [Online] Available at: http://www.worldbank.org/en/topic/universalhealthcoverage/publication/tracking-universal-health-coverage-2017-global-monitoring-report.

United Nations (2014). *World Urbanization Prospects: The 2014 Revision.* [Online] Available at: https://esa.un.org/unpd/wup/publications/files/wup2014-highlights.pdf

Wasi P. (2000). Research in health manpower development. *Human Resources Development Journal*, **4**, 106–10.

World Health Assembly (WHA) (2013). WHA66.23 Transforming health workforce education in support of universal health coverage. [Online] Available at: http://apps.who.int/gb/ebwha/pdf_files/WHA66/A66_R23-en.pdf?ua=1

World Health Organization (WHO) (2006). *The World Health Report: Working Together for Health*. Geneva: WHO.

World Health Organization (WHO) (2010). *Framework for Action on Interprofessional Education & Collaborative Practice [following the CAIPE definition (2002)]*. Geneva: WHO.

World Health Organization (WHO) (2012). SEA/RC/R6 Strengthening health workforce education and training in the region. [Online] Available at: http://apps.who.int/iris/bitstream/handle/10665/128331/SEA-RC65-R7.pdf?sequence=1

World Health Organization (WHO) (2013). *Transforming and Scaling Up Health Professionals' Education.* [Online] Available at:

http://apps.who.int/iris/bitstream/10665/93635/1/9789241506502_eng.pdf
World Health Organization (WHO) (2016). Framework on integrated people-centred health services. Available at: http://www.who.int/servicedeliverysafety/areas/people-centred-care/en/
Zodpey, S., et al. (2018). Assessment of health professional education across five Asian countries —a protocol. *Human Resources for Health*, **16**(1), DOI: 10.1186/s12960-018-0316-6

11.11

Humanitarian emergencies

Craig Spencer and Les Roberts

A challenging moment in history: more people in need, more money, and more difficult tasks

In recent years the humanitarian sector has been the target of intense criticism. The humanitarian community has been seen as driven by wealthy country agendas, inadequately developing and listening to local actors, and not being adequately accountable to international law and humanitarian principals (Agenda for Humanity n.d.). While all of these criticisms have a solid basis of complaint, they do not stem from a humanitarianism in decline. The lack of clarity of purpose of many humanitarian players has arisen in part as a result of the spectacular success of the humanitarian community.

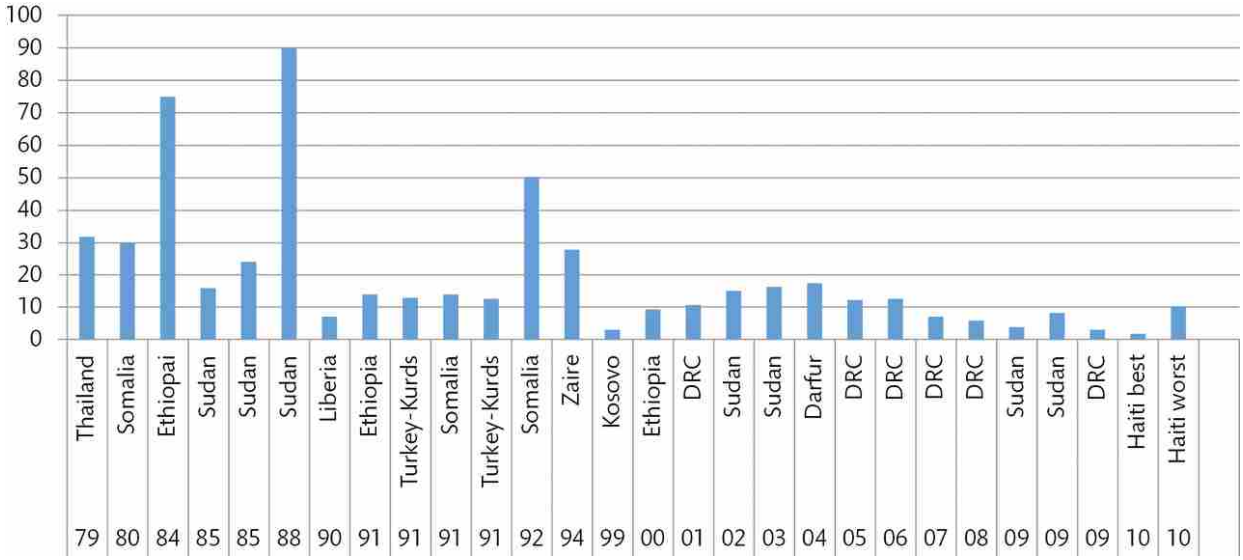


Fig. 11.11.1 Highest mortality periodically recorded over the last 30 years in >100,000 people lasting >1mo. Deaths/1,000/mo.
Source: data from Centre for Research on the Epidemiology of Disasters (CRED) database.

This may partially be explained by a 75% decline in the fraction of people living in poverty since 1981 when current World Bank measures began (Roser and Ortiz-Ospina 2018). Yet perhaps most of the decline is because of the near ubiquitous presence of humanitarian actors during the most acute of crises and their improvement in the science of keeping people alive.

Our present moment in history also involves higher numbers of displaced people than have ever been recorded. The United Nations High Commissioner for Refugees (UNHCR) estimates that more than 60 million people are displaced, with two-thirds of them still in their home country and therefore the responsibility of that government (UNHCR n.d.). One-quarter of today's refugees have fled Syria, a middle income and highly educated country not similar to most of the origins of last century's most acute crises shown in Figure 11.11.1. Those who do manage to cross borders now are not returning for longer periods than those in recent decades (Fig. 11.11.2).

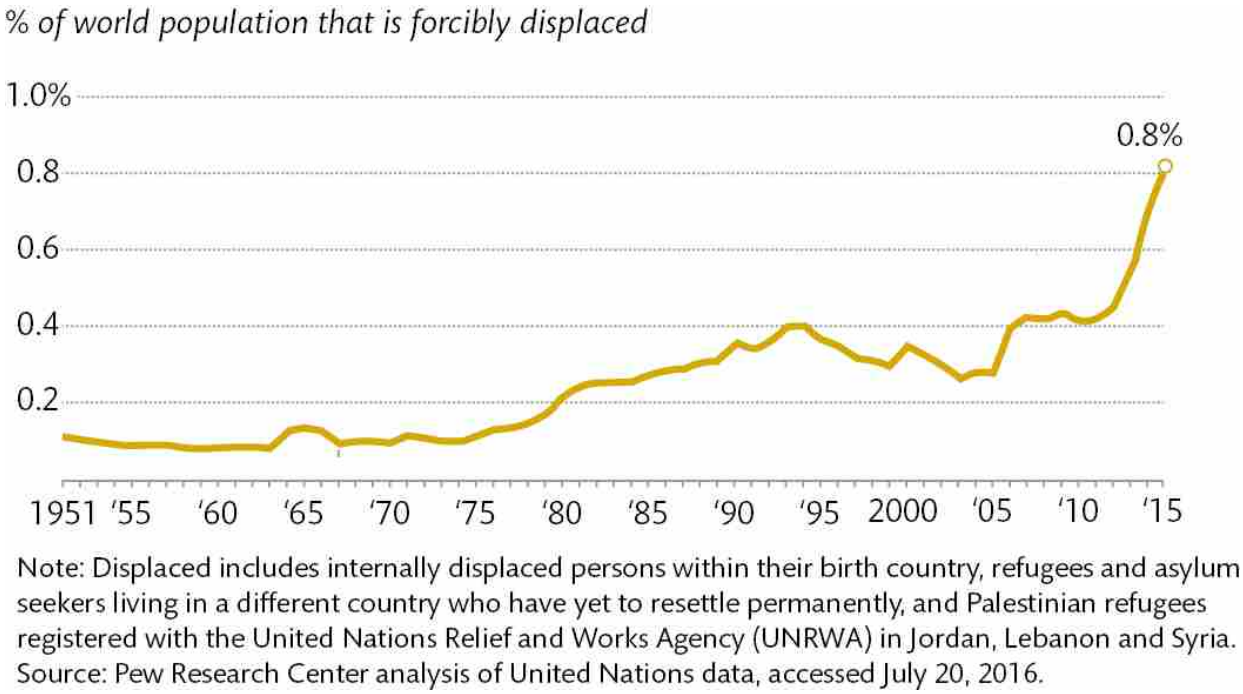


Fig. 11.11.2 Estimate of the percentage of the global population displaced over time. Reprinted with permission from the Pew Research Center 2016. http://www.pewresearch.org/fact-tank/2016/08/03/nearly-1-in-100-worldwide-are-now-displaced-from-their-homes/ft_16-07-28_displaced_world/

Making this time period more complex is the fact that more money is being spent on humanitarian assistance. The 2017 Global Humanitarian Assistance Report suggests that $27.3 US billion was spent in 2016 on humanitarian assistance, more than on development. This is in stark contrast to the 1970s–1990s when OECD countries typically spent more than ten times as much on development aid as on humanitarian assistance (Global Humanitarian Assistance 2000).

Figure 11.11.3 shows the dramatic increase in violent deaths in the WHO Middle East and Northern Africa (MENA) region since 1990 (GBD 2018). In this region, violent deaths have increased in every demographic group, increasing 450% from 1990 to 2015.

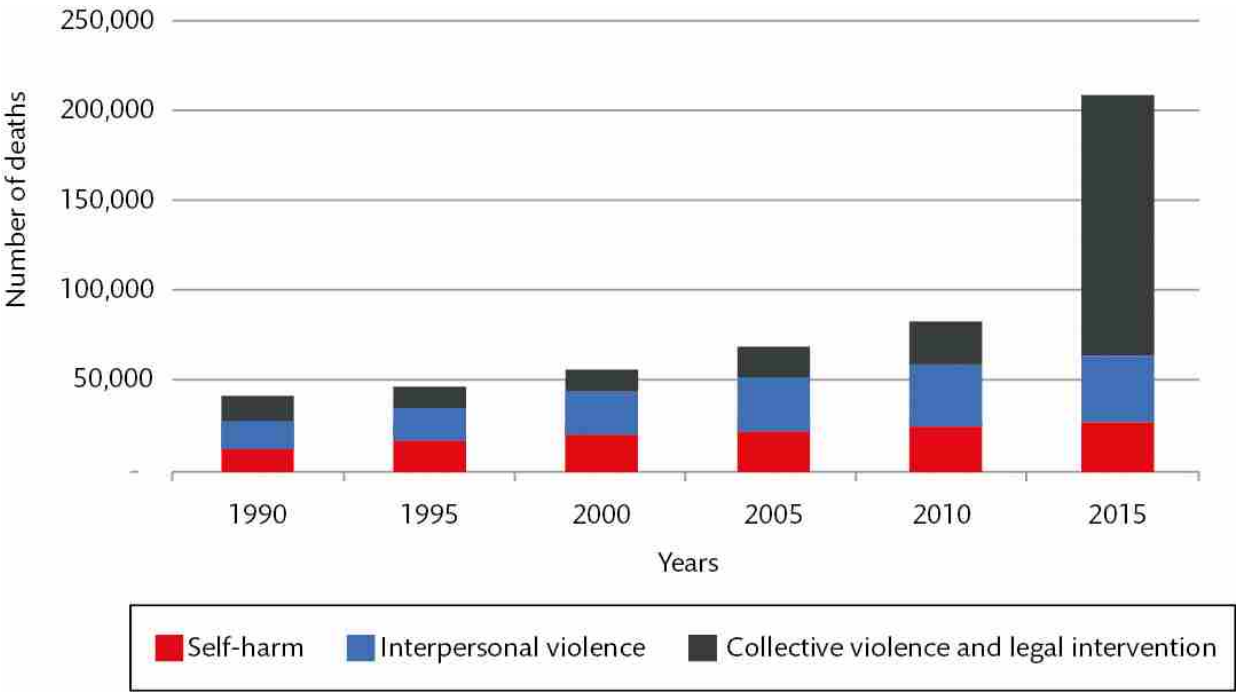


Fig. 11.11.3 Increased numbers of violent deaths MENA region, 1990–2015.
Reprinted from GBD 2015 Eastern Mediterranean Region Intentional Injuries Collaborators (2018) 'Intentional injuries in the Eastern Mediterranean Region, 1990–2015: findings from the Global Burden of Disease 2015 study'. *Int J Public Health*. 2018; 63(Suppl 1): 39–46, under a CC BY 4.0 license.

During this time, violence has become the number one cause of death for children 10 to 19 years old. Thus, the core causes of morbidity and mortality in this region are not easily fixed by non-state actors like humanitarian aid workers. Adding insult to injury, violence against aid workers has become more common and expectations of neutrality and safety for health workers and aid workers has diminished (Jackson 2012). Over 1,000 hospitals and clinics have been attacked during the war in Syria, a level of attacks out of step with historical norms (Briody et al.). On one day in July of 2016, the last four remaining hospitals in the besieged city of Aleppo were bombed, presumably by Russian aircraft (Al-Jazeera 2016a). No United Nations (UN) condemnation or International Criminal Court (ICC) action was taken, and no apology form the Russians was offered. This event, like the invasion of Iraq as a pre-emptive war or the rise of anti-immigrant sentiment across the globe, implies that the humanitarian community is living in an era of incivility that adds significant barriers to their work.

This period involving lower mortality crises means humanitarian assistance in the twenty-first century is less focused on the relatively easy to

treat killers of measles, ARI, diarrhoea, and malaria deaths. Today, aid is more often being provided to more people from middle income countries who want long-term support and have health priorities that involve addressing issues like violence, depression, and unemployment. The Oxford Dictionary defines ‘humanitarian’ as ‘Concerned with or seeking to promote human welfare’. Thus, addressing whatever the most acute needs of a population are is innately the mandate of the humanitarian endeavour. But, the UN, church groups, and non-governmental organizations (NGOs) who tend to lead major humanitarian efforts are innately better suited to preventing measles and malaria deaths than they are to addressing issues that require ongoing engagement and social change.

To summarize, the twenty-first century will be a challenging era for the humanitarian community. There are more people in need, not only for food, water, and medicine to stay alive, but who are also experiencing more complicated and difficult needs in a period of rising expectations. While this sounds difficult, this is in many ways the apex of civilization. The challenges are vast because the expectations of service and compassion are now projected to more than just those at risk of death. The challenges are difficult because we are now expected to operate in environments where aid workers are targeted and where host governments responsible for those at risk are often the source of that risk. The challenges are complex because depression and domestic violence and many of the relief priorities today are issues not easily solved even in the wealthiest of countries. This chapter will review where the humanitarian community has been and will lay out some of the new efforts for dealing with these challenges.

Introduction to emergency public health and humanitarian assistance: where we have been

During the Biafran War (1967–1970) perhaps half of the 2,000,000 population inside the shrinking breakaway enclave died of malnutrition and disease by the end of the conflict (Western 1972). During this crisis, profound advances in modern humanitarian relief arose. Unable to determine which children were starving and which children were just skinny or small, the Friends Overseas Service Committee triggered the development of the Quaker Upper Arm Circumference (QUAC) Stick, the first scheme for

assessing nutritional anthropometry. The Centers for Disease Control and Prevention (CDC) tested the universally accepted theory that malnourished children could not produce an immunological response to being vaccinated by using intervention and control groups. They vaccinated some very malnourished children and followed up weeks later by drawing blood from the study children to see if immunity had developed. In this first known controlled trial in an emergency, it turned out that malnourished children could indeed be effectively vaccinated (Ifekwunigwe et al. 1980). Other developments included the first known public health adaptation of capture-recapture methodologies from wildlife biology in order to estimate the number of deaths (and thus the first real-time conflict mortality measurement), the first guidelines for such things as running feeding centres, and the creation of the medical relief organization Médecins Sans Frontières (MSF).

In the crises that followed, more and more central insights into the science of keeping vulnerable people alive and healthy accumulated at an ever-increasing rate. The East Pakistan Crisis, by chance arising where an international diarrhoeal research centre was based, brought the advent of oral rehydration therapy (ORT) (Sack et al. 1970). Relief for Cambodians on the Thai border demonstrated the value of surveillance and rapid assessments (Glass et al. 1980). The famine and conflict in the Horn of Africa in the early 1980s provided the rationale for the caloric content of rations, a two-stage measles vaccination strategy for children less than 9 months, and a variety of other ‘best practices’ (Toole et al. 1988, 1989). By the 1990s, comprehensive outlines of priorities and best practices were in place based on the accumulation of these various lessons learned and documented (Toole 1990; CDC 1992; MSF 1997).

Most of these guidelines were developed based on experiences in settings of extreme poverty and high mortality, where the excess (above the normal) mortality profiles were dominated by infectious diseases, primarily diarrhoea, respiratory infections, and malaria. While acute respiratory infections are often the leading cause of death in emergencies, it is relatively difficult to prevent infections and emergency response efforts usually focus on getting effective medical treatment to cases. In contrast, the next three major killers in emergencies in decades past, namely malaria, measles, and diarrhoea, can all be prevented by interventions commonly provided as part of humanitarian relief efforts. Figure 11.11.4 next shows the global rate of

death from these three diseases from 1985 to 2015. Data are compiled from multiple sources (Rodolfo et al. 1993; Kosek et al. 2003; Boschi-Pinto et al. 2008; WHO n.d. 2011).

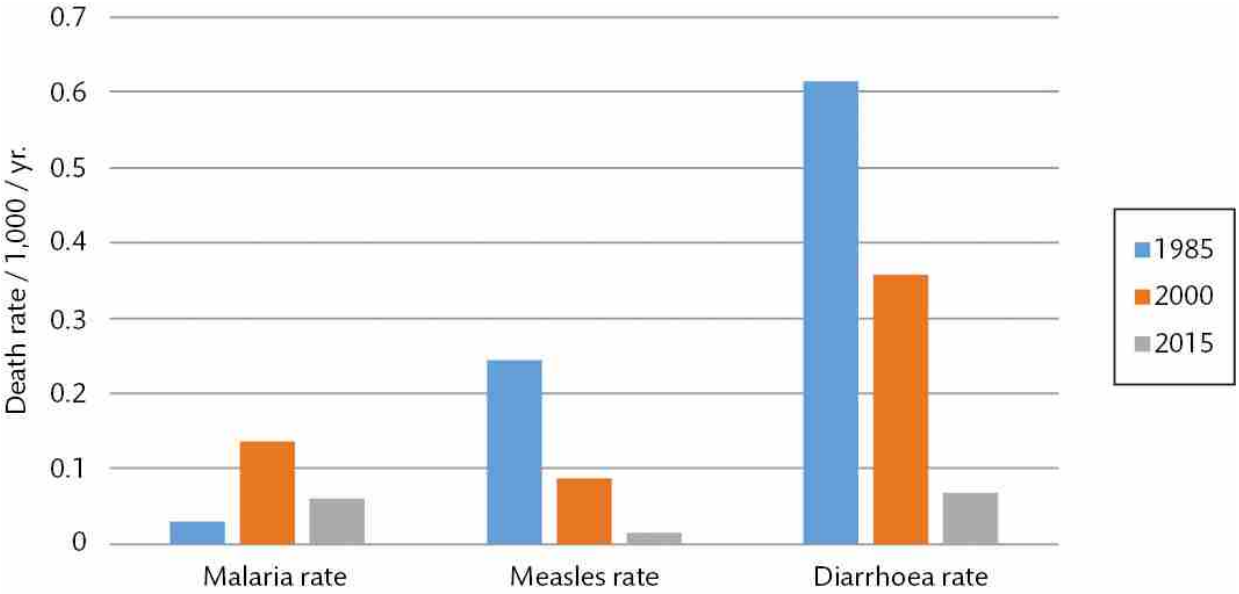


Fig. 11.11.4 Global death rates for malaria, measles, and diarrhoea.
Sources: data from (WHO n.d., 2011, Boschi-Pinto et al. 2008, Kosek et al., 2003, Rodolfo et al., 1993).

In assessing why these dramatic reductions occurred, it is hard to separate what influence the reduced fraction of the global population in poverty has had compared to focused programmes like measles vaccinations and oral rehydration solution (ORS) promotion or newly available technologies like long-lasting insecticide impregnated bednets. It is most likely that all these factors are at work and simultaneously interacting with other improvements such as higher rates of literacy and improving function of health related institutions. In emergencies, people tend to die of the same causes that they die of in stable times, just at a higher rate. Therefore, when compared to the Cold War era, the lower mortality widely seen today may be in part a reflection that our global population is less vulnerable to death in general. The Cold War era crises often continued for many years with large numbers of people displaced and supported in camps for a variety of geopolitical and economic reasons. Crowded camps are ideal environments for spreading

disease. Today, few of the world's displaced stay in camps further diminishing the emphasis on infectious disease control.

To summarize this overview of humanitarian relief evolution over the past half century, the world of humanitarian response today has little in common with the Biafran era. Today, most of the world's displaced have not crossed a border and are displaced within their own country. Many and perhaps most of the world's most vulnerable displaced persons are in urban settings. Finally, likely as a result of the developments of the relief community and greater global wealth, few crises in the twenty-first century to date have been associated with extended periods of high mortality, as were seen in Cambodia, Mozambique, Sudan, Somalia, Biafra, the Democratic Republic of Congo (DR Congo), and elsewhere in the last century. Rising international expectations of support have resulted in expanding humanitarian efforts focused more on issues of suffering, often from chronic or mental health-related disorders.

To address these new realities, this chapter will look at those areas where the twenty-first-century humanitarian endeavour is likely to differ dramatically from the twentieth century, including: preparedness and planning, changing institutional structures and architecture, the growing international attention of disease outbreaks, the evolving public health profile of emergency relief needs, and the role of new and innovative technologies. The roles of emergency planning and relief workers have grown, professionalized, and become politicized at a dramatic rate in recent years and we attempt here to create a framework for understanding these transitions.

Definitions and classification

There are few terms used more divergently in society today than the word 'humanitarian'. Merriam Webster defines humanitarian as: 'a person promoting human welfare and social reform, philanthropist—recognized as a great humanitarian for her work to end world hunger.' The Oxford Online Dictionary says: 'humanitarian (adjective) (1) Concerned with or seeking to promote human welfare. (2) Denoting an event or situation which causes or involves widespread human suffering, especially one which requires the large-scale provision of aid'. In this chapter, and with most humanitarian employers, humanitarian will signify seeking to promote human welfare

while adhering to the Red Cross Code of Conduct, a set of 10 rules designed to promote indiscriminate compassion, impartiality, neutrality, and accountability (ICRC n.d.).

There is currently no universally accepted definition of the term ‘disaster’. This lack of standardization can be attributed in part to the broad range of disciplines involved in disaster management. Nonetheless, common to most definitions is the concept of large-scale loss that exceeds a community’s own capacity to respond effectively, requiring some form of external assistance.

The United Nations International Strategy for Disaster Reduction (UNISDR) defines a disaster as ‘a serious disruption of the functioning of a community or a society involving widespread human, material, economic or environmental losses and impacts, which exceeds the ability of the affected community or society to cope using its own resources’ (UNISDR 2010). This definition is also one of several recognized by the World Health Organization (WHO).

Similarly, there is no standard classification system for disasters, although numerous have been proposed. Generally, disasters are associated with the hazards that cause them. They can be classified as either natural or human generated, with subclassifications of each. Natural disasters are often subclassified according to whether they are acute onset or slow/gradual onset, human-generated disasters as intentional or non-intentional. Generally there are not widely accepted definitions of these subclassifications. Moreover, military, political, and public health professionals have evolved to use words like emergency, disasters, and humanitarian with little functional criteria in common. Further undermining the utility of disaster classifications is the fact that often scholars from different disciplines understand the same crises in different ways. For example, while the 2011 Syrian war is largely painted as politically triggered over a short period by an Arab Spring uprising and a brutal government crackdown, Selby describes the conflict as a slow onset consequence of years of abnormally low rainfall creating migration and political instability (Al-Jazeera 2016b; Selby et al. 2017). Thus, we shall emphasize here epidemiologically based definitions and interpretations, as that will likely be the most useful approach for health professionals.

Epidemiology

The epidemiology of manmade crises (war and most famines) differ dramatically from natural and technological disasters, which aside from drought, tend to have a rapid onset. We shall refer to the latter as disasters and data about their occurrence and mortality effects are far better than with manmade crises. Earthquakes and tsunamis are the deadliest of natural disasters followed by extreme temperature events and floods.

According to the Centre for Research on the Epidemiology of Disasters, during the period from 1996 and 2015, on average approximately 600 natural and technological disasters occurred globally each year resulting in 67,000 annual fatalities (CRED 2017). Over these two decades, the number of annual events has gone up slightly while the annual death toll has not. The most severely affected region was Asia, followed by Africa and the Americas.

Those most vulnerable to the public health impacts of disasters are the poor. While 27% of disasters occurred in countries categorized as having low human development, they accounted for 47% of deaths, because of their limited capacities to mitigate, prepare for, and respond to disasters. Moreover, the economic costs run into billions of dollars and a single disaster can set back a nation's development many years with obvious implications for public health programming and health outcomes. A CDC report cites World Bank estimates that the 2014 Ebola outbreak in Sierra Leone cost the country $2.2 billion in lost gross domestic product (GDP) from 2014 to 2016, roughly a 20% drop (CDC 2016).

Globally, the most frequent types of disasters are transport accidents, followed by floods, windstorms, industrial accidents, earthquakes, and droughts/food insecurity (Table 1). Climate change is expected to affect the frequency and severity of floods, windstorms, and droughts, with their associated public health impacts (World Health Organization and World Meteorological Organization 2012). For example, severe droughts and associated food insecurity have occurred at more frequent intervals over the past decade in the Sahel and Horn of Africa. Monsoon-associated floods in Pakistan in 2010 and 2012 were among the most severe ever recorded.

Table 11.11.1 Frequency, impact, and geographic distribution of major types of disasters, 2006–2015

Disaster	Frequency per year	Average deaths per year	Average number of people affected	Average cost of damage (millions of dollars)	Region with highest frequency	Region with highest deaths
Droughts/food insecurity	24	2025	58,790,500	6,783	Africa	Africa
Earthquakes/tsunamis	25	35,709	8,500,100	47,103	Asia	Americas
Extreme temperatures	25	7,908	8,980,800	3,251	Europe	Europe
Floods	172	5,703	83,095,100	34,270	Asia	Asia
Forest fires	8	75	242,500	2,232	Americas	Europe
Insect Infestation	1	n.a.	510,000	n.a.	Africa	n.a.
Landslides	18	948	303,100	194	Asia	Asia
Volcanic eruptions	6	46	201,800	108	Americas	Asia
Industrial accidents	38	1,169	53,300	2,550	Asia	Asia
Transport accidents	162	4,996	4,300	40	Africa	Africa

Source: data from International Federation of Red Cross (IFRC) and Red Crescent Societies, World Disasters Report, 2016, IFRC, Geneva Switzerland, Copyright ©2016 The International Federation of Red Cross and Red Crescent Societies. All rights reserved.

While such events are usually recorded well and seen as non-controversial, natural disasters have long-term effects such as inducing migration or prolonged interruptions of public services and the actual health effects may be more complex in nature and go unrecorded. An extreme example of this arose when Hurricane Maria struck Puerto Rico in September 2017. Official government estimates placed the death toll at 64 deaths, a number repeated consistently by government officials in the months that followed (Shugarmen 2017). An examination of death certificates by the New York Times three months later placed the death toll at 2,000 and a population-based survey months later conservatively placed the death toll at about 4,600 (Kishore et al. 2018). This is the exception

rather than the rule as acute disasters like transport accidents and earthquakes typically cause trauma deaths over seconds or hours.

There is a growing body of evidence that climate change is interacting with a variety of factors to make populations more vulnerable. Thus it is difficult to distinguish which deaths are associated with climate change, overpopulation, or simply associated with natural disasters. Weather extremes and climate changes interact with other important global trends that threaten to increase the frequency and impact of disasters. Increasing urbanization, poor urban planning, poor land use, unregulated industrialization, new and emerging infectious diseases, and terrorism all have the potential to expose greater numbers of people to hazards (United Nations International Strategy for Disaster Reduction 2010; WHO 2011a; Bhui et al. 2012). The Christchurch earthquake of 2011, Bangkok floods of 2011, and Superstorm Sandy in New York City in 2012 all demonstrated that disasters striking densely populated urban areas have the potential for substantial public health and economic impact. The Nigeria floods of 2012 and on the Indian Subcontinent in 2017 that forced millions from their homes have brought attention to the issues of poor urban planning and settlements in flood-prone zones. The severe acute respiratory syndrome epidemic of 2003 and the H1N1 pandemic of 2009 have demonstrated the need for stronger public health surveillance and response mechanisms. Terrorist attacks across a range of countries have revealed security threats that need to be addressed in disaster planning and mass casualty response plans (WHO 2007). In spite of these complicating factors, because deaths tolls from disasters tend to be collected consistently, evidence suggests that since the 1920s there has been over an 80% reduction in the annual deaths from natural disasters (Richie et al. n.d.). Over this past century, deaths from droughts have dramatically declined while earthquakes have evolved to become the dominant cause of natural disaster deaths. Thus, the relatively consistent documentation of disaster death tolls has recorded the beneficial effects of the world's efforts to attenuate drought and famine effects.

In contrast, the impacts of armed conflicts are far more difficult to quantify. The World Bank estimated that 1.5 billion people live in countries affected by repeated cycles of political and criminal violence, and fragility (World Bank 2011). Most modern-day conflicts are protracted civil wars, in which civilians are often deliberately targeted and frequently displaced from

their communities. Often called ‘complex emergencies’, they are associated with high rates of morbidity and mortality.

In such conflicts, large numbers of deaths can result from violent injury, as in the Rwandan genocide, the Balkan conflicts of the 1990s, and the recent war in Syria. But in many instances, most deaths are due to the so-called indirect health consequences of conflict—increased rates of infectious diseases, acute malnutrition, and complications of chronic diseases that arise due to societal disruption, population displacement, and breakdown in health services. This has been well described in crises in DR Congo, Ethiopia, Somalia, and Sudan (Toole 1997; Coghlan et al. 2006).

The number of ongoing conflicts can be difficult to quantify, as definitions vary. The Uppsala Conflict Database defines a conflict as a ‘contested incompatibility that concerns government or territory or both where the use of armed force between two parties results in at least 25 battle-related deaths in a year. Of these two parties, at least one has to be the government of a state’. Based on this definition there were 52 ongoing conflicts in 2016, of which 2/3rds were intrastate or civil wars (Uppsala Conflict Data Program n.d.). Since the Second World War, there has been a strong tendency for conflicts to become longer in duration and less intense in terms of rates of violence (Garfield et al. 2012). Passive efforts to monitor conflicts are almost always incomplete but tend to be more incomplete in the least socially networked countries. For example, a survey from MSF found that at least 9,400 Rohingya died over a one-month period of ethnic cleansing from August to September 2017 in Myanmar (MSF 2018). Virtually none of those deaths were recorded in the news of by social media. In contrast, five civil society groups in Syria are believed to have recorded the vast majority of violent deaths in that conflict (Price et al. 2014).

Perhaps the most robust and standardized of systems for monitoring deaths from conflict is the previously mentioned Uppsala Conflict Database which attempts to evaluate press, relief agency, and expert opinion to make annual death toll estimates (see ucdp.uu.se). Population-based estimates based on surveys or multiple systems estimations have shown the Uppsala system to be relatively complete in highly networked countries (Our World Data 2017). For example, the Human Rights Data Analysis Group (HRDAG) estimate suggests that the system is 60% complete in Syria (Price et al. 2014 vs. UNDP 2010). Likewise, looking at Algeria from 1997 to 2016 suggests that UCDP is 100% sensitive when compared to the Armed

Conflict Location and Event Data Project (ACLED) Excel database. In contrast, a 2011 comparison of death toll in South Sudan with the exact same ACLED database suggests a sensitivity of only 35%. In the violent and moderately connected setting of Iraq 2003–2010, a comparison of the conflict related deaths recorded in a household survey suggested the Uppsala data set was 12–17% complete (Burnham et al. 2006; Hagopian et al. 2013 vs. UCDP 2010). Further, surveys in CAR in 2009 suggest the UCDP was less than 2% complete there (Vinck and Pham 2011). This suggests that our global data and understanding of deaths from wars is likely of lower quality than our understanding of deaths from natural disasters.

Garfield looked at both the major natural and manmade disasters from 2002 to 2006 (Garfield 2007) and found that while natural disasters tended to affect far more people, conflicts tended to kill far more people. These results are presented in Figure 11.11.5, next.

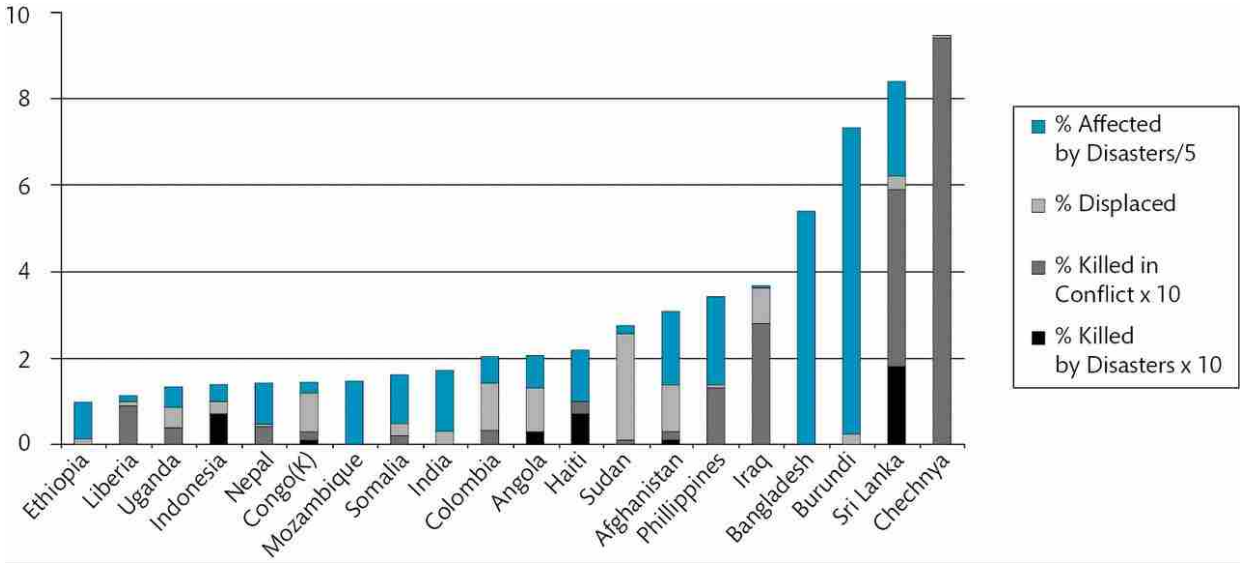


Fig. 11.11.5 Relative measures of populations affected and killed in major emergencies 2002–2006. Reprinted from Garfield R (2007) 'Measuring Humanitarian Emergencies' *Disaster Medicine and Public Health Preparedness* 1(2):110–116, with permission from Cambridge University Press.

In terms of the demands on local societies and the humanitarian community, elevations in mortality rates can persist for years after the formal end of conflict, as documented in Angola, DR Congo, and South Sudan. An important aspect of crises that can expose the affected populations to

increased public health risks is forced displacement. The United Nations Development Program (UNDP) has developed a Human Development Index, in an attempt to articulate how well a society is doing. The index is a composite of life expectancy at birth, an education index, and income. According to the Human Development Index for 2016, nine of the 10 least developed countries on the planet experienced significant levels of armed conflict in recent years. The only one of these countries that has not experienced major conflict, Guinea, is surrounded by conflict-affected countries. Similarly, of the ten countries with the highest maternal mortality, nine have recently endured conflict. According to the World Bank, no low-income fragile or conflict-affected country achieved even one of the eight Millennium Development Goals—three of which are health related.

The process of displacement can increase the risk of infectious diseases (e.g. through housing in unsanitary conditions or exposure to disease-carrying vectors), food scarcity and malnutrition, physical violence, sexual abuse, and psychological trauma. Thus, the disruption associated with crises usually disproportionately affects the health and survival of small children. Access to health services is often limited and ongoing health programmes (e.g. vaccinations) disrupted. Higher disease rates are often documented among internally displaced persons (IDPs), as the well-developed mechanisms to provide protection and humanitarian assistance to refugees do not always apply directly in IDP settings (Salama et al. 2001).

Thus, our knowledge and use of data is rather contrasting between disasters and conflicts. Wars induce far more deaths, and often involve protracted periods of crisis where outside support can be essential for averting deaths. Disasters, in contrast, affect far more people, and tend of create crises of shorter duration. Disaster related epidemiological data is most useful for making preventative plans and policies. This last concept leads us to the topic of disaster planning.

Preparing and planning for disasters

Following the 2010 earthquake in Haiti, the Centre for Research on the Epidemiology of Disasters (CRED) produced an assessment contrasting the previous 10 natural disaster events that struck the island of Hispaniola. Of note, the two countries that comprise this island each had similar populations

of approximately 10 million people. Table 11.11.2 next shows the number of deaths and affected people per event in each country.

Table 11.11.2 Contrast in death rates between the same storms hitting the Haiti and Dominican Republic

Date	Event	Haiti deaths	Haiti affected	DR deaths	DR affected
Sept. 2008	Hurricane Hanna	529	48,000	1	10,745
Aug. 2008	Hurricane Gustav	85	73,006	8	6,257
Aug. 2008	Trop. Storm Fey	10	220	4	20
Dec. 2007	Trop. Storm Olga	3	2,090	33	61,605
Oct. 2007	Trop. Storm Noel	90	108,763	129	79,728
Aug. 2007	Hurricane Storm Dean	9	3,966	1	1,600
May 2007	Flood	2	12,500	9	16,000
March 2007	Flood	14	15.014	3	3,830
Oct. 2005	Trop. Storm Alpha	12	2,192	9	1,000
Sept. 2004	Hurricane Jean	2,754	315,594	11	14,009
Totals		3,508	581,345	208	194,794

Source: data from CRED 2010.

What is striking is that 10 storms hitting the same island by chance affected 2.5 times as many people in Haiti but induced 17 times more deaths in Haiti. There are likely many factors such as crowding and deforestation that work against the safety of the less wealthy Haitian population. But some factors, like construction quality, zoning codes and compliance, and government response plans also likely influenced the sevenfold risk of mortality associated with these storms for residents in Haiti.

Preparing for disasters—health emergency risk management

The most effective means of reducing the public health risks and impacts of disasters is to employ a comprehensive emergency risk management approach. Emergency risk management (ERM) consists of a continuum of synergistic steps that can reduce the risks and consequences of disasters, including prevention (risk reduction and mitigation), preparedness, response, and recovery. These steps are so closely interconnected that for all practical purposes they should not be considered in isolation from each other. Risk assessments and risk reduction activities directly inform disaster preparedness and response, for example. Moreover, the manner by which agencies respond to an emergency will directly impact the recovery process (WHO 2011b).

ERM for health draws on the disciplines of risk management, disaster management, and health systems. The approach emphasizes managing the risks prior to a disaster occurring so that potential public health consequences may be prevented or minimized. Specifically, ERM entails identifying and reducing hazards and associated vulnerabilities, preparing for and responding effectively to disasters, and recovering, rehabilitating, and reconstructing after a disaster has occurred. Experience has demonstrated that countries and communities that have invested sufficient resources in ERM have responded more effectively and have been more resilient to the effects of disasters. Strengthened ERM also has the potential to contribute to sustainable development.

Bangladesh is one of the most natural disaster-prone countries on the planet. Large numbers of people live on a low, flat alluvial delta vulnerable to floodwater coming from the Brahmaputra basin to the north and the typhoon-prone Indian Ocean to the south. A typhoon in 1970 resulted in approximately a quarter of a million deaths (Bern et al. 1993). Over the next decades, the people of Bangladesh built 311 elevated typhoon and flood shelters that could hold 350,000 people and implemented community-based response strategies. A similar 1991 typhoon affecting a far larger population resulted in 138,000 deaths (Mallick et al. 2005). During the 1991 storm, seeking a storm shelter early and being in an official shelter were highly protective against death although only 17% of the affected populations sought shelter in response to the storm warning sirens. Since then, advances in storm warning, public communications and community response have resulted in Southern Bangladesh being held up as an example of disaster preparedness (UNDP 2010).

Guiding principles

In addition to the emphasis on managing risks, ERM is guided by the following key principles (Brennan et al. 2009):

Addressing all hazards. Many of the key functions required to manage risks are common across all hazards, for example, assessments, communications, logistics, etc. Moreover, it is neither efficient nor cost-effective to develop separate plans, procedures, trainings, rosters, and response mechanisms for each individual hazard. While some hazard-specific considerations will rarely be required (e.g. radio-nuclear), an all-hazards approach allows for more efficient management of risks and responses.
Multisectoral approach. Many activities required to protect and preserve health are managed primarily by non-health sectors, for example, water and sanitation, shelter, food security, law and order, etc. Effective mechanisms for multisectoral collaboration and coordination are therefore necessary to ensure optimized ERM in the health sector. These should be established at national and local levels, wherever possible.
Prepared community. Communities are always the first to respond to disasters and can play a critical role in meeting emergent health needs. Community resilience can be strengthened by building their capacity to prevent, mitigate, prepare for, respond to, and recover from disasters. Local governments, civil society organizations, and individual citizens all have the potential of contributing to ERM.

Key components of emergency risk management

Prevention, risk reduction, and mitigation come first.

Table 11.11.2 regarding the impact of the 10 storms on the Dominican Republic vs. Haiti highlights the value of robust building zoning codes, urban planning, flood plain buffer zones, utility repair plans, evacuation plans, pre-recorded radio messages, and all of those things that prevent people from risk exposure which are planned years and months ahead of any disaster. The highest priority of disaster planning should be to keep people and potentially harmful exposures apart. While that statement is very broad (including evacuation orders and do not travel orders and hazardous waste

regulations), it needs to be kept in the forefront of the disaster planner's priorities.

Disaster risk assessments are the lens through which prevention and risk reduction can happen.

In 1993, a truck bomb was exploded in the parking lot under the World Trade Center killing 6 and injuring over 1,000 people (Lambert n.d.). In the aftermath of this explosion, electricity went out and smoke filled the vertical structure. With the electricity out, the elevators did not work. Because the stairwells had no windows and were dark and filled with smoke, it took many hours for people to go down those stairwells in the dark. The CDC sent a team to assess what had gone right and wrong during the event and the team's top recommendation was that battery powered lights be placed in the stairwells in case the power was ever cut off again. In 2001 when the World Trade Center was struck by two airliners 8 years later, thousands and thousands of people were likely saved because those lights were in those stairwells in the minutes after the electricity went out and before the buildings collapsed. This is a particularly dramatic example of the value of a disaster risk assessment.

Ideally, a disaster risk assessment should be multisectorial, multidisciplinary, and attempt to identify risks in advance to develop clear strategies to manage them (International Federation of the Red Cross 2006). The health component of a multisectoral risk assessment aims to determine the likelihood and potential public health impact of future disasters. Possible health consequences include direct impacts on people (e.g. injury, disease) or health facilities, and/or secondary effects due to damaged infrastructure (e.g. water and sanitation systems), population displacement, environmental damage, etc.

A coastal community will have greater vulnerabilities to hurricanes than communities located at a distance from the shoreline. A community on a steep hillside will have different earthquake risks than a nearby village on the flat valley bottomland. Thus, the most important lessons and ideas that will help shape a community's preparation for a disaster will often come from locals and will need to vary from place to place. The task of a disaster planner is to separate out those lessons that are universal (e.g. never drive a car into moving floodwaters, always assume electrical lines are 'hot' even if the electricity has stopped working) and those that are site-specific (e.g. when a tornado comes, go to the church basement).

Strategies and programmes should be based on evidence and endorsed or formed on the level of the implementers.

Disaster risk assessments, based on assessing past disasters, inspecting the environment, and speaking with key partners should be used to develop strategies and programmes to reduce the potential impact on public health by the identified hazards. These may include general health system strengthening measures to improve the coverage and effectiveness of existing health services (e.g. vaccinations, training and supervision of staff, improved drug management). They may also include specific health sector interventions that can mitigate the effects of disasters, such as the successful safe hospital programme in Latin America (Pan American Health Organization 2000, 2002) and early warning alert and response systems for outbreak control. In some places, radio messages or alarms are important to establish. In some situations such as the recent Latin American Zika outbreak, focusing support and efforts to key groups like health workers and pregnant women, may be appropriate. Once a strategy or programme is established, drills or simulations can be an invaluable tool to understand what parts of the response procedure are not working optimally.

Get prepared. The UNISDR definition of preparedness is 'the knowledge and capacities developed by governments, professional response and recovery organizations, communities, and individuals to effectively anticipate, respond to, and recover from, the impacts of likely, imminent or current hazard events or conditions' (UNISDR 2009). Many emergency organizations also use the related term 'readiness' to describe their ability to rapidly and appropriately respond to a disaster. In those cases where emergencies are likely to be predicted or recur (e.g. hurricanes, floods, wildfires), preparedness plans need to be in places on the household level, as well as on the local governance and national levels.

Sound preparedness clearly arises from an effective risk assessment and is linked to early warning systems. It should ensure the required resources and capacities can be rapidly deployed following the onset of a disaster. Clear emergency functions should be specified in advance and rosters of qualified and experienced staff developed. Stockpiles of essential drugs, equipment, and supplies should be established and maintained, including strong logistics procedures. All emergency staff should be trained in the relevant technical standards, performance standards, and response procedures (Sphere Project 2011). Standard operating procedures should address key administrative,

financial, procurement, and communications functions. Coordination mechanisms should be developed in advance, both for improved health sector and intersectoral functioning. Contingency planning may take into account possible scenarios and triggers for response, and further strengthen coordination.

Response is what gets noticed. When Ebola struck in DRC in 2018, WHO was more than 100 times more responsive than they had been during the 2013–2015 outbreak in West Africa in terms of person days delay in responding. WHO had prepared staff, a vaccine had been developed that was utilized in the field within a month of initial reports, and they had protective clothing and materials in stock. It is not by chance that WHO received scathing reviews for years after the 2013 outbreak and almost exclusively good press in 2018. This, in spite of the fact that over both periods responding to Ebola was likely much less than 10% of all that WHO was doing. This is nearly universal: governments and agencies are publicly evaluated on their response to crises more than for the crises they averted or non-emergency related good they do.

This implies that the things needed to respond (knowledge, messages, materials, and people) need to be in place long before a crisis hits. Having preparedness plans, documented operating procedures, a flexible emergency funding strategy, and a rooster of workers are all tasks best achieved far in advance of a crisis, but will usually be what most influences the quality of response.

Recovery is often neglected. Many organizations adhere to the so-called build back better approach. Recovery programmes are increasingly informed by post-disaster needs assessments (PDNA), which provide a measure of loss within each sector and a prioritized plan for recovery and reconstruction (World Bank Global Facility for Disaster Reduction and Recovery 2011). In reality, emergency responders are poorly suited to doing this and local players often do not have the resources. Most emergency funding lasts for months or a year at the longest and reconstruction rarely happens so quickly. While this is widely known and agencies and offices have attempted to address the void between emergency response and reconstruction, this remains a very problematic issue.

In 1993, the US government created a funding office to fill the void between emergency response and development called the Office of Transition Initiatives (OTI). Critics have pointed out how this institution has

done little to end the dropping of health supporting programmes when an acute crisis has passed. In the case of OTI, critics point out how it seems to spend much of its funding on civil society activities like democracy building (Blum 2013). For example, Wikileaks reported that in a 2006 cable, then-Ambassador William Brownfield, outlined a strategy to 'penetrating Chavez's political base', 'dividing Chavismo', 'protecting vital US business', and 'isolating Chavez internationally' and OTI funding was a central part of this (Hart 2013). While this example is from the US, many other countries have interwoven political and humanitarian agendas and it is inevitable that ambassadors and funders will not be able to operate with the humanitarian tenants of impartiality and neutrality. Thus, when recovery and developments get together for a 'Do Development Differently' conference, they conclude that locals are not sufficiently engaged and empowered and that true recovery is often neglected (DDD n.d.). When the UN held a Humanitarian Summit in Istanbul in 2016, they conclude that locals are not sufficiently engaged and empowered and recovery is underfunded or neglected (Agenda for Humanity 2016). Thus, in spite of widespread understanding of the problem and good intent, for the foreseeable future there will be a tendency to provide funding when a crisis is underway; but, in the poorest of settings, for competing priorities to shift the focus from rebuilding post-crisis.

Coordination of response

The coordination of responses occurs by a wide variety of mechanisms that vary widely. There is a tendency for natural disasters to be coordinated by a national emergency response mechanism. Sometimes this coordination is housed in the military, sometimes it is a permanent cabinet level office like in the UK and Japan, and sometimes it is a self-standing office like in Peru and Ghana. In such settings, assistance from abroad is typically coordinated by that entity. In some cases, the coordination structures can evolve or be complex. For example, during the 2014 Ebola outbreak in Sierra Leone, initially, the Ministry of Health directed the focus of response efforts through a National Ebola Response Committee (NERC) which met daily and coordinated most overseas assistance with the WHO. After 3 months, the military took over the national Coordination and the NERC, but the Ministry

of Health continued coordinating the foreign assistance (LR personal observation).

In complex emergencies and including wars, it is often difficult for the host government to play a coordinating role either because it is overstretched or weak, or because humanitarian actors are wedded to their concept of neutrality and are reluctant to cooperate with a warring party. In such settings the UN Cluster System is often established. Figure 11.11.6 is a visual display of this system.

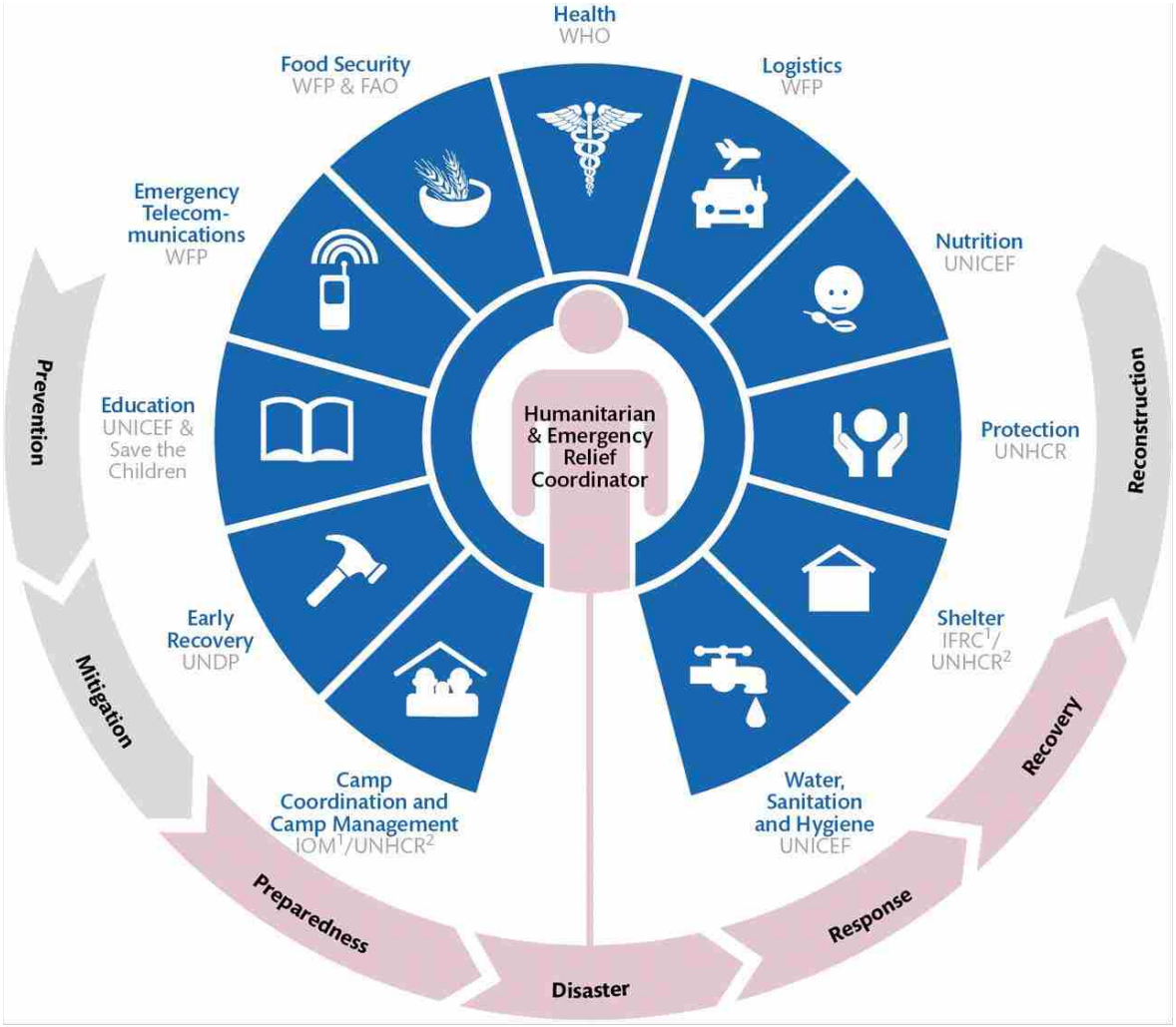


Fig. 11.11.6 UNOCHA depiction of the Cluster System.
Reprinted under a Creative Commons Attribution 4.0 license.

The Cluster System is typically coordinated by the United Nations Office for Coordination of Humanitarian Affairs (UNOCHA n.d.) and operates as a constellation of coordinating bodies. There are 11 sectors that each have a coordination lead, usually a UN agency that attempts to share information, resources and optimize the distribution of assets. In some settings where needed, there are subclusters that attempt to coordinate for a subset of a sector's activities. Typically, members of the humanitarian community gather for regular meetings, often with government agencies present. For a variety of reasons, some humanitarian actors such as MSF tend not to formally participate in Cluster processes. In 2018, 27 countries have 'Health Clusters' operating all led by the WHO. While the success of these clusters is widely debated, they are almost certainly going to be a mainstay of humanitarian coordination for the foreseeable future

Public health needs

Oxygen, water, and food remain the essential priority human needs in almost all settings including the most acute of hardships. The initial handbooks overviewing the public health priorities in high-mortality, low-resource settings remain as relevant today as when they were written (CDC 1992; MSF 1997). Most were lacking in some areas now better appreciated, with obstetric care being perhaps the most dramatic example. Usually those handbooks defined an emergency in terms of a doubling or quadrupling of the baseline mortality rate. In settings of extreme hardship, those displaced and whose livelihoods and social order are disrupted will generally continue to die and suffer from the same ailments they experienced before the crisis but at a higher rate. In some categories of disasters such as natural disasters like earthquakes and floods, death profiles are briefly but dramatically altered but prolonged elevated mortality as seen in the 2017 Puerto Rico example earlier is rare. Thus, in the most typical disruption and displacement scenarios, acute respiratory infections, diarrhoea, and in some places malaria, dominate the prevention and treatment priorities. Yet, these kinds of crises, with mortality rates elevated fourfold or more, are accounting for a smaller and smaller fraction of all humanitarian aid.

Most aid workers today can work an entire lifetime and never see the dramatically elevated mortality of ten or 20 times the baseline rates seen repeatedly in the 1960s through the Rwandan Crisis of 1994. For example, if

we use the threshold of 2% of the population dying each month (ten times the highest of baselines) in a population over 100,000 people to define a humanitarian catastrophe, a review of both published and unpublished mortality reports suggests that from 1979 to 1994, as shown in Figure 11.11.1, there was such a catastrophe on average every 2 years, but none have been reported since that time (Roberts 2013). The camp-based relief setting has also become rare, forcing relief agencies to work in cities and with existing government institutions.

This change away from relief typified by water and food delivery and life-saving primary healthcare approaches is being driven by four forces: the global transition from a disease burden profile typified by acute infectious illness to one dominated by chronic diseases, more aid funding and a downward trend in conflict inducing a lower threshold of what is considered a humanitarian crisis, changing motivations, and the rise of the rights-based framework within humanitarian relief.

Transition from acute infectious diseases to chronic conditions

In recent years, international health crises like the 2013–2016 West Africa Ebola outbreak and the 2015–2016 Zika epidemic have dominated the headlines. In both of these events, the WHO declared a ‘public health emergency of international concern’ to mobilize international support and governments and NGOs led a global response to increase preparedness and funding for these diseases that seemed to threaten global health security.

Yet despite the outsized media coverage and financial contributions to combat them, outbreaks of acute infectious diseases are now responsible for a small proportion of global deaths. Lopez et al. have documented how globally, between 1990 and 2005 the human species transitioned from experiencing disability-adjusted life years (DALYs) lost primarily to infectious diseases to losing DALYs primarily from chronic and non-infectious causes (Lopez et al. 2006). In fact, globally more people will die in the next 3 hours from NCDs than the total number who died from Ebola in the 3-year-long West Africa epidemic (WHO 2018). Yet despite the fact that NCDs are responsible for 60% of all DALYs and 70% of all deaths, they currently receive only 2% of overseas development assistance for health (Allen 2017).

The epidemiological transition occurring in developing countries from infectious diseases to non-communicable diseases as a greater contributor to mortality has been well documented. For example, Lopez et al. contrast the causes of death among adults in various regions based on 1990 and 2001 data respectively (Lopez et al. 2006). Over those 11 years, the non-HIV infectious disease adult death rates declined markedly in Latin America and the Caribbean, the Middle East and North Africa, East Asia and the Pacific, and South Asia. In each locality, this reduction in infectious disease death rates corresponded with lower overall adult mortality. Only in the post-Soviet Europe and Central Asia region and Africa did adult mortality rise over this period; in the former, primarily from cardiovascular diseases and unintentional injuries, and in Africa primarily from HIV. In Africa, overall mortality still decreased over this period because of dramatic reductions in infant and child mortality.

The impact of this epidemiological transition was evident in crises in Bosnia, Kosovo, and Iraq, where Ministries of Health and aid agencies were confronted with morbidity profiles dominated by non-infectious diseases. More recently, the high prevalence of NCDs in Syrians displaced by the ongoing civil war has challenged aid groups and health systems in neighbouring countries. Prior to the Syrian civil war, a 2007 survey found that cardiovascular disease was responsible for 45% of overall mortality in Aleppo, and a situational analysis conducted by the WHO Eastern Mediterranean Regional Office (EMRO) demonstrated that 79% of mortality among Syrians nationally was related to NCDs (Gardemann 2002; Rahim 2014). Providing NCD care for the millions displaced by the conflict has been complicated by the difficulties in humanitarian relief coordination within Syria and by the lack of access to primary healthcare in neighbouring countries where many Syrians have sought refuge.

The growing prevalence of non-communicable diseases in humanitarian settings highlights the disconnect between the health needs of those impacted by crisis and a humanitarian system that has been built to respond to crises characterized by high-mortality from communicable diseases.

Adding to this need to transition medical preparedness for emergencies is the growing appreciation of HIV as a non-fatal chronic condition. An estimated 10.3 million Africans regularly received antiretroviral drugs (ARVs) in 2015, double that of 2011, covering 54% of the total number of Africans who are HIV positive 2011 (WHO et al. 2011; WHO 2016). As a

disproportionate number of humanitarian operations occur in Africa, planning for the chronic disease support aspects of HIV needs to be part of the humanitarian public health portfolio (O'Brien et al. 2010).

The future success of the humanitarian community will depend in part on how well it can develop the capacity to provide quality care for NCDs in humanitarian crises while simultaneously maintaining the ability to respond to outbreaks of acute infectious diseases that will undoubtedly continue to plague our increasingly connected world.

More aid, more in need, and longer armed conflicts

Spending on humanitarian assistance has increased dramatically in recent decades. According to Global Humanitarian Assistance, formal humanitarian assistance spending increased fourfold from 1990 to 2010 (United Nations Children's Fund 2013). Unfortunately, however, more aid is being taken up by the increasing numbers of people displaced by conflict, requiring relief after natural disasters, and to a lesser extent, technological disasters such as the Fukushima nuclear meltdown in 2011. This has the effect of lowering the mortality prioritization with regard to what is considered a humanitarian crisis. It also means since much humanitarian spending is in post-conflict settings, that humanitarian assistance is now inseparable from development work.

At the same time, there was a downward trend in the number of armed conflicts in the world since the end of the Cold War, although this number has been on the rise since 2013, driven largely by ongoing conflict in the Middle East (Gleditsch et al. 2002; PRIO 2017). Although many of the ongoing conflicts are internalized within a state, a growing proportion are internationalized indicating the involvement of external states, which is significant because internationalized conflicts tend to last longer and more difficult to resolve (PRIO 2017). Further complicating the issue, conflict has become more protracted with the average conflict increasing from 19 years in 1990 to 37 years in 2013. Garfield et al. documented how the number of people living in conflict has increased over the past half century but a smaller and smaller portion of those people are in high-intensity conflict settings (Garfield 2012). With prolonged conflicts taking place in multiple geographic regions, the international community faces serious challenges.

Changing motivations and the rights-based framework

Aid meant to induce development may be different in nature from humanitarian aid designed to prevent deaths and suffering in the short term. Donors are continuing to place increasing emphasis on sustainability of humanitarian interventions and demanding local participation and empowerment. These are natural changes in perspective that match a transition from death aversion to priorities involving improved quality of life. A few NGOs such as MSF maintain a philosophy that they work primarily to reduce death and suffering but most others now hold long-term impact and sustainability as central objectives. This creates several new conditions for relief work. Many relief efforts in the 1980s and 1990s defined success by the lowering of the crude mortality rate (CDC 1992). This is rare today. Sustainability efforts usually require training and supporting local officials and institutions, committing to multiyear engagements, and monitoring changes in behaviours. These require different skills from those prevalent in relief agencies in the twentieth century. Reconstruction activities following earthquakes and tsunamis often employ many times more construction and economic experts than health professionals and epidemiologists.

In tandem with this shift in focus from just averting deaths, has been the rise of a rights-based approach to programming. The United Nations Children's Fund (UNICEF) defines a human rights-based approach as ‘a conceptual framework for the process of human development that is normatively based on international human rights standards and operationally directed to promoting and protecting human rights’ (UNICEF 2013). This framework, while not fundamentally contradictory, is very different in focus and in programme implications from the ‘public health’ approach of the last century’s relief operations (CDC 1992; MSF 1997). This public health approach, sometimes called a cost-effectiveness approach, concentrated on how the relief community could reduce the most deaths and suffering with the assets available. This manifested in specific interventions like measles immunization, latrine construction, or primary healthcare clinics which brought broader benefits at low costs.

The rights-based approach has the advantage that it more readily encompasses issues like psychological suffering which are harder to measure, and can project values such as that girls should have equal chances for advancement as boys. But the term ‘normatively based’ in the UNICEF

definition is not well defined or constant. For example, UN resolutions or agency position statements often stress the need to have child re-unification programmes where benefit is difficult to demonstrate (Ager and Metzler 2012). In the public health approach the benefits and the cost-effectiveness of programmes are more readily established, and new ideas can be compared with established programmes. Rights-based programming, in contrast, has no equivalent incentive to demonstrate cost-effectiveness and in fact is more driven by the strength of resolutions or declarations. Thus, if one wants to advocate for the value of counselling rape victims, in a rights-based framework it would be more efficient to attempt to pass a UN resolution than to evaluate the effectiveness of programmes in the field. With the public health approach, the opposite would be true.

In summary, humanitarian public health needs have broadened as the clientele served has begun to span a broader range of development levels, as the need to support people with chronic conditions including HIV has become more important. This may be an indication that the humanitarian endeavour has worked over the past half century. A Biafra-like scenario, where half the population died in a 4-year period, is rare today. Thus, the world's growing sources of compassion and aid can serve more than just those on the verge of death.

New developments for the twenty-first century

A host of new developments make it likely that relief work in the middle of the twenty-first century will look little like efforts in Biafra and South East Asia the century before. The advances of the past decades that will likely expand in the future can be considered in three broad categories: political trends, new approaches, and new technical advances. These categories interact with and influence each other, with technical advances often creating the potential for new orderings of social organization. While not comprehensive, this section seeks to imagine the likely changes to come in the field of humanitarian response.

Political trends

The melding of political and humanitarian agendas

National governmental political agendas and humanitarian relief have never been completely independent. In 1919, the US Congress formed the American Relief Administration (ARA), supposedly an independent NGO-like entity, headed by future President Herbert Hoover, to act as an independent famine relief mechanism in Europe in the years following the First World War (Fisher 1927). Oxfam, among the most famous of NGOs for its politically neutral funding sources, was formed in large part to lobby the British government to loosen the Allied blockade of occupied Europe (Oxfam n.d.). MSF co-founder Bernard Kouchner later became the French Minister of Humanitarian Affairs and the Foreign Minister, allotting funding for his creation, MSF. That humanitarian and political interests overlap is not completely unhealthy if in a democracy the populace is privately choosing to fund humanitarian ventures while simultaneously influencing the priorities of their government to do the same. On the other hand, countries are often in open competition with each other, and the linking of aid to enhance economic, military, or political competitiveness would be antithetical to the humanitarian endeavour. Two documents have attempted to keep humanitarian aid 'neutral' in the eyes of combatants and beneficiaries over the last part of the twentieth century.

Both the Third Geneva Convention of 1949 and the Red Cross Code of Conduct, to which most NGOs and the UN aspire to adhere, actively encourage the separation of humanitarian endeavours and self-interested agendas of individual states. The fourth tenet of the non-binding Red Cross Code of Conduct explicitly states, 'We shall endeavour not to act as instruments of government foreign policy' (International Federation of Red Cross and Red Crescent Societies n.d.). The ICRC and MSF are particularly famous for keeping political and military forces at arm's length even during the acute phases of crises. For example, during the 1992 crisis when Kurds fled northern Iraq but were not allowed into Turkey, MSF refused to move in public or cooperate with any military personnel carrying a weapon. Many military personnel, including the British and American forces who had just invaded Iraq and were thus belligerents in this conflict, operated within the displaced community without weapons so as to work with MSF and other NGO personnel. Thus, until 2001, while funding influenced where relief went and what kinds of relief efforts were funded, an honest effort was made on the part of most donors and most international NGOs to not undertake patently political activities to advance the agenda of one country.

Nonetheless, a remarkable change in the landscape of humanitarian assistance has been the melding of political and humanitarian agendas after the terrorist attacks on New York City on 11 September 2001. Beginning with the invasion of Afghanistan, this narrowing separation of political and humanitarian ventures was often ignored or blurred. In October 2001, US Secretary of State Colin Powell addressed leaders of the American NGO community referring to their organizations as 'force multipliers' (International Federation of Red Cross and Red Crescent Societies n.d.). In Afghanistan, US humanitarian aid was openly discussed as a military tool for winning hearts and minds. In May of 2003, USAID Administrator Andrew Ntasios told the American NGOs that they needed to do a better job of advertising that the goods and services they provide to the Afghan people were coming from the US government. He told the American NGOs that they were an 'arm of the US government' (Klien 2003). Military activities such as distributions and school construction often made it unclear what was being done by combatants and what was being done by NGOs (Falconer 2009). MSF, who had been working in Afghanistan for 33 years, was forced to pull out there and in Iraq after their aid workers were targeted for assassination (Katz and Wright 2004). This blurring of aid and political agendas was again demonstrated when the Central Intelligence Agency hired a physician to conduct a sham vaccination campaign in Pakistan as part of the hunt for Osama Bin Laden in 2011 (Roberts and VanRooyen 2013; Rubenstein 2013). Subsequent to this ploy, the Aid Worker Security Database demonstrated a drastic increase in the number of vaccinators attacked and killed as shown in Figure 11.11.7 next.

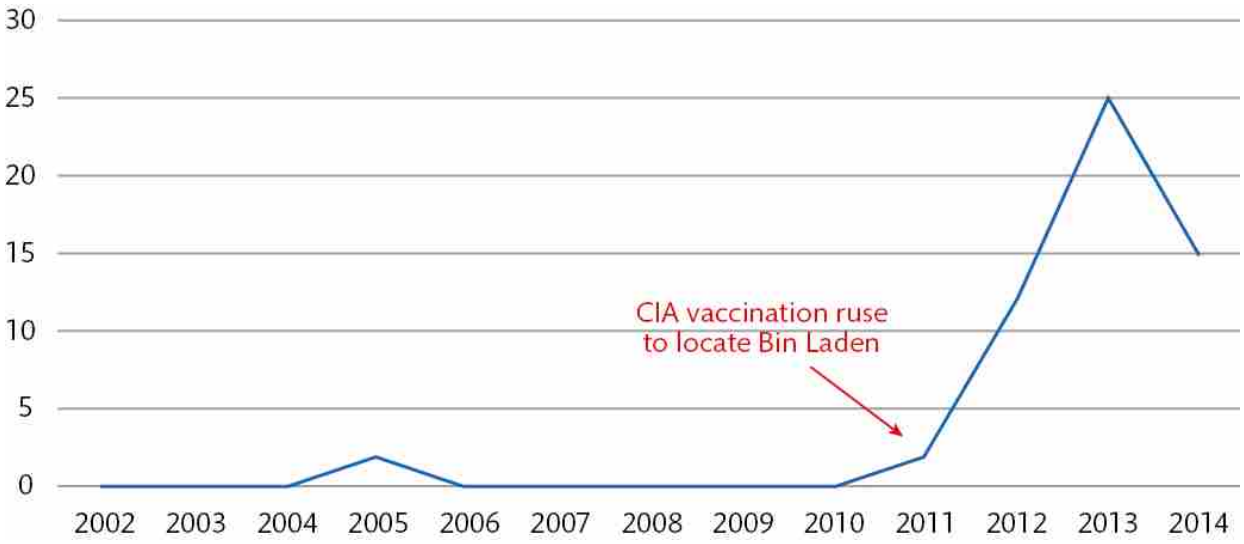


Fig. 11.11.7 Vaccinators attached, 2002–14.
Source: data from Humanitarian Outcomes (2015), Aid Worker Security Database, https://aidworkersecurity.org/Graphic provided courtesy of A. Stoddard.

The responses to natural disasters are not devoid of political overtones or agendas either. Following the 2005 tsunami in a largely Muslim area of the Pacific, the US Secretary of State called the disaster a 'wonderful opportunity' for the United States (Anonymous 2005). Preferential responses to former colonies and near neighbours appear to be an inevitable aspect of bilateral giving following natural calamities.

Attacks on healthcare

As a likely consequence of the blending of political and humanitarian agendas discussed earlier, there is growing worry within the humanitarian community that attacks on healthcare—including workers, structures, and transport—are occurring at an unprecedented rate.

During armed conflicts, international humanitarian law protects healthcare facilities and personnel from being the targets of attack. This protection is enshrined in the Geneva Conventions and was re-iterated by the United Nations Security Council Resolution 2286 as recently as 2016 (UN 2016). Certain high-profile attacks—such as the destruction of MSF's hospital in Kunduz, Afghanistan in 2015 by a US Air Force gunship—have provoked widespread concern among aid workers. Although there have long been documented attacks on healthcare—including Italian attacks on the

Ethiopian Red Cross, and even attacks on Red Cross that incensed Henri Durant in the Franco Prussian War—there is a perception among aid workers of an increase in attacks on healthcare workers and facilities (International Crisis Group 2005; Verma 2017). While anecdotal evidence, especially in Syria suggest an increase, a historical lack of standardized data documenting these attacks makes it difficult to quantify any trends (Briody 2018).

In response numerous organizations have established advocacy campaigns —including ICRC's Health Care in Danger and MSF's Medical Care Under Fire—to raise awareness and document such attacks while comprehensive databases have been established by independent research organizations like the Aid Worker Security Database of Humanitarian Outcomes to record attacks on healthcare workers (WHO 2015).

The WHO-led trauma response during the 2016 Mosul offensive in Iraq provides an example of how healthcare is being provided in potentially more dangerous conditions that may expose aid workers to a greater risk of attack. A coalition of Iraqi and international forces launched military operations to retake the city of Mosul from ISIS but abdicated its responsibility under international humanitarian law to provide medical care to war-wounded combatants. Anticipating a high number of patients with complex trauma, the WHO assumed control of the trauma response and contracted medical NGOs and private medical actors to 'co-locate' or embed within military units on the frontlines of conflict to provide medical stabilization and coordinate a trauma referral system (Fox et al. 2018). Although there were no reported attacks on healthcare workers, there was concern in the humanitarian community that directly working with military forces on the frontlines of a conflict increases the risk to aid workers and compromises the principles of independence and neutrality.

Migration—a record number on the move

An unprecedented 65.6 million people around the world have been forced from their home, and 20 more people are displaced every minute while fleeing violence, poverty, and persecution (Pew Research Center 2016).

Most people on the move today are internally displaced, indicating they have not crossed a border and are still within their home country. Using the example of those displaced by the Syrian civil war, nearly half of the nearly 12 million Syrians on the move are still within that country. The majority of

the rest are living in neighbouring countries—including Turkey (3.4 million), Lebanon (1 million), Jordan (660,000), and Iraq (125,000) (Pew Research Center 2018). Despite the increasingly anti-immigrant and nationalistic rhetoric of politicians in the Western world, relatively few on the people on the move are being hosted in the United States or Europe.

In addition to the large displacement of Syrians since 2011, there are multiple other large populations on the move throughout the world. In Central America, violence in Honduras and El Salvador in 2016 caused 388,000 to flee the region to seek refuge further north in Mexico and the US. Beginning in 2017, nearly 700,000 of Rohingya—a stateless minority group living in Myanmar—fled across the border to Bangladesh to escape persecution. Across sub-Saharan Africa, hundreds of thousands of men, women, and children—primarily from West Africa—have travelled through the Sahel to reach Libya from which many try to reach Europe via the dangerous sea route.

In the course of their journeys and regardless of whether they cross a border, displaced people are subject to numerous challenges, including human rights abuses, lack of access to healthcare, and restricted movement. For example, in Libya countless thousands are held in official and unofficial detention centres where rape, torture and abuse in an attempt to extract ransom are common (Amnesty International 2018). Similarly, many Rohingya report having been the victim of rape and indiscriminate attacks while in Myanmar or while fleeing across the border into Bangladesh.

One aspect of the migration experience that is disturbingly underdocumented is the number who die while on the move. Efforts such as the ‘Missing Migrants’ project by the International Organization for Migration have attempted to quantify the mortality experience of migrants in different contexts, but these are limited by incomplete and unstandardized data collection, limited political funding and support, and the difficult physical environments through which many migrate. Using the Mediterranean Sea route as an example, the Missing Migrant’s Project reports that 6,869 people have died attempting this crossing between 2014 and 2017, making it the most dangerous sea journey in the world (IOM 2018). However, the number is undoubtedly much higher given that this estimate doesn’t include the untold number of people who drowned but whose bodies were never found.

Humanitarian accountability

While aid agencies and the Red Cross Code of Conduct have always purported to support local empowerment and accountability of aid deliverers to beneficiaries and donors alike, many new developments have made that aspiration more common. The third edition of the Sphere Standards (Sphere Project 2011), perhaps the most widely followed guidelines for humanitarian practice, has an expanded emphasis on local participation and accountability to beneficiaries. In 2010, the UN passed an 'Accountability Framework' which may or may not result in more cost-efficient services on the ground, but certainly contributes to an international focus on recording transparently what is achieved by aid efforts and at what cost (UN 2010). Likewise, in 2011, the Inter-Agency Standing Committee (IASC) developed an operational framework on how to ensure accountability to affected populations in humanitarian emergencies which is widely cited among humanitarian players (IASC 2011). Thirdly, the rise of wealthy, highly functional governments in all regions of the world raises the expectations of aid delivery procedures. For example, in 2003, India became a net creditor to the International Monetary Fund (IMF) and the World Food Programme (WFP) (Bijoy 2011). The rise of local NGOs as a primary international delivery mechanism, the prevalence of social networking, and the economic downturn in post-2008 Europe and North America all point towards a future where local and more players have a larger role in what aid is provided and how it will be disbursed.

Growing participation of 'non-traditional' actors in humanitarian response

At its root, today's humanitarian system is fundamentally a Western construct based on European Enlightenment ideals of charity and limitations on the conduct of war. Although some form of the humanitarian impulse is likely as old as mankind, the formal structure we know today was first established in the late nineteenth century and was greatly expanded in the aftermath of World War II with programmes primarily targeted to provide assistance to Europeans displaced by war. The humanitarian community gradually extended its programmes globally to nearly all populations affected by crisis, and its core principles have largely remained unchanged

since their beginnings, despite substantial global geopolitical and financial changes.

However, there are indications that within the global aid community a number of 'non-traditional' actors—including China, Saudi Arabia, and Turkey—are playing a larger role in the humanitarian endeavour. While the majority of funding for humanitarian emergency response comes from 'traditional' government and intergovernmental sources—the US, European Commission and United Kingdom alone contribute more than 50% of the total government humanitarian contributions—recent years have shown greater contributions from outside the rich Western nations. Between 2012 and 2014, the largest increases came from Saudi Arabia and the Gulf States of Kuwait and Qatar (ALNAP 2015) and as shown in Figure 11.11.8, between 2004 and 2016 humanitarian contributions from China have increased fivefold (Hirono 2018).

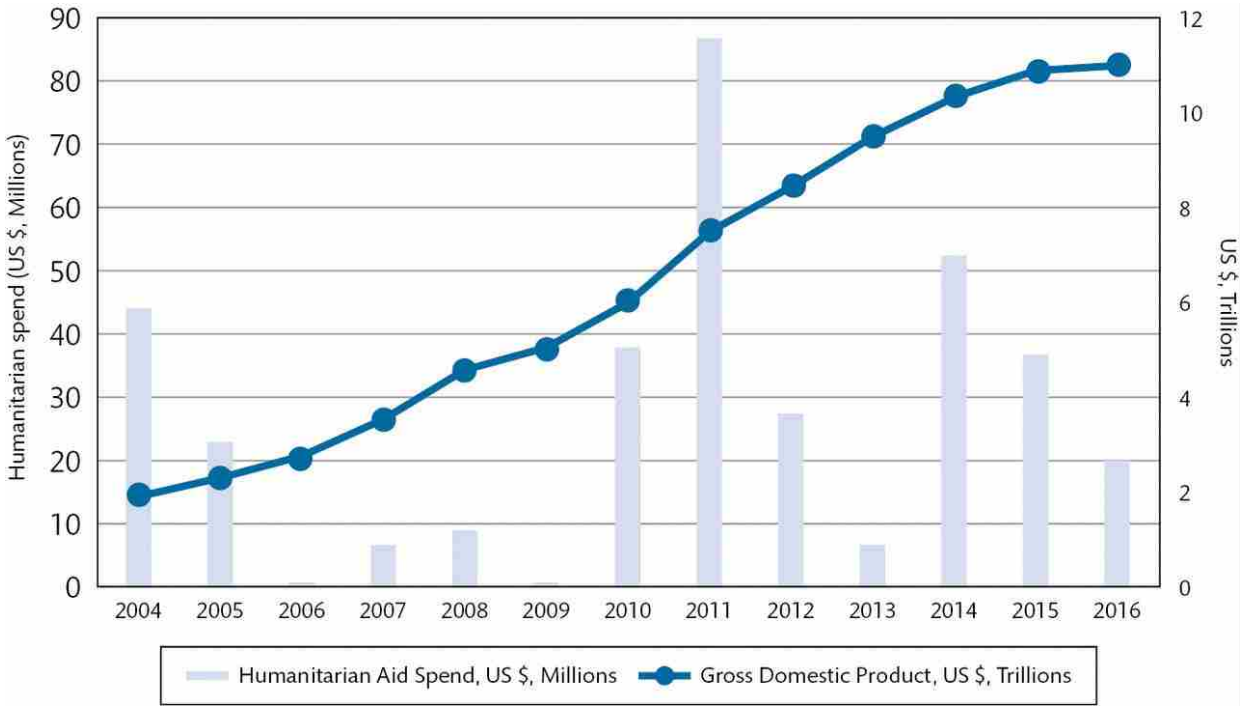


Fig. 11.11.8 Chinese GDP in trillions and humanitarian spending in millions 2004–2015. Sources: UNOCHA FTS, World Bank.

In addition to their growing financial presence, non-traditional actors are also playing a bigger political role in the global humanitarian system. Turkey

has used its growing economic influence and its strategic position between Europe and the conflicts in the Middle East to position itself as a significant contributor to humanitarian assistance worldwide. In recognition of its involvement in Somalia, its role in hosting millions displaced in the Syrian civil war, and its deal with the EU to reduce the number of asylum seekers entering Europe, Turkey was selected to host the first ever World Humanitarian Summit in 2016.

Humanitarian response has largely been shaped by Western ideas and actors since its inception, yet that is currently changing as non-traditional actors constitute a larger part of the financial and political leadership of the humanitarian endeavour. It remains to be seen what impact this growing involvement of non-Western donors will have on the future of humanitarian action.

New approaches

The growing appreciation of local capacities to provide exacting technical services combined with technical advances and greater emphasis on accountability, have paved the way in the past decade for exciting new ways of delivering aid. Among these are the following:

Home-based treatments

In the 1980s, the promotion of ORT was championed aggressively by the UN and aligned agencies because it was seen as a way to overcome the primary barrier to successful treatment of acute illness, namely early detection and treatment. The logic was that if people did not have to go through the effort of transporting the patient and their things to a medical facility, far more people would get treated. This was partially the logic behind the barefoot doctors' programme in post-revolution China and a variety of other initiatives such as mobile clinics (Gish and Walker 1977).

Over the past decade, ambitions to provide antiretroviral treatments in areas of modest medical infrastructure and limited resources have resulted in a variety of models for drug distribution and patient monitoring using community or patient volunteers, or paid community health workers. The evidence suggests that patient management and monitoring schemes that are not physician based, where the monitoring occurs closer to the community level, are associated with better patient retention and survival (Callaghan et

al. 2010). In focused settings of high malaria transmission, home-based treatment of fever has been shown to be an effective life-saving approach (Kidane and Morrow 2000).

In emergency settings, no home treatment strategies have received more attention or inspired as much enthusiasm as home therapeutic feeding. Home therapeutic feeding usually involves training and equipping mothers of malnourished children to cure the child at home with a nutritious ready-to-use food such as eeZeePaste™NUT or Plumpy'Nut® which come precooked and are resistant to bacterial growth. These ready-to-use home treatment products cut down on attrition in feeding centres, are a less expensive means for treating moderately malnourished children, and have been shown to be highly effective in field trials (Patel et al. 2005).

Crowd sourcing and mobile phones

The widespread availability of mobile phones and social media has made it possible to have the whole population participate in the monitoring of crises. One of the most well-known of these efforts is the Ushahidi platform, formed during the violent period following the 2008 Kenyan elections (Ushahidi n.d.).

Perhaps the most dramatic application of the Ushahidi platform was in the February 2010 Chilean earthquake. The 8.8 magnitude quake was followed by a series of aftershocks that damaged thousands of buildings and ultimately killed an estimated 521 people. Within 48 hours of the initial quake, an Ushahidi–Chile platform was established and a team of graduate students in New York began mapping press reports and text messages to create a picture of damage and where people were trapped. One hundred 'incidents' were mapped in the first day of operation and 700 were mapped by the end of 3 days. In the weeks that followed, the Ushahidi platform updated local press members and the general public about food distributions, dangerous buildings, the occurrence of looting, and the progress of children re-attending school (Carlsen et al. 2011). Crowd sourcing strategies have also been used to inform Chinese authorities about acute situations in the aftermath of the Szechuan earthquake, to map the occurrence of sexual violence in Syria, and to map unsafe areas following the Japanese tsunami of 2011.

Mobile phones by themselves are allowing for some exciting innovations. Somalis affected by conflict have received humanitarian aid in the form of

phone credits which can be used to purchase goods in local shops (Global Communities n.d.). After the earthquake in Haiti, the largest mobile phone carrier used cellular networks to compare the movement of people before and after the earthquake to show that 630,000 people in Port-au-Prince left the capital in the 19 days following the earthquake (Bengtsson et al. 2011). In the Democratic Republic of Congo, community members in North Kivu were provided mobile phones to monitor trends in the number of children separated from their families and caregivers in emergencies (Guinote 2012; Rubenstein 2015). Many of these initiatives are exploratory and controlled trials will be needed to establish such procedures as effective protective investments. It is likely mobile-phones will be increasingly utilized by the humanitarian and health community in the years to come.

Monetary aid

As part of the promotion of local and beneficiary empowerment, providing aid that allows beneficiaries to set their own priorities has become increasingly popular. The example of giving displaced Somalis mobile phone credit was discussed earlier. But, giving cash directly to beneficiaries in lieu of rations or material goods has several advantages. People are more likely to end up with what they want and the transfer costs are far lower. It has the potential to support the local economy including creating incentives for future increased local food production. It potentially has downsides such as driving local inflationary pressures and exposing people to increased risks of exploitation and extortion. A recent comprehensive review generally weighs the upsides as being greater than the downsides but suggests that implementers will need to choose the contexts carefully and develop methods for monitoring the success of direct cash assistance (Harvey 2007).

Closely related to monetary aid is the distribution of credit vouchers or coupons. Sometimes these have a specific purpose such as paying for school fees or uniforms. The most common flexible distribution is non-food item (NFI) or food item fairs where a select collection of vendors selling at predetermined prices gather in one place and beneficiaries with coupons spend their coupons at this one place and time. This system has the advantages of giving beneficiaries choice over the aid items they receive, it supports local markets, and leaves beneficiaries less open to local extortion. A recent analysis of such a NFI fair aimed at recently displaced populations in South Kivu Province, DR Congo found that local authorities missed most

IDPs when constructing vulnerable lists and included many locals. At the end of the NFI distribution process, the NFI 'vulnerability score' as defined by UNICEF was unchanged by the distributions (O'Keeffe and Zadravecz 2012). As with so many new practices, implementation and evaluation procedures will need to be developed locally to match each context in order to maximize the efficiency of distributions via fairs and similar voucher systems.

New technical advances

A host of new technical advances or tools also create the potential for transformative advances in humanitarian relief in this century (Interagency Standing Committee 2013). Some technologies, such as the Global Positioning System (GPS), are so essential to relief efforts at present that it is hard to realize that the technology was almost non-existent 20 years ago. It is likely to evolve as a tracking tool for people, goods, and vehicles. One-dollar GPS transmitters will likely result in the tracking of individual cases of vaccines and drugs, and curb looting of warehouses by armed groups. Eye and fingerprint scanning technologies are reducing ration theft and false registrations across the globe. Advances in medicine, in diagnostic techniques, or protective devices such as long-lasting impregnated bednets, are constantly changing the practice of humanitarian relief. Some of the more promising new technical and medical breakthroughs are highlighted here.

Satellite imagery and remote sensing has already changed humanitarian practice. UNHCR has used satellite images to count people moving from Darfur into Chad. For years, remote sensing has been used to predict crop failure and crop yields. The destruction of villages in Darfur and DRC has been documented with before and after remote images. But lately, satellite imagery has been used to undermine the impunity brought by national sovereignty. For example, during the 2017 period government violence against the Rohingya in Rakhine State in Myanmar, Human Rights Watch made extraordinarily detailed and timely documentation of village destruction by remote imagery (HRW 2017).

Cloud computing and internet collaboration hold potential for a host of new ideas and innovations that may assist those responding to emergencies. Groups like Hackers for Humanity attempt to channel the efforts of internet

talents to charities (Hackers for Humanity n.d.). A variety of efforts to solve 'humanitarian problems' have been addressed by crowdsourcing for a solution. A notable example is the use of open-source survey (OpenDataKit) and navigation (OpenStreetMap) software in the 2013–2016 West Africa Ebola outbreak. Using local people with self-owned smartphones, surveyors were able to quickly update geographic and village information used by medical surveillance and response teams at a cost of €3.80 per village surveyed (Nic Lochlainn et al. 2018).

Internet accountability mechanists: Google Earth and the internet are allowing for greater accountability of aid from the perspective of small private donors. A model held up as innovative but already being reproduced is that of Charity Water. Charity Water does not directly implement water supply and sanitation projects, but instead subcontracts with local operational partners to turn Western private donations into water and sanitation infrastructure (Charity Water n.d.). Charity Water provides donors with the latitude and longitude of the specific things people have paid to have built so that over time they can see if indeed a well was built or public tap installed, and they can monitor if it continues to be used over time. They even install monitors so that donors to a specific project can see at any given moment if water is flowing from it. While not all aid is remotely visible, this process of linking donors to beneficiaries visually or by social media will likely grow over time and the process of humanitarian assistance will be better for it (Srinivasan et al. 2012).

Unmanned aerial vehicles: Although unmanned aerial vehicles (UAV) —'drones'- are known primarily for their military uses, there has been increasing recognition of their potential to improve humanitarian and disaster response. UAVs have already delivered small packages to camps in Port-au-Prince after the 2010 Haiti earthquake and helped assess the damage caused by Typhoon Haiyan in the Philippines in 2013 (Oxfam 2013; Reuters 2014). Zipline, a US-based drone manufacturer and operator, has partnered with the Ministry of Health in Rwanda to trial delivery of blood for transfusion to remote areas throughout the country, and other countries have used UAVs to deliver emergency vaccines, HIV medication and antimalarial medication to difficult-to-access areas (Reliefweb n.d.). Confronted by high tuberculosis burden in a very remote location, Doctors Without Borders recently began using an unmanned maritime vehicles (UMV) to transport tuberculosis samples in Papua New Guinea (MSF 2018). Although they may

face challenges from regulatory requirements and local opposition, UAVs may likely play a larger role in future humanitarian and disaster response.

Telemedicine: Telemedicine—medical care provided from a distance—is being increasingly used in developed settings to remotely connect patients and healthcare providers. MSF has been using telemedicine since 2010 to connect field staff with experts based in Europe or the US who can provide advice from afar (Walji 2015). Although its use is still limited, the increasing presence of cellular and wireless connections and the lack of specialized healthcare providers in the most remote areas means telemedicine has promise for greater use in humanitarian response.

Conclusion

The twenty-first century is likely to see an ever growing humanitarian movement that will become more focused on economic, emergency preparedness, and developmental issues. Aid will be more politicized, while simultaneously more locally controlled and with greater accountability for implementation. The post-Biafra humanitarian endeavour has been associated with a dramatic decline in extreme episodes of excess deaths associated with crises and the new face of humanitarian relief, which often looks a lot like development work of old, has largely been built on the successes of twentieth-century humanitarian relief.

REFERENCES

Agenda for Humanity (2016). *Outcome of the World Humanitarian Summit.* [Online] Available at: https://www.agendaforhumanity.org/sites/default/files/A-71-353%20-%20SG%20Report%20on%20the%20Outcome%20of%20the%20WHS.pdf

Ager, A. and Metzler, J. (2012). *Child Friendly Spaces: A Structured Review of the Current Evidence-Base. World Vision*. [Online] Available at: http://reliefweb.int/sites/reliefweb.int/files/resources/CFS_Literature_Review_final_Aug_2012.pdf

Aid Worker Security Database (n.d.). [Online] Available at: https://aidworkersecurity.org/

Al-Jazeera (2016a). Syria civil war: air strikes hit Aleppo Hospitals. [Online] Available at: http://www.aljazeera.com/news/2016/07/syria-civil-war-airstrikes-hit-aleppo-hospitals-160724090834185.html

Al-Jazeera (2016b). Syria's civil war explained from the beginning. [Online] Available at: https://www.aljazeera.com/news/2016/05/syria-civil-war-explained-160505084119966.html

Allen, L. (2017). Non-communicable disease funding. *The Lancet Diabetes & Endocrinology*, 5(2), 92.

ALNAP (2015). Website. [Online] Available at: https://www.alnap.org/
Amnesty International (2018). Libya: number of detained refugees and migrants soars. [Online] Available at: https://www.amnesty.org.uk/press-releases/libya-number-detained-refugees-and-migrants-soars
Anonymous (2005). Condi Rice: tsunami provided 'wonderful opportunity' for US. Agence France Presse, 18 January. [Online] Available at: http://www.commondreams.org/headlines05/0118-08.htm
Armed Conflict Location and Event Data Project (ACLED) Excel Database (n.d.). *ACLED Analysis*. [Online] Available at: http://www.acleddata.com/data/acled-version-7-1997-2016/
Bengtsson, L., et al. (2011). Improved response to disasters and outbreaks by tracking population movements with mobile phone network data: a post-earthquake geospatial study in Haiti. *PLoS Medicine*, **8**, 8.
Bern, C., et al. (1993). Risk factors for mortality in the Bangladesh cyclone of 1991. *Bulletin of the World Health Organization*, 71(1), 73–8.
Bhui, K.S., Hicks, M.H., Lahley, M., and Jones, E. (2012). A public health approach to understanding and preventing violent radicalization. *BMC Medicine*, **10**, 16.
Bijoy, C.R. (2011). India: transiting to a global donor. In: Reality of Aid Management Committee (eds.) *South-South Development Cooperation: A Challenge to the Aid System?* [Online] Available at: http://www.realityofaid.org/userfiles/roareports/roareport_3ce2522270.pdf
Blum W. (2013). The Anti-Empire Report #115. April 2013. https://williamblum.org/aer/read/115
Boschi-Pinto, C.M., Velebit, L., and Shibuya, K. (2008). Estimating child mortality due to diarrhoea in developing countries. *Bulletin of the World Health Organization*, 86(9), 657–736.
Brennan, R.J., Bradt, D.A., and Abrahams, J. (2009). Medical issues in disasters. In: Cameron, P., et al. (eds.) *Textbook of Adult Emergency Medicine* (3rd ed.). London: Churchill Livingstone, pp. 785–92.
Briody, C., Penney, E., Rubenstien, L., Horbar, J., Keenan, W., and Roberts, L. (2018). Review of attacks on health care facilities in six conflicts of the past three decades. *Conflict and Health*, **12**, 19.
Burnham, G., Lafta, R., Doocy, S., and Roberts, L. (2006). Mortality after the 2003 invasion of Iraq: a cross-sectional cluster sample survey. *The Lancet*, **368**, 1421–8.
Callaghan, M., Ford, N., and Schneider, H. (2010). A systematic review of task-shifting for HIV treatment and care in Africa. *Human Resources for Health*, **8**, 8.
Carlsen, J., et al. (2011). *Grant Report for Mozilla Corporation*. [Online] Available at: http://newmediataskforce.files.wordpress.com/2011/02/ushahidi-chile-grant-report-16-jan-2011.pdf
Centers for Disease Control (1992). Famine-affected, refugee, and displaced populations: recommendations for public health issues. *Morbidity and Mortality Weekly Report*, 2441(13), 1–76.
Centers for Disease Control (2016). Cost of the Ebola epidemic | 2014–2016 outbreak West Africa. [Online] Available at: https://www.cdc.gov/vhf/ebola/history/2014-2016-outbreak/cost-of-ebola.html
Charity Water (n.d.). Website. [Online] Available at: https://www.charity-water.org
Coghlan, B., et al. (2006). Mortality in the Democratic Republic of Congo: a nationwide survey. *The Lancet*, **367**, 44–51.
Doing Development Differently (DDD) (n.d.). *The DDD Manifesto 2015*. [Online] Available at: http://doingdevelopmentdifferently.com/
EM-DAT: The Emergency Events Database—Université catholique de Louvain (UCL)—CRED. Brussels [Online] Available at: https://www.emdat.be
Falconer, B. (2009). Military hijacking aid effort in Afghanistan, say NGOs. *Mother Jones*, 3 April.

Fisher, H.H. (1927). *The Famine in Soviet Russia, 1919–1923: The Operations of the American Relief Administration*. New York: Macmillan.
Fox, H et al. (2018). Emergency trauma response to the Mosul offensive, 2016–2017: a review of issues and challenges. [Online] Available at: https://www.humanitarianoutcomes.org/Emergency_Trauma_Response
Gardemann, J. (2002). Primary health care in complex humanitarian emergencies: Rwanda and Kosovo experiences and their implications for public health training. *Croatian Medical Journal*, 43(2), 148–55.
Garfield, R. (2007). Measuring humanitarian emergencies. *Disaster Medicine & Public Health Preparedness*, **1**, 110–16.
Garfield, R.M., Polonsky, J., and Burkle, F.M. (2012). Changes in size of populations and level of conflict since World War II: implications for health and health services. *Disaster Medicine and Public Health Preparedness*, 6(3), 241–6.
Gish, O. and Walker, G. (1977). *Mobile Health Services*. London: Tri-Med Books Ltd.
Glass, R., et al. (1980). Rapid assessment of health status and preventive-medicine needs in newly arrived Kampuchean refugees, Sa Kaeo, Thailand. *The Lancet*, **315**, 868–72.
Gleditsch, N.P., et al. (2002). Armed conflict 1946–2001: a new dataset. *Journal of Peace Research*, 39(5), 615–37.
Global Burden of Disease (GBD) 2015 Eastern Mediterranean Region Intentional Injuries Collaborators (2018). Intentional injuries in the Eastern Mediterranean Region, 1990–2015: findings from the Global Burden of Disease 2015 study. *International Journal of Public Health*, 63(Suppl 1), 39–46.
Global Communities (n.d.). *CHF's Director of Humanitarian Assistance on Drought Relief in the Horn of Africa.* [Online] Available at: http://www.chfinternational.org/node/36995
Global Health Observatory (n.d.). Number of malaria deaths. [Online] Available at: http://www.who.int/gho/malaria/epidemic/deaths/en/
Global Humanitarian Assistance (2000). Commissioned by IASC from Development Initiatives, May 2000. [Online] Available at: http://devinit.org/wp-content/uploads/2017/06/2000-GHA-report.pdf
Guinote, F. (2012). *Calling for Help: How Cell Phones Help Protect Children's Rights in the DRC*. [Online] Available at: http://womensrefugeecommission.org/blog/1292-calling-for-help-how-cell-phones-help-protect-childrens-rights-in-the-drc
Hackers for Humanity (n.d.). Website. [Online] Available at: http://www.hackersforhumanity.com/.
Hagopian, A., et al. (2013). Mortality in Iraq associated with the 2003–2011 war and occupation: findings from a National Cluster Sample Survey by the University Collaborative Iraq Mortality Study. *PLoS Medicine*, 10(10). http://doi.org/10.1371/journal.pmed.1001533
Hart, P. (3 April 2013). WikiLeaks: was Chavez right about U.S. meddling? FAIR. [Online] Available at: https://fair.org/home/wikileaks-was-chavez-right-about-u-s-meddling/
Harvey, P. (2007). *Cash-Based Responses in Emergencies. Overseas Development Group (ODI), Humanitarian Policy Group Briefing Paper 25*. London: Overseas Development Group.
Hirono, M. (2018). Exploring the links between Chinese foreign policy and humanitarian action. HPG/ODI. [Online] Available at: https://www.odi.org/sites/odi.org.uk/files/resource-documents/12015.pdf
Human Rights Watch (2017). Burma: satellite imagery shows mass destruction. [Online] Available at: https://www.hrw.org/news/2017/09/19/burma-satellite-imagery-shows-mass-destruction
ICRC (n.d.). *The Code of Conduct for the International Red Cross and Red Crescent Movement and Non-Governmental Organisations (NGOs) in Disaster Relief*. [Online] Available at: http://www.ifrc.org/Global/Publications/disasters/code-of-conduct/code-english.pdf
Ifekwunigwe, A.E., Grasset, N., Glass, R., and Foster, S. (1980). Immune response to measles and smallpox vaccinations in malnourished children. *American Journal of Clinical Nutrition*, 33(3), 621–4.

Interagency Standing Committee (2011). *End Discrimination and End Discrimination and Ensure Equitable Outcomes for All Outcomes for All. 2010–2011 Accountability Frameworks for AGDM and Targeted Actions*. Geneva: UN.
Interagency Standing Committee (2013). Transformative agenda protocols. [Online] Available at: http://www.humanitarianinfo.org/iasc/page-loader.aspx?page=content-documents-default&publish=0
International Crisis Group (2005). *Ethiopia and Eritrea: Preventing War*. Africa Report No. 101, 22 December 2005. [Online] Available at: http://www.crisisgroup.org/~/media/Files/africa/horn-of-africa/ethiopia-eritrea/Ethiopia%20and%20Eritrea%20Preventing%20War
International Federation of Red Cross and Red Crescent Societies (2012). *World Disasters Report, 2012*. Geneva: IFRC.
International Federation of Red Cross and Red Crescent Societies (n.d.). *Red Cross Code of Conduct*. [Online] Available at: http://www.ifrc.org/en/publications-and-reports/code-of-conduct/
International Federation of the Red Cross (2006). *What is VCA? An Introduction to Vulnerability and Capacity Assessment*. Geneva: IFRC.
International Organization on Migration (IOM) (2018). Missing migrants. [Online] Available at: https://missingmigrants.iom.int/region/mediterranean
International Strategy for Disaster Reduction (UNISDR) (2009). *2009 UNISDR Terminology on Disaster Risk Reduction*. Geneva: UNISDR.
Jackson, A. (2012). Remembering the humanitarians. *Al-Jazeera*, 16 August. [Online] Available at: http://www.aljazeera.com/indepth/opinion/2012/08/201281213836229581.html
Katz, I.T. and Wright, A.A. (2004). Collateral damage—Médecins sans Frontières leaves Afghanistan and Iraq. *The New England Journal of Medicine*, **351**, 2571–3.
Kidane, G. and Morrow, R.H. (2000). Teaching mothers to provide home treatment of malaria in Tigray, Ethiopia: a randomised trial. *The Lancet*, **356**, 550–5.
Kishore, N., et al. (2018). A survey-based estimation of mortality in Puerto Rico after Hurricane Maria. *New England Journal of Medicine*, **379**, 162–70.
Klien, N. (2003). Bush to NGOs: watch your mouths. *The Globe and Mail (Canada)*, 30 June. [Online] Available at: https://www.theglobeandmail.com/opinion/bush-to-ngos-watch-your-mouths/article751269/
Kosek, M., Bern, C., and Guerrant, R. (2003). The global burden of diarrhoeal disease as estimated from studies published between 1992 and 2000. *Bulletin if the World Health Organization*, **81**, 197–204.
Lambert L. World Trade Center bombing of 1993. [Online] Available at: https://www.britannica.com/event/World-Trade-Center-bombing-of-1993
Lopez, A.D., et al. (2006). Global and regional burden of disease and risk factors, 2001: systematic analysis of population health data. *The Lancet*, **379**, 1747–57.
Mallick, D.K., et al. (2005). Case study 3: Bangladesh floods in Bangladesh: a shift from disaster management towards disaster preparedness. *IDS Bulletin*, 36(4), 53–70.
Médecins Sans Frontières (MSF) (1997). *Refugee Health*. London: Macmillan Education Ltd.
Médecins Sans Frontières (MSF) (2018). *'No One was Left': Death and Violence Against the Rohingya in Rakhine State*. Myanmar: Médecins Sans Frontières.
Médecins Sans Frontières (MSF) (2018). Innovating to reach remote TB patients and improve access to treatment. [Online] Available at: http://www.msf.org/en/article/papua-new-guinea-innovating-reach-remote-tb-patients-and-improve-access-treatment
Nic Lochlainn, L.M., et al. (2018) Improving mapping for Ebola response through mobilising a local community with self-owned smartphones: Tonkolili District, Sierra Leone, January 2015. *PLoS One*, 13(1), e0189959.
O'Brien, D.P., et al. (2010). Provision of antiretroviral treatment in conflict settings: the experience of Médecins Sans Frontières. *Conflict and Health*, **4**, 12.

O'Keeffe, J. and Zadravecz, F. (2012). *Bunyakiri NFI Post-Fair Evaluation. Unpublished Report for RHA*. New York: Columbia Mailman School of Public Health.
Our World Data (2017). *World Population Growth*. [Online] Available at: https://ourworldindata.org/world-population-growth
Oxfam (n.d.). *History of Oxfam*. [Online] Available at: http://www.oxfam.org.uk/what-we-do/about-us/history-of-oxfam
Oxfam (2013). *Typhoon Haiyan*. [Online] Available at: https://www.oxfamamerica.org/static/media/files/bn-typhoon-haiyan-philippines-response-071213-en.pdf
Pan American Health Organization (2000). *Natural Disasters: Protecting the Public Health*. Washington, DC: PAHO.
Pan American Health Organization (2002). *Guidelines for Vulnerability Reduction in the Design of New Health Facilities*. Washington, DC: PAHO.
Patel, M.P., et al. (2005). Supplemental feeding with ready-to-use therapeutic food in Malawian children at risk of malnutrition. *Journal of Health, Population and Nutrition*, 23(4), 351–7.
Pew Research Center (2016). A record-high share of the world's population is displaced from their homes. [Online] Available at: http://www.pewresearch.org/fact-tank/2016/08/03/nearly-1-in-100-worldwide-are-now-displaced-from-their-homes/ft_16-07-28_displaced_world/
Pew Research Center (2018). Most displaced Syrians are in the Middle East, and about a million are in Europe. [Online] Available at: http://www.pewresearch.org/fact-tank/2018/01/29/where-displaced-syrians-have-resettled/
Price, M., Gohdes, A., and Ball, P. (2014). *Updated Statistical Analysis of Documentation of Killings in the Syrian Arab Republic. Commissioned by the Office of the UN High Commissioner for Human Rights*. HRDAG. [Online] Available at: https://files.givewell.org/files/shallow/fragile-states/Price,%20Gohdes,%20and%20Ball%202014.pdf
PRIO (2017). Trends in Armed Conflict, 1946–2016. [Online] Available at: https://www.prio.org/utility/DownloadFile.ashx?id=1373&type=publicationfile
Rahim, H.F.A., et al. (2014). Non-communicable diseases in the Arab world. *The Lancet*, **383**, 356–67.
Reliefweb (n.d.). Are drones taking off as the future of African healthcare? [Online] Available at: https://reliefweb.int/report/world/are-drones-taking-future-african-healthcare
Reuters (2014). Drones being developed to deliver medical aid, not bombs. [Online] Available at: https://www.reuters.com/article/us-medical-drones-idUSKBN0GF17I20140815?feedType=RSS
Richie, H. and Rosner, M. (n.d.). Natural disasters. [Online] Available at: https://ourworldindata.org/natural-catastrophes
Roberts, L. (2013). *The Lack of an Evidence Base for Humanitarian Interventions*. Paper presented at Evidence Live, Oxford, UK 24 March. [Online] Abstract available at: http://www.evidencelive.org/2013/programme
Roberts, L. and VanRooyen, M.J. (2013). Ensuring public health neutrality. *The New England Journal of Medicine*, **368**, 1073–5.
Roser, M. and Ortiz-Ospina, E. (2018). Global extreme poverty. [Online] Available at: https://ourworldindata.org/extreme-poverty
Rubenstein, B.L., et al. (2015). Community-based surveillance to monitor trends in unaccompanied and separated children in eastern DRC. *Child Abuse & Neglect*, **19**, 19.
Rubenstein, L. (2013). Unhealthy practice. *Foreign Affairs*, 24 April. Available at: http://www.foreignaffairs.com/articles/139344/leonard-s-rubenstein/unhealthy-practice
Bulatao, R.A. (1993). Mortality by cause, 1970 to 2015. The epidemiological transition: policy and planning implications for developing countries: workshop proceedings. *National Academies Press*. [Online] Available at: https://www.ncbi.nlm.nih.gov/books/NBK236445/table/ttt00026/?report=objectonly

Sack, R.B., et al. (1970). The use of oral replacement solutions in the treatment of cholera and other severe diarrhea. *Bulletin of the World Health Organization*, **43**, 351–60.

Salama, P., Spiegel, P., and Brennan, R. (2001). No less vulnerable: the internally displaced in humanitarian emergencies. *The Lancet*, **357**, 1430–1.

Selby J, Dahi OS, Fröhlich C, Hulme M. (2017). Climate change and the Syrian civil war revisited. *Political Geography*, **60**, 232–44.

Shugerman E. The Independent (3 October 2017). Trump says Puerto Rico should be 'proud' of hurricane death toll: 'Look at a real catastrophe like Katrina's. [Online] Available at: https://www.independent.co.uk/news/world/americas/us-politics/trump-puerto-rico-death-toll-comments-katrina-hurricane-visit-latest-a7981356.html

Sphere Project (2011). *The Sphere Project: Humanitarian Charter and Minimum Standards for Humanitarian Response*. Rugby: Practical Action Publishing.

Srinivasan, V., Lambin, E.F., Gorelick, S.M., Thompson, B.H., and Rozelle, S. (2012), The nature and causes of the global water crisis: syndromes from a meta-analysis of coupled human-water studies, *Water Resource Research*, 48, W10516.

Toole, M.J., Nieberg, P., and Waldman, R.J. (1988). The association between inadequate rations, undernutrition prevalence, and mortality in refugee camps: case studies of refugee populations in Eastern Thailand, 1979–1980, and Eastern Sudan, 1984–1985. *Journal of Tropical Pediatrics*, 34(5), 218–24.

Toole, M.J. Steketee, R.W., Waldman, R.J., and Nieburg, P. (1989). Measles prevention and control in emergency settings. *Bulletin of the World Health Organization*, **67**, 381–8.

Toole, M.J. and Waldman, R.J. (1990). Prevention of excess mortality in refugee and displaced populations in developing countries. *Journal of the American Medical Association*, 263(24), 3296–302.

Toole, M.J. and Waldman, R.J. (1997). The public health aspects of complex emergencies and refugee situations. *Annual Review of Public Health*, **18**, 283–312.

United Nations (UN) (2010). *Agenda Item 130: Review of the Efficiency of the Administrative and Financial Functioning of the United Nations*. 29 January 2010. [Online] Available at: http://www.un.org/en/strengtheningtheun/pdf/A-64-640.pdf

United Nations (2016). Security Council Adopts Resolution 2286 (2016), strongly condemning attacks against medical facilities, personnel in conflict situations. [Online] Available at: https://www.un.org/press/en/2016/sc12347.doc.htm

United Nations Children's Fund (2013). *Human Rights-based Approach to Programming*. Global Humanitarian Assistance. [Online] Available at: http://www.unicef.org/policyanalysis/rights/index.html. See data at: https://docs.google.com/spreadsheet/ccc?key=0AvGCmVxSBN08dFJ2cGRKRkhpYTdWU0liNXgtTy1odlE#gid=24

United Nations Development Programme (UNDP) (2010). *Success Story Leads. Bangladesh: Comprehensive Disaster Management Programme*. [Online] Available at: http://www.undp.org/content/dam/undp/library/crisis%20prevention/Bangladesh%201.pdf

UNHCR (n.d.). Figures at a glance. [Online] Available at: https://www.unhcr.org/en-us/figures-at-a-glance.html

United Nations International Strategy for Disaster Reduction (2010). *Emerging Challenges for Early Warning Systems in Context of Climate Change and Urbanization*. Coppet: Humanitarian & Development Network.

United Nations Office for Coordination of Humanitarian Action (UNOCHA) (n.d.). *What is the Cluster Approach?* [Online] Available at: https://www.humanitarianresponse.info/en/about-clusters/what- is-the-cluster-approach

Uppsala Conflict Data Program (n.d.). *Uppsala Conflict Data Program*. [Online] Available at: http://www.pcr.uu.se/research/ucdp/

Ushahidi (n.d.). Website. [Online] Available at: http://www.ushahidi.com/about-us

Verma, A.A. (2017). Failing to protect humanitarian workers: lessons from Britain and Voluntary Aid Detachments in the Second World War. *Medicine, Conflict and Survival*, **33**, 216–28.
Vinck, P. and Pham, P. (2011). Association of exposure to violence and potential traumatic events with self-reported physical and mental health status in the Central African Republic. *Journal of the American Medical Association*, 304(5), 544–52.
Walji, M. (2015). Bringing telehealth to humanitarian settings. *CMAJ: Canadian Medical Association Journal*, 187(4), E123–4.
Western, K.A. (1972). *The Epidemiology of Natural and Manmade Disasters: The Present State of the Art*. Dissertation for D.T.P.H. London: University of London.
World Bank (2011). *World Development Report 2011: Conflict, Security and Development*. Washington, DC: World Bank.
World Bank Global Facility for Disaster Reduction and Recovery (2011). *Analysing the Social Impacts of Disasters. Volume 1: Methodology*. Washington, DC: World Bank.
World Health Organization (WHO) (2005). *International Health Regulations* (2005) (2nd ed.). Geneva: WHO.
World Health Organization (WHO) (2007). *Mass Casualty Management Systems*. Geneva: WHO.
World Health Organization (WHO) (2011a). *Strengthening Response to Pandemics and other Public-Health Emergencies*. Geneva: WHO.
World Health Organization (WHO) (2011b). Disaster risk management for health: fact sheets. [Online] Available at: http://www.who.int/hac/techguidance/preparedness/factsheets/en/index.html
World Health Organization, UNAIDS, and United Nations Children's Fund (2011). *Global HIV/AIDS Response: Epidemic Update and Health Sector Progress towards Universal Access*. Geneva: WHO. Available at: http://www.who.int/hiv/pub/progress_report2011/en/index.html
World Health Organization and World Meteorological Organization (2012). *Atlas of Health and Climate*. Geneva: WHO.
World Health Organization (WHO) (2015). Documenting attacks on health workers and facilities in armed conflicts. [Online] Available at: https://www.who.int/bulletin/volumes/95/1/15-168328/en/
World Health Organization (WHO) (2016). 17 million people with access to antiretroviral therapy. [Online] Available at: http://www.who.int/hiv/mediacentre/news/global-aids-update-2016-news/en/
World Health Organization (WHO) (2018). Non-communicable diseases. [Online] Available at: http://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases
World Health Organization (WHO) (n.d.). Measles. [Online] Available at: http://www.who.int/news-room/fact-sheets/detail/measles

11.12

Principles of public health emergency response for acute environmental, chemical, and radiation incidents

Virginia Murray, Thomas Waite, and Paul Sutton

Introduction to principles of public health emergency preparation, resilience, and response

Acute environmental, chemical, and radiation incidents and emergencies can have significant effects on people's health, including loss of life. New threats reveal new challenges for managing the health risks and effects of emergencies. Deaths, injuries, diseases, disabilities, psychosocial problems, and other health impacts can be avoided or reduced by skilled disaster management involving health and other sectors.

This chapter considers the various stages of an emergency response and the role of public health at each stage. Case studies are used to illustrate the principles described. Of note is the importance of emergency responders at the scene to save life and protect people (including the emergency responders). This chapter addresses the following:

Incident management arrangements for emergency response including international obligations: including the local emergency response arrangements and how they interact with international obligations such as the International Health Regulations from the World Health Organization (WHO) (see Chapter 11.10).

Public health risk assessment: in any incident response it is essential to assess exposure, as soon as possible, to members of the public, workers, and responders using biological or environmental monitoring or modelling. This will allow assessment of health risks to exposed people and define further environmental and people sampling and modelling to refine risk assessments. There are significant differences in exposure assessment methods between chemical and radiation incidents.

Public Health risk management techniques and countermeasures are important to ensure that advice that is given early enough to be effective to reduce exposure, even at the early stages of an incident when little is known. Countermeasures are broadly similar for all hazards and include decontamination, sheltering, evacuation, specific antidotes, and medical treatments. The chapter touches on mechanisms of recovery and return to normal. Follow-up of affected populations is often essential. Approaches on how to undertake appropriate follow-up of affected populations including ongoing care and epidemiological follow-up are discussed

Public reassurance in emergencies and incidents: it is essential to have clear messages for information and reassurance to cover both affected and unaffected members of the public. These need to build trust in the systems that have been invoked to protect the public.

Identify and publish the lessons and examine the importance of sharing and disseminating learning to decrease the chance of an incident reoccurring and to identify lessons to improve emergency preparedness, response, and recovery.

The chapter offers a conclusion which includes a need for further evidence identifying some of the research gaps and future challenges.

Incident management arrangements for emergency response including international obligations

Incident and emergency risk management for health refers to the systematic analysis and management of health risks, posed by incidents and

emergencies, through a combination of:

Hazard and vulnerability reduction to prevent and mitigate risks
Preparedness
Response
Recovery measures

This is summarized in Figure. 11.12.1.

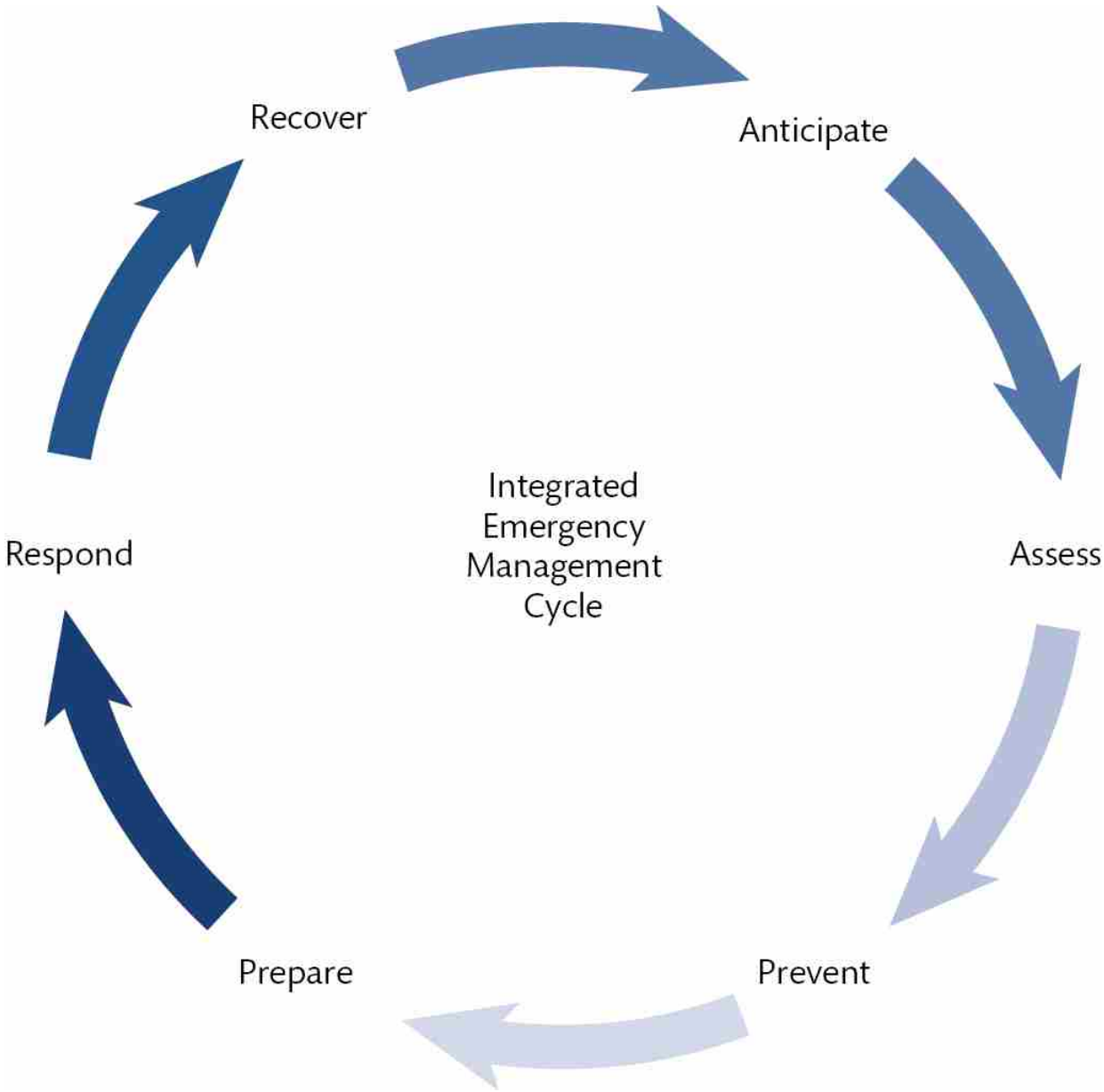


Fig. 11.12.1 The integrated emergency management cycle.

Reproduced with permission from National Health Service (NHS), NHS Commissioning Board Emergency Preparedness Framework 2013, Copyright © 2013, available from https://www.england.nhs.uk/wp-content/uploads/2015/11/eprr-framework.pdf, under the Open Government Licence v3.0.

Public health expertise has a vital role in reducing harm and saving life at every stage in this cycle. The traditional focus of the health sector has been on emergency response however increasingly the challenge is to broaden the focus of incident and emergency risk management to a more proactive approach emphasizing risk identification, reduction, prevention and

mitigation, and the development of community and country capacities to provide timely and effective response and recovery including learning. Resilient health systems based on primary healthcare at community level can reduce underlying vulnerability, protect health facilities and services, and help to scale-up the response to meet the wide-ranging health needs.

Emergencies and other crises may cause ill health directly or through the disruption of health services, leaving people without access to healthcare in times of emergency. They also affect basic infrastructure such as water supplies and safe shelter which are essential for public health.

World Health Organization Emergency Response Framework

The World Health Organization approach to emergency preparation and response is outlined in its Emergency Response Framework 2017 (WHO ERF), an all hazards approach which provides a structured framework for Member States (WHO 2017). The first step of the framework is risk assessment and situation analysis, including the importance of early identification and detection to prevent events from developing into public health emergencies where possible.

The framework then moves to rapid risk assessment and characterization to arrive at a risk level based on the likelihood of the event potential impact on public health. A standard approach to grading incidents and emergencies assists in determining the level of response required from the WHO; three grades are outlined for acute responses and three for incidents which are considered ‘protracted’ or to last beyond six months. Importantly, the initial operational response to any incident should not be delayed by this grading process, as the WHO adopts an important principle of ‘no regrets’ which supports over-resourcing rather than the risk of under-resourcing in the initial stages of a response.

From here the WHO Incident Management Process describes the various components of the WHO response which are common to all emergencies regardless of the underlying nature of the incident. This highlights the move towards structured and predictable response methods for emergency management systems in order to minimize variation and maximize the opportunities for interoperability and standardization. The WHO Incident Management Process is based on what it considers international best practice and is increasingly influencing the development of emergency response

plans for systems around the world. The framework concludes with detailed descriptions of the performance standards and key performance indicators for emergencies which monitor 'process' and output/outcome indicators respectively.

The UK Civil Contingencies Act 2004 (Box 11.12.1) provides an example of the implementation of the framework.

Box 11.12.1 UK Civil Contingencies Act 2004

The UK Civil Contingencies Act and supporting guidance establishes clear roles and responsibilities for emergency preparedness and response for its two categories of responder and imposes legal duties on them (Cabinet Office 2011, 2017). Category 1 responders are those organizations primarily involved in the response to the emergency such as emergency services and local authorities, Category 2 responders are those organizations which are needed to co-operate with Category 1 responders such as the Health and Safety Executive. These categories of responder come together in Local Resilience Forums to prepare, plan, exercise, and review emergency response within their areas. The guidance describes a cycle of emergency preparedness and response as an evolutionary and systematic process by which organizations should prepare for and respond to emergencies.

The cycle starts with direction being set by a risk assessment process. The steps which follow reiterate the importance of the risk assessment phase since the cycle moves on to objective setting, determining actions, and assigning responsibilities. It is the risk identification and assessment phase therefore, which drives the process from here. The process then moves through dissemination and training phases before seeking to validate plans through both actual responses and emergency. The crucial stage of identifying lessons and implementing learning from responses and exercises is followed by review and revision of plans, completing the evolutionary cycle.

UK HM Government guidance (2013) describes six activities within its Integrated Emergency Management approach which set the context for the cycle just described:

- Anticipation
- Assessment
- Prevention
- Preparation
- Response
- Recovery management

This approach provides a comprehensive approach to emergency preparedness and response for the United Kingdom.

It is of note that the importance of anticipation and systematic risk identification, followed by exercising and identifying learning and improving plans, provides a continuous improvement approach to preparing for and responding to public health emergencies.

International consensus views disasters as barriers to progress on health-related development as they often set back hard-earned development gains in

health and other sectors. However the recent synchronous adoption of landmark UN agreements the Sendai Framework for Disaster Risk Reduction, Sustainable Development Goals (SDGs), COP21's Paris Climate Conference, World Humanitarian Summit and Habitat III has created a rare but significant opportunity to build coherence across different but overlapping policy areas to reduce the impacts of disasters on health. The three key 2015 UN landmark agreements are briefly summarized next:

The **Sendai Framework for Disaster Risk Reduction 2015–2030** was endorsed by the UN General Assembly in June 2015 (UNISDR 2015). It puts health at the centre of global disaster risk reduction policy and advocates for action to reduce disaster risks for the next 15 years. It is a voluntary, non-binding agreement with seven global targets, aimed at the reduction of disaster risk and losses in lives, livelihoods and health and links to the critical implementation of the International Health Regulations (2008). It identifies four priorities for action:

- Understanding disaster risk
- Strengthening disaster risk governance to manage disaster risk
- Investing in disaster risk reduction for resilience
- Enhancing disaster preparedness for effective response and to 'Build Back Better' in recovery, rehabilitation, and reconstruction

The **UN Sustainable Development Goals** represent 17 aspirational 'Global Goals' with 169 targets between them, including the universal call to action to end poverty, protect the planet from climate change and ensure that all people enjoy peace and prosperity (UN General Assembly 2015). Target 3d of the health goal is to 'strengthen the capacity of all countries, in particular developing countries, for early warning, risk reduction and management of national and global health risks'.
The international political response to climate change that began at the Rio Earth Summit in 1992 was complemented with the adoption of the Paris Climate Conference in November 2015 (UNFCCC 2015). The **Paris Climate Change Agreement** aims to achieve a legally binding and universal agreement on climate and keeping global warming below 2°C.

Health risks can be understood in terms of hazards, people's exposure and vulnerability to that hazard and the capacity to manage the risk (Cardona et al. 2012). Human vulnerability to emergencies is a complex mix of issues

that includes socioeconomic, health, and cultural factors. In many situations it is not the hazard itself that necessarily leads to an emergency, but the vulnerability and inability of the population to anticipate, cope with, respond to, and recover from its effects. Populations and their health are put at risk from many hazards and the Sendai Framework identifies an all hazards approach, which includes natural, biological, technological, and societal hazards:

Natural: earthquake, landslide, tsunami, cyclones, extreme temperatures, floods, and droughts
Biological: disease outbreaks including human, animal, and plant epidemics and pandemics
Technological: chemical and radiological agent release, explosions, transport, and infrastructure failures
Societal: conflict, stampedes, acts of terrorism, migration, and humanitarian emergencies

Progress has been made at global, regional, national, and community levels, but countries' capacity for risk reduction, emergency preparedness, response, and recovery remains extremely variable.

The burden of emergencies falls disproportionately on vulnerable populations, namely the poor, ethnic minorities, old people, and people with disabilities. Worldwide, the loss of life from climate related emergencies is far higher among the less-developed nations than it is in developed nations (Cardona et al. 2012). Within each nation, including developed nations; poor people are the most affected. Poverty reduction is an essential component to reducing vulnerability to emergencies. High-risk populations must be prioritized in targeted efforts to mitigate human vulnerability including emergency- and disaster-related health outcomes, such as death, injury, illness, and disability.

In the context of emergency risk management, public health programmes build capacities and resilience of individuals and communities to risks, to reduce the impact, to cope with, and to recover from the effects of adversity. They address issues related to disparities in health outcome that arise between the general population and the most vulnerable groups.

International Health Regulations

The International Health Regulations (IHR), agreed by the World Health Assembly in 2005 and revised in 2007 (WHO 2008), were introduced to protect the world from global health threats. The need for such an internationally binding instrument was identified following the emergence of severe acute respiratory syndrome in 2003, which demonstrated how quickly a new disease could spread. As a result, 196 countries acknowledged a responsibility for collective action against infectious diseases, and food and environmental contamination, deemed to be public health emergencies of international concern. Member states are responsible for developing, strengthening, and maintaining the capacity to detect, assess the risks of, and respond to public health emergencies. Furthermore, countries are required to notify the WHO of these threats through their designated National IHR Focal Point within 24 hours. The regulations mandate countries to assess their capacities for disease surveillance and response and report whether these are sufficient to meet their obligations.

Implementation of the IHR and the WHO Emergency Response Framework has resulted in a number of high-risk countries having strengthened their disaster prevention, preparedness, and response systems; in some countries, the health sector has led initiatives developing multisectoral approaches to emergency risk management (WHO/PHE/UNISDR 2017). Multisectoral working is needed to protect the health of the population during and after a major incident or emergency. Wider determinants of health such as water, sanitation, nutrition, and security also need to be adequately addressed. In addition, essential infrastructure such as communications, logistics, energy and water supplies, emergency services, and banking facilities need to be protected through multisectoral working to ensure the continuity of health services.

Public health risk assessment

Risk identification and assessment are the first components of the public health risk management process. The risk assessment stage will take into account the probability of an event occurring and the magnitude of the consequences of the event. In the early stages of an incident, the risk assessment will be subjective but will be refined and quantified as more information is gathered.

Single risk assessment involves four stages: hazard identification, assessing the consequences, assessing their probabilities, and characterizing the risks and uncertainties (Gormley et al. 2011). In an emergency situation, the public health risk assessment can be done through a combination of qualitative and quantitative methods.

One of the greatest challenges of public health—and disaster—risk assessment is prioritization of risks arising among all the hazards in a given geographical location. This is further complicated by the use of different assessment techniques and metrics for understanding risks in different hazard types and thus common, multirisk frameworks are needed. Furthermore, public health emergency planning and response for acute environmental, chemical, and radiation incidents must be cognisant of the frequent and often complex interactions between such hazards.

Such multihazard approaches to understanding risk must start from single-hazard risk assessments and we thus start by outlining this process in generic terms. Natural hazards, such as extremes of rainfall or temperature, or landslides and floods, are particularly prone to cascade or interaction, increasing total risk. Examples include the tsunami in Japan of 2011, or the hurricanes in the Caribbean region of 2017.

In many environmental, chemical, and radiation incidents it is helpful to break down risks by considering the source, pathway, and receptors. (Box 11.12.2)

The source

Information about the source can be weather specific with naturally occurring environmental exposures, such as extremes of temperature, being understood through meteorological assessment, forecasting, or modelling. Chemical data on the source can usually be obtained from the site operator for industrial accidents. Where available, specialized fire and rescue hazardous materials and environmental protection officers can provide invaluable, on-the-scene information in the early phase of a chemical incident. The key requirement is to identify the chemical(s) released and its/their physical form. The chemical’s physico-chemical properties will provide information on its ease of spread throughout the environment. Understanding its toxicological properties and the quantities released is also key to successful risk assessment.

Box 11.12.2 Example of source–pathway–receptor model

Source: chemical release due to road traffic accident involving a tanker.
Pathway: dispersion of the chemical through the air.
Receptor: residents (children and adults).

Most incidents involving chemicals and radiation will require access to specialist advice to interpret the information gathered. It is important to remember that the substances contained in a product inventory are not always those of most public health concern. For example, a fire may lead to the formation of more toxic chemicals and can result in other hazards depending on the temperature and oxygen availability.

The pathway

The pathway describes how the exposure spreads in the environment. If a polluting gas, the main pathway will be through the air and dispersion modelling can be used to predict how it will move spatially and temporally. Liquids, depending on their volatility, may disperse through the soil, water, and air. Solids will disperse more slowly; however, fine solids will form dusts that can disperse quickly in the right conditions. Exposure to a substance can be assessed through a combination of methods: by sampling or monitoring the external environment, by modelling the external environment, or by measuring the substance in the body.

Food can also be part of a complex exposure pathway, for example, some chemicals and even radioactive ones travelling in the air can be deposited on crops which are later ingested leading to exposure.

Exposure to meteorological data may be routinely collected by relevant national agencies, such as the Met Office in the United Kingdom.

Timely information on the type, concentration, and environmental distribution of the pollutants is needed for public health risk assessment. Where available and applicable the concentration of a substance in the environment can be monitored and/or modelled with varying degrees of accuracy and precision depending on time and resources. Unless arrangements have been made at the planning stage for rapid deployment of sampling and measurement equipment, information may be received too late to inform the risk assessment.

A UK example of a planned monitoring system is the Radioactive Incident Monitoring Network (RIMNET), set up in 1988 to monitor the consequences of nuclear incidents abroad. RIMNET comprises 96 monitoring stations around the United Kingdom. A range of organizations can submit monitoring data to RIMNET. Public Health England advises government departments on the public health interpretation of radiation monitoring data.

Environmental matrices containing the substance of interest (e.g. air, food, water, dust, or soils) can be used to aid public health risk assessment. In certain circumstances, real-time measurements can be made, for example, for chemicals in air or monitoring of levels of radiation. Monitoring should be planned as soon as possible. The monitoring strategy should ensure that the samples taken are representative of the general bulk material as substances are rarely homogeneously dispersed in the environmental media. The methods used to take and store the samples should not degrade the substance of interest. The analytical techniques used must be appropriately sensitive, accurate, and reproducible. For direct measurements, monitoring over the right time interval allows comparison with any predefined health-based standards.

Modelling external exposure involves computer simulation to estimate the pollutants' concentration and predict their behaviour through one or more environmental media. Air dispersion models are often used during large-scale incidents. A number of data sources are required, depending on the sophistication of the model, for example, source term, weather data, and topography.

Monitoring and modelling should be considered together in a multiagency group attended by specialist scientific staff as information gained from monitoring could be used to inform the model and model outputs can predict areas to monitor for effective exposure assessments. The uncertainties for this type of external exposure assessment include the accuracy, completeness, and relevance of the data.

The receptors

For the purpose of a health risk assessment, the receptors are the people potentially exposed to environmental, chemical, and radiation incidents. Some population groups exhibit adverse health effects when exposed to extreme temperatures or at lower pollutant levels and may react more quickly than the general population. These are called 'sensitive receptors'.

Children, the elderly, people with pre-existing diseases, and pregnant women may be sensitive receptors, depending on the hazard involved.

The primary exposure routes for chemicals are absorption (contact between the chemical and the skin or eyes), inhalation (breathed into the lungs), ingestion (through contaminated food and/or water), and injection (through breaks in the skin). For radioactive materials, an additional exposure route is external irradiation, where gamma and/or beta radiation particles emitted by materials outside the body penetrate the skin and deposit their energy inside. Alpha particles are a risk to human health only when taken into the body by inhalation, ingestion, or through open wounds as they are easily stopped by the dead layers of the skin.

The duration of exposure is important to assess and estimate health risks. Single or short-term exposures to some chemicals at a sufficient concentration can trigger adverse health effects (e.g. carbon monoxide). Other chemicals require exposure for days and months to cause lasting health effects (e.g. some respiratory irritants).

Once inside the body, chemicals move through the metabolic pathways, are transformed, and either stored or eliminated. Biological markers of exposure can be used to enhance the risk assessment and measure the internal dose. However, the right study design, including the choice of the most appropriate body fluid or tissue to be tested, collection, storage, processing, and analytical protocols require a thorough understanding of the exposure route, absorption, and metabolism of the chemical of interest. Specialist scientific advice should be sought before considering human bio-monitoring to assess exposure.

Risk characterization

This step of the public health risk assessment is usually a quantitative assessment involving a comparison of the environmental, chemical, and radiation incident exposures or concentrations measured against health-based values, where they exist. These values are different for different types of events and even with chemical exposure the same chemicals in different environmental matrices. For example, health guidelines have been set for lead in air, water, and in blood. Health-based values are derived at levels below which adverse health effects are considered unlikely for so-called threshold chemicals. Exposure to some chemicals can lead to health effects

at any level of exposure, these chemicals are called ‘non-threshold chemicals’ and health-based concentrations are set for tolerable risks.

Radiation has both threshold and non-threshold effects. Acute injuries occur above a threshold high dose. At lower doses there are stochastic risks, mainly of cancer induction that are assumed to have a linear no-threshold dose–response function. Public and occupational exposures to radiation aim to be as low as reasonably achievable within the context of specific dose limits to aid emergency planning. Strenuous efforts should be made to avoid acute radiation injuries in emergency situations.

Characterization of the risk posed by natural hazards depends not only on the probability of the event, but also the complex interlinkage with both the infrastructure of modern day living and increasing population vulnerability due to climate change, urbanization communication, and transportation. The knock on consequences of a disaster in a single population or financial centre may have consequences in cities and economies across the world; similarly the victims to a disaster such as the Indian Ocean tsunami in 2004 will not be limited to the countries affected but may carry the individual and public health impacts with them when they return home, potentially many thousands of miles away.

Multirisk assessment methods

The Sendai Framework for Disaster Risk Reduction (UNISDR 2015) promotes the development of multihazard risk assessment of natural hazards. One of four priority actions in the Framework is to better understand disaster risk, with the development of guidelines for regional and national multihazard risk assessments, mapping, and risk profiles recorded as a key deliverable.

In order to prioritize hazards and risks in a given location or region, their quantitative and qualitative characterization must be comparable with one another. Since such characteristics inherently differ by hazard, such as return period, nature, impacts and intensity, and their measurements are necessarily in different units (wind speed, water depth, temperature) this results in great complexity when compiling hazards for a region in order to assess the total public health risk for the population.

Harmonization of risk assessment processes is thus an important first step to overcoming these difficulties. The use of matrices and indices for hazards, risks, and vulnerability have been proposed as a means to address this,

although such approaches do not allow assessment of interacting or cascading risk.

Matrices seek to classify hazards on more than one scale, often impact and intensity or frequency. These characteristics are then grouped categorically (either qualitatively or quantitatively) to allow graphic display and hence comparison of their relative importance.

An example of this is the UK National Risk Assessment (see Box 11.12.3).

Box 11.12.3 The UK National Risk Register of Civil Emergencies 2017 as an example of multirisk assessment methods implementation

The National Risk Register 2017 explains the risks of major emergencies that could affect the United Kingdom in the next 5 years and provides resilience advice and guidance and has been used by UNISDR as a model (UNISDR 2017).

The National Risk Register of Civil Emergencies 2017 edition provides an updated government assessment of the likelihood and potential impact of a range of different civil emergency risks (including naturally and accidentally occurring hazards and malicious threats) that may directly affect the United Kingdom over the next 5 years. Figure 11.12.2 provides some of the assessments used to inform the register.

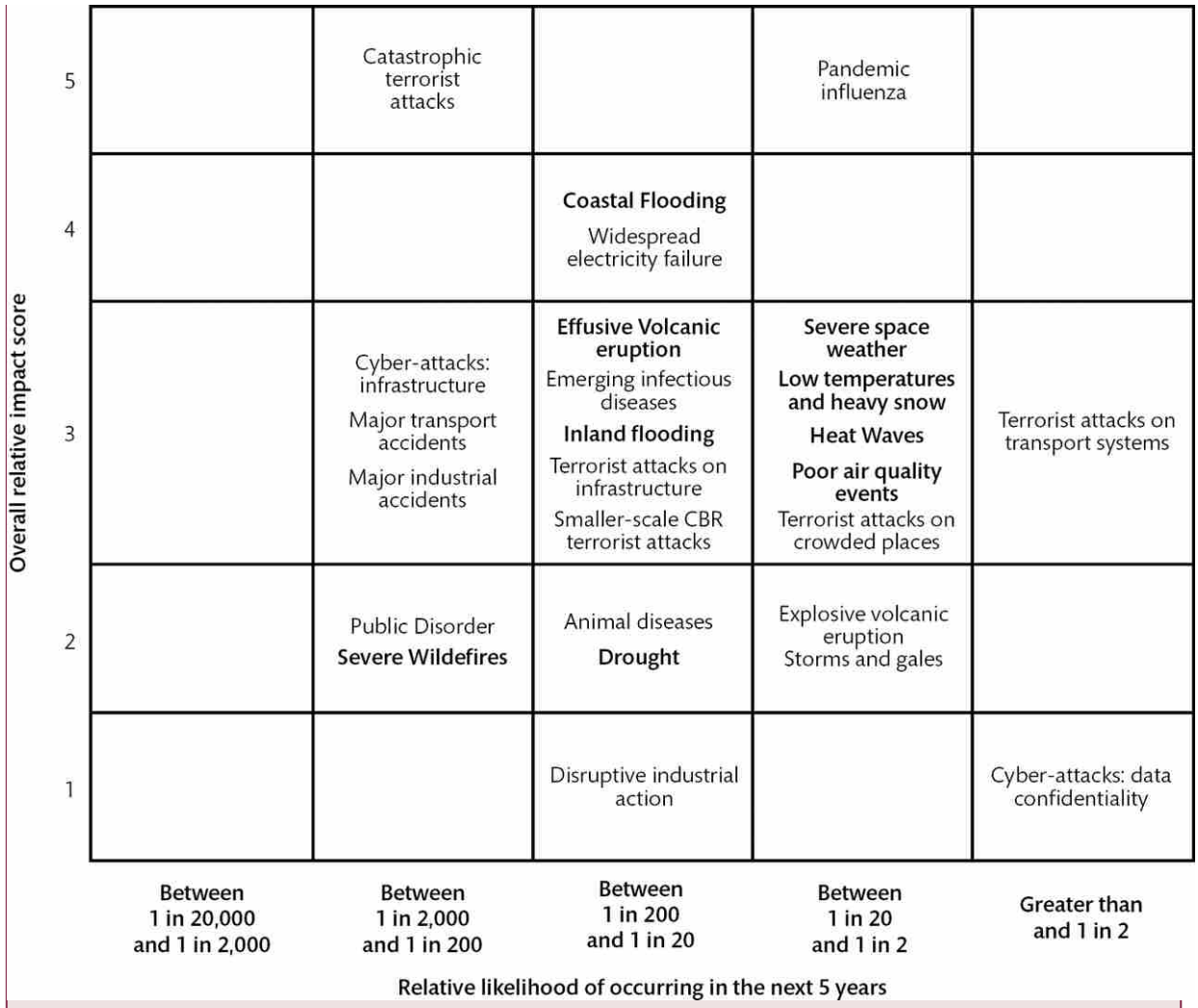


Fig. 11.12.2 Key civil emergency risks published in the 2015 National Risk Register (NRR) from UK NRR).

Reproduced from Hemingway R and Gunawan O. (2018) The Natural Hazards Partnership: A public-sector collaboration across the UK for natural hazard disaster risk reduction International Journal of Disaster Risk Reduction 27 (2018) 499–511, with permission from Elsevier.

In addition to providing information on how the UK government and local responders manage these emergencies, the National Risk Register also signposts advice and guidance on what members of the public can do to prepare for these events. As an open access document regularly updated the Risk Register allows the public to find links and information about how to find the local Community Risk Register.

Indices, or index-based approaches, also seek to allow comparison between diverse risks and hazards. Their composition depends on the selection of indicators—usually composite indicators on the basis of an underlying model of the multidimensional concept that is being measured—

which result in a meaningful, overall quantitative measure. The availability of useful and useable data is often a determining factor in such indices. Sociodemographic data, such as population level metrics of age, income, or sex as well as measures of social disadvantage (such as unemployment) and vulnerability (such as disability or pregnancy) are routinely collected in many countries at census. Composite variables of such data can thus be calculated at different geographic scales, from country level down to small administrative units. Environmental and meteorological data may be available at very different geographies, or with far less precision and thus the age old problem of assigning area level information to individuals or groups must be navigated with great care.

Multiple risk assessments by such means can simply be totalled to obtain an overall hazard level in a geographic location, but this approach overlooks the interaction and cascade between hazards which frequently result in more significant and widespread public health impacts than had the individual hazards occurred at separate points in time. These interactions have been classified into five relationship types (Gill and Malamud 2016):

Triggering relationships (such as rainfall triggering landslides)
Increased probability relationships (such as wildfire leaving an area at risk of landslides should rainfall occur)
Decreased probability relationships (such as rainfall reducing the risk of drought)
Coincidence relationships
Catalysis and impedance relationships (such as volcanic eruption triggering wildfire but impeded by tropical storms)

Qualitative multiple risk assessment methods seek to describe and classify hazard relationships to allow decision makers to select the most important relationships in their region. Semi quantitative methods use matrices to characterize interactions between hazards by the inclusion of multiple hazards in a single matrix, with interactions assigned a score based on judgement of the strength and direction of the interaction (Zschau 2017). Quantitative methods seek to use event or decision trees with underlying Bayesian probability distribution functions to weight particular branches and nodes. This approach has been used in the modelling of scenarios around volcanoes including Mount Vesuvius in Italy (Neri et al. 2008) and the

Kanlaon volcano in the Philippines (Neri et al. 2013) to analyse the risk of eruptions and the consequential hazards they may trigger.

Public health risk management techniques and countermeasures

The next stage of the process considers the options available to prevent or control the risks. It is likely that more than one possible risk management strategy will be available.

For environmental, chemical, and radiation incidents, countermeasures fall into two categories, preventative countermeasures to avoid or reduce people's exposure (sheltering, evacuation, and decontamination) and those used to treat casualties once exposure has occurred (antidotes and medical treatment). It is important to stress that lifesaving treatment should always take priority. Health plans should include arrangements for ambulance transport of casualties, be they contaminated by chemicals or radioactive material with serious injuries, without exposing ambulance crews to significant risk.

A combination of countermeasures can be used depending on the type, scale, and severity of the incident. This should be continuously re-evaluated as the incident evolves and more information is collated to inform the risk assessment. For all of this to happen effectively, multiagency plans need to be developed and exercised so that organizations understand their roles, how they will communicate with each other and members of the public, and the pros and cons of specific countermeasures. The plans should allow responders to scale the response up or down depending on the severity of the release and the outcome of the risk assessment.

For high-risk fixed sites industrial and nuclear installations, plans usually include local sheltering and evacuation strategies based on maps dividing a geographical area into zones, taking into account local infrastructure (e.g. availability of local transport systems, location of rest centres) and effective warning and informing of residents within a certain emergency zone (e.g. using sirens).

For all risk management techniques and countermeasures, vulnerable sections of the community need to be identified early as they may require assistance (e.g. people with special educational needs/disabilities, requiring

medical support, the elderly, the frail, and children) or not be able to understand the messages. Some population groups may be particularly sensitive extreme temperatures or to the substance released; for example, consideration should be given to minimize the exposure of pregnant women in a radiation incident.

Sheltering, shelter in place

Shelter in place (also called shelter *in situ*) is invoked when the safest place to take refuge or cover from an actual or perceived danger is indoors (e.g. in a house, business place, or school). Most buildings can provide significant protection against short-lived environmental, chemical, and radiation releases. In the United Kingdom, the message used by the emergency services, usually through the local media, is ‘go in, stay in, and tune in’ and was developed by an independent National Steering Committee on Warning and Informing the Public (HM Government 2004).

Vehicles, boats, and caravans offer a lower level of protection than residential building, so sheltering in these conditions should be a last resort and for short periods. Closing windows, air conditioning, and vents will help reduce airborne exposures.

Evacuation

Evacuation can range from small-scale, locally managed and coordinated evacuation (one or two streets) to mass (or wide-area) evacuations requiring national coordination (e.g. for wide-area flooding) (HM Government 2014). Evacuation should be conducted in a controlled, systematic way to minimize potential exposure and other risks. Mass evacuation is a complex undertaking as transport, support for vulnerable groups, and the availability of shelter in a designated local authority rest centre (e.g. a school or leisure centre) will be necessary. Longer-term psychological impacts of unscheduled evacuations among vulnerable populations must not be underestimated.

Evacuation may be preferred for longer-lived incidents, especially when it can be implemented before people are exposed (e.g. before a release reaches a community), or before an incident happens.

The fire and rescue services will activate pre-existing plans for evacuation of members of the public and non-essential personnel to a safe distance for defined scenarios, for example, where there is an explosive risk.

Decontamination

Decontamination usually refers to external decontamination of people to reduce further exposure or avoid secondary contamination. As an example of how decontamination can work see Box 11.12.4

Box 11.12.4 Decontamination planning in England and Wales

The management of public mass decontamination is the joint responsibility of the fire and rescue services and the Department of Health. The fire and rescue services are equipped with mobile mass decontamination units including disrobe and re-robe packs. Following the terrorist attacks in the United States on 11 September 2001, new equipment and training procedures, particularly focusing on mass decontamination in public places in the event of a chemical, biological, radiological, or nuclear (CBRN) attack, have been provided (Health Protection Agency 2009).

The UK Department of Health has developed Hazardous Area Response Teams (HART) within ambulance services (NARU 2018). These resources enable specially trained and equipped ambulance professionals to be able to care for patients in the event of major or catastrophic incidents where previously it may not have been safe for them to do so. This includes a CBRN capability, to be able to provide interim decontamination as well as full wet decontamination of casualties.

Where at risk contaminated casualties should be evacuated to a safe environment and individual or mass decontamination should be carried out as soon as possible, preferably within 15 minutes of exposure for maximum benefit. Casualties are usually asked to remove their clothes from head to foot to reduce the risk of inhalation of any contaminant. Recent research in decontamination techniques shows that, in the initial phase of an incident involving non-caustic chemicals, the blotting and rubbing of exposed skin surfaces with dry absorbent material, that is, dry decontamination, is most effective unless the use of water is recommended by expert advisors.

The water decontamination method is the rinse–wipe–rinse method and at its most simple may be just washing hands after any contact with flood water. After removal of chemically or radioactively contaminated clothing, casualties are asked to drench themselves under a lukewarm shower to remove water soluble material and dusts (rinse), apply a detergent solution

using a soft brush to remove organic matter (wipe), and drench again to remove the detergent (rinse).

Non-ambulant casualties and those with open wounds should be decontaminated by trained personnel.

Research continues into decontamination techniques and there are a number of ongoing studies comparing the merits of wet and dry decontamination as well as decontamination of other aspects such as hair, latest guidance and advice should always be sought in this area.

Non-invasive screening of the public for radioactive contamination can be used to triage large numbers of people and assess the need for medication intervention, decontamination, and/or further contamination monitoring. Depending on the radionuclides, walk-through portals and hand-held equipment can be used to detect internal or external contamination. Portable or transportable whole-body monitors should be used to detect the presence of internal contamination and for dose assessment following decontamination.

Antidotes

Treatment for chemical and radiation exposure is mostly symptomatic and supportive. Very few antidotes for individual toxins exist and where they do, they should be administered under medical supervision.

For nuclear reactor accidents where iodine usually forms a significant part of the release, the administration of stable iodine before or in the few hours after exposure, in conjunction with sheltering can form a very effective countermeasure strategy.

Depending on the chemical, antidotes can be used to form an inert complex with the poison, speed up detoxification, reduce the rate of conversion to a more toxic compound, and so on. For example, chelating agents such as dicobalt edetate are recommended to treat moderate to severe cyanide poisoning in adults and 2,3 dimercaptosuccinic acid (DMSA) has been recommended for lead poisoning. Prussian blue (ferric hexacyanoferrate) is used to increase the faecal excretion of caesium and thallium from the body, thus reducing the internal dose. Advice can be obtained in many countries on how to manage chemical and radiation exposure via local agencies such as poisons control centres.

Recovery

Recovery is the process of restoring and rebuilding the community in the aftermath of an incident; this can take years or decades for the most serious events. The recovery from an incident needs to be considered from the outset and most successful recovery strategies begin during the emergency phase of the incident by addressing the environmental, health, infrastructure, and economic impact of the incident. Recovery roles and responsibilities should be planned and rehearsed in a multiagency setting. The appropriate local level of government will usually lead the recovery phase of the incident; in the UK comprehensive recovery handbooks containing guidance have been developed to support evidence-based decision-making (Public Health England 2013, 2015a, 2015b).

Follow-up of affected populations

There are two main reasons to undertake follow-up of populations affected by chemical, radiological, or environmental incidents: part of the long-term response to help and provide services for affected people, and to gather scientific information about the health effects of the exposures that they may have received. In most cases, the decision to set up a follow-up study is done for elements of both reasons. A case study is outlined in Box 11.12.5.

The former may be particularly useful if there is to be ongoing treatment of or compensation for affected people. Compensation usually takes many years to organize, by which time the affected people may have moved away from the local area or died.

There are two major problems with emergency response follow-up studies that distinguish them from other epidemiological research. Firstly, the decision to set up the study may need to be taken very early on in the incident, in order to capture, as fully as possible, the cohort of participants for the study. Secondly, it may be more difficult than in a planned study to determine each individual’s exposure to the noxious agent.

It is not straightforward to decide if a follow-up study is needed in each incident. Done properly, the work will be resource intensive in all but the smallest of incidents. If the cohort does not accurately reflect all those who may have been harmed, then its scientific power and usefulness as a tool to

help patient care may be limited in the long term. Paranthaman et al. (2012) have proposed a decision framework to help responders decide if a follow-up health register study is indicated based on a Delphi consensus study of emergency responders and epidemiologists. The study presented participants with four scenarios and asked responders whether a follow-up study should be set up and why. Although there was a spectrum of views from the range of responders, 11 factors were defined by the participants as important in making the decision about setting up a health register. The factors are not mutually exclusive and their relevance needs to be debated by the team setting up the follow-up study.

Measuring individual or group ‘doses’ to noxious agents may be difficult in emergency situations as resources are rightly focused on more urgent aspects of the response. For some agents, for example intakes of radionuclides, there are accurate biodosimetry methods that can be used in the weeks and months following an incident to retrospectively assess exposure. However, for other agents there is no measurable signal after the event, or the signal decays very quickly. This poses real logistical problems if large numbers of people need to be measured and may well swamp national capability to perform such measurements. Triage arrangements may be needed to ensure that people with the highest clinical need for testing have priority.

In other situations, environmental measurements, combined with information about where people were at the time of the incident, what they were doing, and what injuries they may have suffered allow post hoc assessment of likely ‘doses’. These can be calibrated by actual measurements taken in smaller numbers of people with similar locations and behaviours at the time of the release.

Some of the work to set up a follow-up study is needed in the acute phase of the incident. Therefore emergency plans need to specify how the decision to set up a follow-up study will be taken and plan for the resource and practical implications of this work. Exercising these plans will help to spot rate-limiting steps and identify actions that can be done in advance, for example, detailed understanding of how to collect patient data from hospitals, community services, and rest centres, developing generic data collection, and patient information forms that can be adapted when an incident occurs.

Box 11.12.5 Understanding the impact of flooding and associated disruption on psychological morbidity

After widespread flooding in England in 2013/14, a multiyear National Study of Flooding and Health was established to examine the long-term impact of flooding on the mental health of people living in flood-affected areas.

There are methodological complexities in measuring the mental health impacts of flooding so collecting data on a range of personal factors was essential. Twelve months after flooding, the prevalence of probable psychological morbidity was found to be elevated among both flooded participants (depression 20.1%, anxiety 28.3%, post-traumatic stress disorder (PTSD) 36.2%) and among those who were disrupted but had no floodwater entering their homes (depression 9.6%, anxiety 10.7% PTSD 15.2%) (Waite et al. 2017).

Flooded participants who reported disruption to domestic utilities (such as electricity, gas, or water) or to healthcare had higher odds of all mental disorders than did other flooded participants, as did those displaced from their homes. For example, after adjusting for the depth and duration of floodwater in the home, the odds of probable depression were 1.7 times higher for participants who were displaced compared with those who were not (Munro et al. 2017). The amount of warning received appeared to be a protective factor among those who were displaced: those receiving no warning before flooding tended to report more symptoms of depression and PTSD than those who were forewarned.

Public reassurance in emergencies and incidents

The scientific basis of accurate public risk communication is mature. There is also a considerable literature on the need to build trust in the systems that will be used in an emergency to protect casualties, the public, and emergency responders.

Environmental, radiation and chemical exposures can provoke all of the ‘fright factors’ that are known to be most feared by the public (Slovic 1987; Meara 2002). WHO guidance exists to help risk communicators in emergency situations (WHO 2018a). A case study is presented in Box 11.12.6.

Regulators and organizations with responsibilities for health protection have been criticized for taking a low media profile during emergencies (House of Commons Science and Technology Committee 2012). Scientific organizations can find it difficult to communicate in high-pressure situations where the facts are not known and the uncertainties about what may occur are extensive.

Box 11.12.6 Case study: public risk communication during the polonium contamination in London 2006

In 2006, following enquiries relating to an unexpected death, radioactive polonium was found to have contaminated many public places in London (Rubin et al. 2007; Morgan et al. 2008). Detailed investigation eventually revealed that the public health risks were low, but this was not known at the start of the incident. The investigation needed clear objectives and a prioritized approach to make best use of the environmental and people-monitoring facilities that were available.

During the emergency response, researchers carried out a random telephone survey of 1,000 Londoners and more than 100 potentially exposed people. The survey showed that health protection communications, including media interviews and a daily press release, were generally successful as only 11.7% of those who replied perceived their health to be at risk because of the incident. The number of visitors to London did not decrease.

Daily health protection press releases were considered important to keep people informed. Seventy-one per cent (71%) of responders understood that they could not be contaminated if they had not been in the relevant venues. Fifty-eight per cent (58%) of responders understood that polonium is only dangerous if it enters the body.

The authors suggested that the factors contributing to this effective public communication were:

- The incident was linked to espionage rather than terrorism in the public consciousness. People who thought the motive was terrorism (15% of the sample) were significantly more likely to feel that their health was at risk than the people who attributed the incident to espionage, criminal activity, or spying.
- The public perception was that one person had been targeted. People who thought the incident was intended to harm the general public (14% of the sample) were significantly more likely to feel that their health was at risk than those who thought that it was intended to harm one person only.

Despite these positive outcomes, people who had been in contaminated areas and had been offered urine testing tended not to be happy with the information they received. Some of this displeasure was inevitable because they wanted details related to the criminal investigation that could not be made public. Other people wanted more numerical information about their test results. Uniquely for polonium in urine testing, this was difficult to provide because the same amount of urinary excretion can be associated with a radiation dose that varies by a factor of ten, depending on the assumptions made. For the lowest category of risk, there were insufficient resources to undertake detailed dose assessments, using individual-specific assumptions so people were only told that their results were ‘of no concern’. These complexities are not easy to put across in the midst of an emergency incident.

A small number of people (seven out of 3,837) who contacted the telephone helpline that was set up for the incident failed to be reassured by direct discussion with specialist health protection staff, even if they had not been at any of the contaminated venues (Morgan et al. 2008). Of these, four were contacted by a psychiatrist who found evidence of pre-existing psychological stressors and a history of health-related anxiety. The authors concluded that access to psychiatric help may be useful in incidents such as the polonium contamination in London.

Identify and publish lessons

Identifying lessons from both incidents and exercises is an integral part of the emergency preparedness and response cycle, in order that learning can take place. This requires a structured debrief process to capture learning at various stages of emergency response and recovery. The WHO After Action Review process (WHO 2018b) is a good example of a structured approach to identifying potential learning which is integrated with the response and recovery stages of an emergency. Implementing learning, altering plans, and disseminating the lessons are key responsibilities of responding organizations. Understandably, repetition of lessons identified in previous incidents or exercises, in subsequent emergencies, is increasingly seen as avoidable and the learning component of the planning and response cycle is as critical as the response itself.

This means that plans should be updated where necessary following incidents or exercises and learning should also be disseminated. Reporting incidents is vital to share learning, decrease the chance of the incident reoccurring, and to improve emergency response. However, many incidents and emergencies are very poorly documented. Incident reports may also be useful for members of the public who have been placed at risk, colleagues who are responding to similar incidents, and public health professionals who are called to give evidence in public or legislative inquiries into the incident. Only with detailed documentation can lessons be identified that can then change policy and practice.

A process to ensure consensus on what information will be gathered and how data collection will be organized during and after an incident or emergency should be established as part of the pre-incident emergency planning. Collation of data for reporting should start as soon as possible during the crisis event with plans for continuous updating as more data become available from the multidisciplinary (or potentially international) response.

Organizations, including public health partners, performing incident or emergency responses should jointly develop and document how the quality and validity of information are assured. This needs documented procedures and training for incident responders, including administrative staff, in the methods for information gathering, analysis, and quality assurance.

Kulling et al. (2011) set out a summary of the requirements for reports on health crises and critical health events and recommended issues that should be managed actively at an early stage in a response. These include:

The mandate for the emergency report and that of the respective authority/organization.
Early identification of the authors and their responsibilities.
Ethical approval and quality assurance statements for the report if required.
A summary of the pre-event status if needed, such as baseline disease patterns and user patterns of hospitals and health facilities as well as other relevant previous incidents or disasters described as well as preparedness plans.
A summary of the incident or emergency.
A method statement to include descriptions of hazard and exposure assessments, decision flow, and information management.
A description of the interventions and how they were evaluated in terms of effects, outcomes, and impacts balanced against the goal(s) and objectives of the interventions. It may be useful to distinguish between immediate-, medium-, and long-term impacts.
Coordination and collaboration between organizations from health and other sectors.
Adverse health effects in the responders.
Identify interventions that restore normality to the pre-event level, including how responding organizations intervened to provide resilience and reduce vulnerability in people at risk.
Consider summarizing the lessons identified by describing:

- What went well.
- What went badly.
- What could be done better.

The final task is to publish the report in an appropriate peer-review publication for all to share and learn from and how these can be used to reduce further impacts (Box 11.12.7).

Box 11.12.7 Chemical incident lessons and the Seveso Directives

Since the onset of the industrial revolution, incidents involving chemicals and industry have led to significant loss of life worldwide. Many incidents have occurred and publication of the lessons and learning has resulted in clearer and more implementable legislation (Rees and Waite 2015).

An accident at Flixborough, United Kingdom in 1974, in which a vapour cloud explosion killed 26 people and injured many more almost triggered a ‘domino-effect’ incident at a neighbouring site. In

1976 at Seveso, Italy a chemical plant producing pesticides and herbicides released a large cloud of dioxins. While no fatalities were immediately reported at Seveso, dioxins are known to be lethal in man, even in very small doses. A large area of land was contaminated, 600 people had to leave their homes and up to 2000 were treated for dioxin poisoning. The AZF factory in Toulouse, 2001, had a catastrophic explosion, annihilating the entire compound and instantly killing 29 people. The official investigation indicated that the explosion was the result of mislabelling, which led to compounds reacting that should never have been mixed. Research indicated that better management and improved land use surrounding the plant would have limited the extent of the disaster substantially. In 2005 the Buncefield explosion and fire at an oil storage depot in the United Kingdom was unprecedented at the time, and was the result of an unrecognized explosive mechanism where a spill allowed the formation of a vapour cloud that was confined in such a way by the plant layout. This exploded in a characteristic way; an Open Flammable Cloud Explosion. Within the European Commission the Seveso Directive which addresses Technological Disaster Risk Reduction led to changes in policies included inspections (European Commission 2017). Checks were made to ensure that the policies were being thoroughly implemented, and to broaden and tighten legislation around storage sites to help prevent future unprecedented disasters by effective and flexible planning involving both private and public entities. As these incidents were documented action was taken to introduce a subsequent series of Seveso Directives by the European Commission leading to a significant reduction in the number of chemical accidents in the region, saving potential costs in lives, environment, and finance.

Data from Rees HC and Waite TD. 2015 The Seveso Directives and the UK Control of Major Accident Hazard (COMAH) regulations UNISDR Scientific and Technical Advisory Group Case Studies—2015 Available https://www.preventionweb.net/files/workspace/7935_comahcasestudyera.pdf.

Conclusion

This chapter on the principles of public health emergency response for acute environmental, chemical, and radiation incidents summarizes the stages of response to emergencies using a generic ‘all hazards’ framework and a ‘source–pathway–receptor’ risk assessment model which is adaptable to any situation and conforms with international obligations.

In recent years there have been many environmental, chemical, and radiation disasters that have challenged the emergency preparedness of countries around the globe. Adequate resources for response and pre-planning are vital to deliver a professional, and proportionate response and to define the important factors to consider when establishing follow-up studies after a major incident. The knowledge gained by careful evaluation of emergency actions can greatly improve the efficiency of emergency preparedness and safety of subsequent responses. However, there are great challenges in doing research under such circumstances. A significant number of evidence gaps remain and so detailed documentation and follow-up of

future incidents will continue to be needed to fill those gaps and improve emergency preparedness and response in the future.

ACKNOWLEDGEMENT

With thanks to both Dr Naima Bradley and Dr Jill Meara, Public Health England.

FURTHER READING

United Nations International Strategy for Disaster Reduction (UNISDR) (2017). *National Disaster Risk Assessment Words into Action Guidelines Governance System, Methodologies, and Use of Results 2017. Consultative Version*. [Online] Available at: https://www.unisdr.org/we/inform/publications/52828

United Nations International Strategy for Disaster Reduction (UNISDR) (2015). *Sendai Framework for Disaster Risk Reduction 2015–2030*. [Online] Available at: https://www.unisdr.org/we/coordinate/sendai-framework

World Health Organization (WHO) (2008). *International Health Regulations (2005)* (2nd ed.). Geneva: WHO. [Online] Available at: http://www.who.int/ihr/9789241596664/en/

World Health Organization (WHO) (2017). *Emergency Response Framework* (2nd ed.). Geneva: WHO. [Online] Available at: http://www.who.int/hac/about/erf/en/

World Health Organization (WHO) (2018). *After Action Review.* [Online] Available at: https://extranet.who.int/sph/after-action-review

REFERENCES

Cabinet Office (2011). Emergency planning. In: *Emergency Preparedness, Chapter 5, pp. 16–17.* London: Cabinet Office. [Online] Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/61028/Emergency_Preparedness_chapter5_amends_21112011.pdf

Cabinet Office (2017). *UK National Risk Register of Civil Emergencies*. London: Cabinet Office. [Online] Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/644968/UK_National_Risk_Register_2017.pdf

Cardona, O.D., et al. (2012). Determinants of risk: exposure and vulnerability. In: Field, C.B., et al. (eds.) *Managing the Risks of Extreme Events and Disasters to Advanced Climate Change Adaptation: A Special Report of Working Groups I and II of the Intergovernmental Panel on Climate Change (IPCC)*. Cambridge, UK, and New York, NY: Cambridge University Press, pp. 65–108.

European Commission (2017) Major accident hazards. The Seveso Directive—Technological Disaster Risk Reduction. [Online] Available at: http://ec.europa.eu/environment/seveso

Gill, J.C. and Malamud, B.D. (2016) Hazard interactions and interaction networks (cascades) within multi-hazard methodologies. *Earth System Dynamics*, **7**, 259

Gormley, A., Pollard, S., and Rocks, S. (2011). *Guidelines for Environmental Risk Assessment and Management. Green Leaves III.* [Online] Available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/69450/pb13670-green-leaves-iii-1111071.pdf

Health Protection Agency (2009). *Centre for Radiation, Chemical and Environmental Hazards Chemical Surveillance Report 1st January–31st December 2009*. [Online] Available at: http://webarchive.nationalarchives.gov.uk/20140714100036/http://www.hpa.org.uk/webc/HPAwebFile/HPAweb_C/1284475648621

Hemingway, R. and Gunawan, O. (2018) The Natural Hazards Partnership: a public-sector collaboration across the UK for natural hazard disaster risk reduction. *International Journal of Disaster Risk Reduction*, 27(2018), 499–511.

HM Government (2004). *Preparing for Emergencies: What You Need to Know*. London: HMSO. [Online] Available at: http://webarchive.nationalarchives.gov.uk/20121005013740/http://www.direct.gov.uk/prod_consum_dg/groups/dg_digitalassets/@dg/@en/documents/digitalasset/dg_176618.pdf

HM Government (2014). *Evacuation and Shelter Guidance. Non-Statutory Guidance to Complement Emergency Preparedness and Emergency Response and Recovery*. London: HMSO. Available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/274615/Evacuation_and_Shelter_Guidance_2014.pdf

House of Commons Science and Technology Committee (2012). *Devil's Bargain? Energy Risks and the Public First Report of Session 2012–13 HC 428 [Incorporating HC 1742–i to –iv, Session 2010–12]*. London: HMSO. Available at: http://www.publications.parliament.uk/pa/cm201213/cmselect/cmsctech/428/42802.htm

Kulling, P., Birnbaum, M., Murray, V., and Rockenschaub, G. (2011). (A48) Guidelines for reports on health crises and critical health events—a tool for the development of disaster medicine research. *Prehospital and Disaster Medicine*, 26, s17.

Meara, J.R. (2002). Getting the message across: is communicating risk to the public worth it? *Journal of Radiological Protection*, **22**, 79–85.

Morgan, O.W., Page, L., Forrester, S., and Maguire, H. (2008). Polonium-210 poisoning in London: hypochondriasis and public health. *Prehospital and Disaster Medicine*, 23(1), 96–7.

Munro, A., Armstrong, B., Waite, T.D., Bone, A., Rubin, G.J., and Kovats, S. (2017). Do evacuation and displacement affect the association between flooding and mental health outcomes? A cross-sectional analysis of English Survey Data. *Lancet Planetary Health*, 1(4), e134–41.

National Ambulance Resilience Unit (NARU). Hazardous Area Response Teams. [Online] Available at: https://naru.org.uk/naru-work-activities/hazardous-area-response-teams/

Neri, M., et al. (2008). Developing an event tree for probabilistic hazard and risk assessment at Vesuvius. *Journal of Volcanology and Geothermal Research*, 178(3), 397–415.

Neri, M., Le Cozannet, G., Thierry, P., Bignami, C., and Ruch, J. (2013). A method for multi-hazard mapping in poorly known volcanic areas: an example from Kanlaon (Philippines). *Natural Hazards and Earth System Sciences*, **13**, 1929–2013.

NHS England (2015). *NHS England Emergency Preparedness Resilience and Response Framework*. Leeds: NHS Commissioning Board. [Online] Available at: https://www.england.nhs.uk/wp-content/uploads/2015/11/eprr-framework.pdf

Paranthaman, K., Catchpole, M., Simpson, J., Morris, J., Muirhead, C.R., and Leonardi, G.S. (2012). Development of a decision framework for establishing a health register following a major incident. *Prehospital and Disaster Medicine*, 27(6), 524–30.

Public Health England (2013). *UK Recovery Handbook for Chemical Incidents*. [Online] Available at: https://www.gov.uk/government/publications/uk-recovery-handbook-for-chemical-incidents-and-associated-publications

Public Health England (2015a). *UK Recovery Handbooks for Radiation Incidents 2015*. [Online] Available at: https://www.gov.uk/government/publications/uk-recovery-handbooks-for-radiation-incidents-2015

Public Health England (2015b). *UK Recovery Handbook for Biological Incidents*. [Online] Available at: https://www.gov.uk/government/publications/uk-recovery-handbook-for-biological-incidents

Rees, H.C. and Waite, T.D. (2015). *The Seveso Directives and the UK Control of Major Accident Hazard (COMAH) regulations UNISDR Scientific and Technical Advisory Group Case Studies—2015*. [Online] Available at: https://www.preventionweb.net/files/workspace/7935_comahcasestudyera.pdf

Rubin, G.J., et al. (2007). Public information needs after the poisoning of Alexander Litvinenko with polonium-210 in London: cross sectional telephone survey and qualitative analysis. *British Medical Journal*, 335(7630), 1143.

Slovic, P. (1987). Perception of risk. *Science*, **236**, 280–5.

UK HM Government (2013). *Emergency Response and Recovery: Non Statutory Guidance Accompanying the Civil Contingencies Act 2004 Emergency Response and Recovery. Revised version October 2013*. [Online] Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/253488/Emergency_Response_and_Recovery_5th_edition_October_2013.pdf

UNISDR (2017). *National Disaster Risk Assessment Words into Action Guidelines Governance System, Methodologies, and Use of Results 2017*. [Online] Available at: https://www.unisdr.org/we/inform/publications/52828

United Nations General Assembly (2015). *Transforming Our World: The 2030 Agenda for Sustainable Development Resolution adopted by the General Assembly on 25 September 2015 A/RES/70/1*. [Online] Available at: http://www.un.org/en/development/desa/population/migration/generalassembly/docs/globalcompact/A_RES_70_1_E.pdf

United Nations Framework Convention on Climate Change (UNFCCC) (2015). *Conference of the Parties Twenty-first session Paris, 30 November to 11 December 2015. Adoption of the Paris Agreement FCCC/CP/2015/L.9/Rev.1* [Online] Available at: https://unfccc.int/resource/docs/2015/cop21/eng/l09r01.pdf

United Nations International Strategy for Disaster Reduction (UNISDR) (2015). Sendai Framework for Disaster Risk Reduction 2015–2030. [Online] Available at: https://www.unisdr.org/we/coordinate/sendai-framework

Waite T, et al. (2017). The National Study of Flooding and Health—establishing the cohort and first year cross-sectional analysis of data from a ten year cohort study of the effects of flooding on mental health (2016) *BMC Public Health*, 17(1), 129.

World Health Organization (WHO) (2008). *International Health Regulations (2005)* (2nd ed.). Geneva: WHO. [Online] Available at: http://www.who.int/ihr/9789241596664/en/

World Health Organization (WHO) (2017). *Emergency Response Framework* (2nd ed.). Geneva: WHO.[Online] Available at: http://www.who.int/hac/about/erf/en/

World Health Organization (WHO) (2018a). *Communicating Risk in Public Health Emergencies: A WHO Guideline for Emergency Risk Communication (ERC) Policy and Practice*. Geneva: WHO. [Online] Available at: http://www.who.int/risk-communication/guidance/download/en/

World Health Organization (WHO) (2018b). *After Action Review.* Geneva: WHO. [Online] Available at: https://extranet.who.int/sph/after-action-review

World Health Organization/Public Health England/United Nations International Strategy for Disaster Reduction (2017). *Health Emergency and Disaster Risk Management Fact Sheets*. Geneva: WHO. [Online] Available at: http://www.who.int/hac/techguidance/preparedness/factsheets/en/

Zschau, J. (2017). Where are we with multihazards, multirisks assessment capacities? In: Polijansek, K., Marin Ferrer, M., De Groeve, T., and Clark, I. (eds.) *Science for Disaster Risk Management: Knowing Better and Losing Less*. Luxembourg: Publications Office of the European Union, pp. 99–108.

11.13

Private support of public health

Quarraisha Abdool Karim and Roger Detels

Introduction to private support of public health

Private support for public good may be understood as the cooperation between private individuals, foundations, and/or the commercial sector and the public sector in order to more effectively meet the needs of society locally and globally (Buse and Waxman 2001). This relationship is not necessarily unidirectional, and in several instances, also benefits the giver through increases in positive product demand from company branding through positive media coverage that accompanies giving (Broadbent and Laughlin 2003). The World Health Organization (WHO) defines public–private partnerships (PPPs) in the health sector as the bringing together of a set of actors for the common goal of improving the health of populations based on mutually agreed roles and principles (Kickbusch and Quick 1998). While these principles are based on the founding ideas and values of human ethics such as beneficence, non-maleficence, autonomy, and equity (Buse and Walt 2000a) the reality is that the relationship between giver and recipient is not an equal one.

Over time, while anonymous individual giving has continued, some of these ad hoc institutional contributions have evolved into more defined and legally recognized structures also referred to as philanthropic organizations and includes giving by more wealthy individuals and families, including the Gates and Zuckerbergs, as well as corporate institutions. Giving, particularly by North Americans, dates back to the formation of ‘foundations’ in the earliest years of the twentieth century as a result of wealth acquired through industrialization. In the United States, this idea was incentivized by a ‘tax break’ provided by the US government to private companies and citizens who donated a certain proportion of their wealth to

federal or state government projects or priority initiatives (Butler 2001). Private entities had the autonomy to choose which government or global projects or priorities they would support, and the extent thereof, while deriving the benefit of reduced taxation.

This type of charitable tax deduction by the US government, has allowed foundations to reduce their tax burden by as much as 25% (Strom 2007). *The New York Times* has estimated that charitable deductions cost the US government US$40 billion in lost tax revenue in 2006 (Strom 2007). In effect, this substantial private philanthropic support means that the foundations, and not the US government, are determining where significant public funds are spent and, thus, are having considerable influence over everything from medical research to social policy and politics. Their giving reflects their personal interests and biases. Some of the larger foundations and philanthropists have larger resources than international agencies such as the World Health Organization and the governments of many smaller countries. Philanthropists and foundations take pride in their autonomy in determining what to invest in and traditionally have identified areas that minimize overlap and define their giving identity. Eli Broad, one of the major donors, summarizes the essence and impetus for private giving by asserting that smart philanthropists get more for their money than governments because they know how to invest wisely (Strom 2007).

However, it is important to recognize that foundations often determine their targets and objectives on personally identified and defined priorities rather than an objective assessment of international and national need. Foundations are likely to invest in areas that governments may be reluctant to for political or moral reasons, for example, sexuality education, gender-based violence, fertility control, and harm reduction programmes. Some, such as the Soros Foundation and the Koch Brothers, even invest in promotion of political ideologies which have significant impacts on social issues and public health.

History of giving

One of the earliest known foundations in the United States was the Russell Sage Foundation that was founded in 1906 with a focus on working women and poverty-related illnesses (Hammack and Wheeler 1994). The

establishment of larger and more encompassing foundations began with the founding of the Carnegie Foundation by Andrew Carnegie in 1911 ‘to do real and permanent good in this world’ (Carnegie Corporation of New York n.d.), followed by The Rockefeller Foundation in 1913 (Menasha and Hsig 1915).

These early gifts had a significant impact on institution creation to support and build public health capacity. Public health provided the optimal mechanism through which Rockefeller philanthropy was able to apply scientific findings for the public good (Birn and Fee 2013). The Rockefeller Foundation established the International Health Commission in its first year (Rockefeller Foundation 2013a). This was the first allocation of funds from private sources for international public health activities and had a wide sphere of influence—52 countries on six continents and 29 islands to improve their public health systems. The Rockefeller Foundation simultaneously initiated a 20-year programme to support the Bureau of Social Hygiene which focused on research and education on fertility control, maternal health, and sex education. This foundation played a key role in founding the world’s first schools of public health, first at Johns Hopkins, then at Harvard, and later in other parts of the United States, as well as in 21 other countries that included establishment of the London School of Tropical Medicine and Hygiene in 1921, the China Medical Board in 1914, and the Peking Union Medical College in 1921. The last two institutions have played a pivotal role in the development of public health in China (Rockefeller Foundation 2013b).

The Wellcome Trust, which was established in 1936 in the United Kingdom, is dedicated to supporting research to improve human and animal health. At the time of writing, it is one of the wealthiest foundations in the world (Wellcome Trust n.d.), following the Gates Foundation, the Chan-Zuckerberg Foundation, and the Dutch Stichting INGKA Foundation (Table 11.13.1).

Table 11.13.1 Wealthiest charitable foundations

Organization	Worth
Chan-Zuckerberg Foundation	$58 billion
Bill and Melinda Gates Foundation	$42.3 billion
Stichting INGKA Foundation	$36.0 billion
Wellcome Trust	$25.9 billion
Howard Hughes Medical Institute	$16.9 billion
Ford Foundation	$11.2 billion
J. Paul Getty Trust	$10.5 billion
Kamehameha Schools	$10.1 billion
Mohammed bin Rashid Al Maktoum Fdn	$10.0 billion
Azim Premji Foundation	$9.8 billion

Source: data from Wikipedia https://en.wikipedia.org/wiki/List_of_wealthiest_organizations

A hallmark of this type of giving was the insightfulness of individuals in anticipating, prioritizing, and filling an important gap and creating a strong foundation for others, including governments, to build on. This type of nimble, risk-taking investment is in keeping with the business model based on entrepreneurship that sees and seizes opportunities. These are the principles on which the foundation wealth was created—unfettered by ponderous, risk-averse decision-making, never-ending procedures, and systems that are the hallmark of most government and public sector bureaucracies.

Current giving

Since those early years of giving, private support for public health has increased in magnitude and scale, particularly in the past decade. Today, the Bill & Melinda Gates Foundation, one of the largest private grant-making foundations in the world at the time of writing, supports major programmes worldwide that include a US-based programme to enhance secondary and

post-secondary education, a global development programme which focuses on the alleviation of hunger and poverty, and a global health programme that includes increasing access to childhood immunization programmes, prevention and research in major infectious diseases such as tuberculosis (TB), malaria, and HIV (McCoy et al. 2009b). This foundation has single-handedly redefined the concept of individual social responsibility and has inspired similar giving trends across the globe, including recruiting wealthy individuals in developing countries in support of public health.

In, 2015, Priscilla Chan and Mark Zuckerberg, aged 31, new billionaires from social media, pledged to spend 99% all of their fortune within their lifetimes to 'cure, prevent, or manage all diseases by the end of the century'. In South Africa the Motsepe Foundation was established in 2013 by Patrice Motsepe, one of the country's wealthiest individuals, and listed on the Fortune 500's top ten wealthiest people, committed 50% of his family wealth to philanthropy and $US10 million over the next five years towards several health, education, and job creation efforts (Morvaridi 2012).

Globally, there are now over 40 major foundations, as well as countless smaller ones (Robert Wood Johnson Foundation 2012; Foundation Center 2013) (Table 11.13.1). This increase can be attributed in part to at least three phenomena: substantial growth in personal wealth among a few, global solidarity around the AIDS epidemic and other health threats, and the rise in citizen pressure for corporate accountability/social responsibility. Corporate social responsibility programmes are a new form of giving that adopt a more entrepreneurial approach for product branding, creating brand loyalty and increasing product sales but are not the focus of this chapter.

Globalization of health threats, global solidarity, and the emergence of public–private partnerships

The global burden of disease, especially infectious diseases, disproportionately affects populations in developing countries, due to poverty, weak healthcare delivery systems, inadequate investment in health, and inadequate access to pharmaceuticals, vaccines, and medications (Widdus 2001). Post-industrial revolution eradication of these diseases in developed countries and the growth of a for-profit pharmaceutical industry

have resulted in neglect of research and development in drugs and other medical products for ‘diseases of poverty’, since such products are perceived as being not likely to generate significant return on their investment.

The rise of multi- and extremely drug-resistant strains of TB (MDR- and XDR-TB), outbreaks of emerging diseases causing a severe acute respiratory syndrome, swine flu, avian flu, etc., and the growing antibiotic resistance of previously controlled infections, have served as sharp reminders of the continuing threat posed by these neglected tropical diseases, as well as emerging and re-emerging infectious diseases, regardless of geographical location, in our increasingly globalized world.

Interventions that reduce costs, provide market incentives, and create infrastructure that will allow for product access (Widdus 2001) require the private sector, such as pharmaceutical companies, to collaborate and work with public and not-for-profit entities to ensure a fast and efficient response (Widdus 2001, 2005). The proliferation of global PPPs (Croft 2005) reflects this approach and is rapidly reconfiguring the global public health landscape (Buse and Walt 2000a).

Other examples of private support

Private support of public health given by foundations, industry, pharmaceutical companies, and professional organizations has been directed towards disease control and treatment, poverty alleviation, education, training of health workers, infrastructure development, research, improved agricultural practices, health information dissemination, and sponsorship of the many global health alliances.

Specific examples include the provision of drugs and the reduction of drug prices for developing countries, the earliest example being the Mectizan Donation Program by Merck & Co. Inc. for treatment for onchocerciasis (river blindness) in West Africa in 1988. In 1998, GlaxoSmithKline provided albendazole for treatment of lymphatic filariasis as part of the Mectizan programme. To date, this ongoing programme has reached more than 80 million people (Mectizan Donation Program 2012).

The development, manufacture, and dissemination of the Sanchol oral cholera vaccine is a notable example of the joining together of foundations,

research institutions, governments, and pharmaceutical companies. Contributors to this PPP included Sweden, Vietnam, Korea, and India, the National Institute for Child Health and Human Development (USA), the Gates Foundation, the Global Vaccine Initiative, the National Institute for Cholera and Enteric Diseases (India), the International Center for Diarrheal Disease Research, Bangladesh, Shantha Biotechnica (India), Eubiologics (Korea)m, and VaBiotech (Vietnam) (Human Vaccines 6:9763–66, September 2010, Landes Bioscience; WHO Weekly Epidemiologic Record no. 13, March 26, 2010).

Another important contribution of private agencies has been research. The Howard Hughes Medical Institute, for example, was established by Hughes Aviation to conduct research into human illness (Howard Hughes Medical Institute n.d.), and has established and funded research centres (Howard Hughes Research Institutions) within major academic medical centres in the United States, as well as in Kwa-Zulu Natal in South Africa. There are also the very significant contributions made by pharmaceutical companies to research and development of vaccines and treatments, both independently and in PPPs.

Activism

The role of activism in advancing the scientific agenda and public access to product advances is most apparent in HIV/AIDS. Since the first reported cases of AIDS in the early 1980s in the United States and the discovery of the causative organism, HIV has spread rapidly across the globe, resulting in several million deaths prior to the discovery of effective antiretroviral therapy (Curran and Jaffe 2011; Sharp and Hahn 2011). The transformation of AIDS from a uniformly lethal disease to a chronic, manageable condition has resulted in rapid and phenomenal declines in mortality rates in industrialized countries, and is having an increasing impact on reducing HIV deaths in developing countries, thanks to the support of the Gates and Clinton Foundations and the support of the US government (President’s Emergency Plan for AIDS Relief (PEPFAR)).

The hosting of the XIIIth International AIDS Conference in Durban, South Africa in 2000 was the first time that many in industrialized countries came face-to-face with the plight of the millions of people infected with

HIV in developing countries, who were continuing to die because they were unable to access antiretroviral treatment and high costs were posing a major barrier. While the inequities and disparities in terms of access to life-saving drugs and healthcare between industrialized and developing countries are not new, this conference was a sharp reminder, and catalysed a whole new approach to public giving in healthcare provision and novel partnerships.

The creation of the Global Fund for AIDS, Tuberculosis and Malaria (GFATM) in 2007 addressed three diseases that, in combination, are major contributors to high rates of morbidity and mortality in resource-constrained settings, and which contribute disproportionately to the global burden of disease (Vitoria et al. 2009). Funding from the Bill & Melinda Gates Foundation stimulated additional large donations by Warren Buffett and the William Clinton Foundation.

The GFATM model builds on the emergence of formalized PPPs to address specific public health challenges, including product development stimulated by The Rockefeller Foundation in the late 1990s with their investment in HIV vaccine development through support for the International AIDS Vaccine Initiative (IAVI) at a time when most had given up on a vaccine. IAVI energized the quest for an HIV vaccine through its intramural programme, expanded the numbers of academic institutions engaged in this quest, and encouraged some pharmaceutical companies to invest in vaccine development.

This model was used subsequently by The Rockefeller Foundation to leverage and harmonize resources to increase access to drugs to reduce mother-to-child transmission (MTCT) of HIV through the MTCT+ initiative coordinated by the Elizabeth Glaser Foundation, and access to antiretroviral treatment for AIDS patients in resource-constrained settings (Barr 2007). The disease-and/or product-specific PPP approach builds on the Bayh-Dole Act that was passed by the US Congress in 1980 (Barr 2007) to promote collaboration in the United States between the federal government, universities, and private pharmaceutical companies for development and marketing of new pharmaceutical products.

Public–private partnerships (PPPs)

PPPs may be classified or described according to purpose, organizational structure, scope of activity, or approach. Three categories are described as follows (Buse and Walt 2000b):

Product-based partnerships that enable access: this includes drug or product donation programmes; bulk purchase of products for low-income countries, for example, the albendazole donation programme; and childhood immunization programme vaccine purchases.
Product-development partnerships that enable and incentivize partnerships for discovery and/or commercialization, for example, the Malaria Vaccine Initiative (MVI) established in 1999, the IAVI described earlier, the International Partnership for Microbicides, the development of the Sanchol cholera vaccine and the Tuberculosis Vaccine Initiative.
Systems-based partnerships that support government efforts to strengthen health service delivery through health systems strengthening.

A PPP can involve financing from various sources such as equity contributions, debt contributions, bank guarantees/letters of credit/performance guarantees, bond/capital market financing, mezzanine/subordinated contributions, and intercreditor agreements in some combination of equity and debt (World Bank Group 2011). Sources may include governments, private foundations, individuals, and the business/corporate sector (McCoy et al. 2009a). The wealthiest non-profit charitable foundations globally are the Bill & Melinda Gates Foundation (Vogel 2006), the Chan-Zuckerberg Initiative, the Stichting INGKA Foundation, the Wellcome Trust, and the Howard Hughes Medical Institute (Table 11.13.1). While in previous years giving was largely unregulated, in the light of the post-2009 global financial and economic crisis, much has changed. Countries such as the United States have called for more strict regulations regarding tax credits for charitable giving by private companies. This would cover federal taxes, but the individual states have their own regulations regarding state-controlled taxes. The government of China regulates which non-governmental organizations can operate in their country. The increase in giving to resource-constrained settings by wealthier governments and private donors has elevated health to diplomatic status, wherein good health is viewed as a basic human right and essential for economic development and social stability at a country level. This goal

of ensuring good governance and sustainable societies is facilitated through bi- and multilateral partnerships between those who have and those who do not through the establishment of a global civil society (Vogel 2006) that redefines the meaning of globalization.

Since the inception of PPPs in the United Kingdom (Spackman 2002), over £56 billion has been contributed to over 800 projects, including car parks, schools, tolled highways, and power plants (Garvin and Bosso 2008). Globally, much attention has been given to healthcare, education, social housing (Flinders 2005), and disaster control. Healthcare in particular remains one of the major recipients of PPP funding (Howard 1983, 1991; Novignon et al. 2012). PPP initiatives in health include infectious disease control (Ardian et al. 2007; Kane et al. 2010; Sinanovic and Kumaranayake 2010; Lal et al. 2011; Sturchio and Cohen 2012), drug and vaccine development (Wheeler and Berkley 2001; Nwaka and Ridley 2003; Nwaka 2005; Sanhai et al. 2008), management of hospitals (Goldman et al. 1992; DeVol et al. 2007; Lester 2012), and building of hospitals (Weiner and Alexander 1998; McKee et al. 2006) and institutes of health.

PPPs can function at a global or country level and imply a mutually beneficial, synergistic alliance between the private and public sectors towards a common goal that is shaped by good governance, equity, and efficiency (Goldman et al. 1992). While the benefits of PPPs include mobilization of funds, technology transfer, strengthened delivery systems (Kraak et al. 2012), greater efficiency in terms of time, scope, and improved options, there are also risks involved. PPPs are not the solution to all health challenges, and may not be the most effective or efficient approach in several instances (UK Parliament 2011). For example, the UK Parliament's Treasury Select Committee's review of the PFI scheme in August 2011 concluded that: 'Private Finance Initiative (PFI) funding for new infrastructures such as schools and hospitals does not provide taxpayers with good value for money, and calls for stricter criteria to govern its use'. Careful thought and planning need to be given to evaluate the appropriateness of external funding in each case to ensure it is not counterproductive.

In the health sector, a significant PPP exists between the WHO and the corporate sector (Buse and Waxman 2001). This partnership builds on the United Nations (UN) idea of a 'Global Compact' (Williams 2008) that encourages corporate responsibility in the areas of labour, human rights,

and the environment in response to globalization (Buse and Waxman 2001). What is distinctive in this partnership is the commitment to health and the principles inculcated in its mission and accountability to its constituency, namely its member states (Buse and Waxman 2001). The corporate partnership with WHO has been met with mixed reactions. Some board members have expressed concerns that commercial interests would mean compromising the integrity of the WHO in terms of subordination of its values, re-orientation of its mission, weakened capacity to promote and monitor international regulations, displacement of priorities, and self-censorship (Buse and Waxman 2001). A number of examples cited for this threat of 'institutional capture' include: (1) guidelines for management of hypertension that were unduly influenced by a pharmaceutical company that had stood to benefit from the drugs prescribed, and (2) deliberations on breastfeeding that were considered to have been censored in favour of the WHO's new set of commercial constituencies (Buse and Waxman 2001).

At a country level, delays in procurement processes or changing needs of communities (Fussel and Beresford 2009) can reduce the efficiency and nimbleness of private partners (Hodge 2004; Fussel and Beresford 2009), and rigid commitments can undermine existing government policies and infrastructure, resulting in coexisting government projects being jeopardized and government priorities and agendas being undermined in developing countries.

There were more than 90 international partnerships in health in 2004 (Nishtar 2004). The majority addressed infectious diseases, notably AIDS, TB, and malaria. Some involved individual governments, non-governmental organizations, or both, forming an alliance with the for-profit sector. For example, Sanofi-Aventis has an extensive international programme of humanitarian sponsorship, including a partnership with the Nelson Mandela Foundation and the South African Department of Health to provide free TB treatment in severely affected areas within South Africa (Sanofi-Aventis 2007).

Other notable partnerships are the global health alliances, such as the Global Alliance for Vaccines and Immunizations, Roll Back Malaria, the Stop TB Partnership, and the Global Polio Eradication Program. They may be principally driven by a company, be legally independent, or be hosted by a civil society organization (Nishtar 2004). The Stop TB Partnership, for example, comprises a network of more than 500 partners, including

governments of both developed and developing countries, research and technical health institutes, multilateral organizations, civil society, pharmaceutical companies and other industry partners, and is housed within the WHO (Stop TB Partnership n.d.).

Prioritization of need: competing candidates

One of the major issues facing PPPs is the correct distribution of funds based on the degree of perceived need. Depending on the period in question, as well as the global climate in terms of crisis management, funds are generally distributed according to urgency and magnitude of need. In the healthcare sector, much attention has been given to the global HIV/AIDS pandemic, but recent literature cites a need for a more balanced and better-rounded approach (Bygbjerg 2012). Furthermore, over and above the current agenda of infectious diseases, particularly in low-and middle-income countries, there are issues of economic development, industrialization, urbanization, investment, and ageing, which in turn drive the rising prevalence of non-communicable diseases (NCDs), the majority of which now occur in developing countries (Bygbjerg 2012). Malnutrition, including undernutrition, increased obesity associated with increased economic status, and infection in early life tend to increase the incidence of NCDs in later life (Bygbjerg 2012). In addition, the presence of infectious diseases and NCDs together (Boutayeb 2006) can result in adverse interactions (Bygbjerg 2012). Therefore, attempts to fund each health initiative separately are likely to exacerbate the overall situation and distort the countries' health priorities (Bygbjerg 2012). Joint approaches that target improving the healthcare system as a whole are preferable (Bygbjerg 2012). Towards this end, the United Nations identified 17 sustainable development goals in 2015.

Challenges

While the concept of PPPs might be conceptually appealing, they are associated with many challenges (Hodge and Greve 2005; Asante and Zwi 2007) such as prioritization, sustainability; collaboration between donors,

governments, health departments, and civil society; corruption; and numerous ethical and moral dilemmas. Other existing challenges include balancing private commercial interests with public health interests, managing conflicts of interest, assessing partnership compatibility, and evaluating partnership outcomes (Kraak et al. 2012). In addition, money borrowed from the private sector is often more expensive than money borrowed by governments, and these costs are passed along to the public, although the private sector would assert that they can save money in the end because they can operate with greater efficiency. Others argue that the initial capital outlay required may counter the efficiencies gained. Moreover, these expenditures need to be translated into the correct and most effective set of approaches, strategies, and investments for improving the health of the poor (McCoy et al. 2009a). There are also issues of transparency and government oversight (Fussel and Beresford 2009). Commercial considerations may limit public oversight during procurement (Fussel and Beresford 2009). Contract provisions may limit changes that are necessary later on in the project in order to safeguard public interests as new public policies and challenges evolve (Fussel and Beresford 2009).

Private investments and global health—at a crossroads?

Private giving in all its forms has enormously benefited individuals, institutions in the United States and outside, the recipient countries, and global health has achieved goals which otherwise would have been impossible to attain and importantly within a relatively short timeline. There is no question that the recent surge in giving has reduced health disparities and changed the global landscape in select diseases, and is at an important crossroads. In contrast to private giving in public health in the early twentieth century that was modest, catalytic, and individually determined, the giver today is much more visible, contributions much more substantial, and there is active involvement by more players in shaping the global agenda, national agendas, and in decision-making on setting priorities and resource allocation in international bodies and at a country level. The magnitude of the budgets of a few foundations has in some instances overshadowed the capacity of international organizations such as WHO and individual countries to achieve their own agendas, as well as the

contributions/priorities of smaller philanthropic organizations. Differing priorities of these different players can, in some cases, create uncertainties as to their added value. Thus, skewing by philanthropic giving of the priority setting and investment process globally and nationally has associated costs that may sometimes outweigh the benefits. For example, some perceive a conflict of interest when a foundation or industry plays a pivotal role at this level of decision-making, particularly where there may be some vested interest in the products used or in determining what proportion of the budget goes to which country or health need. The recent founding for a Smoke-Free World supported by the tobacco industry raises some of these issues. Other concerns include the imposition of unduly onerous conditions on the disbursement and use of the funds. While there is a need for accountability in use of funds for their intended purposes, that accountability should enable the desired outcomes to be reached effectively and efficiently, while respecting national sovereignty and avoiding excessive micromanagement.

Corporate social responsibility programmes have a different set of issues compared to foundations (Shah 2011). For example, while Rio Tinto, one of the largest global mining firms, is justifiably proud of its success in confronting malaria in Equatorial Africa, it has at the same time dumped billions of tonnes of toxic waste in Papua New Guinea. Similarly, Coca Cola's global distribution system has enabled more effective distribution of oral rehydration salts for diarrhoea, a major killer of children, particularly in Africa. On the other hand, Coca Cola and other producers of modified, unhealthy fast foods are also contributing to the burgeoning increase in obesity and NCDs in developing countries.

Foundations have always based their giving on the personal perspectives of their founders and subsequent leaders. They too often do not consult public health leaders in determining these goals. With the influx of large sums of money by a few foundations, this approach can very easily distort the public health agenda globally and nationally, as well as the roles of smaller foundations.

Large donations by private industry to international organizations like the WHO create an inherent dilemma for the latter. With over 80% of its budget coming from industry, the WHO has limited flexibility to focus funds on its own priorities. For example, in 2004–2005, WHO ended up spending 61%

of its budget on illnesses that accounted for only 11% of global mortality because of designations of the donated funds.

At a national level, large donations can dwarf the total budget of the recipient country and cannot be matched by the local government in many resource-constrained settings. The concern is that the local public health agenda could end up being determined by external agencies, which is a particular problem when the donor's agenda is not aligned with the country's priorities and diverts scarce resources, such as trained personnel and infrastructure, to lower priorities. Donor programmes often offer higher salaries that draw staff out of the public sector, leaving fewer staff within government to work on other health priorities, which results in deterioration of other important public health programmes and exacerbation of the existing public health human resource crisis extant in most developing countries. For example, in Haiti, the dramatically improved HIV/AIDS situation, facilitated by huge private funding that consumed most of the health workforce, resulted in a concurrent decline in other measures of the population's health, including infant mortality rates (Farmer and Garrett 2007).

In many settings, local government funding typically is less than 5% of total government expenditures. External funding can therefore perpetuate underfunding of health and undermine the responsibility that the government takes for healthcare delivery in their country. In countries with entrenched non-functioning bureaucracies and corruption, donor funds could exacerbate corruption, as funds are diverted from their intended purpose to benefit some government officials.

There are clear advantages to private funding being directed to specific public health challenges and emergencies. This has enabled eradication of smallpox, substantial increases in childhood immunization coverage, and stimulated improved responses in natural disasters and emergency situations. This type of funding works well when there are mutually agreed upon priorities among the various stakeholders, and quick and nimble investments for a short duration while the responsible parties work on plans and systems to address longer-term sustainability issues.

In relation to long-term sustainability in this context, there are at least four inter-related concerns. Firstly, when private donations are not synchronized with global or local priority setting and are targeted at chronic conditions that require a long-term investment, such as AIDS, withdrawal

of funding could result in a sharp worsening of the situation. In the case of AIDS, for example, an interruption in medication supplies could lead to treatment interruptions and promote the emergence of drug-resistant strains of HIV. Hence, significant private funding initiatives with a specific long-term focus, for example, eradication of poliomyelitis or development of a new drug or vaccine, need to be sustained until the goal is achieved for such investments to be effective.

A second area where issues of sustainability are a concern with private funding relates to a foundation having a diverse portfolio. Specifically, the concern here relates to the withdrawal of funding before the goal is achieved or the shifting of funds from one incomplete initiative to a new one. While in the corporate world this type of continuing prioritization and re-prioritization is typical, such an approach could be counterproductive when addressing a multitude of public health challenges that are often different than those confronting corporations. As an illustration, the Gates Foundation provides significant funds to eradicate poliomyelitis, support childhood immunization programmes, develop an AIDS vaccine and microbicides, and pre-exposure prophylaxis studies. In addition, its funds support the Grameen Foundation that uses microfinance loans to empower women in Bangladesh. In partnership with The Rockefeller Foundation, the Gates Foundation supports novel agricultural practices in developing countries to enhance food security. All of these initiatives are important long-term projects with uncertain outcomes. When and how decisions will be made to continue or stop investments will have a critical and wide-ranging impact on countries receiving this aid.

A third area of concern relates to who and how priorities are identified, particularly in settings with multiple challenges. For example, some express concern that the substantial funding in AIDS treatment is at the expense of other health priorities that may have a greater impact on morbidity and mortality, such as diarrhoea and respiratory disease. Others view it as an opportunity to concurrently strengthen healthcare delivery systems and integrate services, notably for TB, maternal health, and infant health, resulting in an overall benefit to all health services.

With the global efforts to treat AIDS patients, a new and major partner has been the pharmaceutical industry. The unprecedented price cuts and negotiation of intellectual property rights have been laudable and quoted as an example for extension to other diseases and products. Capacity building

and technology transfer to less developed countries is viewed as a major opportunity to enable national governments in developing countries to meet the needs of their populations, as opposed to medication donations that meet short-term needs but create longer-term dependence.

A fourth level of concern is the lack of harmonization among donors at the country level. In some countries, there are so many donors, each with their own agenda, that coordination and alignment to national priorities becomes very difficult. The disparity is further exacerbated if corruption is rife. A more critical issue relates to the diversion of critical human resource capacity away from government-identified priorities to donor priorities. The duplication of services by multiple funders also creates redundancy and inefficiencies.

External investment in a country's health should prioritize partnering with improving and supporting existing public health infrastructure, and should recognize that a country's government would generally be in a better position to judge its health priorities than an external organization based in a rich country (Widdus 2003). This of course assumes that the national government is well functioning and capable of determining its priorities.

In some developing countries, in the absence of local acceptable leadership, the UN agencies have assumed a leadership role, and private funding and non-governmental aid is structured to bypass the appropriate governmental agencies, thus creating a situation where the government may retaliate by blocking access to the targets of these charitable organizations.

HIV/AIDS provides another example of harmonization at a country level. All the UN agencies are part of the 'Three Ones' agreement: *one* agreed-upon HIV/AIDS action framework that provides the basis for coordinating the work of all partners; *one* national AIDS coordinating authority with a broad based multisector mandate; and *one* consensus for country-level monitoring and evaluation system (WHO 2004). Similar agreements are needed for donor-supported health projects in general.

China provides a unique example of how government authoritarianism has helped achieve this kind of coordination. Although it was motivated by a different purpose, China insists that large donor agencies have government approval before being allowed to work in-country, and thus, the relevant officials have been able to have a direct role in deciding where and how donor money should be channelled. In South Africa, in contrast, during the Mbeki AIDS denialist era, it was private and donor funding channelled

through non-governmental and academic institutions that ensured access to antiretroviral treatment for the millions dying from AIDS.

Moving forward

The generous giving of the foundations, industry, and private donors has benefited public health greatly, but they have sometimes also had unintended negative consequences. How can the positive impact of private giving be maximized? A few suggestions for strategies that may contribute to this goal include:

Objective assessment of priorities utilizing defined and objective criteria.
Transparency, communication, and coordination with key stakeholders (e.g. local health ministries).
Consideration of the long—as well as the short-term impact of investments.
Building capacity at a national level for sustainability of investments.
Ongoing internal monitoring of progress and periodic evaluation of the impact of the programme by an external group.
Establishment of advisory committees that include both international and local experts.
Coordination of multiple donors and key stakeholders to ensure harmonization and synergy.
Development of transparent governance mechanisms, policies, and procedural frameworks.
Sustained funding commitment for long-term health issues.

Many of these recommendations have already been incorporated into the GFATM. As we move forward and the roles and contributors to private funding continue to grow, it is important that the strengths of private giving, such as focus, nimbleness, innovation, risk-taking, and efficiencies do not get lost in massive, ineffective, and paralysing bureaucratic structures. To achieve the full potential of private giving, it is critical that it is structured to produce the maximum good with the fewest negative consequences.

REFERENCES

Ardian, M., et al. (2007). A public–private partnership for TB control in Timika, Papua Province, Indonesia. *International Journal of Tuberculosis and Lung Disease*, **11**, 1101–7.
Asante, A.D. and Zwi, A.B. (2007). Public–private partnerships and global health equity: prospects and challenges. *Indian Journal of Medical Ethics*, **4**, 176–80.
Barr, D.A. (2007). Ethics in public health research: a research protocol to evaluate the effectiveness of public–private partnerships as a means to improve health and welfare systems worldwide. *American Journal of Public Health*, **97**, 19–25.
Birn, A.-E. and Fee, E. (2013). The Rockefeller Foundation and the international health agenda. *The Lancet*, **381**, 1618–19.
Boutayeb, A. (2006). The double burden of communicable and non-communicable diseases in developing countries. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, **100**, 191–9.
Broadbent, J. and Laughlin, R. (2003). Public–private partnerships: an introduction. *Accounting, Auditing & Accountability Journal*, **16**, 332–41.
Buse, K. and Walt, G. (2000a). Global public–private partnerships: part I—a new development in health? *Bulletin of the World Health Organization*, **78**, 549–61.
Buse, K. and Walt, G. (2000b). Global public–private partnerships: part II—What are the health issues for global governance? *Bulletin of the World Health Organization*, **78**, 699–709.
Buse, K. and Waxman, A. (2001). Public–private health partnerships: a strategy for WHO. *Bulletin of the World Health Organization*, **79**, 748–54.
Butler, S.M. (2001). *Why the Bush Tax Cuts Are No Threat to Philanthropy*. [Online] Available at: http://www.heritage.org/research/reports/2001/03/why-the-bush-tax-cuts-are-no-threat-to-philanthropy
Bygbjerg, I.C. (2012). Double burden of noncommunicable and infectious diseases in developing countries. *Science*, **337**, 1499–501.
Carnegie Corporation of New York (n.d.). Website. [Online] Available at: http://www.carnegie.org
Croft, S.L. (2005). Public–private partnership: from there to here. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 99, S9–14.
Curran, J.W. and Jaffe, H.W. (2011). AIDS: the early years and CDC's response. *Morbidity and Mortality Weekly Report Surveillance Summaries*, **60**, 64–9.
DeVol, R., Bedroussian, A., and Charuworn, A. (2007). *An Unhealthy America: The Economic Burden of Chronic Disease*. Santa Monica, CA: Milken Institute.
Farmer, P. and Garrett, L. (2007). From 'Marvelous Momentum' to health care for all: success is possible with the right programs. [Online] Available at: http://www.foreignaffairs.com/articles/62458/paul-farmer-and-laurie-garrett/from-marvelous-momentum-to-healthcare-for-all-success-is-possible.
Flinders, M. (2005). The politics of public–private partnerships. *The British Journal of Politics & International Relations*, **7**, 215–39.
Foundation Center (2013). *Top Funders* [Online] Available at: http://www.foundationcenter.org
Fussel, H. and Beresford, C. (2009). *Public Private Partnerships: Understanding the Challenge* (2nd ed.). Vancouver: Center for Civic Governance, Columbia Institute.
Garvin, M.J. and Bosso, D. (2008). Assessing the effectiveness of infrastructure public–private partnership programs and projects. *Public Works Management & Policy*, **13**, 162–78.
Goldman, H.H., Morrissey, J.P., Ridgely, M.S., Frank, R.G., Newman, S.J., and Kennedy, C. (1992). Lessons from the program on chronic mental illness. *Health Affairs*, **11**, 51–68.
Hammack, D.C. and Wheeler, S. (1994). *Social Science in the Making: Essays on the Russell Sage Foundation, 1907–1972*. New York: Russell Sage Foundation.
Hodge, G.A. (2004). The risky business of public–private partnerships. *Australian Journal of Public Administration*, **63**, 37–49.

Hodge, G.A. and Greve, C. (2005). *The Challenge of Public–Private Partnerships: Learning from International Experience*. Cheltenham: Edward Elgar Publishing.
Howard, L.M. (1983). International sources of financial cooperation for health in developing countries. *Bulletin of the Pan American Health Organization*, **17**, 142–57.
Howard, L.M. (1991). Public and private donor financing for health in developing countries. *Infectious Disease Clinics of North America*, **5**, 221–34.
Howard Hughes Medical Institute (n.d.). *Origins: 1905–1953*. [Online] Available at: http://www.hhmi.org/about/origins.html.
Kane, S., et al. (2010). Large-scale public–private partnership for improving TB-HIV services for high-risk groups in India. *International Journal of Tuberculosis and Lung Disease*, **14**, 1066–8.
Kickbusch, I. and Quick, J. (1998). Partnerships for health in the 21st century. *World Health Statistics Quarterly*, **51**, 68–74.
Kraak, V.I., Harrigan, P.B., Lawrence, M., Harrison, P.J., Jackson, M.A., and Swinburn, B. (2012). Balancing the benefits and risks of public–private partnerships to address the global double burden of malnutrition. *Public Health Nutrition*, **15**, 503–17.
Lal, S.S., et al. (2011). Global Fund financing of public–private mix approaches for delivery of tuberculosis care. *Tropical Medicine & International Health*, **16**, 685–92.
Lester, G. (2012). The osteoarthritis initiative: a NIH public–private partnership. *HSS Journal*, **8**, 62–3.
McCoy, D., Chand, S., and Sridhar, D. (2009a). Global health funding: how much, where it comes from and where it goes. *Health Policy and Planning*, **24**, 407–17.
McCoy, D., Kembhavi, G., Patel, J., and Luintel, A. (2009b). The Bill & Melinda Gates Foundation's grant-making programme for global health. *The Lancet*, **373**, 1645–53.
McKee, M., Edwards, N., and Atun, R. (2006). Public–private partnerships for hospitals. *Bulletin of the World Health Organization*, **84**, 890–6.
Mectizan Donation Program (2012). 25 Years of the Mectizan Donation Program. [Online] Available at: http://www.mectizan.org/news/25-years-of-the-mectizan-donation-program
Menasha, M.C.E. and Hsig, R.M. (1915). *The Rockefeller Foundation, Annual Report, 1913–1914*. New York: The Rockefeller Foundation.
Morvaridi, B. (2012). Capitalist philanthropy and hegemonic partnerships. *Third World Quarterly*, **33**, 1191–210.
Nishtar, S. (2004). Public–private 'partnerships' in health—a global call to action. *Health Research Policy and Systems*, **2**, 5.
Novignon, J., Olakojo, S.A., and Nonvignon, J. (2012). The effects of public and private health care expenditure on health status in sub-Saharan Africa: new evidence from panel data analysis. *Health Economics Review*, **2**, 22.
Nwaka, S. (2005). Drug discovery and beyond: the role of public–private partnerships in improving access to new malaria medicines. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, **99**, 20–9.
Nwaka, S. and Ridley, R.G. (2003). Virtual drug discovery and development for neglected diseases through public–private partnerships. *Nature Reviews Drug Discovery*, **2**, 919–28.
Robert Wood Johnson Foundation (2012). *Grants: Robert Wood Johnson Foundation Clinical Scholars*. [Online] Available at: http://www.rwjf.org/en/grants/national-program-offices/R/robert-wood-johnson-foundation-clinical-scholars.html
Sanhai, W.R., Sakamoto, J.H., Canady, R., and Ferrari, M. (2008). Seven challenges for nanomedicine. *Nature Nanotechnology*, **3**, 242–4.
Sanofi-Aventis (2007). Sanofi-Aventis joins the fight against tuberculosis in South Africa. [Online] Available at: http://en.sanofi.com/Images/15299_brochure_aam_en.pdf
Shah, S. (2011). How private companies are transforming the global public health agenda. *Foreign Affairs*, 9 November. [Online] Available at: https://www.business-

humanrights.org/en/latest-news/how-private-companies-are-transforming-the-global-public-health-agenda/
Sharp, P.M. and Hahn, B.H. (2011). Origins of HIV and the AIDS pandemic. *Cold Spring Harbor Perspectives in Medicine*, 1, a006841.
Sinanovic, E. and Kumaranayake, L. (2010). The motivations for participation in public–private partnerships for the provision of tuberculosis treatment in South Africa. *Global Public Health*, **5**, 479–92.
Spackman, M. (2002). Public–private partnerships: lessons from the British approach. *Economic Systems*, **26**, 283–301.
Stop TB Partnership (n.d.). Global Laboratory Initiative (GLI). [Online] Available at: http://www.stoptb.org/wg/gli/default.asp
Strom, S. (2007). Big gifts, tax breaks and a debate on charity. *The New York Times*, 6 September. [Online] Available at: http://www.nytimes.com/2007/09/06/business/06giving.html?pagewanted=all&_r=0
Sturchio, J.L. and Cohen, G.M. (2012). How PEPFAR's public–private partnerships achieved ambitious goals, from improving labs to strengthening supply chains. *Health Affairs*, **31**, 1450–8.
The Rockefeller Foundation (2013a). *Our History. Moments in Time: 1913–1919*. [Online] Available at: http://www.rockefellerfoundation.org/about-us/our-history/1913-1919
The Rockefeller Foundation (2013b). Medicine in China. [Online] Available at: http://rockefeller100.org/exhibits/show/education/china-medical-board
UK Parliament (2011). Committee Publishes Report on Private Finance Initiative Funding. [Online] Available at: http://www.parliament.uk/business/committees/committees-a-z/commons-select/treasury-committee/news/pfi-report/.
Vitoria, M., et al. (2009). The global fight against HIV/AIDS, tuberculosis, and malaria current status and future perspectives. *American Journal of Clinical Pathology*, **131**, 844–8.
Vogel, A. (2006). Who's making global civil society: philanthropy and US empire in world society. *The British Journal of Sociology*, **57**, 635–55.
Weiner, B.J. and Alexander, J.A. (1998). The challenges of governing public–private community health partnerships. *Health Care Management Review*, **23**, 39–55.
Wellcome Trust (n.d.). Website. [Online] Available at: http://www.wellcome.ac.uk/
Wheeler, C. and Berkley, S. (2001). Initial lessons from public–private partnerships in drug and vaccine development. *Bulletin of the World Health Organization*, **79**, 728–34.
Widdus, R. (2001). Public–private partnerships for health: their main targets, their diversity, and their future directions. *Bulletin of the World Health Organization*, **79**, 713–20.
Widdus, R. (2003). Public–private partnerships for health require thoughtful evaluation. *Bulletin of the World Health Organization*, **81**, 235.
Widdus, R. (2005). Public–private partnerships: an overview. *Royal Society of Tropical Medicine and Hygiene*, 99, S1–8.
Williams, O.F. (2008). The UN Global Compact: the challenge and the promise. *Leadership and Business Ethics*, **2008**, 229–49.
World Bank Group (2011). *PPP in Infrastructure Resource Center for Contracts, Laws and Regulations*. New York: World Bank.
World Health Organization (WHO) (2004). *WHO: 'Three Ones' Agreed by Donors and Developing Countries*. [Online] Available at: http://www.who.int/3by5/newsitem9/en/

11.14

Global health in the era of sustainable development

Fiona Fleck

Introduction: from millennial to sustainable development goals

The world is experiencing demographic and epidemiological transitions that have resulted in a larger human population living much longer lives. In response to these changes, the World Health Organization (WHO) has expanded its initial focus on infectious causes of death to a much broader agenda, including noncommunicable diseases, violence and injuries, nutrition, antimicrobial resistance, health worker shortages, increasing pollution and decreasing biodiversity, climate change, mass migration and inequitable access to basic medicines and services (Vujicic et al. 2004; WHO 2010b).

The WHO works through national governments and with many other agencies, initiatives, and public–private partnerships to improve human health. The final scorecard of the Millennium Development Goals summarizes the cumulative progress made in recent decades (UN 2000, 2015, 2017; PATH 2017; UNAIDS 2018).

For example, from 1990 to 2015, the maternal mortality ratio declined by 45% and the percentage of births assisted by skilled health personnel increased from 59 to 71% (UN 2015). By 2017, the global under-5 mortality rate had declined to 39 deaths per 1,000 live births from a 1990 rate of 90 deaths per 1,000 live births (UN 2015; UN Inter-Agency Group for Child Mortality Estimation 2018). In 2017, about 85% of children worldwide received at least one dose of measles vaccine, an increase from 73% in 2000 (UN 2015; WHO 2018a).

New human immunodeficiency virus (HIV) infections fell from an estimated 3.5 million cases in 2000 to 1.8 million in 2017 (Jordan et al. 2017). In 2003, 800,000 people living with HIV were receiving

antiretroviral therapy, compared to 21.7 million in 2017 (UN 2015; UNAIDS 2018). Between 1990 and 2015, global malaria incidence fell by an estimated 37% and mortality by 58%, however WHO estimates show that no further progress was made between 2015 and 2017 (WHO 2016a, 2018j). Tuberculosis mortality fell by 42% over between 2000 and 2017 (WHO 2018i).

The WHO's efforts to further improve health outcomes, described in its thirteenth general programme of work 2019−2023 (GPW 13), are closely aligned with the *2030 Agenda for Sustainable Development*. In adopting this agenda, Member States of the United Nations committed to meeting 17 goals and 169 targets, measured by 250 indicators (UN 2015a). Sustainable development is defined as a process that 'meets present needs without compromising the ability of future generations to meet their own needs' (World Commission on Environment and Development 1987).

In the 2030 Agenda, governments agree to reconcile three conflicting demands: improved human welfare, environmental protection, and economic growth. Most high-income countries have tended to fare better on measures of human well-being (Sen 1999), but advances in each dimension of the human development index; life expectancy, education, and gross national income per capita, are associated with a large environmental footprint (Sneddon et al. 2006; Steffen et al. 2015).

The mutually conflicting goals of industrialization (Sustainable Development Goal (SDG) 9, industry, innovation and infrastructure) and climate change mitigation (Goal 13, climate action) and between biodiversity (Goal 14, life below water; Goal 15, life on land) and human welfare (Goal 1, no poverty) are clearly outlined, confronting policy makers with many interactions between and within the different goals. These can be divided into positive interactions (i.e. synergies), in which progress towards one goal contributes towards progress in another, or trade-offs, when progress towards one goal hinders progress in another. A few goals have no demonstrable association in either direction and are thereby unclassified (Worm et al. 2006; Pradhan et al. 2017). The 2030 Agenda challenges governments to uncouple the relationship between economic growth and environmental damage, at a time when most countries will experience a rapid increase in population.

The WHO's focus, Goal 3 (good health and well-being), is synergistic with Goal 1 (no poverty), Goal 5 (gender equality), Goal 10 (reduced

inequalities) and Goal 6 (clean water and sanitation). The health goal has 13 targets, including some of the areas that were covered by the Millennium Development Goals—maternal, newborn, and child deaths, HIV, tuberculosis, and malaria. The new targets are for neglected tropical diseases, hepatitis, water-borne and other communicable diseases, non-communicable diseases and mental health, narcotic drug abuse and the harmful use of alcohol, road traffic deaths and injuries, sexual and reproductive health services, hazardous chemicals, air, water and soil pollution and contamination, tobacco control, research and development on vaccines and medicines, health financing and workforce, and early warning of national and global health risks. The WHO and other health organizations have mapped their respective mandates and set mid-term milestones for targets of SDG 3 in the *Global Action Plan for Healthy Lives and Well-being for All* (WHO 2013a, 2018k).

The WHO constitution established 'the enjoyment of the highest attainable standard of health' as 'one of the fundamental rights of every human being without distinction of race, religion, political belief, economic or social condition' (WHO 1948). The right to health underpins all SDG 3 targets, particularly 3.8: to 'achieve universal health coverage, including financial risk protection, access to quality essential health-care services and access to safe, effective, quality and affordable essential medicines and vaccines for all'.

Based on the 2030 Agenda, the WHO's current five-year strategy presents a vision of: one billion more people benefitting from universal health coverage, one billion more people better protected from health emergencies, and one billion more people enjoying better health and well-being (WHO 2018).

Towards universal health coverage

Universal health coverage (UHC)—defined as ensuring that all people have access to needed health services of sufficient quality without incurring financial hardship—is the subject of sustainable development target 3.8. As a step towards this target, the WHO envisions an expansion of such coverage to an additional one billion people by 2023. In the absence of financial protection, an estimated 800 million people spend at least 10% of

their household budgets on health expenses. For almost 100 million people each year, these expenses are high enough to push them into extreme poverty (WHO/World Bank 2017). At least half of the world's population still lacks access to essential health services.

Progress towards UHC is tracked by measures of coverage of health services, and by the impact that out-of-pocket expenditure on those services has on people's living standards and their ability to spend on other basic needs. These measures reflect the twin imperatives of UHC: that people receive the health services they need without facing financial hardship (Sachs 2012; Wagstaff et al. 2017, 2018).

The WHO and the World Bank evaluate the overall level and the extent to which UHC is equitable because national averages can mask low levels of coverage in disadvantaged groups. For example, only 17% of mothers and children in the poorest fifth of households in low- and lower-middle-income countries received at least six of seven basic maternal and child health interventions, compared to 74% for the wealthiest fifth of households (World Bank 2004; WHO/IBRD 2017).

The International Health Partnership for UHC 2030 has 118 signatories —governments and multilateral, civil society, and philanthropic organizations—to its Global Compact. They have committed to building and expanding equitable, resilient, and sustainable health systems, funded primarily by public finance, and based on primary healthcare, that deliver integrated, comprehensive people-centred and quality health services for all, while taking necessary measures to protect households from financial hazards due to health expenditures.

The WHO defines quality of care as the degree to which health services increase the likelihood of desired health outcomes and are consistent with current professional knowledge (Kieny et al. 2018). The components of services comprise financing, workforce, governance, service delivery, information, medicines, technologies, and infrastructure.

Primary healthcare is an essential component of national health systems that is needed to move countries closer towards UHC. Strong primary healthcare is vital for making progress on the unfinished agendas of communicable diseases, and maternal, newborn, child, and adolescent health, as well as for the growing burden of non-communicable diseases.

Digital health approaches and other system innovations are also playing an important role in strengthening these services.

Each national health system has scope to improve upon the components of its health services. For example, essential medicines and health technologies are often poorly distributed and unfairly priced (FAO 2018a). Repeated shortages continue to affect critical products, including insulin, antihypertensives, intravenous fluid, oral rehydration solution, analgesics, oral antidiabetics, and opioids for palliative care.

Antiretrovirals for HIV, direct-acting antivirals for hepatitis C and second-line treatments for drug-resistant tuberculosis when first marketed were too expensive for many countries. However, some countries have since negotiated price reductions, following voluntary licensing agreements and production of quality-assured generics (Andrieux-Meyer et al. 2015) and UN procurement has brought these medicines within reach of millions of patients.

In some countries, the costs of caring for diabetes alone consume nearly 25% of the national health budget (Zhang et al. 2010). Globally, insulin can be accessed reliably by only half of the 100 million people who need it (Basu et al. 2019). In the United States of America, out-of-pocket expenses for insulin increased by 89% between 2000 and 2010. The US$21-billion global market in insulin is divided between three companies that together account for 99% in terms of value and 96% in terms of volume. Although more than 90 of 132 countries have no tariffs for insulin, the combination of taxes, steep mark-ups, and other supply chain costs have pushed up the drug price and hurt affordability (Health Action International 2015).

The WHO has revised the methods used to measure pricing and affordability of essential medicines—particularly those used for primary care—in the context of the SDG monitoring. The WHO promotes policies and builds technical capacities in low-resourced health systems, develops international standards for the manufacturing and regulation of health products and provides guidance for health systems on how to deliver them safely and cost-effectively.

Better access to medical products is pursued by negotiating price reductions for existing products and incentivizing product development for markets with limited purchasing power. Generic medicines, which do not need to recoup the investment needed for drug discovery, development, and marketing approval, are sold at prices closer to the costs of manufacturing, offering a significant means to reduce healthcare costs in all countries. A

national policy that encourages the use of generic medicines can save an estimated 60% of the costs of medicines (WHO 2010, 2010a).

In 1977, the WHO introduced the first *Model List of Essential Medicines*, its selection of the most effective and affordable medicines available at the time. The model list is intended to be adapted for use in any health system, especially those in low- and middle-income countries. The current list contains 433 products and is used by about 150 countries. In 2007, WHO developed the first *Model List of Essential Medicines* for children under 12 years of age. These lists are updated every two years (WHO 2013b).

The WHO prequalification programme stimulates the supply of quality-assured medicines, including generics, and encourages competition between manufacturers (WHO 2013c). The programme was established in 2001 as a quality-assurance mechanism for products procured by UN agencies, the GAVI Alliance, and the Global Fund. Prequalification has resulted in more reliable supplies, better forecasting, and often substantial price reductions (WHO n.d.). Compliance measures include on-site inspections to ensure adherence to good manufacturing practices. Products can become de-listed if quality is not maintained; the prospect of de-listing has proven to be an incentive for manufacturers to maintain standards.

In addition to financing, infrastructure, and medicines, UHC requires a health workforce of adequate size and skills. However, countries at all levels of socioeconomic development face difficulties in education and training, deployment, retention, and performance of their health workforce (Anand and Barnighausen 2004, 2007; Mullan et al. 2011).

The WHO and the World Bank have developed an ‘SDG index threshold’ of 4.45 doctors, nurses, and midwives per 1,000 population as an indicative minimum density for health workers. In 2016 the global health workforce was over 43 million. The supply projections point out that despite significant growth leading to an aggregate number by 2030 of 67.3 million health workers, the projected global shortage of healthcare workers in 2030 will be more than 14.5 million (WHO 2016).

Disabilities

The WHO estimates that over a billion people, about 15% of the world's current population, have some form of disability. Its *International Classification of Functioning, Disability and Health* defines disability as an

umbrella term for impairments, activity limitations and participation restrictions as the result of interaction between individuals with a health condition and personal and environmental factors. Therefore, the disability experienced by a person with lower limb paralysis in a setting where wheelchairs are available, and where homes, pavements, public transport, and the workplace are adapted to wheelchair access, is very different from the disability experienced by a person who has the same physical condition but who lives in an environment with none of these adaptations.

People with disabilities report seeking more healthcare than people without disabilities and have greater unmet needs. However, all people with disabilities also develop the same health problems that affect the general population, and therefore need access to mainstream healthcare services. Article 25 of the UN Convention on the Rights of Persons with Disabilities reinforces the right of persons with disabilities to attain the highest standard of healthcare, without discrimination. The WHO also provides guidance to policy makers on mental health programmes. Of particular concern are the health services available to people with severe mental disorders; moderate to severe depression, bipolar disorder, and schizophrenia and other psychotic disorders. People with these conditions generally have a life expectancy that is 10–20 years less than the general population. Most of these premature deaths are due to physical health conditions. For the minority who can access healthcare, the care received is often of poor quality and physical health conditions are often overlooked.

Emergency preparedness and response

Since its inception, WHO has coordinated the international response to outbreaks and other cross-border heath threats. Its International Health Regulations oblige Member States to report infectious disease outbreaks and other health incidents such as chemical spills and nuclear incidents, as well as to build core capacities for surveillance and response (WHO 2008a; Heymann et al. 2013). These reports are assessed, and responses are mounted in function of the severity, potential for spread, number of people affected, and the causative event or agent. The strongest rating is a ‘Public Health Event of International Concern’ defined as, ‘an extraordinary event which is determined to constitute a public health risk to other States through

the international spread of disease and to potentially require a coordinated international response' (WHO 2008a, 2018b, 2018c). Since the 2005 revision of the Regulations, there have been four such events: H1N1 influenza (2009), Ebola (2014), polio (2014), and Zika (2016).

This approach to health emergencies was severely tested during the 2013–2016 Ebola virus disease outbreak in West Africa. Problems with the international response have been well documented (WHO 2015a; The Ebola Outbreak Epidemiology Team 2018). There were concerns that neither the affected countries nor the WHO sounded the alarm early enough; and that the rating of the outbreak as a public health event of international concern was overdue. Coordination between the WHO, Ministries of Health, United Nations agencies, and non-governmental partners was weak on many levels. The WHO lacked sufficient staff on location to respond adequately and the dynamics of transmission across the borders of Guinea, Liberia and Sierra Leone highlighted the need to strengthen health security (Moon et al. 2015; WHO 2015a, 2015b; Gostin and Katz 2016).

The International Health Regulations require countries to avoid disruption to travel and trade, but during the 2013–2016 Ebola outbreak more than 40 countries implemented measures that significantly interfered with international traffic, outside the scope of the temporary recommendations issued by WHO. As a result, Guinea, Liberia, and Sierra Leone faced severe political, economic, and social consequences, as well as barriers to receiving necessary personnel and supplies (WHO 2015a).

The lessons learnt from the 2013–2016 outbreak catalysed the establishment of the WHO's Health Emergencies Programme. A more effective response to outbreaks and emergencies is made possible by a unified structure and one line of accountability from headquarters through the WHO's Regional Offices. Administrative and financial adjustments were made, including the creation of a contingency fund for emergencies. An incident management system is used to structure and manage response activities, as a core component of the Emergency Response Framework (WHO 2017a).

The WHO now coordinates more closely with United Nations Office for the Coordination of Humanitarian Affairs, joining other agencies that provide support for humanitarian crises in the field, such as the United Nations Refugee Agency, the World Food Programme, and the United Nations Children's Fund (Nabarro 2015).

As of February 2019, the WHO was responding to 41 graded health emergencies around the world and leading the health cluster in United Nations Inter-Agency Standing Committee cluster system in responses to outbreaks, emergencies, protracted conflicts, and natural disasters in 27 countries and territories.

There are three WHO grades for emergencies, signifying the level of operational response required: Grade 1 (limited response); Grade 2 (moderate response); Grade 3 (major/maximal response) (Inter-Agency Standing Committee 2016). A graded emergency that persists for more than six months becomes a protracted emergency, which is an environment in which a significant proportion of the population is vulnerable to prolonged disruption of livelihoods, disease, and death.

The WHO provides technical assistance to countries to help build their capacity to detect, report, and contain infectious disease outbreaks and other health events in line with the International Health Regulations. This assistance covers early warning alert and response systems, laboratory and surveillance, burden of disease, regulatory capacity building, risk communications, and planning for pandemic product deployment.

The WHO has also developed standardized approaches that countries can take to respond to major epidemic-prone diseases, such as meningitis, yellow fever, and cholera. In 2016, the WHO launched a plan to immunize 1.4 billion people against yellow fever in 40 countries by 2026. In 2018, the World Health Assembly endorsed a goal to reduce cholera deaths by 90% in 47 countries (G20 2018).

The WHO Emergency Medical Teams Initiative assists organizations and Member States to build capacity of clinical response teams and to better coordinate deployment of quality-assured medical teams in emergencies. The WHO also helps to identify dangerous and re-emerging pathogens and provides countries with the technical support they need to strengthen their biosafety, biosecurity, and readiness for outbreaks (Bermingham et al. 2012). To prevent epidemics, countries need resilient and inclusive health systems that extend to rural areas, a sensitive early warning system along with rapid response capacity, and informed communities that trust their health service providers. They also need effective and affordable vaccines, diagnostics, and treatments, reliable laboratory and logistics capacity, and safe treatment facilities, properly distributed, staffed, and equipped.

Preventing epidemics and pandemics

During the 2013–2016 Ebola outbreak, the WHO coordinated field trials for candidate vaccines. In December 2016, the Ebola vaccine rVSV-ZEBOV was shown to confer nearly 100% protection for clinical trial participants in Guinea (Henão-Restrepo et al. 2017). To better anticipate the need for vaccines, diagnostics, and therapeutics, the WHO collects inputs to its R&D Blueprint (WHO 2016b). In 2018, the priority diseases were: Crimean-Congo haemorrhagic fever, Ebola virus disease, Marburg virus disease, Lassa fever, Middle East respiratory syndrome, severe acute respiratory syndrome, Nipah and henipaviral diseases, Rift Valley fever, and Zika. The list also contains 'Disease X', representing the epidemic potential of an unknown pathogen (WHO 2018h). For each of the epidemic-prone diseases, research and development roadmaps, target product profiles, and generic protocols are developed. At the same time, the WHO works to strengthen national regulatory and ethics bodies so that research can be conducted during public health emergencies (WHO 2014).

It is clearly preferable that Member States lead the response and coordination of health emergencies, drawing on the support from international partners as required. Increasingly, the WHO is assisting Member States to establish Emergency Operations Centres and develop their capacities for emergency management and coordination. Partnerships such as the Global Outbreak Alert and Response Network and the Global Health Cluster are vital to effective preparedness and response.

Monitoring and evaluation are difficult in emergency situations but are needed for the improvement and adjustment of public health interventions (Roberts and Brennan 2015). Another challenge for emergency healthcare provision is security, as many outbreaks occur in a conflict setting. WHO is monitoring and compiling data on attacks on health facilities, transport, and staff, as well as engaging in campaigns such as #NotATarget, with Médecins Sans Frontières and the International Committee of the Red Cross (WHO n.d.).

A century after the 1918 influenza pandemic—which killed an estimated 30–50 million people—the WHO is still responding to media coverage, public perception, and government policy on influenza. In 2011, the World Health Assembly approved the Pandemic Influenza Preparedness Framework (WHO 2011). The framework sets out obligations for the

sharing of information and materials relating to influenza viruses, and of medicines and vaccines, once a pandemic is underway. The framework stipulates obligations for the WHO, its Member States, and WHO-appointed influenza laboratories, and the pharmaceutical industry (Gao et al. 2013; CDC 2018).

In 2018, thirty-five countries had the systems needed to detect unusual respiratory disease events, a fivefold increase since 2014. In addition, 34 priority countries had functioning inpatient influenza surveillance: this doubled from 16 countries in 2014 and 29 priority countries have a human–animal interagency coordination mechanism (CDC 2018; WHO 2018l). Moreover, 91 countries shared epidemiological data on influenza, compared to 55 countries in 2014.

However, the continued building of national, regional, and global capacities to prevent, prepare for, respond to, and recover from, health emergencies remains a vital international priority. With the sole exception of disease eradication, gains against infectious diseases are always fragile (Packard 1997). Reductions in transmission can be undone by complacency, diminished control efforts, changes in vector habitats or human behaviour, or by the spread of resistant pathogens.

Antimicrobial resistance prevents effective treatment of a range of infections caused by bacteria, parasites, viruses, and fungi. Without effective antibiotics, trauma, surgery, organ transplants, childbirth, and immunosuppression can become life-threatening events. The high cost of healthcare for patients with resistant infections is due to a longer duration of illness, complications, additional tests, and the need to use more expensive drugs.

Antimicrobial resistance occurs naturally over time, through genetic changes in pathogens. However, the misuse of antimicrobials in medicine and in agriculture is accelerating this process. Antimicrobial resistant-microbes are found in people, animals, food, and the environment. They can spread between people and animals, including from food of animal origin, and from person to person (Ong and Heymann 2007). Poor infection control, inadequate sanitary conditions, and inappropriate food-handling encourage the spread of antimicrobial resistance. Multidrug resistance is a particular concern for major health programmes focused on the control of malaria, tuberculosis, and HIV infection.

The Global Action Plan on antimicrobial resistance was adopted by the World Health Assembly in 2015 and its objectives guide the WHO in supporting Member States to develop and implement national action plans on antimicrobial resistance (WHO 2015b).

The WHO also works with the Food and Agriculture Organization of the United Nations (FAO) and the World Organisation for Animal Health (OIE) in a tripartite 'One Health' approach to promote best practices to avoid the emergence and spread of antibiotic resistance, including optimal use of antibiotics in humans, animals, and the environment.

In 2016, the United Nations General Assembly signalled the world's commitment to taking a coordinated approach to address the root causes of antimicrobial resistance across multiple sectors, especially human health, animal health, and agriculture.

Better health and well-being

In 1800, the global population was less than one billion. Rapid population growth, peaking at 2.6% in 1962, resulted in the global population increasing from 1.5 billion in 1900 to 6.1 billion in 2000. That growth has now slowed, reducing by half since 1960. Life expectancy has risen globally for both men and women, a gain of more than 10 years since 1970, to a global average of 72 years in 2016. The WHO South East Asian region has seen a gain of more than 8 years in the past two decades alone. By 2100, average global fertility is predicted to decline to two live births per woman of reproductive age, average life expectancy to level out at 83 years for men and 85 for women and the global population to therefore stabilize at about 11.2 billion (UN 2017).

Worldwide, the proportion of older people in the total population is increasing at more than three times the overall population growth rate. With some exceptions, populations are now ageing fastest in low- and middle-income countries. A transition towards an older society, which took place over more than a century in Europe, is now taking place in less than 25 years in countries like Brazil, China, and Thailand.

By 2030, for the first time in history, the population of people aged 65 and older are expected to outnumber children under age 10 (1.41 billion versus 1.35 billion). By 2050, older persons are expected to account for

35% of the population in Europe, 28% in North America, 25% in Latin America and the Caribbean, 24% in Asia, 23% in Oceania, and 9% in Africa. These trends have major implications for healthcare, as most people will die from the complications of noncommunicable diseases.

The health impacts of climate change

Quantifiable targets are required because the unprecedented improvements in health outcomes seen at the beginning of the century are fragile and easily reversed, including by the effects of climate change and environmental degradation (WHO 2008b). The WHO expects approximately 250,000 additional deaths per year between 2030 and 2050 to be attributed to the health effects of climate change; 38,000 due to heat exposure in elderly people, 48,000 due to diarrhoea, 60,000 due to malaria, and 95,000 due to childhood undernutrition. The direct costs to health are estimated to be between US$2–4 billion per year by 2030 (GBD 2016 Risk Factors Collaborators 2017; Postigo et al. 2007; WHO 2014, 2018g).

Climate change affects human health directly through extreme weather events and in a more chronic fashion through changes in precipitation, crop yields, and sea levels. More than half of the world's population lives within 60 km of the sea. Rising sea levels and storms destroy homes, medical facilities, and other essential services. Deaths in Europe alone from extreme weather events—heatwaves and cold waves, wildfires, droughts, river and coastal floods, and windstorms—are projected to increase from about 3,000 to about 150,000 per year by 2100 (Forzieri et al. 2017).

Variable rainfall patterns affect the supply of fresh water, a shortage of which compromises hygiene and increases the risk of diarrhoeal disease. Floods contaminate freshwater supplies, heighten the risk of water-borne diseases, and create breeding grounds for insect vectors. They also cause drownings and physical injuries, damage homes, and disrupt health services. Rising temperatures and variable precipitation are likely to decrease the production of staple foods in many of the poorest regions. Crop failures are likely to increase the prevalence of malnutrition and undernutrition, which currently cause 3.1 million deaths every year (USFDA 2013; UNICEF 2018).

Climatic conditions also affect water-borne and vector-borne diseases by lengthening their transmission seasons and altering the geographic range of

vectors. As mosquito habitats change, more people are likely to be exposed to arboviral diseases such as yellow fever, dengue, chikungunya, and Zika, as well as to malaria. After years of declining incidence and deaths, progress in malaria control stalled in 2015, with an estimated 219 million cases and 435,000 related deaths in 2017 (Nayyar et al. 2012; WHO 2018f).

People living near the coast are particularly vulnerable to climate change. Small island states and low-lying coastal countries, home to an estimated 65 million people, are already experiencing the adverse effects of climate change. With one-third of their population living on land that is less than 5 metres above sea level, the threat of sea-level rise, storm surges, and coastal destruction pose existential risks (UNDP 2017). Some small island states may eventually disappear with sea-level rise.

At the 2017 United Nations Climate Change Conference, the WHO and the UN Climate Change secretariat launched a special initiative to protect people living in small island developing states from the health impacts of climate change (Ghebreyesus and Espinosa 2018).

Adaptation measures on these islands are complicated by their unique and fragile biodiversity and their dependence on imported fuels and food. While contributing less than 1% to the world's greenhouse gas emissions, these countries are among the first to experience the worst and most devastating impacts of climate change. Despite the lack of commensurate resources, some small island states are already improving population and health system resilience by diversifying crops, building cyclone-resistant health facilities, and using solar energy to power medical services (UNDP 2017).

Air pollution and a multisectoral approach

Affecting a much larger proportion of the world's population, air pollution is another factor mediating the health impacts of climate change. The WHO estimates that more than 90% of the world's population lives in places where the air quality does not meet the recommended concentrations levels of major pollutants, of which particulate matter is a common proxy indicator. WHO air quality guidelines set limits for particulate matter, nitric dioxide, ozone, and sulphur dioxide—pollutants for which a revision is due in 2020—and other air pollutants. The WHO estimates that 4.2 million deaths from ambient air pollution and 3.8 million deaths from household air

pollution are caused by the contribution of these pollutants to stroke, heart disease, chronic obstructive pulmonary disease, lung cancer, and acute respiratory infections (Prüss-Üstün et al. 2016).

Multisectoral action is vital to global health efforts and central to the SDG agenda because of the many determinants acting upon people's health, such as socioeconomic status, gender, and other social determinants. Such action is thus essential for tackling the various sources of air pollution. For example, when the Olympic Games were held in Beijing in 2008, stringent policies to reduce local and regional emissions were enacted: traffic was reduced, factories were closed, construction halted, and emission standards revised (Su et al. 2015). Air quality improved dramatically in Beijing and adjacent co-hosting Chinese cities. Observational studies showed significantly decreased cardiovascular deaths in the cities affected by the 6-month implementation period and researchers estimated that sustained improvements, consistent with WHO air quality standards, could avert almost 1.57 million premature deaths in China each year (Guojun et al. 2016).

Ambient and household air pollution can be reduced by expanding access to clean household fuels and technologies to meet the domestic energy needs of cooking, heating, and lighting. Ambient air quality can be rapidly improved by the expansion of urban public transit networks, energy-efficient buildings, optimal waste management, and by producing electricity from clean, renewable energy. Policies to reduce air pollution can thus be beneficial for both climate and health, lowering the burden of disease attributable to air pollution, as well as contributing to climate change mitigation. These mutually beneficial policies, akin to the synergistic SDGs, are known as health co-benefits. For example, policies that favour electric vehicles and that promote public and active transport can reduce carbon emissions and traffic congestion, impermeabilization of urban land and noise exposure, while improving the urban environment, people's physical fitness, and ambient air quality (UNEP 2013).

Noncommunicable diseases: determinants and risk factors

The transition from infectious to noncommunicable causes of deaths is indicated by current estimates attributing 71% of all deaths to noncommunicable diseases, 15 million of which occur between the ages of

30 and 69 years. Deaths in adults occurring before the age of 70 are considered premature. In total, cardiovascular diseases account for 17.9 million deaths annually, cancers 9 million, respiratory diseases 3.9 million, and diabetes 1.6 million. These four groups of diseases account for over 80% of premature deaths due to noncommunicable diseases (WHO 2016). Road traffic crashes cause 1.35 million deaths per year and have become the leading cause of death of people aged 5 to 29. Approximately half of the people killed on the roads are pedestrians, cyclists, or motorcyclists (WHO 2018e).

Four risk factors are associated with 70% of the premature deaths due to cardiovascular diseases, cancer, respiratory diseases, and diabetes: tobacco use; the harmful use of alcohol; energy-dense, nutrient-poor diets; and inadequate physical activity. In each area, the WHO gathers data, recommends evidence-based policies, provides technical support, and monitors the impact of interventions. In 2011, 'best buy' interventions were proposed to governments for the prevention and control of noncommunicable diseases, ranging from antihypertensives and antilipidemiants for cardiovascular disease to vaccines for human papillomavirus to prevent cervical cancer. These are interventions that have significant public health impact, and are cost-effective, inexpensive, and feasible to implement (WHO and the World Economic Forum 2011).

To support countries in tackling the increasing burden of road traffic crashes that claim over 1.35 million lives each year, the WHO recommends passing laws to reduce speed and drink-driving, to make seat-belts and helmet mandatory, as well improving road infrastructure with safe pedestrian crossings and cycle lanes, making vehicles safer, improving post-crash care, and investing in public transport.

Recent estimates show that 1.4 billion people worldwide do not meet the WHO recommendations on physical activity (Guthold et al. 2018). A large proportion of this inactivity is a result of the same road traffic conditions that cause the deaths of 351,000 cyclists and pedestrians per year. The WHO therefore recommends the provision of infrastructure, facilities, and public open spaces for walking, cycling, and other physical activity (WHO 2018e).

Nutrition—optimal quantity and quality

Everyone has the right to adequate food and nutrition; however, this human right is continually violated as nearly one in three people suffer from at least one form of malnutrition. In 2017, 462 million adults worldwide were underweight while more than 1.9 billion were overweight, of whom more than 600 million were obese. In the same year, 38.3 million children under the age of 5 years were overweight or obese and 150.8 million were affected by stunting (low height-for-age) and 50.5 million were affected by wasting (low weight-for-height) (FAO 2014, 2018b; WHO 2018).

By 2030, the global community has committed to 'end all forms of malnutrition, including achieving, by 2025, the internationally-agreed targets on stunting and wasting in children under 5 years of age, and address the nutritional needs of adolescent girls, pregnant and lactating women and older persons'.

Poor nutrition continues to cause nearly half of deaths in children under 5, while malnutrition and unhealthy diets are key risk factors for non-communicable diseases such as cardiovascular disease, type 2 diabetes, and certain cancers in adults (Petemadi et al. 2017). People are consuming more processed foods high in calories, fats, sugars, and salt. Nearly two out of three infants under 6 months of age are not exclusively breastfed (Yang et al. 2014).

Many diets are low in whole grains, in fruits, nuts and seeds, or vegetables. Diets are the result of personal factors such as gender, culture, knowledge, skills, and preferences. Diets are also determined by economic, political, and environmental factors including cost and availability of food, marketing, labelling, poverty levels, urbanization, geography, and climate change.

To tackle malnutrition and prevent and control diet-related noncommunicable diseases everyone needs access to safe, affordable, and healthy diets. The Second International Conference on Nutrition Framework for Action provides a set of 60 policy actions and recommendations that governments, can use to improve nutrition (FAO/WHO 2014).

These policy actions need to be translated into concrete actions such as restricting marketing of unhealthy foods to children, taxing sugar-sweetened drinks or banning industrial trans fats, mandating simplified front-of-pack nutrition labelling and introducing nutrition standards for food served or sold in schools, hospitals, and other public institutions

(Hastings et al. 2003; McGinnis et al. 2006; Cairns et al. 2009; De Ruyter et al. 2012; Ebbeling et al. 2012; Qibin et al. 2012; WHO 2015, 2017).

In this 2016–2025 Decade of Action on Nutrition, the response by governments is accelerating; but more action, consistent with the WHO's recommendations, is needed (WHO 2017b).

Minimizing harms from tobacco and alcohol

Tobacco use kills around 8 million people every year, accounting for 17% of adult deaths (GBD 2017). Since 2000, the prevalence of smoking has decreased from a global average of 27% to 20% in 2016. Currently 34% of men and 6% of women smoke, totalling 1.1 billion smokers aged 15 or more globally. While smoking rates have declined since 2000, the number of smokers has not. Over 24 million young people aged 13–15 smoke cigarettes. The number of smokeless tobacco users in 2016 stands at 367 million adults and 13 million young people aged 13–15 (WHO 2013, 2018).

The epidemics of smoking-related diseases—lung cancer, chronic obstructive pulmonary disease, and other respiratory conditions—are decreasing in countries where tobacco control measures have taken effect and the burden of disease is shifting to low- and middle-income countries.

The World Health Assembly adopted its first resolution on tobacco control in 1970. Building on decades of progress, SDG target 3.4, calls upon countries to strengthen their implementation of the WHO Framework Convention on Tobacco Control (WHO FCTC). The WHO FCTC was the first treaty negotiated under the auspices of the WHO and one of the most rapidly ratified treaties in the history of the United Nations (WHO 2012). Provisions in the treaty govern the production, sale, distribution, advertisement, and taxation of tobacco products as minimum requirements for tobacco control (WHO 2013d).

To reduce demand for tobacco, many countries have passed tobacco control legislation restricting tobacco sales by age and type of outlet as well as by increasing taxes. Australia, France, Ireland, Russia, Turkey, the United Kingdom have introduced indoor smoking bans and other measures. In 2012, Australia became the first country to introduce plain packaging (Australian Bureau of Statistics 2012; National Health and Medical Research Council 2013). Plain packaging prohibits the use of logos, colours, brand images, and promotional information on tobacco products

and packaging, other than brand and product names in a prescribed colour and font. Six other countries have since implemented plain packaging laws: France, Hungary, Ireland, New Zealand, Norway, and the United Kingdom. Australia and later France, Norway, and the United Kingdom successfully fought legal challenges to their plain packaging laws. Several other countries are enacting legislation, developing regulations, or approving plain packaging policies.

The development of electronic cigarettes and heated tobacco products has raised a new set of issues for tobacco control, shifting the emphasis to building the scientific evidence and the regulatory capacity needed to quantify the harms of various tobacco and nicotine delivery devices and updating regulations as required. In addition, while many countries have stringent regulations concerning traditional marketing and advertising, many of these activities have shifted to unregulated online channels. The unfinished public health agenda includes building capacity in each country to counter the tactics of the tobacco industry, treating the diseases caused by tobacco use, finding sustainable alternatives to tobacco-growing and educating people about the health risks of smoking and other nicotine delivery devices.

The harmful use of alcohol contributes to 3 million deaths globally every year and over 5% of the burden of disease and injury (WHO 2018d). It has a direct impact on many health-related targets of the SDGs, including those for maternal and child health, infectious diseases (HIV, viral hepatitis, lower respiratory infections, tuberculosis), noncommunicable diseases (cancer, cardiovascular diseases and diabetes), digestive diseases, mental health, violence, injuries, and poisonings. Alcohol per capita consumption per year in litres of pure alcohol is one of two indicators for SDG health target 3.5 —'Strengthen the prevention and treatment of substance abuse, including narcotic drug abuse and harmful use of alcohol' (UN 2015).

Almost all countries have alcohol excise taxes and fewer than half use other price strategies, such as banning below-cost selling or volume discounts. Most countries have a combination of policy measures including taxes, limits on retail and wholesale distribution, advertising restrictions, age limits for consumption and legal limits for blood alcohol concentrations when driving (WHO 2018d).

The WHO launched a new initiative in 2018 to address the lack of progress in reducing the health harms associated with alcohol consumption.

The initiative provides governments with a technical package of five strategies: (1) strengthening restrictions on alcohol availability; (2) advancing and enforcing drink-driving countermeasures; (3) facilitating access to screening, brief interventions, and treatment; (4) enforcing bans or imposing comprehensive restrictions on alcohol advertising, sponsorship, and promotion; and (5) raising prices on alcohol through excise taxes and pricing policies.

Innovation, data, and statistics

When the Millennium Development Goal reporting period ended in 2015, 80% of the world's poor people were living in rural areas, 55% of whom had access to safe water for drinking, compared to 85% of the urban population. While considerable progress had been made, the richest and poorest people did not benefit equally. For example, between 2000 and 2015, skilled birth attendance rose from 80 to 89% in the richest quintile, but only from 33% to 43% in the poorest. Women continued to be more likely to live in poverty and to remain at a disadvantage in the labour market, earning 24% less than men globally. Equity is central to the sustainable development agenda. The world has learnt that national averages can hide gross disparities (Gurria 2011).

Part of the reason for these persistent inequities is that population level statistics have traditionally suffered from incomplete coverage; certain ethnic groups are systematically excluded, institutional populations are not included in household surveys, administrative data only reflect those who can access the services and school-based surveys can only count the children who are able to attend school in the first place (Carrera et al. 2012).

Disaggregated data are needed to track progress for the most vulnerable people, and it costs more to collect and analyse these data, to resolve issues of sample sizes, confidentiality, and transparency, particularly when indicators require multidimensional disaggregation, for example—'child labour in poorest quintile in urban areas'. The challenge is meeting the demand for high-quality, timely, disaggregated data across a very broad agenda. The UN General Assembly was clear: 'Sustainable Development Goal indicators should be disaggregated, where relevant, by income, sex,

age, race, ethnicity, migratory status, disability and geographic location, or other characteristics, in accordance with the Fundamental Principles of Official Statistics' (United Nations 2014).

The WHO welcomes this emphasis, as health systems do not naturally evolve to ensure a fair distribution of services. Decisions aimed at increasing a country's economic growth do not automatically protect the poor or promote their health. International trade agreements do not, by themselves, guarantee food security, job security, health security, or access to affordable medicines. All these outcomes require deliberate policy decisions that place equity at the forefront (Blas and Sivasankara Kurup 2010; Hoegh-Guldberg et al. 2015; FAO 2016) and the ability to measure impact across the diverse groups that constitute a national population.

To quantify its contribution to improvements in health outcomes, the WHO has developed an impact framework with three levels of measurements: health-adjusted life expectancy, an index for each goal and programmatic targets.

The WHO impact framework specifies the baseline data and the indicators by which performance should be measured. Thus, for the first time, all components of WHO's work—from guidelines and technical support, humanitarian response, policy dialogue, and health campaigns—will be measured in terms of the health impact delivered in countries and not merely by whether programmes delivered their planned outputs. This shift from output to impact responds to the demands made by governments and other donors for more transparency and accountability for public funds (OECD 2005, 2008, 2011).

Incentivizing innovation

Innovation can improve any aspect of policies, systems, products and technologies, services and delivery methods that affect people's health and well-being. The WHO is working

towards a more innovative culture in global health that promotes digital technologies for health. In 2018, the organization established an innovation hub with the Programme for Appropriate Technologies in Health (PATH). In partnership, WHO and PATH have developed vaccine vial monitors that show when the cold chain has been breached, vaccines for malaria and Japanese encephalitis, classification and guidelines for digital health

interventions, and the Global Digital Health Atlas, a technology registry platform designed to strengthen impact and facilitate scale of emerging health innovations.

When an unmet health need is identified, an ideal product is defined, and a public–private partnership is formed to develop the product. These projects usually bring together expertise from academia, industry, and public health, with funding from philanthropic sources. Other examples include the Medicines for Malaria Venture, the International AIDS Vaccine Initiative, the Global Alliance for TB Drug Development, and the Drugs for Neglected Diseases Initiative.

The Meningitis Vaccine Project illustrates the benefits of designing a product based on the needs of low-income countries (Meningitis Vaccine Project n.d.; CDC 2012). Ministers of health asked the WHO for a vaccine that could prevent annual epidemics in the 26 countries that form Africa's meningitis belt, and defined its ideal characteristics, including the price they could afford to pay.

In less than a decade, the new vaccine was developed by a consortium of scientists, licensed with support from Canada's regulatory authority, and prequalified by the WHO. The technology for producing a conjugate vaccine was transferred from the United States of America and the Netherlands to the Serum Institute of India, which agreed to manufacture the vaccine at the target price of 50 cents per dose. The first campaigns using the new vaccine began in late 2011; more than 100 million doses had been administered by the end of 2012. Following these campaigns, some hyperendemic countries have recorded their lowest level of meningitis cases in more than a decade.

The economic benefits are expected to be significant. A single case of meningitis can cost a household the equivalent of 3–4 months of income. Mounting a reactive emergency response to epidemics can absorb more than 5% of a country's entire health budget. The project is ground-breaking in another sense. African countries usually wait years, if not decades, for new medical technologies to trickle into their health systems. In this case, Africa was first to receive the best technology that the world could offer.

Technological innovation is needed for other aspects of the medicines supply chain. For example, the problems caused by substandard, spurious, falsely labelled, falsified, and counterfeit medical products have been difficult to resolve by the traditional approach of strengthening national

drug regulatory authorities and quality-assurance laboratories (Cockburn et al. 2005). WHO estimates that only 20% of countries have well-developed systems for regulating medical products—of the remaining countries, around 50% implement drug regulation at varying levels of effectiveness, and the remaining 30% have either no drug regulation in place or a very limited capacity that barely functions. Online sales of medicines further complicate efforts to protect consumers from substandard or falsified medical products (WHO n.d. a).

Past efforts to address the problem of counterfeit and substandard medicines have been hampered by the confusion of public health issues with issues of intellectual property rights and copyright infringements, and high-profile trade incidents in which legitimate generic medicines were seized, in transit, as counterfeits. In many countries, the only legal instrument available to address the public health problems caused by counterfeit and substandard medicines is legislation related to intellectual property rights (Attaran et al. 2012).

The efficiency, effectiveness, quality, sustainability, safety, and/or affordability of products can always be improved. As shown by repeated shortages of insulin (discovered in 1923) and penicillin (in 1942), the fact that a medicine is long-established is not a guarantee of its availability.

Future directions

While much progress has been made in recent decades, global health is a field in need of continued innovation. For the world to sustain a human population of 11.2 billion, major changes are needed in the way that people live, eat, work, and receive healthcare. The SDGs have an ambitious scope and several contradictions that need to be resolved. The health sector faces major challenges from epidemiological and demographic transitions and the difficult task of leaving no-one behind in its provision of high-quality health services. The WHO will need to keep building on its history of success in infectious disease control when working across all sectors to maximize the health benefit of their policies.

Global health’s contribution to sustainable development can help to ensure equitable access to resources, a healthy interface with the natural environment, and negligible variation in life expectancies (FAO 2018c).

The WHO will continue to track changes in the determinants of health and the patterns of disease, identify effective interventions, and measure the degree to which, over time, its recommendations improve health and longevity.

REFERENCES

Anand, S. and Barnighausen, T. (2004). Human resources and health outcomes: cross-country econometric study. *The Lancet*, 364(9445), 1603–9.

Anand, S. and Barnighausen, T. (2007). Health workers and vaccination coverage in developing countries: an econometric analysis. *The Lancet*, 369(9569), 1277–85.

Andrieux-Meyer, I., Cohn, J., de Araujo, E.S., and Hamid, S.S. (2015). Disparity in market prices for hepatitis C virus direct-acting drugs. *Lancet Global Health*, 3, e676–e7.

Attaran, A., et al. (2012). How to achieve international action on falsified and substandard medicines. *British Medical Journal*, 345, e7381.

Australian Bureau of Statistics (2012). *4364.0.55.007—Australian Health Survey: Nutrition First Results—Foods and Nutrients, 2011–12*. [Online] Available at: http://www.abs.gov.au/ausstats/abs@.nsf/Lookup/bySubject/4364.0.55.007~2011-12~MainFeatures~Discretionaryfoods~700

Basu, S., et al. (2019). Estimation of global insulin use for type 2 diabetes, 2018–30: a microsimulation analysis. *Lancet Diabetes & Endocrinology*, 7(1), 25–33.

Bermingham, A., et al. (2012). Severe respiratory illness caused by a novel coronavirus, in a patient transferred to the United Kingdom from the Middle East, September 2012. *Euro Surveillance*, 17(40), 20290.

Blas, E. and Sivasankara Kurup, A. (2010). *Equity, Social Determinants and Public Health Programmes*. Geneva: WHO. [Online] Available at: http://whqlibdoc.who.int/publications/2010/9789241563970_eng.pdf

Cairns, G., Angus, K., and Hastings, G. (2009). *The Extent, Nature and Effects of Food Promotion to Children: A Review of the Evidence to December 2008*. Geneva: WHO. [Online] Available at: https://www.who.int/dietphysicalactivity/Evidence_Update_2009.pdf

Carrera, C., et al. (2012). The comparative cost-effectiveness of an equity-focused approach to child survival, health, and nutrition: a modelling approach. *The Lancet*, **380**, 1341–51.

Centers for Disease Control and Prevention (CDC) (2012). Press briefing transcript. CDC, FDA, Massachusetts Department of Public Health. Joint telebriefing updating investigation of meningitis outbreak. [Online] Available at: http://www.cdc.gov/media/releases/2012/t1011_meningitis_outbreak.html

Centers for Disease Control and Prevention (CDC) (2018). Influenza infections in humans. [Online] Available at: https://www.cdc.gov/flu/swineflu/variant-cases-us.htm#table-infections

Cockburn, R., Newton, P.N., Agyarko, E.K., Akunyili, D., and White, J.H. (2005). The global threat of counterfeit drugs: why industry and governments must communicate the dangers. *PLoS Medicine*, 2(4), e100.

De Ruyter, J.C., Olthof, M.R., Seidell, J.C., and Katan, M.B. (2012). A trial of sugar-free or sugar-sweetened beverages and body weight in children. *The New England Journal of Medicine*, 367(15), 1397–406.

Ebbeling, C.B., et al. (2012). A randomized trial of sugar-sweetened beverages and adolescent body weight. *The New England Journal of Medicine*, 367(15), 1407–16.

Food and Agriculture Organization of the United Nations and the World Health Organization (FAO/WHO) (2014). *Second International Conference on Nutrition*. ICN2 2014/3 Corr.1. [Online] Available at: http://www.fao.org/3/a-mm215e.pdf

Food and Agriculture Organization of the United Nations (FAO) (2016). *The State of World Fisheries and Aquaculture 2016. Contributing to Food Security and Nutrition for All*. [Online] Available at: http://www.fao.org/3/a-i5555e.pdf

Food and Agriculture Organization of the United Nations (FAO) (2018a). FAO Food Price Index declines in June for the first time in 2018. [Online] Available at: http://www.fao.org/news/story/en/item/1143914/icode/

Food and Agriculture Organization of the United Nations (FAO) (2018b). *The State of Food Security and Nutrition in the World 2018*. [Online] Available at: http://www.fao.org/3/I9553EN/i9553en.pdf

Food and Agriculture Organization of the United Nations (FAO) (2018c). *The State of the World's Forests: Forest Pathways to Sustainable Development*. [Online] Available at: http://www.fao.org/policy-support/tools-and-publications/resources-details/en/c/1144279/

Forzieri, G., Cescatti, A., Batista e Silva, F., and Feyen, L. (2017). Increasing risk over time of weather-related hazards to the European population: a data-driven prognostic study. *Lancet Planetary Health*, 1, e200–e208.

G20 Meeting of Health Ministers (2018). Declaration 4th October. Argentina. [Online] Available at: http://www.g20.utoronto.ca/2018/2018-10-04-health.pdf

Gao, R., et al. (2013). Human infection with a novel avian-origin influenza A (H7N9) virus. *The New England Journal of Medicine*, **368**, 1888–97.

GBD 2015 Tobacco Collaborators (2017). Smoking prevalence and attributable disease burden in 195 countries and territories, 1990–2015: a systematic analysis from the Global Burden of Disease Study 2015. *The Lancet*, 389(10082),1885–906.

GBD 2016 Risk Factors collaborators (2017). Global, regional, and national comparative risk assessment of 84 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990–2016: a systematic analysis of the Global Burden of Disease Study 2016. *The Lancet*, **390**, 1345–1422.

Ghebreyesus, T.A. and Espinosa, P. (2018). Health, climate and small island states. *Bulletin of the World Health Organization*, **96**, 78–78A.

Gostin, L.O. and Katz, R. (2016). The international health regulations: the governing framework for global health security. *Milbank Quarterly*, **94**, 264–313.

Guojun, H., Fan, M., and Zhou, M. (2016). The effect of air pollution on mortality in China: evidence from the 2008 Beijing Olympic Games. *HKUST IEMS Working Paper*, **3**, 1–62.

Gurria, A. (2011). *Remarks by OECD Secretary-General Delivered at the Launch of Divided We Stand: Why Inequality Keeps Rising*. Paris: OECD. [Online] Available at: http://www.oecd.org/els/soc/dividedwestandwhyinequalitykeepsrisingspeech.htm

Guthold, R., Stevens, G.A., Riley, L.M., and Bull, F.C. (2018). Worldwide trends in insufficient physical activity from 2001 to 2016: a pooled analysis of 358 population-based surveys with 1·9 million participants. *The Lancet Global Health*, 46(10), E1077–86.

Hastings, G., et al. (2003). *Review of the Research on the Effects of Food Promotion to Children*. Glasgow: University of Strathclyde, Centre for Social Marketing. [Online] Available at: https://www.fabresearch.org/viewItem.php?id=11053

Health Action International (2015). *Inequities and Inefficiencies in the Global Insulin Market. Fact Sheet 1*. November. [Online] Available at: https://haiweb.org/wp-content/uploads/2015/11/ACCISS-Fact-Sheet-1-Inequalities-in-Insulin-Market.pdf

Henão-Restrepo, A.M., et al. (2017). Efficacy and effectiveness of an rVSV-vectored vaccine in preventing Ebola virus disease: final results from the Guinea ring vaccination, open-label, cluster-randomised trial (Ebola Ça Suffit!). *The Lancet*, 389(10068), 505–18.

Heymann, D.L., Mackenzie, J.S., and Peiris, M. (2013). SARS legacy: outbreak reporting is expected and respected. *The Lancet*, 381(9869), 779–81.

Hoegh-Guldberg, O. et al. (2015). Reviving the ocean economy: the case for action—2015. *World Wide Fund for Nature*. [Online] Available at: https://c402277.ssl.cf1.rackcdn.com/publications/790/files/original/Reviving_Ocean_Economy_REPORT_low_res.pdf?1429717323

Inter-Agency Standing Committee (2016). Final IASC system-wide level 3 (L3) activation procedures for infectious disease events. [Online] Available at: https://interagencystandingcommittee.org/principals/documents-public/final-iasc-system-wide-level-3-l3-activation-procedures-infectious

Joint United Nations Programme on HIV/AIDS (UNAIDS) (2018). UNAIDS data 2018. [Online] Available at: http://www.unaids.org/sites/default/files/media_asset/unaids-data-2018_en.pdf

Jordan, M.R., et al. (2017). Human immunodeficiency virus (HIV) drug resistance in African infants and young children newly diagnosed with HIV: a multicountry analysis. *Clinical Infectious Diseases*, 65(12), 2018–25.

Kieny, M-P., et al. (2018). *Delivering Quality Health Services: A Global Imperative for Universal Health Coverage.* World Bank Group. [Online] Available at: https://www.worldbank.org/en/topic/universalhealthcoverage/publication/delivering-quality-health-services-a-global-imperative-for-universal-health-coverage

McGinnis, J.M., Gootman, J.A., and Kraak, V.I. (eds.) (2006). *Food Marketing to Children and Youth: Threat or Opportunity?* Washington, DC: Institute of Medicine, National Academies Press. Available at: http://www.nap.edu/catalog.php?record_id=11514#toc

Meningitis Vaccine Project (n.d.). Website. [Online] Available at: http://www.meningvax.org/

Moon, S., et al. (2015). Will Ebola change the game? Ten essential reforms before the next pandemic. The report of the Harvard–LSHTM Independent Panel on the Global Response to Ebola. *The Lancet*, **386**, 2204–21.

Mullan, F., et al. (2011). Medical schools in sub-Saharan Africa. *The Lancet*, **377**, 1113–21.

Nabarro, D. (2015). WHO steps up its role in health emergencies. *Bulletin of the World Health Organization*, 93(12). [Online] Available at: http://www.who.int/bulletin/volumes/93/12/15-031215/en/

National Health and Medical Research Council (2013). *Australian Dietary Guidelines*. Canberra: National Health and Medical Research Council. [Online] Available at: https://www.nhmrc.gov.au/about-us/publications/australian-dietary-guidelines

Nayyar, G., Breman, J., Newton, P.N., and Herrington, J. (2012). Poor-quality antimalarial drugs in southeast Asia and sub-Saharan Africa. *The Lancet Infectious Diseases*, 12(6), 488–96.

Ong, A.K. and Heymann, D.L. (2007). Microbes and humans: the long dance. *Bulletin of the World Health Organization*, **85**, 422–3.

Organisation for Economic Co-operation and Development (OECD) (2005). *Paris Declaration on Aid Effectiveness*. Paris: OECD. [Online] Available at: http://www.oecd.org/development/effectiveness/34428351.pdf

Organisation for Economic Co-operation and Development (OECD) (2008). *Accra Agenda for Action*. Paris: OECD. [Online] Available at: http://www.oecd.org/development/effectiveness/34428351.pdf

Organisation for Economic Co-operation and Development (OECD) (2011). *Divided We Stand: Why Inequality Keeps Rising*. Paris: OECD. [Online] Available at: http://www.oecd.org/els/soc/dividedwestandwhyinequalitykeepsrising.htm

Packard, R. M., (1997). Malaria dreams: postwar visions of health and development in the Third World, *Medical Anthropology*, 17(3), 279–96.

Petemadi, A., et al. (2017). Mortality from different causes associated with meat, heme iron, nitrates and nitrates in the NIH-AARP Diet and Health Study: population based cohort study.

British Medical Journal, **357**, 1957.
Postigo, A., Haines, A., Neira, M., and Confalonieri, U. (2007). Will increased awareness of the health impacts of climate change help in achieving international collective action? *Bulletin of the World Health Organization*, 85(11), 826–8.
Pradhan, P., et al. (2017). A systematic study of Sustainable Development Goal (SDG) interactions. *Earth's Future*, 5(11), 1169–79.
Program for Appropriate Technologies in Health (PATH) (2017). PATH consolidated financial statements for the years ended 31 December 2017 and 2016. [Online] Available at: https://path.azureedge.net/media/documents/PATH2017-financial-statements.pdf
Prüss-Üstün, A., Wolf, J., Corvalán, C.F., Bos, R., and Neira, M.P. (2016). Preventing disease through healthy environments: a global assessment of the burden of disease from environmental risks. *World Health Organization*. [Online] Available at: https://apps.who.int/iris/handle/10665/204585
Qibin, Q., et al. (2012). Sugar-sweetened beverages and genetic risk of obesity. *The New England Journal of Medicine*, 367(15), 1387–96.
Roberts, L. and Brennan, R. (2015). Emergency public health and humanitarian assistance in the twenty-first century. In: Detels, R., Gulliford, M., Abdool Karim, Q., and Chuan Tan, C. (eds.) *Oxford Textbook of Global Public Health* (6th ed.). Oxford: Oxford University Press.
Sachs, J.D. (2012). Achieving universal health coverage in low-income settings. *The Lancet*, 380(9845), 944–7.
Sen, A. (1999). *Development as Freedom*. Oxford: Oxford University Press.
Sneddon, C., Howarth, R.B., and Norgaard, R.B. (2006). Sustainable development in a post-Brundtland world. *Ecological Economics*, **57**, 253–68.
Steffen, W., et al. (2015). Planetary boundaries: guiding human development on a changing planet. *Science*, 347(6223), 1259855.
Su C., et al. (2015). Assessing responses of cardiovascular mortality to particulate matter air pollution for pre-, during- and post-2008 Olympics periods. *Environmental Research*, **142**, 112–22.
The Ebola Outbreak Epidemiology Team (2018). Outbreak of Ebola virus disease in the Democratic Republic of the Congo, April–May 2018: an epidemiological study. *The Lancet*, 392(10143), 213–21.
United Nations (UN) (2000). *United Nations Millennium Declaration (United Nations General Assembly Resolution 55/2)*. New York: UN. [Online] Available at: http://www.un.org/millennium/declaration/ares552e.pdf
United Nations (UN) (2014). 68/261. *Fundamental Principles of Official Statistics*. [Online] Available at: http://undocs.org/A/RES/68/261
United Nations (UN) (2015). *The Millennium Development Goals Report* 2015. [Online] Available at: http://www.un.org/millenniumgoals/2015_MDG_Report/pdf/MDG%202015%20rev%20(July%201).pdf
United Nations (UN) (2015a). *70/1.Transforming Our World: The 2030 Agenda for Sustainable Development*. United Nations A/RES/70/1. [Online] Available at: https://sustainabledevelopment.un.org/post2015/transformingourworld
United Nations (UN) (2017). Department of Economic and Social Affairs, Population Division. *World Population Prospects: The 2017 Revision*. [Online] Available at: https://population.un.org/wpp/Graphs/DemographicProfiles/
United Nations Children's Fund (UNICEF), World Health Organization, World Bank Group (2018). *Levels and Trends in Child Malnutrition: Key Findings of the 2018 Edition of the Joint Child Malnutrition Estimates*. [Online] Available at: https://www.who.int/nutgrowthdb/estimates2017/en/

United Nations Development Programme (UNDP) (2017). *Rising Tides, Rising Capacity: Supporting a Sustainable Future for Small Island Developing States*. United Nations Development Programme, Alliance of Small Island States. [Online] Available at: https://www.undp.org/content/undp/en/home/librarypage/poverty-reduction/rising-tides-rising-capacity-supporting-a-sustainable-future-for.html

United Nations Environment Programme (UNEP) (2013). *Global Mercury Assessment 2013; Sources, Emissions, Releases, and Environmental Transport*. [Online] Available at: https://wedocs.unep.org/handle/20.500.11822/7984

United Nations Inter-Agency Group for Child Mortality Estimation (2018). *Levels and Trends in Child Mortality: Report*. [Online] Available at: https://www.un.org/en/development/desa/population/publications/mortality/child-mortality-report-2018.asp

United States Food and Drug Administration (USFDA) (2013). *Strategic Plan for Preventing and Mitigating Drug Shortages*. [Online] Available at: https://www.fda.gov/media/86907/download

Vujicic, M., Zurn, P., Diallo, K., Adams, O., and Dal Poz, M. (2004). The role of wages in the migration of health care professionals from developing countries. *Human Resources for Health*, 2(1), 3.

Wagstaff, A., et al. (2017). Progress on catastrophic health spending: results for 133 countries. A retrospective observational study. *Lancet Global Health*, 6(2), e169–79.

Wagstaff, A., Flores, G., Smitz, M-F., Hsu, J., Chepynoga, K., and Eozenou, P. (2018). Progress on impoverishing health spending: results for 122 countries. A retrospective observational study. *Lancet Global Health*, 6(2), e180–92.

World Bank (2004). *World Development Report. Making Services Work for the Poor*. Washington, DC: World Bank.

World Commission on Environment and Development (1987). *Our Common Future*. New York: Oxford University Press.

World Health Organization (WHO) (1948). Constitution of the World Health Organization. Geneva: WHO. [Online] Available at: http://apps.who.int/gb/bd/PDF/bd47/EN/constitution-en.pdf?ua=1

World Health Organization (WHO) (2008a). *International Health Regulations (2005)* (2nd ed.). Geneva: WHO. [Online] Available at: http://whqlibdoc.who.int/publications/2008/9789241580410_eng.pdf

World Health Organization (WHO) (2008b). *Protecting Health from Climate Change—World Health Day 2008*. Geneva: WHO. [Online] Available at: http://www.who.int/world-health-day/previous/2008/en/index.html

World Health Organization (WHO) (2010). *The World Health Report 2010. Health Systems Financing: The Path to Universal Coverage.* Geneva: WHO.

World Health Organization (WHO) (2010a). *Hidden Cities: Unmasking and Overcoming Health Inequities in Urban Settings.* Geneva: WHO.

World Health Organization (WHO) (2010b). *Global Code of Practice on the International Recruitment of Health Personnel*. Geneva: WHO. [Online] Available at: http://www.who.int/hrh/migration/code/practice/en/

World Health Organization (WHO) (2011). *Pandemic Influenza Preparedness Framework for the Sharing of Influenza Viruses and Access to Vaccines and Other Benefits*. Geneva: WHO.

World Health Organization (WHO) (2012). *Global Progress Report on Implementation of the WHO Framework Convention on Tobacco Control*. Geneva: WHO. [Online] Available at: https://www.who.int/fctc/reporting/2012_global_progress_report_en.pdf?ua=1

World Health Organization (WHO) (2013). MPOWER in action: defeating the global tobacco epidemic. Geneva: WHO. [Online] Available at: https://www.who.int/tobacco/mpower/publications/mpower_2013.pdf?ua=1

World Health Organization (WHO) (2013a). Noncommunicable diseases: fact sheet. [Online] Available at: http://www.who.int/mediacentre/factsheets/fs355/en/
World Health Organization (WHO) (2013b). WHO model lists of essential medicines. [Online] Available at: http://www.who.int/medicines/publications/essentialmedicines/en/index.html
World Health Organization (WHO) (2013c). Prequalification of medicine by WHO: fact sheet. [Online] Available at: http://www.who.int/mediacentre/factsheets/fs278/en/index.html
World Health Organization (WHO) (2013d). *Protocol to Eliminate Illicit Trade in Tobacco Products*. Geneva: World Health Organization. [Online] Available at: http://apps.who.int/iris/bitstream/10665/80873/1/9789241505246_eng.pdf
World Health Organization (WHO) (2014). Quantitative risk assessment of the effects of climate change on selected causes of death, 2030s and 2050s. [Online] Available at: https://www.who.int/globalchange/publications/quantitative-risk-assessment/en/
World Health Organization (WHO) (2015). *Guideline: Sugars Intake for Adults and Children*. Geneva: WHO. [Online] Available at: http://apps.who.int/iris/bitstream/10665/149782/1/9789241549028_eng.pdf?ua=1
World Health Organization (WHO) (2015a). *Report of the Ebola Interim Assessment Panel*. Geneva: WHO. [Online] Available at: https://www.who.int/csr/resources/publications/ebola/ebola-panel-report/en/
World Health Organization (WHO) (2015b). *Sixty-eighth World Health Assembly Opens in Geneva*. Geneva: WHO. [Online] Available at: https://www.who.int/mediacentre/news/releases/2015/wha-18-may-2015/en/
World Health Organization (WHO) (2016). *Global Strategy on Human Resources for Health: Workforce 2030*. Geneva: WHO.
World Health Organization (WHO) (2016a). *Eliminating Malaria in the Greater Mekong Subregion: United to End a Deadly Disease*. Geneva: WHO.
World Health Organization (WHO) (2016b). *An R&D Blueprint for Action to Prevent Epidemics*. Geneva: WHO.
World Health Organization (WHO) (2017). *Global Action Plan on HIV Drug Resistance 2017–2021*. Geneva: WHO.
World Health Organization (WHO) (2017a). *Emergency Response Framework* (2nd ed.). Geneva: WHO.
World Health Organization (WHO) (2017b). *United Nations Decade of Action on Nutrition 2016–2025. Workplan*. Geneva: WHO. [Online] Available at: https://www.who.int/nutrition/decade-of-action/workprogramme-doa2016to2025-en.pdf?ua=1
World Health Organization (WHO) (2018). *13th General Programme of Work 2019–2023*. Geneva: WHO. [Online] Available at: http://www.who.int/about/what-we-do/gpw-thirteen-consultation/en/
World Health Organization (WHO) (2018a). *Immunization Coverage*. *Fact sheet*. Geneva: WHO. [Online] Available at: https://www.who.int/news-room/fact-sheets/detail/immunization-coverage
World Health Organization (WHO) (2018b). *Statement of the Nineteenth IHR Emergency Committee Regarding the International Spread of Poliovirus*. Geneva: WHO. [Online] Available at: https://www.who.int/news-room/detail/30-11-2018-statement-of-the-nineteenth-ihr-emergency-committee-regarding-the-international-spread-of-poliovirus
World Health Organization (WHO) (2018c). EXECUTIVE BOARD. EB144/9. 144th session. 26 November 2018. Provisional agenda item 5.3. Polio Eradication. Report by the Director-General. Geneva: WHO.
World Health Organization (WHO) (2018d). *Global Status Report on Alcohol and Health 2018*. Geneva: WHO.

World Health Organization (WHO) (2018e). *Global Status Report on Road Safety*. Geneva: WHO.
World Health Organization (WHO) (2018f). *World Malaria Report*. Geneva: WHO.
World Health Organization (WHO) (2018g). *Fact Sheet: Climate Change and Health*. Geneva: WHO.
World Health Organization (WHO) (2018h). *Annual Review of Diseases Prioritized under the Research and Development Blueprint*. Geneva: WHO.
World Health Organization (WHO) (2018i). *Global Tuberculosis Report*. Geneva: WHO.
World Health Organization (WHO) (2018j). *World Malaria Report*. Geneva: WHO.
World Health Organization (WHO) (2018k). *Towards a Global Action Plan for Healthy Lives and Well-Being for All*. Geneva: WHO.
World Health Organization (WHO) (2018l). *The Pandemic Influenza Preparedness Framework: Partnership Contribution High Level Implementation Plan I Final Report 2014–2017*. Geneva: WHO.
World Health Organization (WHO) (n.d.). Attacks on health care. Geneva: WHO. [Online] Available at: https://www.who.int/emergencies/attacks-on-health-care/en/
World Health Organization (WHO) (n.d. a). General information on counterfeit medicines. Geneva: WHO. [Online] Available at: http://www.who.int/medicines/services/counterfeit/overview/en/index1.html
World Health Organization and the International Bank for Reconstruction and Development/The World Bank (WHO/IBRD) (2017). *Tracking Universal Health Coverage: 2017 Global Monitoring Report.* Geneva: WHO. [Online] Available at: https://www.who.int/healthinfo/universal_health_coverage/report/2017/en/
World Health Organization and the World Economic Forum (2011). *From Burden to 'Best Buys': Reducing the Economic Impact of NCDs in Low-and Middle-Income Countries*. Geneva: WHO. [Online] Available at: https://www.who.int/nmh/publications/best_buys_summary/en/
Worm, B., et al. (2006). Impacts of biodiversity loss on ocean ecosystem services. *Science*, 314(5800), 787–90.
Yang, Q., Zhang, Z., and Gregg, E.W. (2014). Added sugar intake and cardiovascular diseases mortality among US adults. *JAMA Internal Medicine*, 174(4), 516–24.
Zhang, P., et al. (2010). Global healthcare expenditure on diabetes for 2010 and 2030. *Diabetes Research and Clinical Practice*, **87**, 293–301.

Index Vol 1

Notes: Tables, figures and boxes are indicated by an italic *t*, *f* and *b* following the page number

abacavir 153
Abbvie 65
  t
abortion legislation 94
absolute poverty 103–104, 104
  f
  numbers 104
  public health effects 105
ACC *see* American College of Cardiologists/American Heart Association (ACC/AHA)
accidents 15
  t
accountability, healthcare rights 279
Acheson Report 3, 323
acid rain 16
Action on Smoking and Health (ASH) 329
Action Plan for its Global Strategy on Non-communicable Disease 297
Addis Ababa Agenda for Action (AAAA) 141
adolescent mortality 98–99

adults
public health policy in developed countries 324
f, 327
survival in developed countries 323–324, 324
f, 325
t, 326
advanced molecular detection (AMD) 151
adverse drug reactions (ADRs), genetics 152–153
advertising, tobacco industry 220
Advisory Committee on Immunization Practices (US) 263
Aedes aegypti, dengue fever transmission 198
Aeromonas infection 166
Affordable Care Act (2012) 17, 233
Afghanistan 140
endemic polio 35

Africa
health inequalities 140
HIV epidemic 49
water & sanitation improvements 172 *see also* sub-Saharan Africa
African Commission on Human and Peoples' Rights 278
African Health Research System Network 40
age *see also* children; elderly
Agency of Healthcare Research and Quality (AHRQ) 151
agentic interventions, healthcare reduction 133
agento-structural interventions, healthcare reduction 133
Age of Degenerative and Man-Made Diseases 301
Age of Pestilence and Famine 301–302
Age of Receding Pandemics 302
agriculture, climate change and 200
agri-food industry 87, 182
globalization 72
AHA *see* American College of Cardiologists/American Heart Association (ACC/AHA)
AIDS 6
ethical issues 263
legal aspects 297 *see also* HIV infection
air conditioning 202
air pollution 4
f, 15–16, 15
t
air quality 4
alcoholic beverage industry 66
socioeconomic health inequalities 129
alcohol use disorders 13
t
excise taxes 58
b
LMICs, reduction in 41
Alison, W. P. 27
Allende, Salvador 113
allostatic load 129
Alma-Ata Conference (1978) 229, 278
Alzheimer's disease mortality 6
f
AMD (advanced molecular detection) 151
America *see* United States of America (USA)
American Association of Medical Colleges 261
American College of Cardiologists/American Heart Association (ACC/AHA) 331
American Medical Association (AMA) 261
American Public Health Association (APHA) 28
anaemia, undernutrition 180
analytical validity, genetic testing 151
Ancylostoma duodenale infection 166
t, 167
Angola 49
t

animals
antibiotic use 55
reared for human consumption 182
Annales d'Hygine Publique et de Médecine Légal 26

antibiotics
animal use 55
consumption increase 55
resistance to 10–11, 11
b, 55–56
surveillance programmes and 55
Anti-Cancer Council of Vienna 329
antifamine strategies, Qing dynasty (China) 25
antiretroviral therapy (ART) 52
anxiety disorders 13
t
APHA (American Public Health Association) 28
ART (antiretroviral therapy) 52
Ascaris lumbricoides infection (ascariasis) 166
t, 167
transmission barriers 168
ASH (Action on Smoking and Health) 329
Asia-Pacific Health Research System Network 40
associate model of poverty reduction 116
Association for the Taxation of Financial Transaction for the Aid of Citizens (ATTAC) 117
Association of State and Territorial Health Officials (ASTHO) Leadership Institute 254
b
AstraZeneca 65
t
atherosclerosis, *Chlamydia pneumoniae* infection 11
ATTAC (Association for the Taxation of Financial Transaction for the Aid of Citizens) 117

Australia
Aboriginal people smoking rates 88
cardiovascular disease 332
t
lung cancer 330
t
National Heart Foundation 331
premature deaths 325
t
public health history 27–28
Royal Australasian College of Surgeons 329

Austria
  cardiovascular disease 332
    t
  child deprivation 109
    f
  lung cancer 330
    t
  premature deaths 325
    t
Austro-Hungarian Empire 28
autocratic leadership 248–249
automatic protective strategies, behaviour 219–220
Ayushman Bharat programme (India) 316
  b
5-aza-20-deoxycytidine 154, 155
5-azacytidine 154, 155

bacille Calmette–Guérain (BCG) vaccination 34, 35
  f
Baltic countries, tick-borne encephalitis 198
Bandura, Albert 221
Bangkok Charter for Health Promotion (2005) 70

Bangladesh
health inequality reduction 145
b
infant mortality 49
t
BCG vaccination 29–30 *see also* tuberculosis
behaviour 4, 213–227
causal link with health 214
causes of 217–219, 218
f
definition 213
denormalizing behaviour 223
determinants of health 83
disease risk factors 215–217, 216
t
education and 223
enabling factors 218, 219
t
environment and 220–222
genetics and 221–222, 223–224
interventions as 18–19
obesity and 71
population health 215–217, 215
f
poverty and 112
predispositions of 220–222
public health strategies 219–220
reinforcing factors 218–219, 219
t
relationships between 224
scope of 213
social determinants of health 87–88
societal development of 213–215
socioeconomic health inequality reduction 133
socioeconomic status and 222–224
urbanization 50 *see also* complex behaviours

Belgium
cardiovascular disease 332
t
child deprivation 109
f
lung cancer 330
t
nutrition improvements 188
premature deaths 325
t
Belmont Report (1979) 18, 261
Bentham, Jeremy 26
Best Practices for Biospecimen Resources (National Cancer Institute) 268
beta thalassemia 157
BFP (Bolsa Familia Programme) 187
Big Data, right to health 283–284
b
Big Pharma 66
biodegradable pollutants 16
bioethics 261
biofuel industry 182
biological determinants of health 83
biological samples, ethical issues 268
bioterrorism 17
bipolar disorder 13
t
Black report 134
blood donation, HIV infection/AIDS 263

blood pressure
BMI (body mass index) 180
body mass index (BMI) 180
Bolivia 49
t
Bolsa Familia Programme (BFP) 187
Bourbon virus infection 304

Brazil
Family Health Programme 273
health inequality reduction 143, 146
b
infant mortality 49
t
nutrition improvements 187
politics in public health 308
population demographics 50
t
poverty reduction by government reforms 116–117
primary healthcare financing 235
Universal Health Coverage 57
b
BRCA1/2 gene 150
breastfeeding, undernutrition and 180
British Heart Foundation 331
bronchial cancer 6
f
building design, climate change 202, 204
built environment, legal aspects 294

Bulgaria
  cardiovascular disease 332
    t
  child deprivation 109
    f
  lung cancer 330
    t
  premature deaths 325
    t
burial counts, public health history 24
businesses, legal regulation 295

Cameroon 49
  t
Campylobacter infection 166
  t
Campylobacter jejuni infection 166

Canada
cardiovascular disease 332
t
International Confederation of Midwives Triennial Congress (Toronto) 305
lung cancer 330
t
premature deaths 325
t
Public Health Agency of Canada 67
public health policy 326
cancer mortality 216
t
LMICs 37–38
socioeconomic status and 124
capability, health inequality reduction 145
capacity, health policy in developing countries 315–316
carbohydrate intake 181
carbon dioxide (CO_2) 193
obesity and 215
public health policy in developed countries 331, 332
t, 333
voluntary organizations 331
cardiovascular disease, LMICs 37
Caribbean, absolute poverty 104

case studies
developing countries, health policy 310
b
income inequality 99
infectious disease law 296–297
leadership 253–254
b
leadership development 250–251
b
legal aspects 296–299
needle-exchange programmes 95
non-communicable diseases 297–298, 298
t
women's reproductive health 97
cause-specific mortality 124
CEDAW (Committee on the Elimination of Discriminating Against Women) 279
Center for Creative Leadership 250
Centers for Disease Control and Prevention (CDC)
Laboratory Leadership Service 253–254
b
National Leadership Academy for the Public Health 253–254
b
pathogen genomics 151
cervical cancer 11
LMICs 37
CESCR (Committee on Economic, Social and Cultural Rights) 279
Chadwick, Edwin 26–27, 112
challenges management 256–257
Charities Act (1601) 24
CHD *see* coronary heart disease (CHD)
chemical food additives 181
chemical pathogens, water & sanitation 165
Chicago, public health history 28
Child Growth Reference (WHO) 180
child labour 17
child marriage cessation of 273–274
childrearing practices, developed countries 327

children
deprivation of 109
f
growth & nutritional status 180
health inequalities 140
health trends in developed countries 327
mortality & income inequality 98–99
mortality & maternal behaviour 217
right to health & climate change 281–282
b
safe disposal of faeces 172
survival in developed countries 323, 324
f
Children's Fund (UN) 74

China
food supplies 183
greenhouse gas emissions 201
politics in public health 308
population demographics 50
t
Chlamydia pneumoniae infection 11
Chlamydia trachomatis infection 36–37, 166
t, 167
chloroquine-resistant *Plasmodium falciparum* 305
cholera 69
control/elimination 34
Christian church, public health history 24–25
chromosomal abnormalities 149

chronic diseases
ethical issues 262
prevalence in low- and middle-income countries 232
prevalence of 6
chronic heart disease (CHD)
chronic lower respiratory disease 216
t
chronic obstructive pulmonary disease (COPD) 6
f
chronic respiratory disease 37
CHWs (community-based health workers) 311–312
civil litigation 295
civil rights protection 292
civil society, globalization and 117
climate change 16, 54–56, 193–212
adaptations 201–205
burden of disease 199–200, 199
f
children's right to health 281–282
b
direct health effects 54
f, 55
forecasting 198–201
health policy in developing countries 304
impediments on action 208
indirect health effects 54
f, 55
mitigation of 205–207
observational studies 196–198, 197
f
observed & simulated 195
f
present impacts 198
quiz questions & answers 194
b, 207–208
size of 196
b
social determinants of health 88
urbanization 50 *see also* environment
Clinical Decision Support Systems (CDSS) 314
clinical trials, ethical issues 267
clinical utility/validity, genetic testing 151
CLTS (community-led total sanitation) 173
clustered regularly interspaced short palindromic repeats (CRISPR) 156
advantages/disadvantages 157
Cas9 156
clinical trials 156–157
ethics 157–158
CMNN (communicable, maternal, neonatal, and nutritional disease) 50
Cochrane, Archie 231

Codex Alimentarius Commission 182
coercive strategies, behaviour 220

cognition
behaviour and 217–219, 218
f
change & globalization 64
cognitive bandwidth 129
collaborative leadership 247–248, 248
b
collective action, non-communicable diseases 70
Colombia 49
t
colonialism, low- and middle-income countries 33

commercialization
health policy in developing countries 308
USA public health 335
Commission on Intellectual Property Rights, Innovation and Public Health (WHO) 75
Commission on Social Determinants of Health (WHO) 113, 123, 131, 143
Social Exclusion Knowledge Network 110
Committee on Economic, Social and Cultural Rights (CESCR) 279
Committee on the Elimination of Discriminating Against Women (CEDAW) 279
communicable diseases 9–11
health policy in developing countries 301
mortality from 7–8
newly identified 10
b
reduction in 6
communicable, maternal, neonatal, and nutritional disease (CMNN) 50
communication, leadership 245
community-based health workers (CHWs) 311–312

community factors
climate change interventions 202
disease burden change 232
individual rights *vs.* 261–262
malnutrition prevention programmes 185
community-led total sanitation (CLTS) 173
community socioeconomic deprivation (CSD) 68
competencies, leadership 252
t

complex behaviours
  definition 213
  synergistic causal pathways 214–215
confidence, leadership 246
conflicts *see* war
CONSEA (National Council on Food and Nutrition Security) 187
consent, ethical issues 268
conservative model of poverty reduction 115
contemporary health issues 5–7
contingency approach, leadership 244–245
continual learning, leadership 256

contraception
accessibility & availability 96
United States Supreme Court 97
Convention on the Rights of the Child (CRC) 274, 276, 298
coordination, health inequality reduction 144
COPD (chronic obstructive pulmonary disease) 6
f
coronary heart disease (CHD)
physical activity and 215
tobacco smoking and 214
Council for International Organization of Medical Science (2002) 18
counter advertising, tobacco industry 72–73, 73
f
country health planning (CHP) 39
CRC (Convention on the Rights of the Child) 274, 276, 298
CRISPR *see* clustered regularly interspaced short palindromic repeats (CRISPR)
critical period model 130

Croatia
cardiovascular disease 332
t
lung cancer 330
t
premature deaths 325
t
crop growing periods, climate change 196
cross-boundary effects, health determinants 89
Cryptosporidium infection 166
Cryptosporidium parvum infection 166
t
CSD (community socioeconomic deprivation) 68
CTX001-121 157

Cuba
climate change 203, 203
t
politics in public health 308
primary healthcare financing 235
Cullen, William 25
cultural explanations, poverty 112
cycling 207
Cyclone Nargis (2008) 202–203
CYP2D6 gene 153

Cyprus
cardiovascular disease 332
t
child deprivation 109
f
lung cancer 330
t
premature deaths 325
t
cytarabine 155
cytochrome P450 family 153

Czech Republic
cardiovascular disease 332
t
child deprivation 109
f
lung cancer 330
t
premature deaths 325
t

DALYs *see* disability-adjusted life years (DALYs)
Daniels, Norman 233

data
collection, ethical issues 266
quality in developing countries 308–309
storage, ethical issues 158–159
DBM *see* double burden of malnutrition (DBM)
death *see* mortality
decentralization, developing countries 314–315
decisiveness, leadership 246
Declaration of Helsinki (2002) 18
Declaration on the Rights of Indigenous Peoples (DRIP) 278
Declaration on the Right to Development (1986) 114
definition of public health 3
Delhi (India)
greenhouse gas emissions 207
Mohalla clinics 236
delivering as one (DaO) 39
dementia 6
f
democratic leadership 248
demographic dividend 49

demography
climate change projections 200
income and 48–49, 49
t, 50
t
shift in 7, 8
f
transition in developing countries 301
dengue fever 198, 304–305

Denmark
cardiovascular disease 332
t
child deprivation 109
f
lung cancer 330
t
premature deaths 325
t
denormalizing behaviour 223
dental health 14
depression 13
t
climate change 198
deprivation amplification hypothesis 129–130
deregulation, legal aspects 296
determinants of health 83–91
behavioural determinants 83
biological determinants 83
cross-boundary effects 89
environmental determinants 83
multiple scales & multiple determinants 89
political challenges 89–90
social determinants *see* social determinants of health (SDH)
temporality of 88–89

developed countries
definitions 323
public health policy *see* developed countries, public health policy
survival trends 323–326 *see also* high-income countries
developed countries, public health policy 323–338
adult health trends 324
f, 327
cardiovascular disease 331, 332
t, 333
child health trends 327
institutions 326
knowledge 326–327
large scale failures 333–335
strategy globalization 335–336
tobacco smoking 327–329
traffic injuries 329, 331

developing countries
causes of death 216
health inequality reduction *see* developing countries, health inequality reduction
health policy *see* developing countries, health policy
developing countries, health inequality reduction 139–148
Agenda 2030 141–142
context definitions 142–143
growth in inequalities 143–144
methods for 144–145
social participation 145–146
urban transition 142
developing countries, health policy 301–321
actors in 307–308
capacity building 315–316
case study 310
b
current challenges 304–305
decentralization 314–315
evidence-based decision-making 314
finance 309, 315–316
governance 308–309
health transition 301–302
improvement strategies 312–317
inequality disparities 316–317
information 309–310
infrastructure 310–311, 315–316
logistics 310–311
maternal & child health 305–306
medical supplies 310–311
nutrition 306–307
out of programme experience 312–313
political & economic effects 308
private–public partnership 313–314
risk pooling 312–313, 313
f
service delivery 311
Sustainable Development Goals 317
technology 310–311
training 315–316
universal health care 317
workforce 311–312
development agencies, developing countries 315

diabetes mellitus
  disability-adjusted life years 180
  LMICs 37
  mortality from 6
    f
  type 1 215
  type 2 215 *see also* United Kingdom Prospective Diabetes Study (UKPDS)
diabetic retinopathy 155
diagnostic genomic testing 150
diarrhoeal disease 165–166
  climate change effects 201
  failure to treat 173
  infections 166
    t
  mortality 163
  mortality from 6
    f
  transmission barriers 168, 169
    t
Die Medicinische Reform 27
diet 180–181
  fat intake 181
  patterns of 180–181, 214
  socioeconomic health inequalities 129 *see also* nutrition
Difference Principle (Rawls) 265–266
digital health technologies 58
dignity, healthcare rights 279
directly observed treatment short course (DOTS) 18
disabilities 126–127
disability-adjusted life years (DALYs) 7–8, 9
  f
  malnutrition 180
  urbanization 50

diseases
climate change effects 200
occurrence measurement 7–9
socioeconomic health inequalities 126–127
disease vectors 196
DMD (Duchenne muscular dystrophy) 157
DNA methylation 154
DNA methyltransferase (DNMT) enzyme inhibitors 154, 155
Doha Declaration (2001) 75
domestic law, infectious diseases 297
DOTS (directly observed treatment short course) 18
double burden of malnutrition (DBM) 179–180, 306
interventions for 184–185, 186
f
Down's syndrome 158
DPT (diphtheria, pertussis and tetanys) vaccine 34, 35
f
dracunculiasis 37, 165
drinking water 38, 163
DRIP (Declaration on the Rights of Indigenous Peoples) 278
drug development, pharmacogenomics 153–154
Duchenne muscular dystrophy (DMD) 157
duties of public health 291–292
dysentery 165

EAE (experimental autoimmune encephalomyelitis) models 155
early child development, social determinants of health 85
East Africa 49

East Asia
absolute poverty 104

Eastern Europe
adult survival rates 324, 326
public health failures 333
eating disorders 13
t
Ebola epidemic (2014-2015) 69, 263–264
ECDC (European Centre for Disease Control) 67
ecological systems, climate change 195–196
economic exclusion 109
economic factors 48–50
climate change 200, 201
determinants of health 89–90
developing countries 308
transition & public health 47–62
economic globalization 64, 140
health policy in developing countries 304 *see also* globalization
Economist Intelligence Unit (EIU) 234

education
behaviour and 219
social determinants of health 85
EGAPP (Evaluation of Genomic Applications in Practice and Prevention) 151

elderly
disease occurrence 13
increase in 17
LMICs 42
population growth 13
Electronic Patient Registries (EPR) 314
El Niño Southern Oscillation (ENSO) 193
emerging diseases, globalization 69
emotional intelligence (EI) 246–247, 247
b
employment, social determinants of health 85–86
empowerment, human rights & public health 276

enabling factors
behaviour 218, 219
t
education & behaviour 223
socioeconomic status & behaviour 222–223
ENAP (Every Newborn: an action plan to evade preventable health) 305
energy efficiency, housing 206, 207
t
enforcement of public health 3–4
Engels, Friedrich 112
England 125
t see also United Kingdom (UK)
Ensemble Prévenons l'Obésité Des Enfants (EPODE) 187–188
ENSO (El Niño Southern Oscillation) 193
Entamoeba histolytica infection 166
t

enteric pathogens
climate change 203
transmission 171
enteropathy 173
enterotoxigenic *Escherichia coli* producing heat-stable toxin (ST-ETEC) 166
environment 4, 15–17
air pollution *see* air pollution
behaviour cause 217–219, 218
f
behaviour, influence on 221
determinants of health 83
health concerns 66–67
health hazards 132
repair of 20
rescue of 16–17 *see also* climate change
environmental (tropical) enteropathy 173
EPHFs (essential public health functions) 67, 67
b
EPI (Expanded Program of Immunization) 34
epidemics 301–302
ethical issues 263–265
epidemiologic transition, developing countries 301
epigenetic therapy 154–155
advantages/disadvantages 155–156
future work 155–156
epigenomic technologies 155
epigenomic-wide association studies (EWAS) 155
EPMM (strategies towards ending preventable maternal mortality) 305
EPODE (Ensemble Prévenons l'Obésité Des Enfants) 187–188
EPR (Electronic Patient Registries) 314

equality
healthcare rights 279
promotion in inequality reduction 144
Ericka Grimes v. Kennedy Krieger Institute, Inc (2001) 267
Escherichia coli infection 166
t
enterotoxigenic *Escherichia coli* producing heat-stable toxin 166
essential public health functions (EPHFs) 67, 67
b

Estonia
cardiovascular disease 332
t
child deprivation 109
f
lung cancer 330
t
premature deaths 325
t
ethical issues 18, 261–272
Ethiopia 37
genetic information and 221
genetic research 268
genomics 158–159
global epidemics 263–265
health promotion and 262
history of 261–262
infectious diseases 263
non-research activities of public health 266
pandemic preparedness 263
prevention of disease 262
public health research 266, 267–268
ethnicity 333–334
European Centre for Disease Control (ECDC) 67
European Heart Network 331
European Union (EU) 335–336
Evaluation of Genomic Applications in Practice and Prevention (EGAPP) 151
evaluation of health issues 67–68
Every Newborn: an action plan to evade preventable health (ENAP) 305
evidence-based decisions 308, 314
EWAS (epigenomic-wide association studies) 155
excise taxes 58
b
exclusionary forces, health inequality 145
excreta disposal 164
exercise *see* physical activity
experimental autoimmune encephalomyelitis (EAE) models 155
Expert Committee on Drug Dependence (WHO) 95

facilitation, leadership 245
Faculty of Public Health Good Practice Framework (UK) 251
b
faecal–oral disease
diarrhoeal disease transmission 165–166
infection route 164
transmission barriers 167–168, 167
f
Family Health Programme (Brazil) 273

FAO (Food and Agriculture Organization), double burden of malnutrition 179
farm sizes 182
Farr, William 26
fasting plasma glucose (FPG) 180
FCTC *see* Framework Convention on Tobacco Control (FCTC)
FDA (Food and Drug Administration) 97
federal funding, needle-exchange programmes 96
fertility, health inequalities 142
fetal programming, growth patterns 130–131

finance
speculation in food industry 182–183
Universal Health Coverage 57 *see also* income

Finland
cardiovascular disease 332
t
child deprivation 109
f
lung cancer 330
t
mortality rates & socioeconomic status 125
t
North Karelia programme 88
premature deaths 325
t
fluoridation, water supplies 219
foetal growth 180
Food and Agriculture Organization (FAO), double burden of malnutrition 179
Food and Drug Administration (FDA) 97
food environment 181

food industry
  companies in 182
  financial speculation 182–183
Food Marketing to Children and Youth: Threat or opportunity (Institute of Medicine) 71–72

foods
cardiovascular disease 331, 333
insecurity & climate change 55
nutrient content 58
prices 182
processed *see* processed foods
production increase 181
formulation, public health issues 3–4
FPG (fasting plasma glucose) 180
Framework Convention Alliance 72
Framework Convention on Climate Change (1992) (UN) 207
Framework Convention on Tobacco Control (FCTC) 75, 87, 298 *t*, 335
LMICs 40
Protocol to Eliminate Illicit Trade in Tobacco Products 72

France
cardiovascular disease 332
t
child deprivation 109
f
comparative epidemiology 28
history of social health 25–26
lung cancer 330
t
mortality rates & socioeconomic status 125
t
nutrition improvements 187–188
post-Napoleonic era 26
premature deaths 325
t
public health history 27
public health policy 326
sugary soft drink bans 188
Frank, Johann Peter 28
freedom restrictions, public health 292
The Function of Public Health (Institute of Medicine) 67
functions of public health 3–5, 3
b
future of public health 20
The Future of the Public Health in the 21st Century (Institute of Medicine) 67, 335

Gaffney, Adam 233
GASP (Group Against Smoke Pollution) 329
Gates Foundation 29

gender
behaviour & socioeconomic status interactions 223–224
healthcare in developing countries 316
social determinants of health (SDH) 86
water & sanitation 170–171
General Agreement on Tariffs and Trade (GATT) 65
General Medical Council (GMC) (UK)
ethical issues 261
leadership 246
genetics 215
f
behaviour cause 217–219, 218
f
behaviour, influence on 221–222
diversity 154
ethical issue of research 268
information & ethics 221
pathogen genomics 151
predisposition in behaviour 220–222
samples & ethics 267
socioeconomic health inequalities, high-income countries 130
variation 149 *see also* genomics

genetic testing
ethical issues 268
evaluation 151–152
preimplantation genetic testing 150–151
prenatal genetic testing 150–151
Genetic Testing Evidence Tracking Tool (GETT) 151
Genetic Testing Registry (GTR) 150
GeneXpert MTB/RIF test 151
genome editing 156–158
genome-wide association studies (GWAS) 149
drug suitability 154
genomics 149–162
diagnostic testing 150
drug development 153–154
epigenetic therapy 154–155
ethical implications 158–159
genetic testing evaluation 151–152
genome editing 156–158
hereditary breast & ovarian cancer 150
newborn screening 150
pathogen genomics 151
pharmacodynamics 153
pharmacogenomic testing 152–153
pharmacokinetics 153
predictive genomic testing 150
predictive testing 150
preimplantation & prenatal testing 150–151 *see also* genetics
genotyping, drug suitability 154
geographical factors, health policy 311

Germany
cardiovascular disease 332
t
child deprivation 109
f
lung cancer 330
t
premature deaths 325
t
public health history 27, 29
Ghana 49
t
GHGs (global health groups) 75
GHO (Global Health Observatory) 68
Giardia lambia infection 166
t
Gilead 65
t
Gini index 98, 98
t, 139, 140
GlaxoSmithKline 65
t
Global Access to Healthcare: Building Sustainable Health Systems (Economist Intelligence Unit) 234
Global Action Plan for the Prevention and Control of Non-Communicable Diseases (2013-2020) (WHO) 72
Global Alliance for Vaccines and Immunization (GAVI) 34
global disease burdens 232
Global Eradication of Poliomyelitis (1988) 34–35
global financial crisis 75
global food production, increase 181
Global Gag Rule (Mexico City) 96–97, 97
f
Global Health and Foreign Policy (2012) (UN) 234
Global Health Expenditure database (WHO) 12
global health groups (GHGs) 75
Global Health Observatory (GHO) 68
global health promotion 70–74
globalization 63–76
civil society and 117
definition 63–67
drivers of 64
economics 65–67
environmental health concerns 66–67
essential public health functions 67, 67
b
health policy in developing countries 302, 304
health status 67–68
illegal products 67
information & communication technologies 73–74

information technology 73–74
social determinants of health 87
Global Onchocerciasis Control Program 37
global public health 74–75
Global Strategies in Diet, Physical Activity, and Health (WHO) 72
global village 63–64
GMC *see* General Medical Council (GMC) (UK)
GNI (gross national income) 33
GNP (gross national product) 25
Godber, George 328
Going Universal (World Bank) 56–57
Good Professional Practice Framework 253
b
governance, health policy in developing countries 308–309
government policies, socioeconomic health 133
Grant, James 244
Great Britain *see* United Kingdom (UK)
great leader theory 244

Greece
cardiovascular disease 332
t
child deprivation 109
f
lung cancer 330
t
premature deaths 325
t
public health history 29
greenhouse gases 193, 194
emission reduction 89, 201, 204, 205–206
political interventions 19
green revolution 181
Gregory, John 261
gross national income (GNI) 33
gross national product (GNP) 25

growth
fetal programming 130–131
undernutrition 180
GTR (Genetic Testing Registry) 150
guiding principles, leadership 256
GWAS *see* genome-wide association studies (GWAS)

Haemophilus influenzae (Hib3) vaccination 34, 35
f, 36
harm-reduction strategies
politics of public health 94–96 *see also* needle-exchange programmes (NEPs)
Havasupai Trive v. Arizona Board of Regents 267
Haygarth, John 25
HBOC (hereditary breast and ovarian cancer) 150
HCC (hepatocellular cancer) 11
HDAC (histone deacetylase) inhibitors 154
HDI (Human Development Index) 106

health
income and 48
f
inequalities 4, 140–141 *see also* developing countries, health inequality reduction
information limitations 68
outcomes in climate change 198
policy in developing countries *see* developing countries, health policy
promotion & ethical issues 262
related behaviour in high-income countries 129–130
social identity as risk 108
socioeconomic health inequalities 126
f
stages of transition 301–302, 302
t, 303
f
Health Action Zones 134
t
health belief model 88

healthcare
accountability of providers 233
analysis of issues 67–68
citizens' right to 232–234
data role 236–237
equity in 230–231, 230
t
financing in developing countries 309
government role & responsibility 234
human rights-based approaches 276–277
Italian reforms 274
performance measures 236–237
rights in practice 280
right to 277–280
social determinants of health 85
socioeconomic health inequalities, high-income countries 130
workforce 235–236, 311–312
healthcare access 17, 229–239
ethical issues 265
policy importance 234–235
health education programmes, HIV infection 214
Health Field Concept 215
f
health information systems (HIS) 308
Health Law Digest (WHO) 293
Health Metrics Network (HMN) 68
Health Quality and Safety Commission (New Zealand) 237
Health Relief Models 217
health sector action, climate change and 204
b

health status
globalization 67–68
income *vs.* 127
healthy life expectancy 127
heart disease, mortality 216
t
heat exposure effects 55
Hennock, E. P. 27
hepatitis B (HepB) vaccination 34, 35–36, 35
f
hepatitis C 96
hepatocellular cancer (HCC) 11
hereditary breast and ovarian cancer (HBOC) 150
high-income countries
climate change 197
perinatal–juvenile mortality rates 216–217
socioeconomic health inequalities *see* socioeconomic health inequalities, high-income countries
see also developed countries
Hippocratic Oath 261
histone deacetylase (HDAC) inhibitors 154
histone methyltransferases (HMTs) inhibitors 154
histone modifications 154
histone-modifying enzyme inhibitors 154
history of public health 23–32
16th to 19th century 24–26
food security 23–24
modern system 26–27
19th Century 27–29
20th Century 29–30
HIV infection 6
antiretroviral therapy 52
behaviour effects 214
developing countries 304
diarrhoeal disease 165
ethical issues 263
middle-income countries 49
needle-exchange programmes 96
pharmacodynamics 153
prevention with needle-exchange programmes 94–95
sub-Saharan Africa 57
water & sanitation 171
HIV Preventative Trial Network (HPTN) 267
HMN (Health Metrics Network) 68
HMTs (histone methyltransferases) inhibitors 154
homeless people 14
homicide 14
hookworm infections 166
t
transmission barriers 168
horizontal equity, healthcare 230

household water treatment and safe storage (HWTS) 172

housing
climate change 206
energy efficiency 206, 207
t
HRH (Human Resources for Health) crisis 311
Human Development Index (HDI) 106
Human Genome Diversity Project (1991) 267
Human Genome Project 149
behaviour and 221
Human Resources for Health (HRH) crisis 311
human rights 273–290
climate change effects 200
controversies 277
definition 274–275
entitlements & obligations 276–277
international laws 277–278
legality of 275–276
national laws 278–279
poverty 114
protection of 18
public health and 275
Human Rights Council, UN General Assembly 276
humidity 196

Hungary
cardiovascular disease 332
t
child deprivation 109
f
lung cancer 330
t
premature deaths 325
t
Hurricane Katrina (2005) 203
HWTS (household water treatment and safe storage) 172
Hygiène publique 27

IACHR (Inter-American Commission on Human Rights) 278

Iceland
cardiovascular disease 332
t
child deprivation 109
f
lung cancer 330
t
premature deaths 325
t
ICESR *see* International Covenant on Economic, Social, and Cultural Rights (ICESR)
identification, public health issues 3
IHD *see* ischaemic heart disease (IHD)
IHP (International Health Partnerships) 39
IHR (International Health Regulations) 69, 296–297
illegal products, globalization 67
Illich, Ivan 231
ILO (International Labour Organization), Health and Safety Programme 17
IMF (International Monetary Fund) 182
immunization *see* vaccines
implementation science 5
improvement strategies, developing countries 312–317
INCLEN Trust 40

income
demography and 48–49, 49
t, 50
t
health and 48, 48
f
health status *vs.* 127
social determinants of health 85, 85
f see also finance
income inequality 98–100, 99
f
case study 99
growth in 98, 98
t
partisan divide 99–100
solutions to 100
US political context 99
Independent High-Level Commission on Noncommunicable Diseases (WHO) 70
independent nationality determined contributions (INDCs), greenhouse gas emissions 207

India
Ayushman Bharat programme 316
b
household wealth 85
f
income & health 48
f
infant mortality 49
t, 85
f
litigation aspects 295–296
My Health, My Voice 273
population demographics 50
t see also Delhi (India)
indigenous peoples, social determinants of health 86
individual rights, community interests *vs.* 261–262

Indonesia
infant mortality 49
t
population demographics 50
t
indoor air pollution 16
industries 20
developing countries 301
infant mortality, Brazil 116

infections
cause of death 6
t
control of 295
reduction in needle-exchange programmes 94–95
infectious disease, law case study 296–297
Infectious Disease Notification Act (1899) 29

infectious diseases
behaviours 214
climate change effects 197, 197
b
health policy in developing countries 304–305
inequalities 140–141
middle-income countries 51, 51
f
influenza A (H1N1) pandemic 304
legal aspects 297
influenza pandemic (1918-1919) 69
ethical issues 263

information
campaigns of 294
developing countries 309–310

informational environment
  alteration of 297
  legal aspects 294
information and communication technologies (ICTs), globalization 73–74
informed consent, ethics 158

infrastructure
developing countries 310–311
investment 315–316
injuries 14
control law 298–299
future of 20
intentional injuries 299
LMICs, prevention in 41
middle-income countries 50, 51, 51
f
mortality rates & socioeconomic status 124 *see also* accidents; unintentional injuries
inspiration, leadership 245
institutions, developed countries 326
Institut Pasteur 28
insulin dependent diabetes 215
integrity, leadership 246, 246
b
Intellectual Property Rights and Public Health, Doha Declaration (2001) 75
intentional injuries 299
Inter-American Commission on Human Rights (IACHR) 278
Intergovernmental Panel on Climate Change (IPCC) 194

international assistance
healthcare rights 279
right to health 284–286
b
International Association of National Public Health Institutes 256
International Confederation of Midwives Triennial Congress (Toronto) 305
International Covenant on Economic, Social, and Cultural Rights (ICESR) 234, 274, 276
healthcare access 232
human rights 277
Optional protocol (2013) 276
International Health Partnerships (IHP) 39
International Health Regulations (IHR) 34, 69, 296–297
International Labour Organization (ILO), Health and Safety Programme 17
international law, human rights 277–278
International Monetary Fund (IMF) 182
International Sanitary Conferences 74
international trade in food 182
interpersonal influence, leadership 244
intersectoral collaboration 58
interventions of public health 18–19
in vitro fertilization 149
iodine deficiency 307
IPCC (Intergovernmental Panel on Climate Change) 194

Ireland
cardiovascular disease 332
t
child deprivation 109
f
lung cancer 330
t
premature deaths 325
t
iron deficiency 307
ischaemic heart disease (IHD)
mortality 6
f
socioeconomic status 124

Israel
cardiovascular disease 332
t
lung cancer 330
t
premature deaths 325
t
issues of agency, developed countries 327

Italy
cardiovascular disease 332
t
child deprivation 109
f
health system reforms 274
lung cancer 330
t
mortality rates & socioeconomic status 125
t
premature deaths 325
t
public health history 29

Japan
  cardiovascular disease 332
    t
  lung cancer 330
    t
  premature deaths 325
    t
  public health 336
  public health history 29
Johnson & Johnson 65
  t
Joint Monitoring Programme (JMP) 164, 165
  f
Jordan, infant mortality 49
  t
judicial approach, human rights & public health 276
justice theories, poverty 113–114

kala-azar (visceral leishmaniasis) 3
key performance indicators (KPI) 236
knowledge 326–327
Korea 29
Kuznets, Samuel 327

labelling, mandatory in food 72
labelling standards, global public health 74
The Lancet
  Commission on Migration and Health (2018) 53
  Global Health Commission (2018) 315–316
  Global Health Commission on High Quality Health Systems 57
  Maternal and Child Nutrition Series (2008) 306
  Nutrition Series (2008) 183

Latin America
absolute poverty 104
food supplies 183
litigation aspects 295–296
social institutions 144

Latvia
cardiovascular disease 332
t
child deprivation 109
f
lung cancer 330
t
premature deaths 325
t
law enforcement authorities, needle-exchange programme opposition 95
leadership 243–260
across lifecourse 255
actions 245, 245
b
case study 253–254
b
challenges management 256–257
competencies 252
t
confidence 246
continual learning 256
decisiveness 246
definition 243
development 250, 250–251
b
development steps 258
b
emotional intelligence 246–247, 247
b
future 256
guiding principles 256
integrity 246, 246
b
journey development 257
leadership theory evolution 244–245
life-long investments 256
management *vs.* 243–244
managerial leadership 248
mentors 256
observational studies 258
optimism 246
organizational needs 258
resonant leadership 249–250, 250
b
styles of 247–250, 247
b
values 256
Leadership Framework (LF) (NHS) 251, 252
f
learning health systems 56

legal frameworks, global health 75
legal issues 291–300
  built environment 294
  case study 296–299
  deregulation 296
  healthcare rights 279
  human rights 275–276
  informational environment 294
  injury control law 298–299
  lawmaking process 293
  personal, professionals & business regulation 295
  power to tax & spend 293–294
  public health law enforcement 296
  public health statutes 292–293
  social justice 292
  socioeconomic environment 295
  tort system 295–296
Le-Nest initiative (Shanghai) 236
leprosy 36
Lesotho 49
  t
LGBTQ+ social exclusion 109
liberalization, capital flows 65
LICs *see* low-income countries (LICs)
life course perspective, high-income countries 130–131

life expectancy
LMICs 39
socioeconomic groups *vs.* 126
life-long investments, leadership 256
lifestyle promotion, ethical issues 262

Lithuania
cardiovascular disease 332
t
child deprivation 109
f
lung cancer 330
t
mortality rates & socioeconomic status 125
t
premature deaths 325
t
living conditions, health inequalities 144–145
LMICs *see* low- and middle-income countries (LMICs)
Local Government Act (1858) 27
logistics, developing countries 310–311
long non-coding RNA (lncRNA) 156
long-term strategies, health inequality reduction 144
Louis, Pierre 26
low- and middle-income countries (LMICs) 33–43
antibiotic use 56
childhood disease elimination 34–36
chronic illness prevalence 232
classification of 33
climate change and 38
demands for food 183
disease control & elimination 33–34, 35
f, 36
f
double burden of malnutrition 179
environmental health promotion 38
future work 41–42
healthcare expenditure 232
health information 67–68
health policy, planning & financing 39–40
health systems reform 38–40
health workforce (2030) 40
history of 33–38
legal reforms 293
migration and 53, 66
multinational retailers 184
non-communicable diseases 37–38, 232
nutrition capacity 184
political economy 183
population of 33
processed food intake 181
research & development 40
social determinants of health 40–41
social inequalities 139
tropical disease control 36–37
universal healthcare provision 235

vaccination strategies 34–36, 35
f
lower respiratory tract infections 6
f
low-income countries (LICs)
economic growth in 89–90
environmental determinants of health 83
non-communicable diseases 297
perinatal–juvenile mortality rates 216–217
low-risk transitions, climate change 206

lung cancer
epigenetic therapy 156
mortality from 6
f
years of life lost 330
t
lupus 155

Luxembourg
cardiovascular disease 332
t
child deprivation 109
f
lung cancer 330
t
premature deaths 325
t
Lyme disease 198
lymphatic filariasis (LF) 37
Lysenko affair 333

Mahler, Hafdan 244
MAI (Multilateral Agreement on Investment) 117

malaria
chloroquine-resistant 305
climate change effects 199
malnutrition 167
causes of 307
definition 179
diarrhoeal diseases 167
elimination of 183
Lancet Nutrition Series (2008) 183
private sector role 183
public health history 23
water & sanitation 170

Malta
cardiovascular disease 332
t
child deprivation 109
f
lung cancer 330
t
premature deaths 325
t
Malthus, Thomas 26

management
definition 243–244
leadership *vs.* 243–244
managerial leadership 248
Manson, Patrick 197–198
A Manual of the Diseases of Warm Climates (Manson) 197–198
manufacturing industries 17
Mar del Plata Declaration (1977) 163
marginalized groups 230
Marshall Plan 144, 145
b
Marx, Karl 112
material factors, high-income countries 128
materialist factors, poverty 112
maternal and child health (MCH), LMICs 38, 305
maternal and child undernutrition (MCU) 184
maternal & child health, developing countries 305
Mauritius 144, 145
b
McDermott, John 234
McGregor Burns, James 248
McKeown, Thomas 231, 327
MCU (maternal and child undernutrition) 184
MD1 (myotonic dystrophy type 1) 157
MDGs *see* Millennium Development Goals (MDGs)
mDiabetes 58
MDT (multidrug therapy) 36–37
MDTF (multi-donor trust fund) 39
M&E (monitoring and evaluation) 308

measles
control & elimination 35
outbreak (2015) 264–265
vaccination 35
f
medical supplies, developing countries 310–311
Medicinische Polizey 28
meganucleases 156
Mendelian randomization studies 37
mental health 12, 13
t
middle-income countries 52–53
migration 53–54
poverty and 52
f
social determinants of health 86
treatment & care gaps 53
mentors, leadership 256
Merck 65
t
MERS (Middle East Respiratory Syndrome) 10, 304
methodology development 5

Mexico
  Oportunidades CCT programme 58
  population demographics 50
    t
Mexico City, Global Gag Rule 96–97, 97
  f
mHealth 58
microRNAs (mRNA) 155–156
MICs *see* middle-income countries (MICs)
micronutrients, LMICs 41
middle class, income inequality 99

Middle East
Middle East Respiratory Syndrome (MERS) 10, 304
middle-income countries (MICs) 47–62
  antibiotic use 56
  definition 47
  demographic transition 48–49, 50
    t
  health challenges 50–56
  health systems 56–58
  HIV epidemic 49
  infant mortality 49
    t
  medical infrastructure 310
  population of 47
  urbanization 53
middle-income-trap 47
midlevel health workers (MLWs) 311–312

migration
climate change effects 196
developing countries 304
health inequalities 142
mental health 53–54
middle-income countries 53–54
mortality rate 53
phases of 53
f
urbanization and 50
Millennium Development Goals (MDGs)
absolute poverty 104
child health 217
introduction of 141
LMICs 42
maternal & child health 306
water & sanitation 163, 171
Mill, John Stuart 262
mixed-lineage leukaemia (MLL)-rearranged leukaemias 155
MLL (mixed-lineage leukaemia)-rearranged leukaemias 155
MLWs (midlevel health workers) 311–312
Mohalla clinics, Delhi 236
MomConnect 58

monitoring
  health issues 67–68
  public health issues 3
monitoring and evaluation (M&E) 308
morbidity 126–127, 126
  f
morcellement 25–26
Morocco 85
  f

mortality
causes of 6
f
cause-specific mortality 124
climate change effects 200–201
disease occurrence measurement 7
health inequalities 142
migration 53
poverty-related causes 6
t
socioeconomic health inequality 124, 125
t
socioeconomic health inequality measurement 124, 126
mosquitoes 196
MPI (multidimensional poverty index) 106–107, 107
b
MPOWER 40
MRFIT (Multiple Risk factor Intervention Trial) 214
MS (multiple sclerosis) 155
multidimensional poverty index (MPI) 106–107, 107
b
multi-donor trust fund (MDTF) 39
multidrug therapy (MDT) 36–37
Multilateral Agreement on Investment (MAI) 117
multimorbidity 52
b
contrasting patterns of 52
middle-income countries 51–52
multinational retailers, middle-income countries 184
multiple determinants of health 89
Multiple Risk factor Intervention Trial (MRFIT) 214
multiple sclerosis (MS) 155
Myanmar 49
t
My Health, My Voice (India) 273
myotonic dystrophy type 1 (MD1) 157

NAFTA (North American Free Trade Agreement) 182
Namibia 49
t
nanoparticles 16
Nathanson, Constance 326
National Bioethics Advisory Commission (2001) 266, 268
National Cancer Institute (NCI), Best Practices for Biospecimen Resources 268
national constitutions, right to health 291
National Council on Food and Nutrition Security (CONSEA) 187
National Health Service (NHS)
history of 29–30

Leadership Framework 251, 252
f
leadership qualities framework 253
b
Sustainable Development Unit 205, 205
f
National Heart Act (USA) (1948) 331
National Heart Foundation (Australia) 331

national laws
human rights 278–279
public health laws 298–299
national stunting rates 187
National Vaccine Advisory Committee (US) 263
natural selection, social selection *vs.* 111–112
NBS (newborn screening) 150
NCDs *see* non-communicable diseases (NCDs)
ncRNA (non-coding RNA) 154, 155–156
NDM-1 (New Delhi-metallo-beta-lactamase) 10–11
Necator americanus infection 166
t, 167
needle-exchange programmes (NEPs) 94–96
attacks on science 95–96
case study 95
fear tactics 95
neglected tropical diseases (NTDs) 36–37
neighbourhood design 294

neoliberalism
  globalization 64
  poverty reduction 115
NEPs *see* needle-exchange programmes (NEPs)

Netherlands
cardiovascular disease 332
t
child deprivation 109
f
lung cancer 330
t
premature deaths 325
t
net household income, poverty and 105
newborn screening (NBS) 150
New Delhi-metallo-beta-lactamase (NDM-1) 10–11
New Poor Law (1834) 28–29
disproportionate payments 26
New York (USA) 28

New Zealand
cardiovascular disease 332
t
climate change 206–207, 207
t
healthcare workforce 235–236
Health Quality and Safety Commission 237
lung cancer 330
t
premature deaths 325
t
next-generation sequencing (NGS) 151
NGOs *see* non-governmental organizations (NGOs)
NGS (next-generation sequencing) 151
NHANES *see* National Health and Nutrition Examination Survey (NHANES)
NHS *see* National Health Service (NHS)
nicotinamide 155

Nigeria
endemic polio 35
infant mortality 49
t
population demographics 50
t
NIPT (non-invasive prenatal testing) 150–151
non-coding RNA (ncRNA) 154, 155–156
non-communicable diseases (NCDs) 5, 6
f, 11
case study 297–298, 298
t
climate change effects 198
deaths from 7
f
domestic strategy legislation 297
Eastern Europe 333
global burden of 70
health policy in developing countries 305
middle-income countries 50, 50
t
mortality 12
f
prevalence in low- and middle-income countries 232
undernutrition and 179
non-discrimination, healthcare rights 279
non-governmental organizations (NGOs) 19–20
women's reproductive health politics 96
non-invasive prenatal testing (NIPT) 150–151
North American Free Trade Agreement (NAFTA) 182
North Star Alliance, sub-Saharan Africa 236

Norway
cardiovascular disease 332
t
child deprivation 109
f
lung cancer 330
t
mortality rates & socioeconomic status 125
t
premature deaths 325
t
Novartis 65
t
NTDs (neglected tropical diseases) 36–37
Nussbaum, Martha 107
nutrient content, foods 58
nutrition 13–14, 179–192
action challenges 184–187
animal protein 182
deficiencies as cause of death 6
t
food environment 181
food system changes 181–183
health and 180
LMICs 41
political economy 183–184
status 180
success stories 187–188
transition in developing countries 301 *see also* diet

Obama, Barack 97, 99–100

obesity
cardiovascular disease 215
increase in 70–71, 71
f
non-communicable diseases 11
numbers of 179
socioeconomic health inequalities 129
type 1 (insulin dependent) diabetes 215
USA public health 334 *see also* overweight

observational studies
  climate change 196–198
  leadership 258
occupational health and safety laws 17, 299
ODA (official development assistance) 141
OECD *see* Organisation for Economic Co-operation and Development (OECD)
Office International d'Hygiène Publique 74
official development assistance (ODA) 141
onchocerciasis (river blindness) 37
OOPE (out of programme experience) 312–313
opioids 334
Oportunidades CCT programme (Mexico) 58
opportunity development, health inequalities 144
opportunity equality, outcome equality *vs.* 113
opposition, needle-exchange programmes 95–96
optimism, leadership 246
oral cancers 14
oral health 14
oral rehydration therapy (ORT) 165
Organisation for Economic Co-operation and Development (OECD) 323
  income inequality 99
  relative poverty 105, 106
    f
organizational needs, leadership 258
ORT (oral rehydration therapy) 165
outcome equality, opportunity equality *vs.* 113
out-of-pocket payments (OOP), LMICs 39
out of programme experience (OOPE) 312–313
overnutrition increase 70–71

overweight
increase in 71
f
numbers of 179 *see also* obesity

Pakistan 49
t
endemic polio 35
pandemic preparedness, ethical issues 263
parasitic diseases 6
t
paratyphoid fever 166
t, 167
transmission barriers 168
Parent-Duchâtelet, Alexandre 26
Paris Agreement (2015) 201, 207
participation, healthcare rights 279
pathogen genomics 151
Patient Protection and Affordable Care Act (2010) 30
PCR (polymerase chain reaction) 154
Pellegrino, Edmund 262
People's Health Movement (PHM) 117
Percival, Thomas 261
performance measures, healthcare services 236–237
perinatal–juvenile mortality rates, high- *vs.* low income countries 216–217
Personal Genome Project 268
personal qualities, leadership 255
personal regulation, legal aspects 295
Peru 85
f
Pfizer 65
t
PHAC (Public Health Agency of Canada) 67
pharmaceutical industry 65–66, 65
t
pharmacodynamics 153

pharmacogenomics
  drug development 153–154
  testing 152–153
pharmacokinetics 153
Philippines 313
  b
PHM (People's Health Movement) 117
PHN (public health nutrition) 184
PHSKF (Public Health Skills and Knowledge Framework) 250

physical activity
coronary heart disease and 215
patterns of 214
socioeconomic health inequalities 129
physical systems, climate change 195–196
plague 34
Plasmodium infections *see* malaria
Pogge, T. 113–114

Poland
cardiovascular disease 332
t
child deprivation 109
f
lung cancer 330
t
premature deaths 325
t
public health history 29

policies
human rights & public health 276
malnutrition 185
socioeconomic health inequalities, high-income countries 132–133
policymakers 94
polio vaccinations 34, 35
f, 36
f
Political Declaration on the Prevention and Control of Non-Communicable Disease 297
politics of public health 93–102
analysis of 93
f
challenges 89–90
developing countries 308
evaluation of 93–94
evidence-based policies 98–100
harm-reduction strategies 94–96
see also needle-exchange programmes (NEPs)
interventions 19
motivations 94
policymakers 94
protection of rights 292
women's reproductive health 96–98
pollution, biodegradable 16
polymerase chain reaction (PCR) 154
Poor Law (1598) 24
Poor Law (1601) 24, 25
f
Popular Opinion Leader model 19

population
ageing of 17
climate change effect 201
ethics of research 267
growth of 5
mobility in globalization 66
projections of 12–13
rising age of 7
size in globalization 69
population health 215–217, 215
f

Portugal
cardiovascular disease 332
t
child deprivation 109
f
lung cancer 330
t
premature deaths 325
t
post-communist countries 336
poverty 5, 6
t
absolute *see* absolute poverty
capability lack as 106–108
communicable disease 9
cultural/behavioural explanations 112
definition 103–104, 139
deprivation as 108
developing *vs.* developed countries 105
equality of opportunity *vs.* outcome 113
ethical poverty line 104–110
health effects 110–112
human rights 114
interventions 112–115
justice theories 113–114
materialist explanations 112
mental health and 52
f
reduction & nutrition 187
relative poverty 105
social exclusion as 108
structural explanations 112

poverty reduction
  examples of 116–117
  public policies 115–116
powers of public health 291–292
predictive genomic testing 150
predisposing factors 217–218
  socioeconomic status & behaviour 222
preimplantation genetic testing 150–151
premature death, developed countries 324
prenatal genetic testing 150–151
prenatal screening, ethics 158
Presidential Commission on the Human Immunodeficiency Virus Epidemic (1988) 263
Preston curves 29
Preston, Samuel 48

primary healthcare
access to 231, 231
t
antibiotic use 56
definition 229–230
extended value of 234
financing of 235
guiding principles 230
key performance indicators 236
secondary healthcare *vs.* 230
private–public partnership, developing countries 313–314
private sector role, malnutrition 183
private support of public health 19–20
processed foods 181
advertising and 184
production changes, globalization 66
professionals, legal regulation 295
programmatic innovations, water & sanitation 172–173
PROGRESS 143
ProMED mail 69, 70
b

promotions
  public health 3–4
  tobacco industry 220
Protocol to Eliminate Illicit Trade in Tobacco Products 72
provision of healthcare 17
psychosocial factors, high-income countries 128–129
PTMT (thiopurine methyltransferase) 153
Public Health Act (1848) 26, 27, 302
Public Health Act (Sweden, 1874) 28
Public Health Agency of Canada (PHAC) 67
public health nutrition (PHN) 184

public health policy
  developed countries *see* developed countries, public health policy
  nature & scope of 323
Public Health Skills and Knowledge Framework (PHSKF) 250
public policy, health determinants 90
public sanitation 172
public sector coordination, water & sanitation 174
public spending, legal aspects 293–294

Qing dynasty (China), antifamine strategies 25

quality healthcare services
  developing countries 315
  middle-income countries 57–58
quarantine, public health history 24

rainbow model (2006) 89
rain, climate change and 194–195
Ravallion, Martin 105
Rawls, John 113, 233, 265–266
re-emerging diseases, globalization 69
regional human rights law 278

regulatory frameworks
global health 75
public health 292

reinforcing factors
behaviour 218–219
education & behaviour 223
relational aspects, social exclusion 109–110
relational justice theory 113–114
relationship management, emotional intelligence 246–247, 247
b
relative poverty 105
reproductive health politics 96–98
research 5
genetic research 268
population ethics 267
public health 68–69, 266, 267–268
resource-poor settings 267
resilient health systems, middle-income countries 56
resonant leadership 249–250, 250
b
resource constraints 4
healthcare rights 280
research & ethics 267
respect, healthcare rights 279–280
respiratory infections 6
t
rheumatoid arthritis (RA) 155
rich–poor disparities 20
risk control in public health 68–69
risk multipliers, climate change and 203
risk pooling, developing countries 312–313, 313
f
river blindness (onchocerciasis) 37
road injury mortality 6
f
Robinson, Kenneth 329
Roche 65
t
The Rockefeller Foundation 112–113
Rogers, Edward S. 221

Romania
cardiovascular disease 332
t
child deprivation 109
f
lung cancer 330
t
premature deaths 325
t
rotavirus infection 166, 166
t
Royal Australasian College of Surgeons 329
rural areas, developing countries 311
Rwanda 146
b

Sade, Robert 233
SAFE strategy (surgery, antibiotics, facial cleansing, and environmental improvement) 36–37
safe water consumption adherence 171–172
SAGE (Study on global AGEing and adult health) 51–52
Salmonella infection 166
t
climate change 203
sample storage, ethics 159

sanitation
definition 164
LMICs 38 *see also* water & sanitation
Sanofi 65
t
SAPs (structural adjustment programmes) 182
SARS *see* severe acute respiratory syndrome (SARS)
Say, J.-B. 27
scaling up nutrition (SUN) initiative 183–184
Scandinavia 198
scarcity theory 129
SCD (sickle cell disease) 157
schistosomiasis 166, 166
t
transmission barriers 168
schizophrenia 13
t
schooling improvement 170
science, technology, engineering, and mathematics (STEM) 5

Scotland
mortality rates & socioeconomic status 125
t
The State Hospital 273 *see also* United Kingdom (UK)
SDGs *see* Sustainable Development Goals (SDGs)
SDH *see* social determinants of health (SDH)
secondary healthcare 230
access to 231–232, 231
t
primary healthcare *vs.* 230
security, water & sanitation 170–171
self-awareness, emotional intelligence 246–247, 247
b
self-management, emotional intelligence 246–247, 247
b
Sen, Amartya 106, 107
Senegal, mDiabetes 58
service delivery, developing countries 311
service investment, climate change 204
SES *see* socioeconomic status (SES)
severe acute respiratory syndrome (SARS) 9–10, 20, 67, 304
epidemic (2002-2003) 69, 304
legal aspects 297
SGLT-2 inhibitors 37
Shanghai, Le-Nest initiative 236
shared sanitation 172
Shigella infection 166, 166
t
short-lived climate pollutants (SLCPs) 206
sickle cell disease (SCD) 157
Simon, John 28–29

Singapore
cardiovascular disease 332
t
lung cancer 330
t
premature deaths 325
t
public health 336
public health history 29
single-agent communicable diseases 214
single nucleotide polymorphisms (SNPs) 149
cardiovascular disease risk factor 37
Sin Tax, Philippines 313
b
Siracusa Principle 276, 278
SIS (Statistical Information System) (WHO) 68
SJS (Stevens–Johnson syndrome) 153
SLCPs (short-lived climate pollutants) 206

Slovakia
cardiovascular disease 332
t
child deprivation 109
f
lung cancer 330
t
premature deaths 325
t

Slovenia
cardiovascular disease 332
t
lung cancer 330
t
mortality rates & socioeconomic status 125
t
premature deaths 325
t
smallpox vaccine 18, 34
Smith, Adam 26, 64
Smith, Richard 231
SNPs *see* single nucleotide polymorphisms (SNPs)
social activism 20
social awareness, emotional intelligence 246–247, 247
b
social circumstances 4
Social Cognitive Theory 217, 221
social democratic model, poverty reduction 115–116
social determinants of health (SDH) 83–88, 84
f, 85
t
behaviours 87–88
climate change 88
early child development 85
employment conditions 85–86
gender 86
globalization 87
healthcare services 85
income, education & employment 85, 85
f
indigenous peoples 86
mental health 86
socioeconomic factors and 84
urbanization 86–87

social development
climate change projections 200
health inequality reduction 144

social exclusion
  LGBTQ+ 109
  relational aspects 109–110
Social Exclusion Knowledge Network, Commission on Social Determinants of Health 110
social group inequalities 265
social identity 108
social interventions 18
socialist model, poverty reduction 116
social justice theory 113, 265–266, 292
social learning theory 88

social media
  globalization 74
  health policy in developing countries 307–308
social medicine 41, 113
social–normative theory 223
social participation, health inequality reduction 144, 145–146
social selection, natural selection *vs.* 111–112
social support 129
Society of Public Health Educators 262
sociocultural environment 130

socioeconomic factors
legal aspects 295
life expectancy *vs.* 126
social determinants of health 84

socioeconomic health inequalities
diseases & disabilities 126–127
generic health indicators 126
f
high-income countries 123–138
English reduction strategies 134, 134
t
morbidity 126–127, 126
f
mortality 124, 125
t, 126
overarching ideas 130–131
reduction strategies 131–134
selection *vs.* causation 127–128
specific causal pathways 128–130, 128
f
measurement of 123
normative assessment 131–132
socioeconomic status (SES)
behaviour and 222–223
behaviour & gender interactions 223–224
causal impact on health 127–128
climate change 196–197
measurement of 111, 123
social determinants of health 85
soil pollution 16
soil-transmitted helminth infections 167
transmission barriers 168

South Africa
circular migration 54
human rights laws 278–279
infant mortality 49
t
MomConnect 58
politics in public health 308
population demographics 50
t
quadruple burden of disease 51
Treatment Action Campaign 66, 273
South Asia 172
southeast Africa 49

South East Asia
water & sanitation improvements 172
Southeast Pacific 140

South Korea
cardiovascular disease 332
t
lung cancer 330
t
premature deaths 325
t
public health 336

Spain
cardiovascular disease 332
t
child deprivation 109
f
lung cancer 330
t
mortality rates & socioeconomic status 125
t
nutrition improvements 188
premature deaths 325
t
public health history 29
spatial change, globalization 63–64
Special Rapporteur on the Right of Everyone to the Enjoyment of the Highest Attainable Standard of Physical and Mental Health 276
sponsorship, tobacco industry 220
The State Hospital (Scotland) 273
state social insurance, Germany 29
Statistical Information System (SIS) (WHO) 68
status identity factor 223
STEM (science, technology, engineering, and mathematics) 5
ST-ETEC (enterotoxigenic *Escherichia coli* producing heat-stable toxin) 166
Stevens–Johnson syndrome (SJS) 153
strategies towards ending preventable maternal mortality (EPMM) 305
mortality 6
f, 216
t
structural adjustment programmes (SAPs) 182

structure
  health inequality reduction 145
  interventions 19
  poverty 112
Study on global AGEing and adult health (SAGE) 51–52
stunted growth 307
  national rates 187
  undernutrition 180
sub-Saharan Africa
  absolute poverty 104
  HIV epidemic 57
  North Star Alliance 236
  water & sanitation 163
substance use disorders 13
  t
Sudan, dracunculiasis 37
sudden unexpected death in infancy (SUDI) 235–236
sugar intake reduction 181
sugary soft drink bans, France 188
suicide 14
SUN (scaling up nutrition) initiative 183–184
support, leadership 245
Supreme Court (USA), contraceptive mandate 97
surface temperatures, climate change 193

surveillance
  antibiotics and 55
  public health 68–69
Sustainable Development Goals (SDGs)
  drinking water & sanitation 38
  ethical poverty line 104
  health policy in developing countries 317
  introduction of 234
  LMICs 42
  malnutrition targets 185, 187
  maternal & child health 305
  water & sanitation 164, 171
Sustainable Development Unit (NHS) 205, 205
  f
sustainable policies, health inequality 144

Sweden
cardiovascular disease 332
t
child deprivation 109
f
lung cancer 330
t
mortality rates & socioeconomic status 125
t
premature deaths 325
t
public health history 28
sweetened beverages, excise taxes 58
b

Switzerland
cardiovascular disease 332
t
lung cancer 330
t
mortality rates & socioeconomic status 125
t
premature deaths 325
t
system leadership 249, 249
b, 250
b

TACs (transnational alcohol companies) 66, 87
TALENS (transcription activator-like effector nucleases) 156

taxation
  global public health 74
  intersectional collaboration 58
  legal aspects 293–294
Tax Cuts and Job Act of 2017 100
TBE (tick-borne encephalitis) 198

technology
  globalization 64
  health inequality reduction 144
  health policy in developing countries 310–311
  water & sanitation 172–173
temperature increases, climate change 193–194
temporal change, globalization 64
temporality, determinants of health 88–89
TEN (toxic epidermal necrolysis) 153
tetanus toxoid vaccination 35
Thailand 57
thalassemia 157
theory of justice & fairness (Rawls) 233
thiopurine methyltransferase (PTMT) 153
tick-borne diseases 198
tick-borne encephalitis (TBE) 198
TNCs (transnational corporations) 65, 87

tobacco industry
advertising 220
excise taxes 58
b
globalization 72–73
legal control 295–296
packaging & advertising regulation 294
political support for 19
promotions & sponsorship 220
Tobacco Master Settlement Agreement (USA) (1998) 329

tobacco smoking
control strategies 220
coronary heart disease and 214
denormalizing behaviour 223
developed countries 327–329
LMICs 40–41
mortality and 216, 329
prices 59
f
socioeconomic health inequalities 129

tobacco smoking prevention
antismoking campaigns 19
history of 329–330
tort system 295–296
Townsend, Peter 105
toxic epidermal necrolysis (TEN) 153
tracheal cancer 6
f
trachoma 36–37, 166
t, 167
transmission barriers 169, 170
trade, developing countries 304
Trade-Related Aspects of Intellectual Property Rights Agreement (TRIPS) 87

traffic injuries
LMICs 41
  public health policy in developed countries 329, 331
  USA public health 334–335
training, developing countries 315–316
transcription activator-like effector nucleases (TALENS) 156
transformational leadership 248
transnational alcohol companies (TACs) 66, 87
transnational corporations (TNCs) 65, 87
Transport Action Commission (1986) 329

transportation
climate change 206
communicable disease 9–10
Transtheoretical Model of Stages of Change 217
trastuzumab 221
Treatment Action Campaign (South Africa) 66, 273
trichostatin A (TSA) 155
Trichuris trichuria infection 166
t, 167
transmission barriers 168
triple burden of disease, middle-income countries 50–51
TRIPS *(Trade-Related Aspects of Intellectual Property Rights Agreement)* 87
tropical diseases 11
mortality 6
t
tropical storms, climate change 202–203
Trump, Donald 97, 100
TSA (trichostatin A) 155

tuberculosis
economic inequalities 140–141
increase in 69
mortality 6
f see also BCG vaccination

Turkey
health reform 310
b
infant mortality 49
t
population demographics 50
t
Turkmenistan 85
f
type 1 (insulin dependent) diabetes 215
type 2 diabetes 70
typhoid fever 166
t, 167
transmission barriers 168

UDHR *see* Universal Declaration of Human Rights (UDHR)
Uganda 85
f
UHC *see* Universal Health Coverage (UHC)
UK *see* United Kingdom (UK)
UKPDS *see* United Kingdom Prospective Diabetes Study (UKPDS)
UN *see* United Nations (UN)

undernutrition
definition 179
maternal & child health 306
treatment of 307

unhealthy products
removal from market 297, 298
t
taxation of 297
UNICEF Conceptual Framework 307
f
UNICEF Innocent Research Centre 108, 109
f
relative poverty 105
unintentional injuries 14
mortality 216
t see also injuries
United Kingdom (UK)
cardiovascular disease 332
t
lung cancer 330
t
19th Century trading 26
public health policy 326
child deprivation 109
f
premature deaths 325
t
United Nations (UN)
Children's Fund 33, 74
Climate Summit (2019) 29
Declaration of Human Rights 114
Environment Programme 206
Framework Convention on Climate Change (1992) 207
Global Health and Foreign Policy (2012) 234
LMIC classification 33
World Population Ageing (2015) 17
United Nations Development Programme (UNDP) 33, 74, 106
United Nations General Assembly (UNGA)
High Level Meetings 70
Human Rights Council 276
UN Program on HIV/AIDS (UNAIDS) 42
United States of America (USA)
adult survival rates 324, 326
cardiovascular disease 332
t
lung cancer 330
t
Lyme disease 198
National Heart Act (1948) 331
National Vaccine Advisory Committee 263
needle-exchange programmes 95
New York 28
poverty 5

premature deaths 325
t
public health failures 333–335, 334
f
public health history 27–28
public health policy 326
social security 29
Supreme Court, contraceptive mandate 97
Tobacco Master Settlement Agreement (1998) 329
tobacco smoking health policy 328
urinary tract stones 199
women's reproductive health politics 96
Universal Declaration of Human Rights (UDHR) 274, 277
healthcare access 232
Universal Health Coverage (UHC) 313–314, 314
f
components of 317
f
health policy in developing countries 317
middle-income countries 56–57, 57
b
sustainable development goals 47
Universal Periodic Review (2008) 276
Urban Heat island effect 50
urbanization 49–50
climate change 202
food supplies 183
health and 86–87
health inequalities 142
middle-income countries (middle-income countries) 53
planning 297
slums 49–50
urinary tract stones 199
USA *see* United States of America (USA)
US Preventative Services Task force (USPSTF) 151

vaccines 9, 18
valproic acid 155
vector elimination 18
Venezuela 235
vertical disease control, LMICs 39, 41–42
vertical equity, healthcare 230
Vibrio cholerae O1 infection 166
Vibrio infection 166
t
Vibrio parahaemolyticus infection 204
Victorian Tobacco Act (1987) 329

Vietnam
infant mortality 49
t
politics in public health 308
Villermé, Louis-René 26
violence 5, 14, 141
future of 20 *see also* war
Virchow, Rudolf 27, 112
visceral leishmaniasis (kala-azar) 37
vitamin A deficiency 180, 307
voluntary interventions, ethical issues 262
voluntary organizations, cardiovascular disease 331
vulnerable populations 14–15

Wales 125
t
war 5, 17
climate change effects 196 *see also* violence
wasting, undernutrition 180

water
associated definitions 164
availability improvement 170
pollution 16
service ladder 164–165 *see also* water & sanitation

water quality
definition 164
diseases and 55
water-related vectors 164
water & sanitation 163–178
challenges in 173–174
community-led total sanitation 173
developments in 171–173
diseases and 165–167, 166
t, 167–170
literature 164
LMICs 38
programmatic innovations 172–173
sanitation marketing 173
schooling improvement 170
security & gender equality 170–171
technology 164, 172–173
terminology 164–165
water availability improvement 170
wealth, accumulation *vs.* distribution 139
welfare state 131
welfare wall 116
Western Pacific 140
Whitehead, M. 132
Whitehead's action spectrum 132, 132
f
WHO *see* World Health Organization (WHO)
whole genome sequencing (WGS) 149
women's reproductive health politics 96–98
workforce 5
work-related injuries 299

World Bank
absolute poverty estimation 104
financing for global health 74
Going Universal 56–57
structural adjustment programmes (SAPs) 182
World Development Report (2006) 115

World Development Report
2006 115
2010 & climate change 206
World Health Organization (WHO)
challenges to role 74
Child Growth Reference 180
childhood obesity recommendations 185
climate change impact 16
Commission on Intellectual Property Rights, Innovation and Public Health 75
Commission on Social Determinants of Health *see* Commission on Social Determinants of Health (WHO)
communicable diseases 8–9
Expert Committee on Drug Dependence 95
Framework Convention on Tobacco Control 40, 87
Global Action Plan for the Prevention and Control of NCDs (2013-2020) 72
Global Health Expenditure database 12
Global Strategies in Diet, Physical Activity, and Health 72
healthcare systems 315, 315 *f*
healthcare workforce 235, 236
health inequalities 123
Independent High-Level Commission on Noncommunicable Diseases 70
injuries as mortality cause 14, 298
International Health Regulations (IHR) 69
public health history 29
right to health 277
SAFE strategy 36–37
sanitation guidelines 171
Statistical Information System 68
Strategies towards ending preventable maternal mortality 305
Study on global AGEing and adult health 51–52
World Health Report (2006) 40
World Mental Health Surveys 53
World Population Ageing (2015) (UN) 17
World Trade Organization (WTO)
establishment of 65
Trade-Related Aspects of Intellectual Property Rights Agreement (TRIPS) 87

yaws 37
years of productive life lost (YPLL) 7–8, 9 *f*
communicable diseases 8–9
yellow fever (YF) 34
Young, Iris Marion 114
Yugoslavia 29

Zambia, infant mortality 49
t
zebularine 154
Zika epidemic (2016) 10
ethical issues 264
Zimbabwe, malaria & climate change 199
zinc deficiency 180, 307
zinc finger nucleases (ZENS) 156
zoning laws 294

Index Vol 2

Notes: Tables, figures and boxes are indicated by *t*, *f* and *b* following the page number

absence of dsssisease 338
absolute risk judgements 486
absorption, toxicology 475
ACASI (audio-computer-assisted self-interviewing) 56

acceptability
  complex interventions 358–359
  surveillance system evaluation 270
accountability, health governance 396, 397
  t
ACGIH *see* American Conference of Governmental Industrial Hygienists (ACGIH)
acoustic neuroma 441
active case finding, outbreak investigation 65–66
active surveillance 262–263
actual disease transmission 252
acute care hospital setting 341
acute diseases, natural history 424
adaptive trial designs 45
administration, occupational health risk management 467
administration censoring 90
administration databases 4
adult mortality estimates, Global Burden of Disease study 233
advancing methodology, systematic reviews 202
advertising overload, tobacco smoking prevention 133
Advice Leading to Public Health Advancement (ALPHA group) 355
advocacy, *Ottawa Charter for Health Promotion* 338

aetiology
national multipurpose surveys 49
outbreak investigation 64–65
aflatoxins, hepatocellular cancer 432
age-adjusted cancer mortality study 426
agency, qualitative research 101
b
agency theory 399–400
Agenda for Sustainable Development (UN, 2030) 354
agent–environment–host interactions 40, 40
f
age-specific death rates 323–324
age-specific fertility rate (ASFR) 323
b
age-specific mortality rate (ASMR) 324
b
age structure, stable population theory 327–329
AGREE II (Appraisal of Guidelines for Research and Evaluation II) 151
AIDS 307–309
mortality rate effects 318–319
social movement 309 *see also* HIV infection
Airs, Waters, Places (Hippocrates) 498
alcohol consumption, hepatocellular cancer 432
all-cause mortality, Global Burden of Disease study 234
allergens, dose–response curves 474

allocation
  definition 143
    t
  efficiency in economic appraisal 367
Alma-Ata Conference (1978) 341
ALPHA group (Advice Leading to Public Health Advancement) 355
alpha particles *see* ionizing radiation
ALS (amyotrophic lateral sclerosis) 439–440
alternative sexualities 311–312
Alzheimer's disease, tobacco smoking and 54
America *see* United States of America (USA)
American Conference of Governmental Industrial Hygienists (ACGIH) 464–465
  skin absorption notation 475
American Public Health Association (APHA) 499
American Sociology Association, Medical Sociology Section 292
Americans with Disabilities Act 462
AMSRAR-2 critical appraisal tool 200
amyotrophic lateral sclerosis (ALS) 439–440
analogy, causation 186

analysis error units
  meta-analysis of systematic reviews 198
  risk communication 493
analysis of variance (ANOVA) 218
  assumptions of 218
analytical steps, Global Burden of Disease study 232–236
animal responses, toxicology 475
annual risk of death 250
ANOVA *see* analysis of variance (ANOVA)
antiretroviral therapy (ART) 383–384
APHA (American Public Health Association) 499
APHEA (European Agency for Accreditation in Public Health Education) 301
Appraisal of Guidelines for Research and Evaluation II (AGREE) 151
Arnett, John L. 299
ART (antiretroviral therapy) 383–384
artificial intelligence 202
asbestosis 459
ASFR (age-specific fertility rate) 323
  b
ASMR (age-specific mortality rate) 324
  b
assignment 143
  t
ASSIST 355–356
Astana Declaration (2018) 341
asymmetric power dynamics 311
at-risk populations 66–67
attributable risk estimation, case–control studies 82, 82
  t
audio-computer-assisted self-interviewing (ACASI) 56
audiometry 465

Australia
  Medicare Benefits Schedule and Pharmaceutical Benefits Scheme 155
  National Health and Medical Research Council (NHMRC) 103
  The North Coast Project 136
  People's Health Movement 103–104
autobiographies 106
automation 466
availability bias 486–487
availability heuristic 385

balancing equations, migration measurement 325–326
Bangkok Charter (2006) 353
Bangladesh, People's Participatory Planning (PPP) 108
Bartlett's test 218
baseline imbalance, cluster randomized trials 115
baseline visits, cohort studies 87
Bayesian meta-regression 234–235
Bayesian methods 162
Becker, Marshall H. 299

behaviour
  communication technologies 30
  environmental health 423
  research in HIV infection/AIDS 307–308
  risk communication 493
behavioural economics 381–389
  availability heuristic & salience bias 385
  definition 381
  future work 386–388
  insights & interventions 382–386
  limited attention 385
  loss aversion & endowment effect 384
  present bias & incentives 383–384
  social norms & social influence 385–386
  status quo bias & defaults 384–385
Behavioural Insights Team 387
Behaviour Risk Factor Surveillance System (BRFSS) 266
BEI (Biological Exposure Limits) 465

beneficial effects
infrared radiation 443
ultraviolet radiation 444
benefit-cost ratio 379
benefit payments, economic appraisal 366
benefits, economic appraisal 373
Bernoulli, Daniel 244
best practices, tobacco smoking prevention 133
beta particles *see* ionizing radiation
between-cluster correlation 113
between-cluster variability 114
between-like design, implementation science 412
between-site comparison design, implementation science 413, 414
t

bias
  causal inference 167–179
  cohort studies 96–97
  cross-sectional studies 54–56
  evaluation in life course epidemiology 282
  minimization of 142
  prediction problems 163–167
  samples 53, 208
  systematic reviews 193
bias risk evaluation 196–197
bias towards the null 173
big data, surveillance systems 267
Bill & Melinda Gates Foundation 404
biological age 281
biological components, environmental health 422–423
Biological Exposure Limits (BEI) 465
biological gradient, causation 185
biological mechanisms, life course epidemiology 279–281
biological monitoring, occupational exposure limits 465–466
biological sample collection 51

biomarkers
cohort studies 90
environmental health 431, 431
t, 432
t
life course epidemiology 281
occupational exposure limits 465
biomedical model of urbanization 499–500
birth cohort studies 284
birth defects, surveillance systems 265–266
birth rate, fertility measure 322
birth registries 320

Blackfoot disease studies
biomarkers 431–433
cancers in cohort members 429
community-based cross-sectional survey 427–429, 427
t
community-based health survey 428–429, 428
t
geographical clustering 425–426, 425
t, 426
t
longitudinal cohort studies 429
public water supply system 430
blogs 7
body mass index (BMI), mortality *vs.* 188–189
bottom-up costing 372
botulism outbreaks 72–73
box plots 210
BRFSS (Behaviour Risk Factor Surveillance System) 266
British Sociology Association, Medical Sociology Research Network 292
Bucharest Early Intervention Project 276
Building Healthy Online Communities 33
built environment 500

CALIBER programme 5
Campbell Collaboration 151
LMIC research capacity 158
Canadian Institutes of Health Research (CIHR), Model of Knowledge Translation 410

cancer
extremely low frequency fields 439
ionizing radiation 449–452
occupational health 459
radiofrequency radiation 441–442
static radiation fields 438
toxicology 476–477
ultraviolet radiation 444

capacity building
behavioural economics 387
health governance 397
t
CARDIA study 86
cardiovascular disease (CVD)
community-based interventions 134–135
community health promotion 135–137
extremely low frequency fields 440
London drivers 54
physical activity and 282
CARES programme (Philippines) 384
case–control monitoring 178
case–control studies 44, 75–84
applications 75
attributable risk estimation 82, 82
t
bias risk assessment 153
t
case definition 77–78
cohort studies nesting 76
cohort studies *vs.* 83
confounding control 80
confounding minimization 79–80
control definition 78–82
cumulative incidence of disease 75–76
disease aetiology 82–83
exposed *vs.* non-exposed people 75, 75
t, 76
t
information bias minimization 80–82
interviews & questionnaires 76–77
physical & laboratory measurements 77
randomized trials *vs.* 82–83
records 77
relative-risk estimation 175–176
retrospective ascertainment of exposure status 76
selection bias minimization 78–79
special issues 175
validity conditions 176

case definitions
  outbreak investigation 65
  surveillance systems 262
case information, outbreak investigation 66
case-selection validity 177

case studies
precautionary principle 482, 482
b
qualitative research 105–106
cataracts 452
catastrophic potentials, risk perception 488
categorical variables 209, 209
t, 210
f
causal inference 183–191
cause & effect 187
confounding & measures of association 187–188
feasible interventions 189–190
life course epidemiology 277–279, 281–283
outcome removal 188
policy & beyond 190
potential outcomes & counterfactual causality 186–187
practical uses 188–190
risk factor removal 188–189
validity & bias 167–179
causal models, host–agent interactions 423–424
causation 183–191
modern theories of 183
cause & effect 183–184
causal inference 187
cause establishment, outbreak investigation 70

cause of death database
demography 320–321
Global Burden of Disease study 234
Cause of Death Ensemble model (CODEm) 234
CBA (cost-benefit analysis) 369–370
CBR (cost-benefit ratio) 379
CCA (cost-consequence analysis) 370
CDC *see* Center for Disease Control (CDC)
CDMA (code-division multiple access) 442
CDSR (Cochrane Database of Systematic Reviews) 151
CEA (cost-effectiveness analysis) 366, 370–371, 371
f, 378
CEAC (cost-effectiveness acceptability curve) 375–376, 376
f
cellular processes, toxicology 474–475
censoring 222–223
censuses 20–21, 21
b
Center for Disease Control (CDC)
Epidemic Intelligence Service 61
Framework for Programme Evaluation in Public Health 401
Morbidity and Mortality Weekly Report 269
Prevention Office on Smoking & Health 132
surveillance systems 260
Central Line Theorem 214
CEPH (Council on Education for Public Health) 301
CER (cost-effectiveness ratio) 370
CERQual 202
CERTCs (Clinical Epidemiology Research Training Centres) 158
CEUs (Clinical Epidemiology Units) 158
CFIR (Consolidated Framework for Implementation Research) 411
Chadwick, Edwin 344, 498
chains of risk model (trajectory model) 275
charting, qualitative data analysis 107

chemicals
interactions in toxicology 476
occupational health 459–460
safety assessment 476
Chernobyl (Ukraine, 1986) 449, 450–451
child labour, occupational health 462
child mortality, health metrics 25, 25
b
China, National Center for AIDS/HIV Control and Prevention 22–23
China Infectious Disease Automated-alert and Response System (CIDARS) 265
China's HIV/AIDS Comprehensive Response Information Management System (CRIMS) 22–23, 23
b
chi-squared distribution 216
chi-squared test 215–217, 216
t
chronic diseases 424
chronic obstructive pulmonary disease (COPD) 459
CIDARS (China Infectious Disease Automated-alert and Response System) 265
cigarette smoking, hepatocellular cancer 432
citizens, outbreak detection 60–61

civil society
health governance 395–396
sexuality 313
climate change, urbanization and 503–504
Climate Change Action Plan (Quito, Ecuador) 504
clinical epidemiology 149–159
applications 149
definition 149
evidence synthesis 150–151
research evidence location 149–150, 150
t
Clinical Epidemiology Research Training Centres (CERTCs) 158
Clinical Epidemiology Units (CEUs) 158

clinical information
outbreak causes 70
outbreak investigation 66

clinical manifestations
  epidemiology 41
  outbreak investigation 66
clinical practice guidelines (CPGs) 151
Clinical Practice Research Datalink (UK) 4, 5
clinical reference, synthesized summaries 151
cluster-level analysis 120–122
cluster randomized trials (CRTs) 113–128, 157
  analysis choice 124, 124
    b
  analysis strategies 120
  baseline imbalance 115
  between-cluster variability 114
  cluster-level analysis 120–122
  covariate adjustment 122–123
  design considerations 115
  design effect 114
  design strategy choice 116–117, 117
    b
  ethical issues 126
  experimental & observational units 114–115
  impact evaluation 118
  individual-level analysis 123–124
  logarithmic transformation 122
  matched-pair design 125–126
  matching 115–116
  methodology 113–114
  power considerations 120
  randomized controlled trials *vs.* 114–115
  reporting of 126–127
  restricted randomization 116
  sample size 114, 118–119, 120
    b
  statistical estimates 119–120
  stratification 116, 126

clusters
sampling 52, 208
trials and research randomization 8
types/size 115
CMA (cost-minimization analysis) 370
CMO (context-mechanism-outcome) hypotheses 357
Cochrane Collaboration 151, 203
effective practice and organization of care 158
LMIC research capacity 158
randomized controlled trials 196
ROBINS-1 tool 196
Cochrane Database of Systematic Reviews (CDSR) 151
Cochrane Handbook for Systematic Reviews of Interventions 203
Cochrane Library 149, 203
code-division multiple access (CDMA) 442
CODEm (Cause of Death Ensemble model) 234
coding, mortality causes 321
coefficient of variation 211
coefficients of association 219, 220
f
coffee drinking, myocardial infarction and 168
coherence, causation 185
cohort analysis, period analysis *vs.* 322
cohort studies 44, 85–99
analysis 90–96
applications 97
baseline visit 87
bias risk assessment 153
t
biomarker changes 90
case–control studies nesting 76
case–control studies *vs.* 83
data collection 85
design issues 85–88
exposure measures 89
exposure type & measures 88–89
follow-up investigations 87–88
intervention effects 88
joint analysis of survival & longitudinal data 96
longitudinal data analysis 95
matching 178
multivariate analysis 93–95, 94
t
outcome data 90–91, 90
t, 91
t
person-time analysis 91–92
population effectiveness 88–89
risk factors 85

study population 86–87, 87
t
survival analysis 92–93
survival data 222
timescales & outcome types 89–90
time to event 90
types of 85
validity 96
collider-stratification bias, life course epidemiology 277–278, 278
f
colonialism, urbanization 499
Commission on the Social Determinants of Health (WHO) 335
commodity ordering 31–32
common source outbreaks 68–69, 68
f
Communicating Risks and Benefits: An Evidence-Based User's Guide (FDA) 492

communication
health governance 397
t
risk *see* risk communication
two-way 486
f
communication technologies 29–36
data speed 29–34
drawbacks to 32–33
global *vs.* local trends (USA) 29
literacy 34
opportunities for 34
preferences 34
public health intentions 30–32
research field 33
social media and 30, 34
social networking 30
young people 30
communities 15, 129–130
health governance 396
health promotion, coronary vascular disease 135–137
organizing theory 134, 134
t
Ottawa Charter for Health Promotion 340
sexuality 313
transformation phase 135
community-based interventions
community organizing theory 134, 134
t
coronary vascular disease 134–135
health communication-behaviour change theory 134
tobacco smoking prevention 132
community diagnosis 15–19, 41
community engagement 17, 17
b
data management 19
future work 19
primary data collection 17–18
process of 15–16, 15
b, 16
b
secondary data collection 18–19, 18
b

community engagement
community diagnosis 17, 17
b
health governance 396, 397
t, 403
implementation science 413, 415
community intervention trials, HICs 129–140
developments 130
history of 129, 130
t
community-placed programmes 129–130
comorbidities, Global Burden of Disease study 235–236
Companion to Social Theory (Turner) 344
comparative analysis 214–219
group comparisons 214, 215
t
Pearson chi-squared test 215–217, 216
t
proportion/percentage comparisons 214–215, 215
t
comparison matching, cardiovascular disease community-based interventions 135

comparison validity
causal inference 168–169
prediction problems 164
compartmental infection transmission models 245
complex interventions 351–363
definitions 351
development stages 354
evaluation of 357–358
future work 360
health inequalities 354
implementation issues 359–360
logic model development 356–357, 356 *f*
policy & practice partners 355, 355 *f*
population health 353–354
process evaluation methods 358–359
public involvement 355
quantitative & qualitative evidence 354–355
research perspectives 353
research to practice 360
standardized *vs.* flexible 352–353
within system interventions 353–354
theory importance 355–356
complex sampling 208–209
comprehensiveness imperative, tobacco smoking prevention 132
comprehensive sexuality education (CSE) 310
comprehensive surveillance systems 263
compulsory surveillance systems 263
computed tomography (CT) 447
computer packages 107
Comte, August 292
Conducting Biosocial Surveys 301
Conference on Housing and Sustainability (Habitat III) (UN) 497
confidence intervals 213–214
matched-pair cluster randomized trials 125
means of 213–214
proportions for 209
t, 214
confirmability, qualitative research assessment 108
conflicts, community diagnosis 18
conflict theory 293
confounders 188

confounding
causal inference 169, 187–188
control in case–control studies 80, 82
t
definition 143
t
minimization in case–control studies 79–80
prediction problems 164–165
randomization 170
random sampling 165
variables 223
variables in case–control studies 79–80
confounding bias 277
congenital malformations, radiofrequency radiation 442
consistency, causation 184
Consolidated Framework for Implementation Research (CFIR) 411
Consolidation of Standards for Reporting of Trials (CONSORT) guidelines 126
CONSORT (Consolidation of Standards for Reporting of Trials) guidelines 126
construct error 166
construct validity 487
contamination, cluster randomized trials 115
content data analysis 106
context, complex interventions 359
context-mechanism-outcome (CMO) hypotheses 357
continuous common source outbreaks, epidemic curves 68–69
continuous models, infection transmission 246
continuous time models, infection transmission 248–250, 248
f
continuous variables 210
t
contraceptive use failure 325

control
case–control studies 78–82
definition 143
t
occupational health risk management 466
outbreaks 61
controlled trials, environmental health 430
control-selection validity 177–178
conversation data analysis 107
Cooper, C. R. 299
coordination, health governance 395
COPD (chronic obstructive pulmonary disease) 459
coroner reports 264
correlation 219
analysis assumptions 221
cost assignment, life course epidemiology 283–284
cost-benefit analysis (CBA) 369–370
cost-benefit ratio (CBR) 379
cost-consequence analysis (CCA) 370
cost-effectiveness
acceptability 371, 371
f
surveillance system evaluation 270
cost-effectiveness acceptability curve (CEAC) 375–376, 376
f
cost-effectiveness analysis (CEA) 366, 370–371, 371
f, 378
cost-effectiveness ratio (CER) 370
cost measurement 372–373
cost-minimization analysis (CMA) 370
costs, occupational health 458
cost-utility analysis 371–372, 371
b
Council on Education for Public Health (CEPH) 301
counselling, communication technologies 30
counterfactual 143
t
counterfactual causality 186–187
counterfactual framework 186

covariates
adjustment 122–123
Global Burden of Disease study 232–233
COVID-19 pandemic 339, 391
coxibs (cyclooxygenase-2 inhibitors) 8
Cox proportional hazards regression 225, 226
t
CPGs (clinical practice guidelines) 151
credibility, qualitative research 107
CRIMS (China's HIV/AIDS Comprehensive Response Information Management System) 22–23, 23
b
critical deconstructivism, health psychology 299
critical period model, life course epidemiology 275–277
crossover design, cluster randomized trials 117
cross-sectional studies 43–44, 47–58
analysis of 56–57
bias 54–56
bias risk assessment 153
t
biological sample collection 51
community diagnosis 16
definition 47
design issues 49–50
environmental health 426–429, 427
t
item non-response 52
mixed designs 47
national multipurpose surveys 47–49
non-responders 50–52, 52
f
sampling methods 52–53
sampling & response rates 50
uses 47–49
web-based research & wireless technologies 53–54
CRTs *see* cluster randomized trials (CRTs)
crude birth rate 323
b
crude death rate 324
b
CSE (comprehensive sexuality education) 310
culture, health sociology 294
cumulative incidence of disease 75–76
cumulative logit regression 226
cumulative risk model 275, 276
f, 489–490
cumulative sum techniques 430
CVD *see* cardiovascular disease (CVD)
cyberbullying 32–33
cyclooxygenase-2 inhibitors (coxibs) 8

DACA (Deferred Action on Childhood Arrivals) 146
DAGs (directed acrylic graphs) 275–276, 279, 283
DALYs *see* disability-adjusted life years (DALYs)

data
aggregation 268
analysis *see* data analysis
collection *see* data collection
content type 263
extraction, systematic reviews 196
Global Burden of Disease study 232, 233–234
insufficiencies 14
linkage 268–269
management *see* data management
quality *see* data quality
reporting 269
security in electronic health records 9
speed communication technologies 29–30
transmission 269
visualization in communication technologies 32

data analysis
community diagnosis 16
demography 319–320, 321–322
practice of 107
purpose of 106–107
surveillance systems 268
databases, electronic *see* electronic databases

data collection
cohort studies 85
communication technologies 32
community diagnosis 15–16, 16 *b*
sources 263–264
surveillance systems 262–268

data management
community diagnosis 19
surveillance systems 269

data quality
  electronic health records (EHRs) 8
  qualitative research 105
  surveillance system evaluation 270
DCA (Disease Choice Experiment) 373
DCEA (distributional cost-effectiveness analysis) 377
DDREF (Dose and Dose-Rate Effectiveness Factor) 448
death *see* mortality
death certificates 320
  public health information 3
death registries 320
deceased controls, case–control studies 79
DECIPHer public health research centre 355
decision analysis, economic appraisal models 375, 376
decision-making
  economic theories 381–382
  scarcity theory 386
decomposition of overall population change 326
Deferred Action on Childhood Arrivals (DACA) 146
degrees of freedom (df) 214
delivery of complex interventions 358
Demographic and Health Survey Programme (DHS) 321
demographic transition, stable population theory 329–330
demography 317–333
  cause of death database 320–321
  data analysis 319–320, 321–322
  definition 317
  global issues 317–319, 317
    f, 318
    b
  life tables 324–325
  migration measurement 325–326
  model life tables 325
  mortality measurement 323–324
  outbreak investigation 66
  population dynamics 326–327
  recent trends 331
  reproduction rates 322–327
  stable population theory *see* stable population theory
De Morbis Artificum Diatriba (Ramazzini) 457
Denner, J. 299
denormalizing 131–132, 133
dependability, qualitative research assessment 108
dependent error, prediction problems 166
dependent variables 220
Descartes, René 292

descriptions
  outbreak investigation 66–69
  qualitative data 104
descriptive statistics 56–57
design effect (DEff) 114
design strategy, cluster randomized trials 116–117, 117
  b
detail supply, qualitative assessment 489
determinants framework, implementation science 411
deterministic infection transmission models 245–246
Developmental Origins of Health and Disease hypothesis 276–277, 280
development assistance, public health programmes 391
DFLE (disability-free life expectancy) 230
DHS (Demographic and Health Survey Programme) 321

diagnosis
community diagnosis 16
verification in outbreak investigation 64–65
diagnostic radiography 447

diagnostic test accuracy studies
bias risk assessment 153
t
evidence synthesis 201–202
dialogue, health governance 396
diaries 106
diarrhoeal disease 77
DiD (difference in differences) 144
difference equation 247–248
difference in differences (DiD) 144
differential error, prediction problems 166
Diffusion of Innovation theory 33
dimensions of risk, risk perception 488–489, 488
f
diphtheria 254, 254
f
directed acrylic graphs (DAGs) 275–276, 279, 283
disability-adjusted life years (DALYs) 373
calculation 236
cost-utility analysis 371–372
Global Burden of Disease study 230–231
health gain benefits 373
risk factors and 236
disability-free life expectancy (DFLE) 230
disability metrics 25
b
disability weights (DWs) 231, 235
disabled workers 462
discount rates 372–373
discourse analysis 107
discrete infection transmission models 246
discrete time models 246–248, 250–251, 251
f
discrete variables 210
t

disease
burden in occupational health 458–463
characteristics in surveillance systems 262
epidemiology, Global Burden of Disease study 234
history, epidemiology 41
knowledge in outbreak investigation 61
notification of 468
precursors in epidemiology 42
prevalence, Global Burden of Disease study 234–235
spectrum of 41
surveillance 3
transmission mechanisms 42–43
Disease Choice Experiment (DCA) 373
Disease Control Priorities Project 296
disease control specialists 63
disease registries 265–266
disease-specific information systems 22
DisMod-MR 234–235
dissemination, community diagnosis 16
distribution, toxicology 475–476
distributional cost-effectiveness analysis (DCEA) 377
distribution cost-effectiveness analysis 377–378
divorce, life table analysis 325
DNA damage, ionizing radiation 448
document-related qualitative research 106
domains of stewardship (WHO) 401
domination-deprivation model 293
dose, complex interventions 358–359
Dose and Dose-Rate Effectiveness Factor (DDREF) 448
dose–response relation
risk assessment 477, 477
b, 478
b
toxicology 473–474, 474
f
dropout, community diagnosis 17–18
drugs, discontinuation and non-compliance 7
Du Bois, W. E. B. 499
Durkheim, Emile 291, 292–293
Dutch Hunger Winter (1944-1945) 281–282
DWs (disability weights) 231, 235
dynamic modelling 241, 242
t, 245 *see also* infection transmission models

EA *see* economic appraisal (EA)
EASD (European Association for the Study of Diabetes) 301
Eastern Mediterranean Regional Office's ten governmental principles (WHO) 401–402
EBM (evidence-based medicine) 347

Ebola epidemic (2014-2015) 402–403, 403
t
ECEA (extended cost-effectiveness analysis) 377
ECHA (European Chemical Agency) 475
ecological imperative, tobacco smoking prevention 132–133
ecological studies 43
environmental health 425–426
ecological wheel model 424
economic appraisal (EA) 365–380
benefit measurement 373
cost-benefit analysis 369–370
cost-consequence analysis 370
cost-effectiveness analysis 370–371, 371
f
cost measurement 372–373
cost-utility analysis 371–372, 371
b
distribution cost-effectiveness analysis 377–378
economics of 365–369
efficiency 367–368
equity and 368–369, 368
t, 376–378
extended cost-effectiveness analysis 377
financial appraisal *vs.* 365
formulation of 369
health gain 373–374, 373
f
investment appraisal 370
marginal costs & benefits 366–367
models of 374–376
opportunity cost 365–366
public health intervention 378–379, 378
b
social costs & benefits 366
types of 369–374
economic efficiency 367–368
economic imperative, tobacco smoking prevention 131

economics
appraisal *see* economic appraisal (EA)
behaviour and *see* behavioural economics
core predictions 382
decision-making 381–382
health and 500
occupational health 458
ECRHS (European Community Respiratory Health Survey) 49, 51–52

education
community diagnosis 18
occupational health risk management 467
tobacco smoking prevention 132
effectiveness perspective 353
effective practice and organization of care (EPOC) 158
efficiency 353
economic appraisal 367–368
eHealth 29
EHRs *see* electronic health records (EHRs)
EI (emotional intelligence) 103
elderly, proportion in population 328–329
electric fields *see* static radiation fields
electromagnetic spectrum 437, 438
f

electronic databases
  indexing bias 193–194
  systematic reviews 195
electronic health records (EHRs)
  cohort studies 86–87
  data quality 8
  development stages 5
    t
  ethical issues 9
  hospital referrals and 6
  information systems, HICs 4
  links between 5
  pandemics and 6
  research challenges 8–9
electronic medical records (EMRs) 22
ELF (extremely low frequency) fields 439–440
eligibility rules, systematic review screening 195–196
embodiment, life course epidemiology 279–280
emergency response planning 468
emotional intelligence (EI) 103
empowerment, health promotion 337
empowerment liberatory health education 342–343
EMRs (electronic medical records) 22
enabling, health promotion 337
enclosure, engineering controls 466

endemic diseases
  environmental health 422
  geographical clustering 425
endowment effect, behavioural economics 384
engineering controls, occupational health risk management 466

environment
building in tobacco smoking prevention 133
monitoring in occupational health 464–465
outbreak causes 70
outbreak investigation 69–70
environmental health 421–436
biomarkers 431, 431
t, 432
t
cross-sectional surveys & longitudinal studies 426–429, 427
t
ecological studies 425–426
historical perspective 421–422
host–agent interactions 422–425, 424
f
individual susceptibility 431–433
interventions 433
intervention studies 429–431
risk assessment 433 *see also* Blackfoot disease studies; environmental risk analysis; toxicology
Environmental Impact Assessment Act (Taiwan) 480

environmental levels
extremely low frequency fields 439
infrared radiation 443
intermediate frequency radiation 440
ionizing radiation 445–448
radiofrequency radiation 441
static radiation fields 437–438
ultraviolet radiation 444
environmental metrics 25
b
environmental models, health psychology 299
environmental risk analysis 473

epidemic curves
common source outbreaks 68
continuous common source outbreaks 68–69, 68
f
propagated source outbreaks 69
Epidemic Intelligence Service (CDC) 61

epidemics
  unknown aetiology 42 *see also* outbreaks
epidemiological inference 161–162
epidemiological research validity & bias 161–182
  inference & validity 161–162
  statistical misinterpretations 162–163
epidemiological transition, stable population theory 330–331
epidemiological triangle model 424
epidemiology 39–46
  applications of 43
  clinical epidemiology *see* clinical epidemiology
  definition 39–41
  ionizing radiation 451
  methodological issues 45
  occupational health 458–463
  outbreak causes 70
  proof provision 40–41
  public health support 41–43
  research validity & bias *see* epidemiological research validity & bias *see also* clinical epidemiology; life course epidemiology
EPIET (The European Programme for Intervention Epidemiology Training) 73
epigenetics 280
epilepsy, extremely low frequency fields 439
EPOC (effective practice and organization of care) 158

equity
  economic appraisal and 368–369, 368
    t, 376–378
  impact analysis 368
  trade-off analysis 368
ERR (excess relative risk) 451
errors, rolling censuses 320
Escherichia coli O104:H4 infections 69, 70
  t
estimation, Global Burden of Disease study 231–232

ethical issues
  cluster randomized trials 126
  electronic health records 9
  surveillance systems 270–271
ethnography 104
ethnomethodology 293–294
EU-ADR project 6
European Agency for Accreditation in Public Health Education (APHEA) 301
European Association for the Study of Diabetes (EASD) 301
European Chemical Agency (ECHA) 475
European Community Respiratory Health Survey (ECRHS) 49, 51–52
European Medicine Agency 155
The European Programme for Intervention Epidemiology Training (EPIET) 73
European Union (EU), occupational health legislation 458

evaluation
  implementation science 411
  public health programme management 400–401
event-based surveillance 263
event definition, qualitative assessment 489
evidence access in LMICS 157–158
evidence-based medicine (EBM) 347

evidence indirectness
GRADE 152
LMICS 158

evidence quality
GRADE 151–152, 154
systematic reviews 199–200
evidence synthesis 201–202
clinical epidemiology 150–151
evolutional spiral model 424–425, 424
f
excess relative risk (ERR) 451
excretion, toxicology 476
exercise, cardiovascular disease and 282
existence assessment, outbreak investigation 63
ExpandNet/WHO Framework 413
experimental epidemiology 129
experimental evidence, causation 185–186
experimental studies 44–45
experimental units, cluster randomized trials 114–115
exponential decline of population 248, 248
f
exposed people, nonexposed people *vs.* 75–76, 75
t, 76
t
exposure limitation, ultraviolet radiation 444–445

exposure measures
cohort studies 88–89
definition 143
t
occupational health 464–466
toxicological risk assessment 477–478, 477
b, 478
b

exposure status
  cohort studies 88–89
  confounding in case–control studies 80
  incomparable assessment 81–82
  retrospective ascertainment 76
extended cost-effectiveness analysis (ECEA) 377
external comparisons, univariate analysis of cohort studies 91–92, 92
  t
external records, cohort studies 87
external validity, cohort studies 96
extremely low frequency (ELF) fields 439–440
eye, infrared radiation and 443

factorial design
  cluster randomized trials 117
  implementation science 412–413
failure time data 222
familiarization, qualitative data analysis 107
family psychotherapy 300
family related behavioural change 330
Farr, William 320
FDA *see* Food and Drug Administration (FDA)
federal governments 394, 404
feedback, health governance 396
FELTP (Field Epidemiology and Laboratory Training Programme) 73
femininity, sexuality and 311

fertility
decline *vs.* economic threshold 329
definition 323
b
measures of 322, 323
b
proximate determinants of 330
reduction in 328
trends in 329–330
FETP (Field Epidemiology Training Programme) 73
fidelity, complex interventions 358
field epidemiologists 63
Field Epidemiology Training Programme (FETP) 73
field operation planning, outbreak investigation 64

finance
  appraisal *vs.* economic appraisal 365
  health governance 394
Fisher, Ronald 130
5CP (Stanford Five-City Project) 136–137
five governance principles (United Nations Development Programme) 401
fixed effects models 144
flexibility, surveillance systems 270
focus groups, qualitative research 105
follow-up investigations
  cohort studies 87–88
  outbreak investigation 72
follow-up validity
  causal inference 168, 171–172
  prediction problems 164, 165
Food and Drug Administration (FDA)
  Communicating Risks and Benefits: An Evidence-Based User's Guide 492
  health resource optimization 155
  Voice of the Patient initiative 489
formative evaluation, tobacco smoking prevention 132
Framework for Programme Evaluation in Public Health (CDC) 401
framework of health governance 393–394, 393
  f
Framingham Heart Study 86, 87
  t
  cardiovascular disease 136
Frank, John 344
freedom of information, health governance 402

frequency
  categorical variables 209
  quantitative variables 209, 210
    t
Freud, Sigmund 298–299
Fundamental Aspects of Governance (World Bank) 401

Galton, Francis 499
gamma rays *see* ionizing radiation
GAMs (generalized additive regression models) 227
Gatehouse whole-school intervention 356
GAVI (Global Alliance for Vaccines and Immunizations) 398
GBD *see* Global Burden of Disease (GBD) study
GBD Compare 236
GEAS (Global Early Adolescent Study) 310
GEEs *see* generalized estimating equations (GEEs)
gendered experience, sexuality 309–310
general fertility rate 323
  b
general fertility ratio 322
General Household Survey (UK) 266

generalization
  causal inference 171
  cross-sectional studies 49
generalized additive regression models (GAMs) 227
generalized estimating equations (GEEs) 227
  cluster randomized trial analysis 123–124

genetics
  confounding in case–control studies 80
  environmental health 423
  epidemiology 43
genetic tests, occupational exposure limits 466
geographical clustering, endemic diseases 425
geographic information systems 24
geometric mean 209

germ theory
environmental health 421
urbanization 498–499
GFR (gross reproductive rate) 322–323
Global Alliance for Vaccines and Immunizations (GAVI) 398
Global Burden of Disease (GBD) study 229–240
abbreviations 230
t
analytical steps 232–236, 233
f
communicable diseases 237, 239
data synthesis 232
b
diseases & injuries 230
t
estimation principles 231–232
future work 239
key results 237, 237
f, 238
f, 239
key terms 232
b
population surveys 231
result exploration 236–237
value choices & population health metrics 230–231
Global Early Adolescent Study (GEAS) 310
global health governance 392
Global Observatory for eHealth (GOe, WHO) 29
Global Outbreak Alert and Response Network (WHO) 73
Global Public Health Intelligence Network (GPHIN) 267
Global Report on Urban Health (WHO/UN Habitat) 497
Global Solar UV Index (UVI) 444
GOe (Global Observatory for eHealth, WHO) 29
goitre, Taiwan 430–431

governance
public health programmes *see* health governance
urbanization and health 501
governments, health governance 394–395
GPHIN (Global Public Health Intelligence Network) 267
GRADEPro 203
GRADE Working Group 203
Grading of Recommendations Assessment, Development and Evaluation (GRADE) 151–152, 154
bias risk 152, 153
t
evidence indirectness 152
evidence quality 199–200
impression 152, 154, 154
f
publication bias 154
public health guidelines 154–155
result inconsistency 152
strength of recommendations 154
Graunt, John 319
gross reproduction rate (GRR) 323
b
gross reproductive rate (GFR) 322–323
group behaviours, health and 500–501
group comparisons 214, 215
t
GRR (gross reproduction rate) 323
b
GUEST study (Guidance for Feasibility Studies of Complex Public Health Interventions) 357
Guidelines for Drinking-Water Quality (WHO, 1993) 433
guideline transferability, LMICS 158

HAALSI (Health and Aging in Africa: A Longitudinal Study of an INDEPTH Community in South Africa) 284
HAART (highly-active anti-retroviral therapy) 88
Habitat III (Conference on Housing and Sustainability, UN) 497
HAC (Humanitarian Action for Children) 402–403
HADS (Hospital Anxiety and Depression Scale) 301
HALE *see* health-adjusted life expectancy (HALE)
Handbook for Guidance Development (WHO) 202

harmful effects
infrared radiation 443
ultraviolet radiation 444
hazard function 92, 93
f

hazards
elimination/substitution 466
occupational health risk assessment 464
toxicological risk assessment 477, 477
b
HBV *see* hepatitis B virus (HBV)
HCC *see* hepatocellular cancer (HCC)
HCV (hepatitis C virus) 429
HCWs *see* healthcare workers (HCWs)
head-to-head randomized implementation trial design 412

health
definition in health promotion 337–338
economic appraisal 373–374, 373
f
education 342–343
equity *see* health equity
expectancies in Global Burden of Disease study 230
extremely low frequency fields 439
governance *see* health governance
inequalities in complex interventions 354
infrared radiation 443
intermediate frequency radiation 440
ionizing radiation 448–449
problems in surveillance systems 261
promotion *see* health promotion
radiofrequency radiation 441–442
static radiation fields 438
ultraviolet radiation 444
urbanization 500–501 *see also* population health
health-adjusted life expectancy (HALE) 232
Global Burden of Disease study 230
Health and Aging in Africa: A Longitudinal Study of an INDEPTH Community in South Africa (HAALSI) 284
The Health and Physique of the Negro American (Du Bois) 499
Health and Retirement Study (USA), International Partner Studies 284
Health and Safety at Work Act (UK, 1974) 458

healthcare systems
delivery metrics 25
b
health sociology 296
information systems 267–268
providers networks 265
healthcare utilization metrics 25
b
healthcare workers (HCWs)
occupational health infections 460
outbreak detection 60
training 61
health communication-behaviour change theory 134

health equity
qualitative research 102
trade-offs 368
urbanization 500
health governance 391–408
actors in 394
t, 398
t
assessment framework 401–402
civil society 395–396
community engagement 403
concepts of 392
framework of 393–394, 393
f
government roles 394–395
levels of 397–401
local governments 403
national governments 403–404
national level 399
private sector 395
public health organizations 399–400
public–private partnerships 404–405
supranational governance systems 402–403
supranational level 397–399
theories of 392–393
tools of 396, 397
t
health impact assessment (HIA) 340
Health in All Policies (HiAP) movement 335, 346, 399, 504
Health Inequalities Assessment Toolkit 354
Health Information Technology for Economic and Clinical Health (HITECH) Act 267–268
health metrics 24–26, 25
b
morbidity metrics 26
mortality metrics 24–25
health policy metrics 25
b
health promotion 335–350
beliefs & assumptions 336–337
definition 335, 336
enabling & empowerment 337
evaluation of 346–347
health definition 337–338, 338
t
health education 342–343
health service re-orientation 341–342
personal skill development 340–342
political economy 345–347
politics & philosophy 344
process of 336

public health 343–344
research in 346–347
social justice 345
social movement 344–345 see also *Ottawa Charter for Health Promotion*
Health Promotion Glossary 345
health records, electronic *see* electronic health records (EHRs)
health-related quality of life (HRQOL) 373
Health Relief Model 299
health resource optimization 155–157
evidence application to individuals 155–156, 156 *f*
evidence application to settings 156–157
Health Sociology Review 292
Health Survey in England 47

health systems
evaluation of 157
provision 396
re-orientation 341–342
strengthening of 396
health technology assessment (HTA) 365 *see also* economic appraisal (EA)
health volunteers, outbreak detection 60–61
health worker effect 54
Healthy Cities programme 500
healthy life expectancy 232
Global Burden of Disease study 236
healthy worker effect (HWE) 458
hearing loss, occupational health 460
hepatitis B virus (HBV) 429
infection risk factors 432–433
Taiwan vaccination programme 430
hepatitis C virus (HCV) 429
hepatocellular cancer (HCC)
environmental health 424–425
multifactorial aetiology 429
risk factors 432
herd immunity 244
heritable defects, ionizing radiation 452
heterogeneity of effect 171
heteronormativity 310
HIA (health impact assessment) 340
HiAP (Health in All Policies) movement 335, 346, 399, 504
Hidden Cities: Unmasking and Overcoming Health Inequities in Urban Setting (WHO/UN Habitat) 497
high-income countries (HICs)
community intervention trials *see* community intervention trials, HICs
health behaviour 385
information systems *see* information systems, HICs
psychotherapy 299
registration systems 320
highly-active anti-retroviral therapy (HAART) 88
Hill, Austin Bradford 184
Hill's considerations 184–186
Hippocrates 421, 498
Hiroshima atomic bombing (1945) 449, 451
HIV infection 307–309
cumulative risk 489–490
epidemiological studies 40
epidemiology of 308
incidence & sexually-transmitted infection 384
mortality rate effects 318–319
prevention trials 415
surveillance 266, 267 *see also* AIDS
HMG-CoA reductase inhibitor drugs (statins) 8
Hobbes, Thomas 292

horizontal equity, healthcare 368–369
Hospital Anxiety and Depression Scale (HADS) 301
Hospital Episode Statistics 4

hospitals
acute care setting 341
referrals & electronic health records 6
host–agent interactions 422–425, 424
f
host components, environmental health 423, 423
t
host factors, cohort studies 88
household surveys, information systems, LMIC 22, 22
b
HRQOL (health-related quality of life) 373
Humanitarian Action for Children (HAC) 402–403
HWE (healthy worker effect) 458
hybrid designs, implementation science 414
t
hyperthermia, infrared radiation 443

hypothesis testing
outbreak investigation 69, 70
t
surveillance systems 260–261

IARC *see* International Agency for Research on Cancer (IARC)
IAVI (International AIDS Vaccine Initiative) 404
ICC (intracluster correlation coefficient) 114
ICD-10 *see* International Classification of Diseases and Related Health Problems (ICD-10)
ICNIRP *see* International Commission on Non-Ionizing Radiation Protection (ICNIRP)
I-Corps (Innovation Corps) initiative (National Science Foundation) 416
ICPD (International Conference on Population and Development) 313
ICRP (International Commission on Radiological Protection) 452
IDS (integrated disease surveillance) 263
IF (intermediate frequency) radiation 440
IHME *see* Institute for Health Metrics and Evaluation (IHME)
IHR (International Health Regulations) 260
ILI (influenza-like illness) surveillance 266
ill controls, case–control studies 79
Illness Perception Questionnaire 301
ILO *see* International Labour Organization (ILO)
immediate risk model, life course epidemiology 275
immunization registries 4
impact evaluation, cluster randomized trials 118
impact inventory, economic appraisal 378–379
impairment level cross-validation, Global Burden of Disease study 235
implementation, public health programme management 400
Implementation Outcomes Framework (IOF) 411
implementation science 409–418, 414
t
definitions 409–410
epidemiology 43
frameworks 410–411, 411
t
future work 415–416
public health programme management 400
scale-up frameworks 411–412
stakeholder & community engagement 413, 415
terminology 410
impression, GRADE 152, 154, 154
f
improvement schemes, urbanization 498–499
IMRs (infant mortality rates) 324, 324
b
inaccurate judgements, risk perception 486
incidence, morbidity metrics 26
incidence assessment errors, causal inference 174

incidence–prevalence bias (Neyman bias) 54, 55
f
incident cohort 86
INCLEN (International Clinical Epidemiology Network) 158
INCLUSIVE intervention 356
f, 357, 358
incomplete observation, cohort studies 89–90
independent conceptual basis, health promotion 343
independent error 166
independent variables 220
in-depth interviews, qualitative research 104–105
index cases, outbreak investigation 67
indexing, qualitative data analysis 107
indexing bias 193–194
indicator-based surveillance 263
indirect effects, causal inference 171
individual-based models, infection transmission 245
individual-level analysis 123–124
individual-level regression 125
individual patient analysis 201
individual susceptibility, environmental health 431–433
Industrial Revolution, occupational health 457–458
industrial sources, ionizing radiation 447
infant mortality, diarrhoeal disease 77
infant mortality rates (IMRs) 324, 324
b

infections
fall in mortality 331
occupational health 460
infection transmission models 241–258, 245
b
building blocks of 250–254
examples 254–255, 254
f
fundamental models 246–250
fundamentals of 242–244, 243
b
historical overview 244–245
infection types 241–242
model world 243–244, 243
b
problems with 241–242
susceptible infectious compartment models 252–254, 253
f
temporal growth/decline 246–250
types of models 245–246
Infectious Diseases of Humans: Dynamics and Control (Anderson & May) 245
inflation adjustments 372
influenza-like illness (ILI) surveillance 266

information
flow in surveillance systems 269
formatting of 491–492
health governance 397
t
outbreak investigation 63, 66
provision in communication technologies 30
quality in cohort studies 87–88
registration systems 320
selection, communication 491
surveillance systems 260, 268–269

information bias
case–control studies 80–82
systematic reviews 194
information systems, HICs 3–11
administrative databases 4
disease surveillance 3
electronic health records 4
future developments 6–7
intervention assessments 8
population surveys 3
public health research randomization 7–8
recent developments 4–6, 5
t
registries 4
vital statistics 3 *see also* electronic health records (EHRs)
information systems, LMIC 13–28
census 20–21, 21
b
community diagnosis *see* community diagnosis
future work 26–27
geographic information systems 24
health metrics *see* health metrics
household surveys 22, 22
b
multisectoral collaborations 27
private–public partnerships for health 26–27
routine health information systems 22
sentinel surveillance systems 24
verbal autopsy 21–22
vital registries 21
informed consent 126
informed electorate imperative, tobacco smoking prevention 132
infrared (IR) radiation 443
initiation, community diagnosis 15
injecting drug users, HIV infection/AIDS 308
injury analysis, Global Burden of Disease study 235
Innovation Corps (I-Corps) initiative (National Science Foundation) 416
Institute for Health Metrics and Evaluation (IHME) 229–230
mortality estimates 261, 261
t
instrumental variables (IVs) 88–89, 145, 146
integrated disease surveillance (IDS) 263
intermediate frequency (IF) radiation 440
intermediate variables 170–171
internal comparisons, univariate analysis of cohort studies 91, 91
t

internal validity
causal inference 168
cohort studies 96
cross-sectional studies 49
International Agency for Research on Cancer (IARC) 477
extremely low frequency fields 440
International AIDS Vaccine Initiative (IAVI) 404
international civil society organizations 403
International Classification of Diseases and Related Health Problems (ICD-10) 232
death certificates 320
occupational health 458 *see also* Global Burden of Disease (GBD) study
International Clearinghouse for Birth Defects Surveillance and Research 265–266
International Clinical Epidemiology Network (INCLEN) 158
International Commission on Non-Ionizing Radiation Protection (ICNIRP) 439
ultraviolet radiation 445
International Commission on Radiological Protection (ICRP) 452
International Conference on Population and Development (ICPD) 313
International Health Regulations (IHR) 260
International Labour Organization (ILO)
migrant workers 463
occupational health definition 457
International Nuclear Workers Study (INWORKS) 450
International Partner Studies, Health and Retirement Study (USA) 284
International Society for Infectious Diseases, Program for Monitoring Emerging Diseases 267
International Sociological Association, Research Committee on Medical Sociology 292
International Standards for Drinking Water (WHO, 1958) 433
International Study of Asthma and Allergies in Childhood (ISAAC) 49
sampling methods 53
third phase of 50

internet
  cross-sectional studies 53–54
  information systems, HICs 7
  outbreak detection 60
  sexuality and 313
  surveillance systems 267
interrupted time series (ITS) 145
interval scale, health sociology 297
intervention assessments 376
  information systems, HICs 8

interventions
behavioural economics 387
cardiovascular disease community-based interventions 135
cohort studies 88
definition 143
t
environmental health 433
epidemiology 42

intervention studies
bias risk assessment 153
t
environmental health 429–431
interview-based studies
case–control studies 76–77
fatal diseases 81
qualitative research 104–105
intracluster correlation coefficient (ICC) 114
intrinsic rate of natural increase 327
inverse prevention law 354
investment appraisal 370
in vivo models, infection transmission models 246
INVOLVE, National Institute for Health Research 355
INWORKS (International Nuclear Workers Study) 450
IOF (Implementation Outcomes Framework) 411
ionizing radiation 445–452
accidental sources 450–451
health effects 448–449
medical procedures 447–448
policy & prevention 452
protection principles 452
sources & environmental levels 445–448
units of 446
f, 446
t
IR (infrared) radiation 443
ISAAC *see* International Study of Asthma and Allergies in Childhood (ISAAC)
isolation, engineering controls 466
item non-response, cross-sectional studies 52
ITS (interrupted time series) 145
IVs (instrumental variables) 88–89, 145, 146

James, William 298
Journal of Health and Social Behaviour (JHSB) 292

Kant, Emmanuel 292
Kaplan–Meier curves
cohort studies 91
survival analysis of cohort studies 93, 93
f
Kermack, William 244–245
kinship networks, sexuality 310
knowledge action gap, LMIC 409
Knowledge to Action (K2A) Framework 410

laboratories
measurements in case–control studies 77
outbreak causes 70
outbreak detection 60
surveillance systems 266
urbanization science models 499
laboratory technicians 63

Lalonde report
health definition 338
A New Perspective on the Health of Canadians 343
lamivudine 430
Laplace, Pierre-Simon 213
LASI (Longitudinal Study of Aging in India) 284
Last, John 39–40
late entries, cohort studies 89
Lazarsfeld, Paul 297
LBP (low back pain) 461

legislation
occupational health 458
outbreak investigation 61–62
sexuality 312–313
LET (linear energy transfer) 445
life course epidemiology 275–288
biological mechanisms 279–281
causal inference 277–279
conceptual models 275–277, 276 *f*
future directions 284
goals of 277
methodological innovation 281–284
sequence analysis 283–284

life expectancy
at birth 325
increase of 421
socioeconomic groups *vs.* 230 *see also* healthy life expectancy
Life Expectancy and Probability of Death tool 237
lifestyle, health sociology 294–295
life tables 324–325, 324
b
demography 324–325
Global Burden of Disease study 230, 233
life years lost 487
limited attention, behavioural economics 385
linear energy transfer (LET) 445
literacy 34
literature reviews, systematic reviews *vs.* 194
t
LMICs *see* low- and middle-income countries (LMICs)
local contextual analysis 415

local governments
health governance 403
outbreak investigation 63–64

location
  outbreak investigation 67–68
  statistical methods 209–210
Locke, John 292
logarithmic transformation 122
log-binomial regression 226
logic model development 356–357, 356
  f
logistic regression 223–224, 224
  t
lognormal models 94
logos 102
London drivers, cardiovascular disease 54
longitudinal data analysis, cohort studies 95, 96, 96
  f
longitudinal studies 426–429, 427
  t
Longitudinal Study of Aging in India (LASI) 284
loss aversion, behavioural economics 384
low- and middle-income countries (LMICs) 14
  b
  environmental health 423
  evidence access 157–158
  health behaviour 383, 385, 387
  information systems *see* information systems, LMIC
  knowledge action gap 409
  population ageing 318
  population growth 317–318
  surveillance systems 271
low back pain (LBP) 461
lung diseases, occupational health 459

macro-costing 372
MACS *see* Multicentre AIDS Cohort Study (MACS)
magnetic fields *see* static radiation fields
magnetic resonance imaging (MRI) 437, 438–439
management of public health programmes 400–401 *see also* health governance
Management Systems International (MSI), Scaling Up Management Framework for Practitioners 413
man-made sources, ionizing radiation 447

mapping
  Global Burden of Disease study 233–234
  qualitative data analysis 107
marginal benefits, economic appraisal 366–367
Marmot, Michael 335
married fertility rates 322
Martignacco Project (1977) 136
Massachusetts Department of Public Health (USA) 5
mass media 71
matched-pair cluster randomized trials 125–126

matching
  cluster randomized trials 115–116
  cohort studies 178
maternal mortality ratio (MMR) 25
McIntyre, Charles 292
McKendrick, Anderson 244–245
MDGs *see* Millennium Development Goals (MDGs)
mean 209, 211
  f
  comparison 217, 217
    t
  confidence intervals 213–214
mean change 219

measles
infection transmission models 254
outbreak (2015) 254
f
measured confounders 169–170
measurement error, separation of effects 174
measurement errors, life course epidemiology 278–279, 278
f

measurement validity
causal inference 168, 172
prediction problems 164, 165–166
measure of effect 187
measures of association, causal inference 187–188
Mèdecins San Frontières (MSF), Ebola epidemic 403
media 60
median 209
mediation, *Ottawa Charter for Health Promotion* 339
medical examination, reports 264
medical examinations 467–468
Medical Expenditure Panel Surveys (USA) 235–236
medical procedures, ionizing radiation 447–448, 448
t
Medical Sociology Research Network, British Sociology Association 292
Medical Sociology Section, American Sociology Association 292
medical sources, ionizing radiation 447
Medicare Benefits Schedule and Pharmaceutical Benefits Scheme, health resource optimization 155
Medicines for Malaria Venture (MMV) 404
MEMA kwa Vijana (MkV) trial (Tanzania) 121–122, 122
t
menstruation onset 310
mental processes, qualitative assessment 490
men-who-have-sex-with-men (MSM) 311–312
men-who-have-sex-with-men-and-women (MSMW) 311–312
MERS (Middle East respiratory syndrome) 460
message characteristics, tobacco smoking prevention 133–134, 134
t
meta-analysis 45
systematic reviews 197–198
metabolism, toxicology 476
meta-regression 198
method of least squares 220
methodological innovation, life course epidemiology 281–284
methodological issues, epidemiology 45
Mexico, *Opportunidades* programme 146
MHHP (Minnesota Health project) 137
miasma theory 421
micro-costing 372
Middle East respiratory syndrome (MERS) 460
migrant workers 463
measurement of 325–326
military personnel, census weaknesses 320
Millennium Development Goals (MDGs) 13–14
effectiveness of 14
progress monitoring 14
women workers 462 *see also* Sustainable Development Goals (SDGs)
mining big data 202
Ministry of Health, health governance 398
Minnesota Health project (MHHP) 137

miscarriages 440
misclassification, causal inference 173
mixed health systems, health governance 395
MMR (maternal mortality ratio) 25
MMV (Medicines for Malaria Venture) 404
mobile telephones 29
  radiofrequency radiation 441
mode 210, 211
  f
model 243
  b
model life tables, demography 325
Model of Knowledge Translation (Canadian Institutes of Health Research) 410
model world 243
  b
moderate-to-vigorous intensity physical activity (MVPA) 282
molecular determinants, epidemiology 43

Monte Carlo simulation studies
economic appraisal models 375
life course epidemiology 283
Montesquieu, Charles 292
morbidity, occupational health 458
Morbidity and Mortality Weekly Report (CDC) 269
morbidity metrics 25
b, 26

mortality
age-adjusted cancer mortality study 426
all-cause mortality 234
body mass index *vs.* 188–189
causes of 321
child mortality 25, 25
b
children 233
decreases in 326
estimates 261, 261
t
estimation and risk factors 236
Global Burden of Disease study 233, 234
levels 318–319
measurement of 323–324
measures of 324
b
morbidity metrics 26
occupational health 458
probability of 247, 487
rates by age 319, 319
f
Mortality and Causes of Death visualization, Global Burden of Disease study 236
mortality metrics 24–25, 25
b
MOST (multiphase optimization strategy implementation trial) 413
motor vehicles, pedestrianism and 500
Moving to Opportunity (MTO) for Fair Housing Demonstration program 275–276
MS (multiple sclerosis), extremely low frequency fields 439
MSI (Management Systems International), Scaling Up Management Framework for Practitioners 413
MSM (men-who-have-sex-with-men) 311–312
MSMW (men-who-have-sex-with-men-and-women) 311–312
MTO (Moving to Opportunity) for Fair Housing Demonstration program 275–276
Multicentre AIDS Cohort Study (MACS) 86, 87
t, 88
longitudinal data analysis 95, 96
f
multivariate analysis 94–95, 95
f
multicomponent multilevel community interventions, tobacco smoking prevention 133
Multi-dimensional Health Locus of Control scale 301
multimodal regression 226
multiphase optimization strategy implementation trial (MOST) 413
multiple sclerosis (MS), extremely low frequency fields 439
multisectoral collaborations 27
multivariate analysis 223
cohort studies 93–95, 94
t
univariate analysis of cohort studies 92

multiwalled carbon nanotubes (MWCNTs) 479, 479
b
MVPA (moderate-to-vigorous intensity physical activity) 282
MWCNTs (multiwalled carbon nanotubes) 479, 479
b
Mycobacterium tuberculosis infection 460
myocardial infarction, coffee drinking and 168
mythos 102

Nagasaki atomic bombing (1945) 449, 451

narrative analysis
  data analysis 107
  systematic reviews 198–199
narrative data 104
national authorities, outbreak investigation 63–64
National Cancer Data Repository (UK) 5
National Cancer Institute (NCI), Surveillance, Epidemiology, and End Result (SEER) data 282
National Center for AIDS/HIV Control and Prevention (NCAIDS, China) 22–23
national governments 398, 403–404
National Health and Medical Research Council (NHMRC, Australia) 103
National Health and Nutrition Examination Survey (NHANES) 47
  clinical practice guidelines 151
  sampling methods 53
national health governance 399
National Health Interview Survey (USA) 266
National Institute for Health and Clinical Excellence (NICE)
  clinical practice guidelines 151
  cost-consequence analysis 370
  GRADE 154
  health resource optimization 155
National Institute for Health Research, INVOLVE 355
National Institute of Environmental Health Sciences (NIEHS) 440, 442
National Institute of Health (US), PROMIS initiative 489
national multipurpose surveys 47–49
National Science Foundation, Innovation Corps (I-Corps) initiative 416
natural background radiation 446–447
natural disasters 422
natural experimental studies 141–148
  design of 142
naturally-occurring sources, ionizing radiation 445–447, 447
  t
NCAIDS (National Center for AIDS/HIV Control and Prevention, China) 22–23
NCI (National Cancer Institute), Surveillance, Epidemiology, and End Result (SEER) data 282
negative-binomial regression 226
negative controls 282
negative correlation 221
  f
neo-functionalism 293
nested case–control studies 44
net equity impacts 368
Netherlands, PHARMO Record Linkage System 4
net reproductive rate (NRR) 323, 323
  b
network governance 393
network meta-analysis 201
Newcastle Ottawa Scale (NOS) 196
New Economic Foundation 103
newly-emerged diseases 59
A New Perspective on the Health of Canadians (Lalonde report) 343
newspapers 60

Neyman bias (incidence–prevalence bias) 54, 55
f
NGOs *see* non-governmental organizations (NGOs)
NHANES *see* National Health and Nutrition Examination Survey (NHANES)
NHMRC (National Health and Medical Research Council, Australia) 103
NICE *see* National Institute for Health and Clinical Excellence (NICE)
NIEHS (National Institute of Environmental Health Sciences) 440, 442
nominal levels, health sociology 297
non-compliance, drugs 7
non-concurrent cohort study 86
non-differential errors 166
non-exposed people, exposed people *vs.* 75–76, 75
t, 76
t
non-governmental organizations (NGOs)
Ebola epidemic 403
tax exemptions for 395
non-linear correlation 221
f
non-medical interventions 353–354
non-parametric correlation coefficient 221
non-parametric tests 218
non-persuasive communications 492
non-probability sampling 208
non-randomized controlled trials 153
t
non-randomized studies, bias risk evaluation 196–197
non-responders, cross-sectional studies 50–52, 52
f
non-threshold model, toxicology 474, 474
f
normal distribution 212–213, 212
f
normal theory regression 223
The North Coast Project (Australia) 136
North Karelia (Finland), cardiovascular disease rates 136
NOS (Newcastle Ottawa Scale) 196
notifiable diseases, surveillance systems 264–265
NRR (net reproductive rate) 323, 323
b
nuclear weapon testing 451
Nudge Unit Team 387
numeric probabilities 487
nutritional metrics 25
b

obesity
phthalates and 54
prevalence of 48, 48
f
observables, controlling for 143–144
observational cohort studies 85
observational epidemiology 129
observational studies 141–142
bias risk assessment 153
t, 196–197
exposure conceptualization 189
life course epidemiology 281

observations
  cluster randomized trials 114–115
  qualitative data 104
  qualitative research 104
observer bias 97
occupational diseases 422
  registries 461
occupational hazard analysis 295–296
occupational health 457–472
  child labour 462
  costs & economics 458
  definition 457
  disabled workers 462
  disease burden & epidemiology 458–463
  evolution of 457–458
  exposure assessment 464–466
  future work 469
  historical aspects 457–458
  legislation 458
  major types 458–460
  migrant workers 463
  risk assessment 464
  risk management 466–469
  shift workers 463
  special worker populations 461–463
  total workplace safety & health 463–464
  under recognition of illness 461
  women workers 462
Occupational Safety and Health Act (USA, 1970) 458
occupational skin diseases (OSD) 459

occupations
  studies in radiofrequency radiation 442
  ultraviolet radiation exposure limitation 445
  work-related asthma and 48–49
OECD (Organization for Economic Co-operation and Development) 321
official disease notification systems 60
omissions, rolling censuses 320
OMOP initiative 6
one-sample *t*-test 219
ongoing studies 195
on-site reporting 70–71
Open Working Group Proposal for Sustainable Development 395
Opportunidades programme (Mexico) 146
opportunity cost 365–366
ordinal measurement 297
Oregon Health Plan Lottery 143
Organization for Economic Co-operation and Development (OECD) 321
organ-specific factors, distribution 476
OSD (occupational skin diseases) 459
Ottawa Charter for Health Promotion 335, 338–339, 353
  community action 340
  health definition 338
  health promotion definition 336
  impact of 343
  public policy building 339
  supportive environments 339–340
Ottawa Statement on the Ethical Design and Conduct of Cluster Randomized Trials 126

outbreak investigation
active case finding 65–66
case definition 65
case information 66
cause establishment 70
criteria for 59–60
detection 60–61
diagnosis & aetiology verification 64–65
environmental studies 69–70
follow-up investigations 72
hypothesis testing 69, 70
t
implementation issues 63–64
improvement of 71
international cooperation 72–73
on-site reporting & recommendations 70–71
outbreak confirmation 64
outbreak description 66–69
purposes 61–62
risk communication 71, 72
f
steps of 64–72, 64
b
team components 62–63
outbreaks 59–74
confirmation 64
control 71
definition 59
future prevention 71
investigations *see* outbreak investigation *see also* epidemics
outcome-affected measurement, causal inference 176–177
outcome data, cohort studies 90–91, 90
t, 91
t
outcome removal, causal inference 188
outcome types, cohort studies 89–90
outlier cases 67
overuse 155
Oxford Centre for Evidence-based Medicine Level 199

paired *t*-test 219
pandemics, electronic health records 6
panel studies 50
PAR (participatory action research) 108–109
parasitic diseases 331
parenthood, sexuality and 311
Pareto efficiency 367
Pareto improvement 367

Pareto optimum 367
parity progression ratios 322
  definition 323
    b
Parsons, Talcott 293

participants
data analysis 201
risk magnitude assessment 486
participatory action research (PAR) 108–109
participatory processes, *Ottawa Charter for Health Promotion* 340

partnerships
cardiovascular disease community-based interventions 134–135
health governance 395, 397
t
passive smoking 40
passive surveillance 262–263
PATCH programme 133
pay-for-performance programmes 6–7
Pearson chi-squared test 215–217, 216
t
limitations of 216–217, 216
t
Pearson correlation coefficient *(r)* 219–220, 220
f
significance 220
Pearson, Karl 215
pedestrianism, motor vehicles and 500
People-Centred Health Promotion (Raeburn & Rootman) 340
People's Health Movement (PHM) 103–104
People's Participatory Planning (PPP) 108

percentages
categorical variables 209
comparisons 214–215, 215
t
perfect positive correlation 221
f
perinatal mortality rate 324
b
period analysis, cohort analysis *vs.* 322
permissible levels, occupational exposure limits 464
permission, outbreak investigation 63–64
personal protective equipment (PPE) 467
persons, surveillance systems 262
person-time analysis, cohort studies 91–92
persuasive communications 492
pertussis (whooping cough) 254, 254
f
PHARMO Record Linkage System (Netherlands) 4
phenomenology, health sociology 293–294
Philippines, CARES programme 384
philosophy, health promotion 344
PHLAME wellness programme 463–464
PHM (People's Health Movement) 103–104
photokeratitis, ultraviolet radiation 444
phthalates, obesity and 54
physical activity, cardiovascular disease and 282
physical components, environmental health 422
physical environment, built environment 500
physical measurements, case–control studies 77
PICO-ST structure 194–195
piloting studies, complex interventions 357–358, 357
t
place, surveillance systems 262
plausibility, causation 185
plausible range methods 375
point estimates, matched-pair cluster randomized trials 125
point-of-care tool 151
point prevalence, cross-sectional studies 56
Poisson regression 224–225, 226
t

policy
development in health governance 394
extremely low frequency fields 440
formulation 397
t
implementation 143
infrared radiation 443
intermediate frequency radiation 440
ionizing radiation 452
radiofrequency radiation 442
shaping in health governance 396
static radiation fields 438–439
ultraviolet radiation 444–445
policymakers 102
policy partners 355, 355
f
polio 254–255, 254
f

politics
economy in health promotion 345–347
health promotion 344
outbreak investigation 61–62
polychlorinated biphenyls (PCBs), banning of 481

population
ageing of 318
age structure 323
case–control studies 77–78
cohort studies 88–89
determinants of size 326
distribution, national multipurpose surveys 48–49
elderly proportion 328–329
growth 317, 317
f, 327
heterogeneity, causal inference 190
infection transmission models 244
projections 327
proportion of elderly 317
screening 157
statistical methods 207
urbanization and 497
young age structures 327
population censuses 3, 319–320
population change, decomposition of 326
population dynamics 326–327

population health
  complex interventions 353–354
  health education 343–344
  measurement of 229–240 *see also* Global Burden of Disease (GBD) study
  promotion 344
  qualitative research 102
  surveys 266
population-level infection transmission models 246
population momentum 327
population pyramids 327, 328
  f
population surveys 3
positive definition of health 338
positron emission tomography (PET) 447
possible case definition 262
potential outcomes 186
  causal inference 186–187
  models 187

poverty
  psychology and 386
  temporal discounting and 386
power considerations, cluster randomized trials 120
power dynamics, asymmetric 311
PPE (personal protective equipment) 467
PPP (People's Participatory Planning) 108
PPPHs (private–public partnerships for health) 26–27
PPPs (public–private partnerships) 404–405
practice, qualitative research 102
practice partners 355, 355
  f
practitioners, risk communication 493
precautionary principle 488
  case study 482, 482
    b
  toxicological risk assessment 481–482
PRECEDE-PROCEED 133
Precursors Cohort study 91
prediction problems, validity & bias 163–167
predictive value positive (PVP) 270
pre-employment examination 467
pre-exposure prophylaxis (PrEP), HIV infection/AIDS 308
preferences, health behaviour 382
present bias 383–384

prevalence
cross-sectional studies 56
morbidity metrics 26, 26
b
national multipurpose surveys 49
studies of 43–44
prevalence incidence studies 153
t
prevalent cohort 86

prevention
extremely low frequency fields 440
infrared radiation 443
intermediate frequency radiation 440
ionizing radiation 452
outbreaks 61
radiofrequency radiation 442
static radiation fields 438–439
ultraviolet radiation 444–445
Prevention Office on Smoking & Health (CDC) 132
The Prevention of Malaria (Ross) 244
prevention paradox 480
b
prevention programmes, outbreak investigation 61
prices, health behaviour 382
primary cases, outbreak investigation 67

primary data
community diagnosis 17–18
health sociology 296
primary healthcare, reorientation 341
primary prevention, occupational health risk management 466
principal agent theory 392
principle of mass action 252
Principles of Psychology (James) 298
PRISMA 200
privacy, electronic health records 9
private–public partnerships for health (PPPHs) 26–27
private sector, health governance 395
probability 207–208
probability distribution 211–212
probability judgements, quantitative assessment 487
probability models 226–227
probability of death 247, 487
probability per unit time of death 250
probability sample 208
probable case definition 262
problem identification, cardiovascular disease community-based interventions 134
procedural error, prediction problems 166
process evaluation methods, complex interventions 358–359
procreation, sexuality and 311
productivity, health psychology 300
PROGnosis RESearch Strategy (PROGRESS) partnerships 156
prognostic studies, bias risk assessment 153
t
Program for Monitoring Emerging Diseases (ProMed) 60, 267
PROGRESS acronym 354
progressive universalism 354
Project Northland (USA) 352, 360
Project Talent Aging Study 284
PROMIS initiative, National Institute of Health (US) 489
Promoting Planetary Health and Sustainable Development for All 339
propagated source outbreaks 69
proportional hazards assumption 226
proportional sampling 208

proportions
comparisons 214–215, 215
t
confidence intervals 209
t, 214
propriety scores 143–144
prospective meta-analysis, evidence synthesis 201
PROSPERO 194
Protecting Participants and Facilitating Social and Behavioural Sciences Research 301
PROTECT project 6
proved personal, social, and health education (PSHE) 357
proximate determinants of fertility 330
proxy-variable error
causal inference 172–173
prediction problems 166
PSHE (proved personal, social, and health education) 357
psychology in public health 298–302
history of 299–300
poverty and 386
psychological theories 298–299
research & services 299–301
training and 301–302
psychotherapy 299
puberty onset 309

publication bias
GRADE 154
meta-analysis of systematic reviews 198, 199
f

publications
health sociology 292
language and selection bias 193
public concerns, outbreak investigation 61–62

public health
  administrators 63
  economic appraisal 378–379, 378 *b*
  evaluation of programmes 42
  governance of organizations 399–400
  health promotion 343–344
  surveillance *see* surveillance systems
  surveillance systems 261
Public Health Act (1848) 421

public involvement
complex interventions 355
qualitative research 103
public–private partnerships (PPPs) 404–405
public subsidies 382
published reports 195
PulseNet 266
PVP (predictive value positive) 270

QALYs *see* quality-adjusted life years (QALYs)
QDA (qualitative data analysis) 107
QIF (Quality Implementation Framework) 410
QSurveillance 6
qualitative assessment 489–491
cumulative risk 489–490
detail supply 489
event definition 489
mental processes 490
risk beliefs to risk decisions 490–491 *see also* quantitative assessment
qualitative data 104
qualitative data analysis (QDA) 107
qualitative evidence 354–355
qualitative research 101–111, 101
b
case studies 105–106
complex interventions 358–359
definition 101
document-related research 106
further research preparation 102
future directions 109
health psychology 301
health sociology 297
interview-based studies 104–105
policy & practice 102
population health & health equity 102
quality assessment 107–108
sampling 106
social justice theory 102–103
study design/collection methods 104
textual data analysis 106–107
qualitative reviews 202
quality-adjusted life years (QALYs) 373
cost-utility analysis 371–372
health gain benefits 373
f
Quality Assessment of Diagnostic Accuracy Studies (QUADTAS) tool 201
Quality Implementation Framework (QIF) 410
Quality of Life index 297

quantile 210
quantitative assessment 486–489
  probability judgements 487
  risk definition 487–489
  risk magnitude assessment 486–487
  standards evaluation 487 *see also* qualitative assessment

quantitative research
  complex interventions 354–355
  health psychology 301
  health sociology 297
quantitative variables 209
  measurement 215
    t, 218–219, 219
    t
quasi-experiments 141–148

questionnaires
  case–control studies 76–77
  cohort studies 89
Quick, J. C. 300
Quick, J. D. 300
Quito (Ecuador), Climate Change Action Plan 504
QUOROM statement 200

R_0 (reproduction number) 245
radiation 437–456
  extremely low frequency fields 439–440
  infrared radiation 443
  intermediate frequency 440
  ionizing radiation *see* ionizing radiation
  radiofrequency radiation 441–442
  static fields *see* static radiation fields
  ultraviolet radiation 444–445
radiation therapy 448
radio, radiofrequency radiation 441
radiofrequency (RF) radiation 440, 441–442
radiographers 450
radiologists 450
radiotherapy, long-term health effects 450
radon 446, 450
Ramazzini, Bernardino 422, 457
random components of error, prediction problems 166–167
random digit dialling (RDD) 50–51
  case–control studies 78
random effect regression models 123–124, 227

randomization
confounding 170
information systems, HICs 7–8
policy implementation 143
randomized controlled trials (RCTs) 85, 113
bias risk assessment 153
t, 196
case–control studies *vs.* 82–83
cluster randomized trials *vs.* 114–115
economic appraisal 378
GRADE 152
origins of 130
population size 129
randomization in 141
random samples 52, 208
confounding 165
statistical methods 207–208
range of variables 211
rapid reviews, evidence synthesis 202
rates 249
rational reconstruction 299
Rawls, John 345
Rayyan QCI 203
RCTs *see* randomized controlled trials (RCTs)
RD (regression discontinuity) 145–146
RDD *see* random digit dialling (RDD)
reach, complex interventions 358–359
Reach, Effectiveness, Adoption, Implementation, and Maintenance (RE-AIM) framework 359–360, 411
REACH regulations 482
RE-AIM (Reach, Effectiveness, Adoption, Implementation, and Maintenance) framework 359–360, 411
realist perspective 353
real-world randomization 142–143, 243
b
recall bias 282
recommended exposure limits (RELs) 465
redistribution, Global Burden of Disease study 233–234
reflexivity, qualitative research 101
b
Regional Office for Europe's TAPIC framework (WHO) 402

registries
birth and death 320
disease registries 265–266
immunization registries 4
information systems, HICs 4
lack of 321
occupational disease registries 461
regression 219, 220–221, 220
f
regression adjustment 143–144

regression analysis
  assumptions 221
  cohort studies 91
regression coefficient 220
regression discontinuity (RD) 145–146

regulation
health governance 394, 397
t
toxicological risk assessment 479–480
rehabilitation, occupational health risk management 468
relative-risk estimation
case–control studies 175–176
risk factor–disease pairs 236
relative risk judgements 486

reliability
health sociology 298
qualitative research assessment 108
RELs (recommended exposure limits) 465
repeated measures, causal inference 174
reporting completeness 265
representativeness, surveillance systems 270
representative sampling 50
reproduction number *(R_0)* 245
reproduction rates 322–327
reproduction ratios 323
b

reproductive health
occupational health 459–460
static radiation fields 438
reproductive rights 313–314

research
  communication technologies 33
  implementation science 412–413
  qualitative research 102
  translation into patient care 155
research capacity, LMICS 158
Research Committee on Medical Sociology, International Sociological Association 292
research evidence location, clinical epidemiology 149–150, 150
  t
resource dependency theory 400
resource stewardship, health governance 394–395
respondent-driven sampling 52–53
response rates, cross-sectional studies 50
restorative practice (RP) 357
restricted randomization 116
result inconsistency, GRADE 152
retrospective ascertainment, causal inference 176
retrospective cohort studies 179
return on investment (ROI) 379
return to work (RTW) 468
REVEAL-HBV cohort study 429, 433
reverse causation, life course epidemiology 278, 278
  f
REVMAN 203
RHIS (routine health information systems) 22
right-censored subjects, cohort studies 90
Rio Declaration on Environment and Development (1992) 481

risk
absolute risks 486
assessment *see* risk assessment
communication *see* risk communication
comparisons 489
definition in quantitative assessment 487–489
health behaviour 382
magnitude assessment 486–487
perception *see* risk perception *see also* risk factors

risk assessment
environmental health 433
occupational health risk assessment 464
toxicological risk assessment 477
b, 478–479
risk beliefs, qualitative assessment 490–491
risk communication 485–495
evaluation 492
expertise in 485
information formatting 491–492
information selection 491
outbreak investigation 71, 72
f
process management 493
toxicological risk assessment 480–481
risk decisions, qualitative assessment 490–491
risk factor–disease pairs 236

risk factors
cohort studies 85
environmental health 421–422
epidemiology 42
exposure estimation 236
national multipurpose surveys 48
outbreak investigation 66
removal, causal inference 188–189
risk perception 485–495
comparison of risks 489
definition of risk 487–489
dimensions of risk 488–489, 488 *f*
inaccurate judgements 486
qualitative assessment *see* qualitative assessment
quantitative assessment *see* quantitative assessment
risk per unit time 250
ROBINS-1 tool 196
ROBIS tool 200
Rochester Epidemiology Project 4
ROI (return on investment) 379
rolling censuses 320
Rose, Geoffrey 354
Ross, Ronald 244
Rousseau, Jean Jacque 292
routine health information systems (RHIS) 22
RP (restorative practice) 357
RTW (return to work) 468
The Rules of Sociological Method (Durkheim) 293
rule specification, infection transmission 244

de Saint Simon, Henri 292
salience bias, behavioural economics 385

sampling
  cluster randomized trials 114, 120
    b
  cross-sectional studies 50, 52–53
  health sociology 298
  proportional to size 208
  qualitative research 106
  statistical methods 207
sampling distribution 213
sampling frame 208
sampling variability 213
Sanders, M. R. 300
SARI (severe acute respiratory infection) surveillance 266
SARS *see* severe acute respiratory syndrome (SARS)
SARs (structure–activity relationships) 474
scale-up frameworks, implementation science 411–412
Scaling Up Management Framework for Practitioners (Management Systems International) 413
SCAPIS (Swedish CArdioPulmonary BioImage Study) 459
scarcity theory, decision-making psychology 386
scatter plots 219, 220
  f
science perspective interventions, behavioural economics 387
SCIH (Swiss Centre for International Health) 404
scoping reviews 201
SD (standard deviation) 211, 211
  f
SDC (Swiss Agency for Development and Cooperation) 404
SDGs *see* Sustainable Development Goals (SDGs)
SDH *see* social determinants of health (SDH)
seasonal migrants 320

secondary data
community diagnosis 18–19, 18
b
health sociology 296
secondary prevention, occupational health risk management 466, 467
second demographic transition, stable population theory 330
second generation surveillance 267
security, water & sanitation 270
SEER (Surveillance, Epidemiology, and End Result) data 282
segregation, engineering controls 466

selection bias
  case–control studies 44, 78–79
  cohort studies 96
  systematic reviews 193
selection validity, causal inference 177
self-administered questionnaires 56
self-identification, sexuality 308
self-reporting, health psychology 301
sensitive period, life course epidemiology 275–276
sensitive time period model 275–276

sensitivity analysis
economic appraisal models 375
meta-analysis of systematic reviews 198
surveillance system evaluation 270
sentinel surveillance systems 263, 266
information systems, LMIC 24
separation of effects, measurement error 174
sequence analysis, life course epidemiology 283–284
sequential multiple assignment randomized implementation trial (SMART) 413
service access, communication technologies 32
service location identification, communication technologies 30–31, 31
f
SES (socioeconomic status) 295
severe acute respiratory infection (SARI) surveillance 266
severe acute respiratory syndrome (SARS) 460
multi-country outbreaks 72
outbreak investigation 67, 67
f
propagated source outbreaks 69, 69
f
severity distributions, Global Burden of Disease study 235
sexual diversity 311–312
sexual health 313–314
sexuality 307–316
categories of 311–312
gendered experience 309–310
heterosexual relations 310–311
HIV infection/AIDS 307–309
sexual diversity & alternative sexualities 311–312
sexual health & sexual rights 313–314
social construction of 309
social regulation of 312–313
youth and 309–310
sexually-transmitted infections (STIs)
cumulative risk 489–490
HIV infection incidence 384
surveillance 266
sexual rights 313–314
sex workers 312
HIV infection/AIDS 308
SG (standard gamble) 374
sharps injuries 460
Shaw Billings, John 292
shift workers, occupational health 463
Shigella sonnei infection 68–69, 68
f
significance tests, matched-pair cluster randomized trials 125
silent mutations 476
silicosis 459
Simmel, Georg 293

Simons, John 344
simple interventions 351
simple random sampling 208
simplicity, surveillance system evaluation 270
single infectious persons 253
skewed data 210, 211
  f
skin cancer 444
skin diseases 459
slums 499–500
SMART (sequential multiple assignment randomized implementation trial) 413
Smith, Southwood 291
SMRs (standardized mortality ratios) 91, 169, 323, 324
  b
snowball sampling 52–53
  qualitative research 106
Snow, John 141, 344, 421
social benefits, economic appraisal 366
social class, health sociology 295
social cognitive theory 134

social construction
qualitative research 101
b
sexuality of 309
social costs, economic appraisal 366
social desirability bias 56
social determinants of health (SDH) 480
urbanization 498
social efficiency 368
social environment, urbanization 500–501
social factors, sexuality 308–309
social influence, behavioural economics 385–386
socializing norms, sexuality 309

social justice
health promotion 345
qualitative research 102–103

social media
communication technologies and 30, 34
future trends in 33–34
information systems, HICs 7
outbreak detection 60
theories of 33
social mental health metrics 25
b
social movement, health promotion 344–345
social networks 30, 295
sexuality 313

social norms
behavioural economics 385–386
tobacco smoking prevention 131–132
social regulation of sexuality 312–313
social reproductive span 330
Social Science & Medicine (SSM) 292
social stratification, health sociology 295
social structures, qualitative research 101
b
social theory, health promotion and 346

socioecological approaches
complex interventions 353
t
Ottawa Charter for Health Promotion 339
socioeconomic metrics 25
b
socioeconomic status (SES) 295

sociology in public health
history 291–294
research and 294–301
sociological theories 292–294
training and 301–302
Sociology of Health and Illness 292

sources
extremely low frequency fields 439
infrared radiation 443
intermediate frequency radiation 440
ionizing radiation 445–448
radiofrequency radiation 441
static radiation fields 437–438
ultraviolet radiation 444
South Africa, cardiovascular disease community health promotion 136
sparse-data bias 163
spatial segregation, urbanization 501
special control groups, causal inference 178

specification validity
causal inference 168, 174–175
prediction problems 164, 167

specificity
  causation 184
  toxicology 474–475
specimen collection, outbreak investigation 63, 65
SSPE (subacute sclerosing panencephalitis) 41
stability, surveillance system evaluation 270
stable population theory 327–331
  age structure 327–329
  definition 327
  demographic transition 329–330
  epidemiological transition 330–331
  proximate determinants of fertility 330
  second demographic transition 330
staggered entries, cohort studies 89

stakeholders
health governance 397
t
implementation science 413, 415
public health organization governance 400
standard deviation (SD) 211, 211
f
standard gamble (SG) 374
standardization 188
standardized mortality ratios (SMRs) 91, 169, 323, 324
b

standards evaluation
quantitative assessment 487
surveillance systems 270
Stanford Five-City Project (5CP) 136–137
Stanford Three Community Study (3CS) 135–136
Starfield, Tony 243, 243
b
static models, infection transmission models 245
static radiation fields 437–439
health effects 438
policy & prevention 438–439
sources & environmental levels 437–438
statins (HMG-CoA reductase inhibitor drugs) 8
stationary population 327
statistical inference 213
health psychology 300
statistically significant difference 216
statistical methods 207–227
analysis of variance 218
association & correlation types 220, 221
f
censoring 222–223
coefficients of association 219, 220
f
comparative analysis *see* comparative analysis
complex sampling 208–209
confidence intervals 213–214
correlation & regression 219, 221
Cox proportional hazards regression 225, 226
t
location measures 209–210
logistic regression 223–224, 224
t
mean comparison 217, 217
t
misinterpretations 162–163
modelling 242
t
multivariable analysis 223
normal distribution 212–213, 212
f
Pearson correlation coefficient 219–220, 220
f
Poisson regression 224–225, 226
t
population 207
probability 207–208
probability distribution 211–212
probability models 226–227
proportional hazards assumption 226

quantitative variable measurement 215
t, 218–219, 219
t
random sampling 207–208
regression 220–221, 220
f
sample size of cluster randomized trials 118–119
sampling 207
sampling distribution 213
scatter plots 219, 220
f
sensitivity analysis 375
survival data 222, 222
f
time-dependent variables 226
two-sample *t*-test 217–218
variability measures 211
variables 209

statistical significance
causation 186
misinterpretations of 162–163
status quo, behavioural economics 384–385
statutory periodic medical examinations 467–468
stepped wedge design 157
Stetler Model, implementation science 410
stewardship theory 400
STIs *see* sexually-transmitted infections (STIs)

stochastic models
infection transmission models 245–246
ionizing radiation 448
stratification, cluster randomized trials 116
stratified cluster randomized trials 126
stratified sampling 208
strength, causation 184
strength of recommendations, GRADE 154
stress management, health psychology 300
stress-related ill-health 461
striped wedge trials 117
structural functional theory, health sociology 293
structure–activity relationships (SARs) 474
study population, cohort studies 86–87, 87
t
study searches, systematic reviews 195–196
subacute sclerosing panencephalitis (SSPE) 41
subgroup analysis 197–198
subject matter experts, risk communication 493
sub-Saharan Africa, HIV infection 319
Suicide (Durkheim) 293
sunburn 444
SUPPORT 157–158
support work, public health programme management 400
supranational health governance 397–399
Surgeon General's Report on Tobacco and Health (1965) 132
surveillance, active surveillance 262–263
surveillance bias 96–97
Surveillance, Epidemiology, and End Result (SEER) data 282
surveillance imperative, tobacco smoking prevention 132
surveillance systems 259–274
data analysis 268
data collection methods 263–268
data collection type 262–263
data content type 263
data management 269
data reporting 269
definition 259–260
disease registries 265–266
effectiveness 61
establishment of 260
ethical issues 270–271
evaluation 270
healthcare information systems 267–268
health problems 261
history of 259–260
information dissemination 269
information sources 268–269
laboratory surveillance 266
low- and middle-income countries 271

objectives of 260–262
organizational structure 270–271
syndromic surveillance 264–265
target populations 261–262
surveys, migration measurement 326
survival analysis, cohort studies 92–93

survival data
cohort studies analysis 96
statistical methods 222, 222
f
survival function estimation 92–93
susceptible infectious compartment models 252–254, 253
f
sustainability, complex interventions 359–360
Sustainable Development Goals (SDGs) 19–20, 20
b
data revolution 19–20
development of 19
health governance 395
implementation science 410
urbanization 497 *see also* Millennium Development Goals (MDGs)
Swedish CArdioPulmonary BioImage Study (SCAPIS) 459
Swiss Agency for Development and Cooperation (SDC) 404
Swiss Centre for International Health (SCIH) 404
Switzerland, cardiovascular disease community health promotion 136
symbolic interactions, health sociology 293
syndromic surveillance 264
definition 6
synthesized summaries 151
synthetic controls 144–145
systematic components of error 166–167
systematic reviews 151
bias 193
bias risk evaluation 196–197
data extraction 196
data synthesis 197–199
definition 193–194
evidence quality 199–200
future work 202
literature reviews *vs.* 194
t
meta-analysis 197–198
narrative analysis 198–199
practice guidelines 202
protocol development 194
quality appraisal 200–201
question formulation 194–195
reporting of 200
software & resources 202–203
study searches 195–196
systems perspective 353
systems science, complex interventions 352

tailoring imperative, tobacco smoking prevention 132

Taiwan
endemic goitre 430–431
Environmental Impact Assessment Act 480
hepatitis B vaccination programme 430
Tanzania, MEMA kwa Vijana (MkV) trial 121–122, 122
t
target populations 208
life course epidemiology 279
surveillance systems 261–262

tax initiatives
  non-governmental organizations 395
  tobacco smoking prevention 131, 382
TDF (Theoretical Domain Framework) 411
technical efficiency, economic appraisal 367
technology-based risk communications 491
television, radiofrequency radiation 441
temporal discounting, poverty and 386
temporal growth/decline, infection transmission models 246–250
temporality, causation 184–185

teratogenicity
ionizing radiation 452
radiofrequency radiation 442
tertiary prevention, occupational health risk management 466, 468–469
test-negative controls, case–control studies 79
textual data, qualitative research 106–107
TFR *see* Total Fertility Rate (TFR)
thematic framework identification, qualitative data analysis 107
Theoretical Domain Framework (TDF) 411
theoretical minimum risk factor level 236
theory importance, complex interventions 355–356
A Theory of Justice (Rawls) 345
theory of network governance 392–393
theory of preventative stress management (TPSM) 300
thorium-232 446
3CS (Stanford Three Community Study) 135–136
3-D (dirty, dangerous and demanding) jobs 463
threshold limit values (TLVs) 464–465
threshold model, toxicology 473–474, 474
f
time-dependent variables 226
time–place–person interactions, epidemiology 40
timescales, cohort studies 89–90

time to event
cohort studies 90
survival analysis of cohort studies 92
time trade-off (TTO) 374

timing
health behaviour 382
national multipurpose surveys 48
outbreak investigation 68–69
surveillance system evaluation 270
surveillance systems 262
tissue reactions, ionizing radiation 448
TLVs (threshold limit values) 464–465
tobacco smoking, Alzheimer's disease and 54
tobacco smoking prevention 131
f
best processes & best practices 133
comprehensiveness imperative 132
denormalization imperative 131–132
ecological imperative 132–133
economic imperative 131
education 132
formative evaluation & tailoring imperative 132
history of 130, 131–137, 131
f
informed electorate imperative 132
message characteristics 133–134, 134
t
social normative imperative 131–132
surveillance imperative 132
synthetic control methods 145
taxes 131, 382
top down costing, cost measurement 372
Total Fertility Rate (TFR) 318, 323
b
definition 323
b
limitation of 322
Total Worker Health (TWH) 463–464
total workplace safety 463–464
toxicological risk assessment 476–482
communication 480–481
components 477
b
dose–response evaluation 477, 477
b, 478
b
exposure evaluation 477–478, 477
b, 478
b
future work 479–480
hazard identification 477, 477
b
precautionary principle 481–482

risk characterization 477
b, 478–479
toxicology 473–484
chemical interactions 476
definition 473
exogenous agent pathways 475–476
laws of 473–475
risk assessment *see* toxicological risk assessment
toxic stress/stressors, urbanization 501, 502
f, 502
t, 503, 503
f
TPSM (theory of preventative stress management) 300
trajectory model (chains of risk model) 275
transferability, qualitative research assessment 108
transformations 221
analysis of variance 218
qualitative research 103
TRANSFoRm project 7–8
transgender individuals 312
transparency, qualitative research 101
b
transportation, static radiation fields 437
trauma, occupational health 460
Die Traumdeutung (Freud) 298
trend analysis, surveillance systems 268
trials 141–142
behavioural economics 387
triangulation, qualitative research assessment 107–108
trusted third parties (TTPs) 4
t-test 217–218
TTO (time trade-off) 374
TTPs (trusted third parties) 4
tuberculosis 460
TWH (Total Worker Health) 463–464
two-group, two-period differences in differences 144
two-sample *t*-test 217–218
two-stage sampling 208
type 1 error, Pearson chi-squared test 216, 216
t
type 2 error, Pearson chi-squared test 216, 216
t

UK *see* United Kingdom (UK)
Ukrainian Swiss Perinatal Health Programme 404
ultraviolet (UV) radiation 444–445
UN *see* United Nations (UN)
uncontrolled trials, environmental health 430

underenumeration, census weaknesses 320
UNDP *see* United Nations Development Programme (UNDP)
UNESCO, precautionary principle 481–482, 481
b

UN Habitat
Global Report on Urban Health 497
Hidden Cities: Unmasking and Overcoming Health Inequities in Urban Setting 497
unitary governments 394
United Kingdom (UK)
Clinical Practice Research Datalink 4, 5
General Household Survey 266
Health and Safety at Work Act 458
health sociology/psychology 301
National Cancer Data Repository 5
Workplace Exposure Limits 465
United Nations (UN)
Agenda for Sustainable Development 354
Conference on Housing and Sustainability (Habitat III) 497
demography data sources 321
supranational governance systems 402–403
United Nations Development Programme (UNDP)
five governance principles 401
governance/management of public health programmes 392
United States of America (USA)
communication technologies 29
Deferred Action on Childhood Arrivals 146
health sociology/psychology 301
Massachusetts Department of Public Health 5
Medical Expenditure Panel Surveys 235–236
National Health Interview Survey 266
National Institute of Environmental Health Sciences 440
Occupational Safety and Health Act (1970) 458
Project Northland 352, 360
United States Preventative Services Task Force 199
univariate analysis, cohort studies 91–92, 93, 93
f

unmeasured confounders
causal inference 170
prediction problems 165
UN Mission for Ebola Emergency Response (UNMEER) 402–403
unpublished reports, systematic reviews 195
uranium-238 446
urbanization 497–505
biomedical model (1950s-1990s) 499–500
climate change and 503–504
colonialism 499
germ theory 498–499
governance and health 501
health equity (21st century) 500
health influences 500–501
historical aspects 498
modern health governance 503
population levels 497
social environment 500–501
toxic stress 501, 502
f, 502
t, 503
toxic stressors 503
f
USA *see* United States of America (USA)
US Environmental Protection Agency (USEPA) 489
User-Experience (UX) design 492
User-Interface (UI) design 492

vaccines, efficacy measures 42

validity
  case–control studies 176
  causal inference 167–179
  cohort studies 96
  health sociology 298
  internal *see* internal validity
  prediction problems 163–167
vapours, gases, dusts, and fumes (VGDF) 459
variability measures 211
variables 209
  range of 211
VAS (visual analogue scales) 374
vascular risk assessment programmes 8
verbal autopsy 21–22
verbal qualifiers 487
Veterans' Administration (VA) studies, Global Burden of Disease study 233
VGDF (vapours, gases, dusts, and fumes) 459
village leaders, outbreak detection 60–61
Villermé, Louis-René 498
Virchow, Rudolf 292
visual analogue scales (VAS) 374
vital registration 320
  information systems, LMIC 21

vital statistics
information systems, HICs 3
surveillance systems 264, 264
f
Voice of the Patient initiative (FDA) 489
voluntary surveillance systems 263

wearable communication technologies 33–34
web-based research, cross-sectional studies 53–54
Weber, Max 291, 293
Weekly Epidemiology Record (WHO) 269
Weibull models 94
weighted means 209
weight-of-evidence approaches 477
well-being, occupational health 463
WELs (Workplace Exposure Limits, UK) 465
Western individualism, sexuality 313
WFS (World Fertility Survey) 321
WHAS (Worcester Heart Attack Study) 222
WHO *see* World Health Organization (WHO)
wireless technologies, cross-sectional studies 53–54
within-cluster correlation 113
within-site comparison design 413, 414
t
within site design 412
women-who-have-sex-with-women (WSW) 311–312
women-who-have-sex-with-women-and-men (WSWM) 312
women workers, occupational health 462
Worcester Heart Attack Study (WHAS) 222
workers' compensation 468–469
Workplace Exposure Limits (WELs, UK) 465
work-related asthma (WRA) 48–49, sss459
work-related diseases 460–461
work-related musculoskeletal disorders (WRMSD) 460–461
work-related skin diseases (WRSD) 459
workstation redesign 466–467

World Bank
fundamental aspects of governance 401
governance/management of public health programmes 391
World Fertility Survey (WFS) 321
World Health Assembly 259
World Health Organization (WHO)
cardiovascular disease sponsored projects 137
clinical practice guidelines 151
Commission on the Social Determinants of Health 335
community definition 15
community participation 129
constitution 13
b
coronavirus case definitions 65
b
demography data sources 321
Disease Control Priorities Project 296
domains of stewardship 401
Eastern Mediterranean Regional Office's ten governmental principles 401–402
eHealth 29
extremely low frequency fields 440
formation 13
Global Observatory for eHealth (GOe) 29
Global Outbreak Alert and Response Network 73
Global Report on Urban Health 497
governance/management of public health programmes 391, 392
GRADE 154
Guidelines for Drinking-Water Quality (1993) 433
Handbook for Guidance Development 202
health governance 394, 398, 401
Hidden Cities: Unmasking and Overcoming Health Inequities in Urban Setting 497
HIV surveillance 266
host–agent interactions 422
International Agency for Research on Cancer 477
International Health Regulations 260
International Standards for Drinking Water (1958) 433
mortality estimates 261, 261
t
occupational health definition 457
online interactive maps 71
Regional Office for Europe's TAPIC framework 402
sexual *vs.* reproductive rights 313–314
STI surveillance 266
Weekly Epidemiology Record 269
World Health Organization Quality of Life (WHOQOL) index 297
WRA (work-related asthma) 48–49, 459
written data 104
WRMSD (work-related musculoskeletal disorders) 460–461
WRSD (work-related skin diseases) 459
WSW (women-who-have-sex-with-women) 311–312

WSWM (women-who-have-sex-with-women-and-men) 312

X-rays *see* ionizing radiation

years lived with disability (YLDs) 236
years of life lost (YLL) 232
  calculation 236
  risk factors and 236

youth
  communication technologies 30
  sexuality and 309–310

ZAMSTAR trial 116, 117
zero-inflated regression models 226
zero-reporting methods 269

Index Vol 3

Notes: Tables, figures and boxes are indicated by *t*, *f* and *b* following the page number

AA *see* Alcoholics Anonymous (AA)
AAAH (Asia-Pacific Action Alliance on Human Resources for Health) 716
AARP (American Association of *Retired* Persons) 513
AATD (alpha-1 antitrypsin deficiency) 47
abdominal aortic aneurysms 635
ABI (ankle-brachial index) 12

abortions
maternal health 423
mortality 425
absolute quarantine 610
absolute risk, tobacco smoking 320
Abstinence, Be faithful, and Condom use (ABC) campaigns 205
ACASI (audio-computer assisted self-interviewing) 342
ACC *see* American College of Cardiology/American Heart Association (ACC/AHA)
accelerometers, physical activity 77
ACCORD (Action to Control Cardiovascular Risk in Diabetes) 7
accountability, adolescent health 463–464
accuracy, medical screening 626, 627
ACE (Acute Care for Elders) model 518
achieved status 470
ACIP (Advisory Committee on Immunological Practices) 603
ACLED (Armed Conflict Location and Event Data Project) 723
ACO (asthma-COPD overlap) 45
ACTG (AIDS Clinical Trials Group) Protocols 203
Actinobacter infection 168
t
Action on Smoking and Health (ASH) 668
Action Plan for the Global Strategy for the Prevention and Control of Non-Communicable Diseases (WHO) 664
Action Plan on Antimicrobial Resistance (WHO, 2015) 281
Action Plan on mental health (WHO) 97
Action to Control Cardiovascular Risk in Diabetes (ACCORD) 7
active case finding (ACF), tuberculosis 219
activism, private support of public health 753
activities of daily living (ADLs) 514
activity restriction, infection control measures 611

acute care
injury prevention 370–375
older people healthcare 518
Acute Care for Elders (ACE) model 518
acute infectious diseases, humanitarian emergencies 728
acute lower respiratory infections 598
acute necrotizing ulcerative gingivitis (ANUG) 115
acute respiratory infections (ARIs) 174–177
aetiology 175, 175
t
antibiotic treatment 176–177
child health 440–441
clinical guidelines 441
diagnosis 175–176, 441
epidemiology 174–175
evidence-based interventions 176
forced migrant mortality 532
integrated management of childhood infections 177
presentation 175–176
prevention 177
public health implications 177
second-hand smoking 322
acute watery diarrhoea 172
AD *see* Alzheimer's disease (AD)
ADA *see* American Diabetes Association (ADA)

adenocarcinoma
colorectal cancer 24
oesophageal cancer 27
adenomatous polyposis coli *(APC)* gene 24
adenosine 297
ADHD (attention-deficit hyperactivity disorder) 546
ADLs (activities of daily living) 514
administrative data, substance use disorders 342
adolescent health 453–467
community interventions 461–486
conceptual framework 456
f
demography 455–456, 455
f
disease burdens 457, 459
education 460–475
epidemiological transitions 455–456
growth & development 453–455
healthcare barriers 459–460
health services 459
HIV infection 456
investment case & accountability 463–464
investments 452–453
mental health 456, 459
mortality 457
non-communicable diseases 456, 457
peers, media, & marketing 461
physical health 456
pregnancy 434, 459
psychosocial development 454–455
region by 456–457
risk factors 457–458
structural interventions 452–453
youth engagement & empowerment 463

adolescents
concepts of 453, 454
f
demography 455
f
developmental context 460
suicide epidemiology 96–97
adoption studies, complex multifactorial diseases 299–300

adults
  life expectancy 418
  obesity definition 57–58, 58
    t
  transition delay in families 406–407
ADVANCE 7
advertising restrictions, alcohol consumption prevention/control 355
Advisory Committee on Immunological Practices (ACIP) 603

advocacy
forced migrants in developed countries 537
gender healthcare 426
Aedes aegypti
Dengue fever transmission 593
urbanization 273
yellow fever transmission 592
Aetiology of Neonatal Infections in South Asia (ANISA) 167
Affordable Care Act (USA) 561
aflatoxin contamination 23

Africa
  delayed marriage 407
  public health strategies 666
  refugees 528
  South *see* South Africa
  sub-Saharan *see* sub-Saharan Africa
Africa Centres for Disease Control 666
African Genome Variation Project (AGVP) 311
African Neonatal Sepsis Trial (AFRINES) 167
After Action Review process (WHO) 748
age, obesity trends 60–66, 60
  f
age-friendly cities 519
ageing *see* older people
ageing support ratios 409, 409
  f
ageism 513

agency
  definition 470
  forced migrants 539
  gendered healthcare 424
  ideas of 567–568
Agenda for Sustainable Development (UN, 2030) 464, 759
agent–host relationship, injuries 362
Age of Degenerative and Man-Made Diseases 687
Age of Pestilence and Famine 687
Age of Receding Pandemics 687

aggression
genotypes 381
prevention 103
agriculture, anti-infective drugs 278–282
AGVP (African Genome Variation Project) 311
AHA *see* American College of Cardiology/American Heart Association (ACC/AHA)
AIDS 197–211
child health 443–444
forced migrant mortality 532
gender 419–420, 420
b
global distribution 202–203
historical aspects 197–198
substance use disorders 340–341
treatment 201–203
AIDS Clinical Trials Group (ACTG) Protocols 203
AIM-High Trial 9
airborne transmission 600
isolation recommendations 610

air pollution
asthma risk factor 50
health impacts 764
monitoring in environmental health 646
war 394
airport malaria 279
Airs, Waters and Places (Hippocrates) 639
alanine aminotransferase (ALT) 251
albendazole 182
alcohol companies 576
sub-Saharan Africa 577
alcohol consumption 349–360
adolescent health 457
adverse effects 349–350
alcoholic liver disease 259–260
breast cancer 25
cancer aetiology 19
chronic alcohol consumption 349–350
control in Hong Kong 668
cumulative effects 350–351
disabled people 493
drink driving legislation 353, 369
historical aspects 349
HIV infection & substance use disorders 345
increase in 574
integrated policies 356
interpersonal violence 382–383
laryngeal cancer 29
lower lip cancer 116
mental health 350
oesophageal cancer 26
oral cavity cancer 28
pancreatic cancer 28
partner violence 387
positive effects 350
prevention/control effectiveness 352–354, 765
prisoners 547
public health and 351
risk factors 474 *see also* drinking behaviour
alcoholic cirrhosis 22, 258
alcoholic liver disease 257–259
comorbidities 259
epidemiology 258
f
morphology 258
natural history 258
prevention 259
risk factors 258–259
societal costs 257–258
Alcoholics Anonymous (AA) 355–356

12 Steps of Alcoholics Anonymous 345
alcoholic steatohepatitis 258
alcohol use disorders 96
alcohol use-values 349
aldolase 233
aldosterone system blockers 13–14
alleles 298

allergies
  asthma and 48
  asthma risk factor 49
Alma-Ata Conference (1978) 641, 663
alpha-1 antitrypsin deficiency (AATD) 47
ALT (alanine aminotransferase) 251
Alum Shinrikyo 292
Alzheimer's disease (AD) 149–151
  dementia 148
  genetics 150
  incidence & prevalence 149
  risk factors 149
American Association of *Retired* Persons (AARP) 513
American College of Cardiology/American Heart Association (ACC/AHA)
  diabetes mellitus definition 7
  Guideline for the Management of Blood Cholesterol 13
  Guideline on Cardiovascular Risk Assessment (2014) 11
  Guideline on Lifestyle Management 13
American Diabetes Association (ADA)
  cardiovascular disease prevention 13
  diagnosis 86
    t
  disease definition 85
American Medical Association (AMA) 59
American Public Health Association (APHA)
  Control of Communicable Diseases Manual 597
  injuries 362
American Refugee Committee International 536
American Relief Administration (ARA) 729
American Thoracic Society (ATS)
  asthma guidelines 44
  COPD definition 43
aminoindanols 339
amodiaquine 235
amphetamine 338
amphetamine-type stimulants (ATS) 338
AMR *see* antimicrobial drug resistance (AMR)
Ancylostoma ceylanicum infection 605
Ancylostoma duodenale infection
  chemoprophylaxis 604
  diagnosis & treatment 182
    t
androgens, sexual violence 381
angina pectoris 3–4
Angola, yellow fever 2015 outbreak 273
ANHER (Asian Network on Health Professional Education Reform) 717
animal husbandry, anti-infective drugs 278–282
animals, infectious diseases control 601, 611–612
ANISA (Aetiology of Neonatal Infections in South Asia) 167
ankle-brachial index (ABI) 12

Anopheles mosquito life cycle 230–231
ANRS 075 trial 204
antenatal care 438

anthrax
bioterrorism 277
emerging/re-emerging infections 276
Sverdlovsk (1979) 277
anthropogenic activities, emerging/re-emerging infections 275
anthropometric markers, obesity 57

antibiotic resistance
emerging infectious diseases 617 *see also* antimicrobial drug resistance (AMR)
antibiotic-resistant *Neisseria gonorrhoeae* 189, 194

antibiotics
duration of 173
meningitis 178
neonatal meningitis 170
t
neonatal sepsis 170
t
over- an under-prescribing of 278
typhoid fever therapy 180
antibody responses, vaccination 602
antidotes, emergencies 746
anti-immunoglobulin E treatment 52
anti-infective drugs, animal husbandry/agriculture 278–282
anti-inflammatory medications, Alzheimer's disease 150
anti-interleukin 5 treatment 52
antimalarial drugs 234–235
travellers 239
antimicrobial drug resistance (AMR) 618, 762–763
emerging/re-emerging infections 277–278
infectious diseases 182–183 *see also* antibiotic resistance
antimicrobial-resistant *Neisseria gonorrhoeae* 192–193
antioxidants 262
antipersonnel landmines 397
antiplatelet therapy 12
anti-retroviral therapy (ART) 165
childhood HIV/AIDS 444
HIV infection 201–202
t
humanitarian emergencies 732
long-term use problems 203
neonates 444
non-AIDS defining illnesses 203
prevention of mother-to-child transmission of HIV 204
ANUG (acute necrotizing ulcerative gingivitis) 115

anxiety disorders
older people healthcare 517
prevention 103
APC (adenomatous polyposis coli*)* gene 24
APHA *see* American Public Health Association (APHA)
appendectomy 605
ARA (American Relief Administration) 729
areca nut chewing 26
Argentina, Less Salt, More Life campaign 672
ARIC (Atherosclerosis Risk in Communities) 12
ARIs *see* acute respiratory infections (ARIs)
Armed Conflict Location and Event Data Project (ACLED) 723
Arms Trade Treaty (UN) 395
ART *see* anti-retroviral therapy (ART)
ART (atraumatic restorative treatment) 114
artemisinins 234
resistance to 278
arthropathies 129
artificially acquired passive immunization 604
artificial sweeteners 121
Ascaris lumbricoides infection (ascariasis)
child health 443
diagnosis & treatment 182
t
mortality 165
t
ASD (autism spectrum disorder) 492
ASH (Action on Smoking and Health) 668

Asian countries
delayed marriage & cohabitation 406
East Asia *see* East Asia
family changes 403–404
hepatitis B 250–251
late marriage 406
LGBT+ legislation 463
older people support ratios 409
South Asia *see* South Asia
Asian Network on Health Professional Education Reform (ANHER) 717
Asia-Pacific Action Alliance on Human Resources for Health (AAAH) 716
aspartate transaminase (AST) 251
Aspergillus flavus infection 22
Aspergillus parasiticus infection 22
asphyxia, perinatal 438–439
aspirin 12
ASPR (Special Assistant to the President for Biodefense) 288
asset building, Health Needs Assessment 569
association, causation *vs.* in tobacco smoking 321
association studies, gene discovery 301–302
AST (aspartate transaminase) 251
asthma 44–45
burden of 48
categorization 45
chronic obstructive pulmonary disease overlap 45
clinical assessment 44–45
epidemiological surveys 45
epidemiology 48
future prevalence 48
prevalence 49
f
prevention 51–52
risk factors 48–49
treatment 52
asylum seekers 525–526, 526
t
mental health in developed countries 534–535
numbers of 528
vaccination programmes 595
atherosclerosis, second-hand smoking 322
Atherosclerosis Risk in Communities (ARIC) 12
atopy, asthma risk factor 48–49
atraumatic restorative treatment (ART) 114
atrial fibrillation 4
stroke and 10
at-risk populations, malaria 240–241
ATS *see* American Thoracic Society (ATS)
ATS (amphetamine-type stimulants) 338
attachment, interpersonal violence 381
attention-deficit hyperactivity disorder (ADHD) 546

audio-computer assisted self-interviewing (ACASI) 342

Australia
disability prevalence 491
Health Care of Asylum Seekers and Refugees 538
indigenous people as prisoners 545
Institute of Health and Welfare 499
Law Reform Commission 545
National Disability Insurance Schemes 499–500
Physical Activity and Sedentary Behaviour Guidelines 75
prisoner numbers 545
Productivity Commission 499–500
stereotypes of older people 514
tobacco smoking prevention 765
tobacco smoking warnings 324
Australia, Bureau of Statistics, older prisoners 548
authentic learning, transformational learning 716
autism spectrum disorder (ASD) 492
availability, alcohol consumption prevention/control 354
awareness, interpersonal violence 387
awareness raising, war 398
azidothymidine (AZT) 201

Bacillus anthracis infection 614
Bacillus Calmette–Guérin (BCG) 585–587
child vaccination routine 586
t
revaccination 222

bacterial infections
asthma risk factor 51
cancer aetiology 20
t
chain of infection 598
food poisoning 172
Bangkok Charter (2005) 663
Bangkok floods (2011) 723
Bangladesh, humanitarian emergencies 724
bans, tobacco smoking 327
BARDA (Biomedical Advanced Research and Development Authority) 288
bariatric surgery 262
barriers, host infection control measures 607
baseline activity 74
t
bathing, host infection control measures 606
bath salts 339
BCG *see* Bacillus Calmette–Guérin (BCG)
BCS (Blantyre coma score) 232
bedaquiline 220

behavioural changes
host infection control measures 605–607
infected human control measures 611
non-alcoholic fatty liver disease/non-alcoholic steatohepatitis 262
benefits, medical screening 630, 631
benign tumours 17
beta blockers 13–14
Beveridge, William 563
BHR (bronchial hyperresponsiveness) testing 45
Bhutan, tobacco sale prevention 331
bias, rare-variant searches 303
bicycle *safety* 369
Bill & Melinda Gates Foundation 752, 754
Disease Control Priorities Network 368
tuberculosis economics 223
bioinformation, tuberculosis diagnosis 217
Bio-Intelligence Center 288
biological agents, war 396–397
Biological and Toxin Weapons Convention (BWG) 397

biological controls
malaria prevention 237
vectors 608
biological disasters 741
biological HIV vulnerability 419, 419
f
biological interventions, gender health 424
biological propagation, infectious diseases 604
biological risk factors, child mortality & morbidity 434–435
biological screening, infectious diseases control 614
Biomedical Advanced Research and Development Authority (BARDA) 288
BioSense 289
bioterrorism 287–295, 397, 618–619
acronyms 289
t
agent–*perpetrator*–victim connection 292, 292
f
damage estimates 294
emerging/re-emerging infections 277
evolution 287–288
future work 293–295, 294
t
global concerns 293
historical aspects 289–290
microbial agents 287
t
preparedness against 292–293, 293
f
programmes 288
f
relevance 287–288
risk of 287–288
treaties, initiatives, directives & legislation 288–289
BioWatch 289
birth-ascribed status 470
birth registries 447
bisexuals *see* LGBT+ people
Black and Minority Ethnic Elders Study 510
Blantyre coma score (BCS) 232
Blastomyces dermatitidis infection 605

blood pressure
control 12–13
measurement 632
blood smear microscopy 233
blood transfusions, emerging/re-emerging infections 274
bloody diarrhoea 172
Bloomberg Initiative to Reduce Tobacco 330
Bloomberg School of Public health 372
BMI *see* body mass index (BMI)
Boceprevir 257
BOD *see* Burden of Disease (BOD)
body mass index (BMI)
musculoskeletal disorders 137
obesity definition 57, 58
t
Bone and Joint Decade's Task Force on Neck Pain and It's Associated Disorders 131
Bordetella pertussis infection *see* pertussis (whooping cough)
Borrelia burgdorferi infection 605
bovine spongiform encephalopathy (BSE) 183
re-emergence (UK, 1990-1998) 271, 274
BRCA1/2 gene
breast cancer 25
prostate cancer 30
breast cancer 24–26
age-specific incidence 28
f
endogenous hormones 25
incidence 25
mortality from 24–25
prevention 26
screening in disabled people 496–497
survival from 25–26

breastfeeding
diarrhoea prevention 441
gastrointestinal tract infections 174
HIV transmission 203
host infection control measures 606
initiation of 168–169
passive immunization 604
breath testing, tuberculosis diagnosis 219
bronchial hyperresponsiveness (BHR) testing 45
bronchodilators, COPD 44
BSE *see* bovine spongiform encephalopathy (BSE)
bullying, mental health 102
buprenorphine 343
Burden of Disease (BOD) 419, 625–626
healthcare services 687–688
obesity 62–63, 62
t, 63
t
burns prevention 370
Bush, George W. 288, 289

Calymmatobacterium granulomatis infection (Donovanosis) 188
t
Cambodia, Socio-Economic Survey 494
Campbell Collaboration, injury prevention 368
camps, refugees 530–531
Campylobacter infection 188
t

Canada
Changing Directions, Changing Lives 667, 667
b
disabled people exclusion 494
environmental health officers 653
indigenous people as prisoners 545
National Trauma Registry 363
population ageing 509
public health strategies 667, 667
b
tobacco smoking warnings 324
canalization 307
cancer 17–38
aetiology 19–20
benign tumours 17
carcinogenesis 17
chemoprevention 33
epidemiology & distribution 18–19, 18
t
global burden 20–21
incidence changes 19
indigenous peoples' health 480–481
malignant tumours 17
mortality 634, 635
f
primary prevention 31–33
secondary prevention 33
sexually-transmitted infections 192
cancer screening 624
disabled people 496–497, 497
b
overdiagnosis 634–635
preclinical phase 625
proof of 628
candidate gene association studies 300, 301
cannabis 338
CANTOS trial 12
CANVAS trial 7
capability building, Health Needs Assessment 569
capability theory 568
capacity building, environmental health impact assessment 648–649
Capacity Building for Preventing Injuries and Violence: Strategic Plan 2009–2013 (WHO) 371
CAPRISA 004 trial 205
carbapenem-resistant *Enterobacteriaceae* (CRE) 278
carcinogenesis 17

carcinogens
identification of 18–19
oesophageal cancer 26–27
cardiovascular disease (CVD) 3–16
Alzheimer's disease 150
definitions 3–5
diabetes mellitus 7
f, 86
distribution 3–5
gender 417, 420–421, 421
b, 424
geographical distribution 3, 4
f
geographical mortality 5
f
healthcare & gender 421
incidence 3–5
indigenous peoples' health 480
mortality 3, 4
older people healthcare 516
preventative strategies 12–14
prevention & alcohol consumption 350
risk assessment 9–12, 10
f, 632–633
risk factors 5–10, 6
f
types of 3
Cardiovascular Health Study (CHS) 12
career paths, healthcare workers 707
carers, decline in 409
Carnegie Foundation 751
carotid intima medial thickness (CIMT) 12
carpal tunnel syndrome 152
cART (combination antiretroviral therapy) 201, 202
Cartagena Declaration 528–529
case–control studies
Alzheimer's disease 149–150
cancer 18
disease gene discovery 301
caste 469–477
race *vs.* 474
t
cathinones 339

causation
association *vs.* in tobacco smoking 321
emerging/re-emerging infections 282
environmental health assessment 645, 645
b
CBGR (Common But Differentiated Responsibilities) principle 640
CBR (community-based rehabilitation), disabled people 499
CBT *see* cognitive behaviour therapy (CBT)
CCR5 co-receptor antagonists 202
t
CCT (conditional cash transfer programmes) 462
CDC *see* Centers for Disease Control and Prevention (CDC)
CEDAW (Convention on the Elimination of All Forms of Discrimination Against Women) 462
cell culture, typhoid fever diagnosis 179–180
Census of Fatal Occupational Injuries, United States Bureau of Labor Statistics 362
Centers for Disease Control and Prevention (CDC)
bioterrorism 288, 290–291, 290
b, 397
blood lead monitoring 646
Category B agents 290
Category C agents 290
COPD prevention 48
b
Handbook Series in Zoonoses 611
Health Needs Assessment definition 559
Healthy Hospital Choices strategy 671
injuries 362
isolation recommendations 610
malaria recommendations 229–230
malnourishment 720
social determinants of health 471
tobacco smoking 321
traumatic brain injury 145
central administration, health services 704
Centre for Research Excellence in Disability and Health 500, 501
Centre for Research on the Epidemiology of Disasters (CRED) 722
t, 724, 725
t
cerebral malaria 231
cerebrospinal fluid examination 178
cerebrovascular disease 4

cervical cancer
human papillomavirus infection 591
screening in disabled people 496–497
CETA (Common Elements Transdiagnosis Approach) 387
CETP (cholesterol ester transferase protein) inhibitors 9
CFZ (clofazimine) 220–221
Chadwick, Edwin 639
Chagas disease 275
chain of infection 561–598
chancroid 205
Changing Directions, Changing Lives (Canada) 667, 667
b
Chan, Priscilla 752
Chan-Zuckerberg Initiative 754
charitable foundations 751–752, 752
t
Chartered Institute of Environmental Health 653
CHD *see* coronary heart disease (CHD)
chemical incidents 748
b
chemical vector control 607–608
chemical weapons, war 396
Chemical Weapons Convention (CWC) 396
chemoprevention, cancer 33

chemoprophylaxis
  animal control measures 612
  host infection control measures 604–605
chemotherapy *see* pharmacotherapy
Chikungunya virus, 2005 outbreak 273

child abuse
interpersonal violence 378, 381–382
violence 385
childcare, psychological toll 422
Child Growth Standards (WHO) 435, 439
child health 431–452
acute respiratory infection 440–441
at-risk newborn care 438–439
care during infancy 439
determinants & indicators 434–435
diarrhoeal disease 441–442
early development 435
enteric infections 442–443
global initiatives 447–448
health system building blocks 445–446
HIV/AIDS 443–444
humanitarian emergencies 446–447
integrated service delivery 445
key supportive actions 435–436
malaria 443
malnutrition 439–440
maternal & newborn care 436, 437
t
micronutrient deficiencies 440, 440
t
neglected tropical diseases 442–443
progress measurement 447–448
school-age children 433–434, 433
f
suboptimal development/disability prevention 444–445
Sustainable Development Goals 448, 448
b
under-fives 431–433, 432
f, 433
t
Child Health Days 445
Childhood Obesity Plans 675

child marriage
health effects 462
prevention of 474

children
burns prevention strategies 370
drowning prevention strategies 369–370
environmental health assessment 644
environmental health services 653–654
infectious diseases 163, 166, 166
t
interpersonal violence interventions 387
life expectancy 418
mortality 163, 166
t, 565
obesity definition 58
obesity trends 60, 60
t, 61
f
parasitic diseases 182
Child Survival Call to Action 447

China
Elderly Rights Law 519
Global Adult Tobacco Use Surveys 325, 330
one-child families 404–405
private support of public health 756–757
public health strategies 660
Chlamydia infection 31
Chlamydia trachomatis infection 187
t
morbidity 191–192
mother-to-child transmission 191
chloroquine 234–235
chloroquine-resistant *Plasmodium falciparum* 235, 278
CHNA (Community Health Needs Assessment) 559
Choice of Termination Act (1996 South Africa) 424
cholecystectomy 24
cholera 171
t
child vaccination routine 586
t
emerging/re-emerging infections 276
population growth and 274–275
Tanzania outbreak, 1998 271
vaccination 592–593
vaccine 174
cholesterol control, cardiovascular disease prevention 13
cholesterol ester transferase protein (CETP) inhibitors 9
chondropathies 130
chromosomal instability (CIN), colorectal cancer 24
chromosome variation 298
chronic alcohol consumption 349–350
chronic daily headache 145
chronic diseases, infectious diseases *vs.* 623
chronic hepatitis 308
chronic hyperglycaemia 152
chronic infections, humanitarian emergencies 728
chronic kidney disease 635
chronic malnutrition 440
chronic obstructive pulmonary disease (COPD)
asthma-COPD overlap 45
burden of 46
clinical assessment 43–44
definitions 43
epidemiological assessment 44
epidemiology 45–46
future prevalence 46
indigenous peoples' health 481
prevalence 46
f
prevention 48

risk factors 46–47
second-hand smoking 322
treatment 52
chronic traumatic encephalopathy (CTE) 145
CHS (Cardiovascular Health Study) 12
CHWs (community health workers) 436
Ciba Guest Symposium, COPD definition 43
CIMT (carotid intima medial thickness) 12
CIN (chromosomal instability), colorectal cancer 24
ciprofloxacin-resistant *Shigella* 442

cities
public health strategies 671 *see also* urbanization
Civil Contingencies Act (2004, UK) 740, 740
b
civil disturbances, emerging/re-emerging infections 275–276
civilians, mortality in war 393–394
CJD (Creutzfeldt–Jakob disease) 148, 183
CLAS (Healthy Latin America Coalition) 579

classroom behaviour management
  maternal health 423
  mental health 103
cleaning, infectious diseases control 615

climate change
child health 446
emerging/re-emerging infections 275, 617
health impacts 763
humanitarian emergencies 722–723
indigenous peoples' health 484
Climate Change Conference (2017, UN) 763
Clinical Readiness Initiative (CRI) programme 289
clinical recognition indicators, bioterrorism 290
clinicians, environmental health services 653–654
clofazimine (CFZ) 220–221
closed injuries, traumatic brain injury 145
Clostridium tetani see tetanus
cloud computing, humanitarian emergencies 733
cluster headaches 144
Cluster System (UN) 727, 727
f
cluster trials, medical screening 630
CM (contingency management) 344
CMRs (crude mortality rates) 531
CMV (cytomegalovirus) infection 188
t
CNVs (copy number variants) 298
Coalition for Epidemic Preparedness Innovation 288
coalitions, non-communicable disease 579
coastal communities, humanitarian emergencies 724–725
cocaine 338–339

Cochrane Collaboration
  healthcare quality & waste reduction 689
  injury prevention programmes 367–368
Cochrane Library, tobacco smoking cessation 330
Cochrane reviews, Health Needs Assessment 560
cognitive behaviour therapy (CBT)
  post-traumatic stress disorders 103
  substance use disorders interventions 344
cohabitation 403, 405–406
  children, effects on 408–409

cohort studies
  cancer 18
  child chronic malnutrition 440
  disease gene discovery 301
  tobacco smoking and mortality 320–321
Collaborating Centres network (WHO) 282
collective violence *see* war
College of Community Medicine (Hong Kong) 677
colonial processes, indigenous peoples' health 484
colorectal cancer 23–24
  screening in disabled people 497
combination antiretroviral therapy (cART) 201, 202
combination drugs, malaria treatment 234
Commission on Chronic Illness (USA) 624
Commission on Microeconomics and Health (WHO, 2001) 281, 282
Commission on the Social Determinants of Health (CSDH) 565
  dental health inequalities 118, 119
    b, 123
Committee on Refugees (USA) 536
Common But Differentiated Responsibilities (CBDR) principle 640
Common Elements Transdiagnosis Approach (CETA) 387
common risk factor approach (CRFA), dental disease 117

communicable diseases
child health 448
forced migrant mortality 533
refugee camps 530–531
communication, healthcare services 688–689

communities
adolescent health interventions 461–462
alcohol consumption policies 356
alcohol consumption prevention/control 355
dental services 123
disabled people living 494
Health Needs Assessment 561
infectious diseases control measures 602
injuries 367
interpersonal violence 384–386, 387
mental health 102–103, 106
older people healthcare 517–518
physical activity 79
b
preparation in humanitarian emergencies 724
protection by vaccination 603
suicide prevention 105
Communities that Care (CTC)
adolescent health 462
mental health 102–103
community-based rehabilitation (CBR), disabled people 499
community-engaged health professionals 711
Community Health Needs Assessment (CHNA) 559
community health workers (CHWs) 436
comorbidities, non-communicable diseases 576
comparative needs, Health Needs Assessment 560
complete quarantine 610
complex multifactorial diseases 299–300
complex system computational models 473
complicated malaria 231–232
composite indicators, environmental health 647
Comprehensive Nuclear-Test-Ban-Treaty (CTBT) 395
conceptual models, physical activity 77
conditional cash transfer programmes (CCT) 462

condoms
HIV infection prevention 205
host infection control measures 606, 607
conduct disorders, prevention 103

conflicts
interpersonal violence 383
life expectancy 418–419 *see also* war

confounding
genetic association studies 303–304
medical screening 629
congenital rubella syndrome (CRS) 591
reduction by vaccination 584
t
Congo, Ebola virus case study 661
consent, informed 310
b
contact precautions, isolation 610
contaminated water avoidance 611
contingency management (CM) 344
control measures, infectious diseases 599
Control of Communicable Diseases Manual, American Public Health Association 597
conventional weapons 395
Convention on the Elimination of All Forms of Discrimination Against Women (CEDAW) 462
Convention on the Rights of People with Disabilities (CRPD, UN) 97, 501
Convention on the Rights of the Child (CRC) 436, 453
Convention on Trafficking (UN) 530
cooling, infectious diseases control 615
COPD *see* chronic obstructive pulmonary disease (COPD)
copy number variants (CNVs) 298
core groups, sexually-transmitted infections 193
coronary artery calcium 12
coronary heart disease (CHD)
diabetes mellitus mortality 7
f
incidence 3–4
lipoprotein (a) and 306–307
prevalence 4
Correlate of Risk Targeted Intervention Study (CORTIS) 214, 216
corticosteroids 178
CORTIS (Correlate of Risk Targeted Intervention Study) 214, 216
Corynebacterium diphtheriae see diphtheria
COSH (Council on Smoking and Health) (Hong Kong) 668
cosmetic dentistry 124
cost-effectiveness, obesity prevention 68
t

costs
epilepsy 148
headache 145
medical screening 630–631
prisoners 550 *see also* economics
cost sharing, healthcare services 693

cost utility
Health Needs Assessment 562
quality-adjusted life years and 567
cough, infected human control measures 611
Council on Smoking and Health (COSH) (Hong Kong) 668
Countdown to 2030 initiative 448–449
counterfeit medicines 766
country comparisons, physical activity 75–76
coverage, healthcare services 686
CRC (Convention on the Rights of the Child) 436, 453
CRE (carbapenem-resistant *Enterobacteriaceae*) 278
C-reactive protein (CRP) 11–12
CRED (Centre for Research on the Epidemiology of Disasters) 722
t, 724, 725
t
Creutzfeldt–Jakob disease (CJD) 148, 183
CRFA (common risk factor approach), dental disease 117
crime causation theories 545–546
CRI (Clinical Readiness Initiative) programme 289
crisis perceptions, forced migration 523–524
CRISPR, bioterrorism 294
Crohn's disease 24
crowd sourcing, humanitarian emergencies 732–734
CRP (C-reactive protein) 11–12
CRPD (Convention on the Rights of People with Disabilities) 97, 501
CRS *see* congenital rubella syndrome (CRS)
cruciferous vegetables 21–22
crude mortality rates (CMRs) 531
cryptogenic cirrhosis 261
CSDH *see* Commission on the Social Determinants of Health (CSDH)
CTBT (Comprehensive Nuclear-Test-Ban-Treaty) 395
CTC *see* *Communities that Care* (CTC)
CTE (chronic traumatic encephalopathy) 145
Culex mosquitoes, Japanese encephalitis 591

culture
healthcare services 699
b
interpersonal violence 384–385
parents in families 410
tuberculosis diagnosis 217–218
current care delivery models 688
curricula reform, transformational learning 715
CVD *see* cardiovascular disease (CVD)
CWC (Chemical Weapons Convention) 396
Cyclospora, infected foods 279
cytomegalovirus (CMV) infection 188
t
cytosine 297

d4T *(stavudine)* 201
DAAs (direct acting antiretrovirals) 345
daily living, dental health inequalities 119
DALE *see* disability-adjusted life *expectancy* (DALE)
DALYs *see* disability-adjusted life years (DALYs)
Danish twin studies, multiple sclerosis 154
Dark Winter 289

data
disabled people 499
global health sustainability 764–765
processing 310
b
dating violence 387
ddI (didanosine) 201
DDT (dichlorodiphenyltrichloroethane) 237
decayed, missing, and filled teeth (DMFT) index 114
decentralization, developing countries healthcare 690–691
deciduous teeth trauma 117
decision-making power, non-communicable diseases 576–577
Declaration of Territorial Asylum (DTA) 529
Declaration on Promoting Health (2016) (Shanghai) 663–664

decontamination
emergencies 745–746, 745
b
radioactive incidents 746
definition of public health 710
deforestation, emerging/re-emerging infections 275
dehumanization 567
delamanid 220
delayed marriage 405–406
delinquency, interpersonal violence 383–384
dementia 148–149
older people 516
dementia with Lewy bodies (DLBs) 148, 151
demographic transition, older people 511–512

demography
adolescent health 455–456, 455 *f*
emerging infectious diseases 617
healthcare services 687
Dengue fever 180–181
child health 443
child vaccination routine 586 *t*
mortality 165 *t*
vaccination 593–594
denosumab 307–308
dental caries 114–115, 115 *f*
dental health 114–127
access to 122
deprofessionalization of 124
dietary sugars 121
evidence base 124
fluoride and 119–120
future development 123
LMICs 123
oral cleaning 121–122
scope of 114
technology 124 *see also* dental caries; oral disease
dental health inequalities 118
daily living conditions 119
power/money/resources 119
reduction in 118–123
dentures 29
Department of Health and Human Services (US)
Physical Activity Guidelines for Americans 75
Surgeon General's Report on Physical Activity and Health 75
Department of Violence and Injuries Prevention (WHO) 368
Department of Violence and Injury Prevention and Disability (WHO) 371

depression
  interpersonal violence 383
  magnitude & costs 96
  older people healthcare 517
  prevention 103
de-scaling, periodontal diseases 116
descriptive indicators, environmental health 647
deterrence, alcohol consumption prevention/control 353
detoxification, substance use disorders interventions 344

developed countries
dental caries 114
refugees' public health 530 *see also* high-income countries (HICs)

developing countries
healthcare services 687, 704 *see also* low- and middle-income countries (LMICs)
development displace 526
t, 527
legal frameworks 530
Diabetes Control and Complications Trial of Type 1 diabetes 6–7
diabetes mellitus 85–94
cardiovascular disease and 6–7, 13
challenges 88–90
classification 85
colorectal cancer 24
complications 90
countries by 88
f
definition 85, 85
f
diagnosis 86
economic cost of caring 760
evidence-based care 91
gestational *see* gestational diabetes
global burden of disease, part of 86, 88
global changes in 90
healthcare costs 88
incidence 87
f
lifespan increase 90
liver cancer 22
LMIC *vs.* HIC 91
management 91
mono- & polyneuropathies 152
mortality from 7
f, 88
multimorbidity 90, 91
non-alcoholic fatty liver disease 262
prevalence 87
f
prevention 91
research in 90–91
social impacts 88
type 1 diabetes 85
type 2 *see* type 2 diabetes
type 2A diabetes 90
type 2B diabetes 90
worldwide prevalence & trends 86, 87
f
Diabetes UK 675
b
Diagnostic and Statistical Manual of Mental Disorders (DSM)
alcohol consumption 350
dementia 148

substance use disorders 340

diarrhoeal disease
acute watery diarrhoea 172
bloody diarrhoea 172
burden & mortality 170, 171
f
child health 441–442
clinical features 172
t
environmental health 643
forced migrant mortality 531
global burden 598
inflammatory diarrhoea 172
mortality 165
t, 721
f
persistent diarrhoea 172
severe community-acquired diarrhoea 173
travellers' diarrhoea 172
under-fives child health 433
dichlorodiphenyltrichloroethane (DDT) 237
didanosine (ddI) 201

diet
asthma risk factor 50
cancer aetiology 19
cardiovascular disease prevention 13–14
individual factors 64
laryngeal cancer 29
obesity 64
obesity recommendations 66
oesophageal cancer 27
pancreatic cancer 28
Parkinson’s disease 153–154
sugars & dental health 121 *see also* nutrition
dietary abstinence, gastrointestinal tract infections 173
Dietary Approaches to Stop Hypertension (DASH) 13
digestive vacuoles (DVs), *Plasmodium* lifecycle 229
digital health technologies 760

diphtheria
1990 outbreak 273
artificially acquired passive immunization 604
mortality 165
t
reduction by vaccination 584
t
vaccination 588
direct acting antiretrovirals (DAAs) 341
direct costs, oral disease 117
direct fee-for-service systems, dental services 122–123
direct transmission, infectious diseases 599–600
Dirty Ashtray Award 329
disability-adjusted life *expectancy* (DALE)
adolescent health 457
gender-based violence 422
gender effects 419
disability-adjusted life years (DALYs)
alcohol consumption 350–351
cancer 21
environmental health 643, 647
gender effects 419
humanitarian emergencies 728
infectious diseases 166
injuries 364, 364
f
mental health 95
musculoskeletal disorders 135, 138
substance use/misuse 337
disability prevention, child health 444–445
disability-specific initiatives 496
disabled people 489–506
conceptualization evolution 489–491
disability measurement 490–491
discrimination 496
global health sustainability 761
health of 491–493
health promotion 496–497
health-related rehabilitation 498–499
health service access 497–498
housing 494
inclusive programmes 496
interpersonal violence 378–379
measurement of 490–491, 499
musculoskeletal disorders 130
older people 511–514
prevalence 491, 491
f
public health 496–497
research 500

risky behaviours 492–493
services & supports 499–500
social determinants of health 493–496
stigma 496
violence 494–495

disadvantaged populations
  mental health 99, 100
  suicide 97
disaster zones, Health Needs Assessment definition 559
discrimination, disabled people 496
disease burdens, adolescent health 457, 459
Disease Control Priorities Network 368
Disease Priorities Handbook 368
diseases, medical screening effectiveness 625–626
disinfection, infectious diseases control 615
disposable equipment, infectious diseases control 615
district hospitals, developing countries 706
diversity generation, genetics 298
divorce 407–408
  children, effects on 408–409
dizygotic (DZ) twins, complex multifactorial diseases 299
DLBs (dementia with Lewy bodies) 148, 151
DMFT (decayed, missing, and filled teeth) index 114
DNA 297–298
Dobell, Horace 624
documentation, war 397–398
Do Development Differently conference 726
Donovanosis *(Calymmatobacterium granulomatis* infection) 188
  t
dorsopathies 129
dose–response environmental health assessment 646
D-P-S-I-R (driving forces-pressure-state-impact-response) model 647
dracunculiasis 616
drink driving legislation 353, 369

drinking behaviour
  alcoholic liver disease 259 *see also* alcohol consumption
driving forces-pressure-state-impact-response (D-P-S-I-R) model 647

droplet nuclei
  disease transmission 600
  isolation recommendations 610
drowning 362, 364
  prevention strategies 369–370
drug-resistant tuberculosis 213–214
Drugs for Neglected Diseases Initiative 766
drug use/abuse *see* substance use disorders (SUDs)
DSM *see Diagnostic and Statistical Manual of Mental Disorders* (DSM)
DTA (Declaration of Territorial Asylum) 529
DTP containing vaccine 586
  t
Duffy antigen, *Plasmodium vivax* 228
Dunning Committee, Netherlands 694–695
dust, infectious diseases 600
DVs (digestive vacuoles), *Plasmodium* lifecycle 229
dynamic approach, social justice 568
dysentery, child diarrhoea 442
dysfunctional parenting, interpersonal violence 381
dyslipidaemia, cardiovascular disease 8–9
DZ (dizygotic) twins, complex multifactorial diseases 299

early adolescence 454

early life
child health 435
experiences & prisoners 546
interpersonal violence 381–382
mental health 98–99
early onset sepsis (EOS) 167
EASD (European Association for the Study of Diabetes) 13

East Asia
divorce 407–408
non-marital childbearing & divorce effects 408
Eastern Mediterranean Regional Office (EMRO) 728

eating behaviour
  disorder prevention 103
  host infection control measures 606
Ebola haemorrhagic fever 270, 618
  Congo case study 661
  emerging/re-emerging infections 275, 276
  epidemic (2014–2015) 762–763
  vaccination 595
Ebola Outbreak Epidemiology Team (2018) 761
Echinococcus granulosus infection 612
ecological studies, hepatocellular carcinoma 22

economic development
  emerging infectious diseases 617
  environmental health 642–643
economic empowerment interventions 387
economic migration 525
Economic Research Service (ERS) 137

economics
alcohol consumption and 351
barriers to adolescent healthcare 459–460
child health 445–446
forced migration 523–524
global health sustainability 766
healthcare services 686
Health Needs Assessment 562–563
incentives in environmental health 651
injuries 365
interpersonal violence 379
non-communicable diseases 573, 576–577, 579–580
obesity 62–63
oral disease 117
transformational learning 715 *see also* costs

ecosystems
  emerging/re-emerging infections 282
  healthcare services 691
  indigenous peoples' health 484
ECRHS (European Community Respiratory Health Survey) 48
ecstasy (methylenedioxymethamphetamine) 338

education
adolescent health 460–475
alcohol consumption prevention/control 352
Alzheimer's disease and 150
child health 447
disabled people exclusion 494–495
healthcare services 686
transformative learning 712
war 398
effectiveness evidence, alcohol consumption prevention/control 352–353
efficiency, Health Needs Assessment 562–563
eHealth, HIV infection prevention 206
EHOs (environmental health officers) 652–653
EHRs (electronic health records), genomics 310
EIA (environmental impact assessment) process 648
Eisenhower, Dwight D. 394–395
elder abuse 517
elderly people *see* older people
Elderly Rights Law (China) 519
electrodiagnostic testing 151
electromyography 151
electronic cigarettes 765
electronic health records (EHRs), genomics 310
ELF (enhanced liver fibrosis) panel 260
Elizabeth Glaser Foundation 753
El Niño Southern Oscillation (ENSO) 275
emergencies 739–750
antidotes 746
bioterrorism management 293, 293
f
decontamination 745–746, 745
b
evacuation 745
lesson identification & publishing 747–748
management of 740
f
population follow-up 746–747, 746
b
preparedness 761–763
public reassurance 747
recovery 746
types of 741
Emergency Committee (IHR) 282
Emergency Medical Teams Initiative (WHO) 761
Emergency Response Framework (WHO) 740–741, 761
emergency risk management 724–725
key components 724–725, 725
t
emerging/re-emerging infectious diseases 269–286
antimicrobial drug resistance 277–278

control of 617–618
economic impact 271–272, 272
f
epidemiology 269
future work 281–282, 283–284
geographic distribution 270–271, 271
f
human behaviour 276–277
international spread 279–280
international systems 281–282
limited human–human transmission 269–270
national systems 281–286
non human–human transmission 269
public health security 279–280
risk factors 273–274
safety procedure failures 274–277 *see also* neglected tropical diseases (NTDs)
emotional abuse, interpersonal violence 381–382
empowerment structures, local action 674–675, 674
b
EMRO (Eastern Mediterranean Regional Office) 728
enablers, healthcare services 697–699
encephalitis 165
t
ENCODE project 306
b
endogenous hormones, breast cancer 25
End TB Strategy 213

England
public health strategies 667, 678
sugar reduction strategies 672 *see also* United Kingdom (UK)
enhanced liver fibrosis (ELF) panel 260
ENSO (El Niño Southern Oscillation) 275
Entamoeba histolytica infection 188
t
enteric infections 442–443
Enterobacter infections
age-of-onset 168
t
antibody-resistance 277–278
Enterobius vermicularis infection 182
t
entertainment, violence 385
entomological inoculation rate, malaria 242
entomological surveillance indicators, malaria 242
entry inhibitors, HIV infection treatment 202
t
environmental health 639–657
burden of 643–644, 643
b, 643
t
child health 446
concept & scope 641
decision-making tools 646–649
enteric dysfunction 434–435
financial incentives 651
genetics and 300
global initiatives 640
global trends 641–642
historical aspects 639–640
impact assessment 647–649
infectious diseases control 601, 612–614
international regulations 651
intersectoral action & partnerships 649–650
issues of 642–643
multiple sclerosis 154
musculoskeletal disorders 130
non-communicable disease prevention 574
non-communicable diseases 574–578
policies 650–651
public health strategies 669–675
urban *vs.* rural areas 642
vector control 608
environmental health assessment 644–646
conduction of 645
epidemiology & toxicity 644–644
health outcomes 644
multifactorial exposure 644

nature of 644
risk assessment 645–646
vulnerable groups 644
environmental health officers (EHOs) 652–653
environmental health services 651–654, 652
b
children 653–654
practitioners 652–653, 652
b
environmental impact assessment (EIA) process 648
environmentally displaced persons 526, 526
t
legal frameworks 530
environmental measurements, emergencies 747
Environmental Protection Agency (EPA) 289
EOS (early onset sepsis) 167
EPA (Environmental Protection Agency) 289
EPI *see* Expanded Programme on Immunization (EPI) (WHO)

epidemics
prevention of 762–763
zoonotic control measures 612

epidemiological studies
hypertension 633
mental health 422–423
war documentation 397

epidemiological transition
adolescent health 455–456
older people 511–519
under-fives child health 431

epidemiology
  cancer primary prevention 32
  environmental health assessment 644–645
  Health Needs Assessment 560–561
epigenetics, prisoners 546
epilepsy 147–149
  costs and burden of 148
  genetics 148
  incidence 147
  infections 148
  mortality 147–148
  prevalence 147
  seizure classification 147
Epstein–Barr virus infections
  cancer aetiology 19
  multiple sclerosis 154
equality, equity *vs.* 417
  f
equal opportunities, mental health 102
equipment design, food sanitation 614

equity
dental services 123
efficiency trade-off 563
equality *vs.* 417
f
Health Needs Assessment 562–563
erectile dysfunction 322
ERS (Economic Research Service) 137
ERS (European Respiratory Society) 43
erythrocytes, malaria 231
Escherichia coli infection
age-of-onset 168
t
child health 442–443

ethical issues
  forced migrants 539
  genetic epidemiology 310, 310
    b
  health equity/health equality 566–567
  prisoners 549
ethnic admixture 304
ethnicity 469–477
  asthma risk factor 49
  dental health inequalities 118
  diabetes mellitus 89
  maternal health 423
  obesity trends 61
  suicide 97
  terminology 470
Euro-centric knowledge systems, indigenous peoples' health 484

Europe
delayed marriage & cohabitation 405
non-marital childbearing & divorce effects 408
public health strategies 665
European Association for the Study of Diabetes (EASD) 13
European Code Against Cancer 36
t
European Community Respiratory Health Survey (ECRHS) 48
European Respiratory Society (ERS) 43
European Union (EU), public health strategies 665
evacuation, emergencies 745
Everybody Active Every Day 674
Every Newborn Action Plan 447
Every Woman Every Child's Global strategy for Women's and Children's Health 447
evidence, dental health 124
excise taxes, tobacco taxation 327–328
Executive Order on Children's Environmental Health and Safety (1997) 653–654

exercise
cardiovascular disease prevention 13
definitions 74
t
older people 515
osteoarthritis 133 *see also* physical activity; sedentary behaviour
exocrine pancreatic insufficiency related diabetes 85
exome sequencing 303
Expanded Programme on Immunization (EPI) (WHO) 583
polio vaccination 587

exposure
  environmental health assessment 646
  routes in bioterrorism 290
external exposure modelling 742
external investment, private support of public health 756
extremely drug-resistant (XDR) *Salmonella typhi* 179
extremely drug-resistant tuberculosis (XDR-TB) 278, 752

facility design, food sanitation 614
faeces control, infectious diseases 613
Fagerstrom test 330
falls 364
  older people healthcare 516
  traumatic brain injury 145
false positives 628
familial aggregation, complex multifactorial diseases 299
familial hypercholesterolemia (FH) 6
familial relative risk (FRR) 309
  f
families 403–413
  adult transition delay 406–407
  association studies 300
  delayed marriage 405–406
  development & health 410–411
  divorce 407–408
  gender roles 404
  global changes 403–404
  grandparents–grandchildren relationship 410
  key institution as 410
  late marriage 406–407
  members as prisoners 548
  mental health 102
  older people and 409–411, 519
  reproductive health issues 410
  size & host infection control measures 607
  smaller sized 404–405
  substance use disorders 339 *see also* cohabitation

family history
  cardiovascular disease 5–6
  oesophageal cancer 26
  pancreatic cancer 28
family knowledge, gastrointestinal tract infections 174
Family Planning Summit 447
fasting glucose 86
fasting plasma glucose (FPG) 86
Fatal Accident Reporting System, United States Department of Transportation 362
FCA (Framework Convention Alliance) 329
FCTC *see* Framework Convention on Tobacco Control (FCTC)
FDA (Food and Drug Administration), tobacco smoking warnings 325
FDC (fixed-dose combination) drugs 202
feasibility, infectious diseases control 615
Federal Highway Safety Administration 368
Federal Highway Traffic Safety Act (1966) 368
female genital mutilation 533
female homicide 378
FEM-PREP trial 205
fentanyl 345

fertility
  decline & family *size* 404–405
  population ageing 509
FEV1 (forced expiratory volume in 1 second) 43–44
FH (familial hypercholesterolemia) 6
FIB-4 index 260
finances *see* economics
finite resources, Health Needs Assessment 561
Finland, Tobacco Act (2010) 331

firearms
  legislation 386
  suicide 104
fire burns 364
Five Year Forward View on Mental Health, Mental Health Taskforce 667
fixed-dose combination (FDC) drugs 202

flooding
emerging/re-emerging infections 276
understanding of 746
b
fluid avoidance, gastrointestinal tract infections 173
fluoride, dental health 119–121, 121
t
fluoride milk 120
fluoride mouth rinses 121
fluoride tablets 120
fluoride toothpaste 120–121
fluoride varnishes/gels 121
Fogarty International Center 371–372
Food and Drug Administration (FDA), tobacco smoking warnings 325
food-borne infections, emerging infectious diseases 617
food-borne infections, child health 442
food industry 576
food preparation/processing, infection control measures 611, 617
food *preparation*/processing, emerging infectious diseases 617
food sanitation 613–614
food storage, infectious diseases control 613
forced displacement 731
emerging/re-emerging infections 275–276
war 394
forced expiratory volume in 1 second (FEV1) 43–44
forced migration 523–542
categories of 525–527, 526
t see also asylum seekers; development displace; environmentally displaced persons; internally-displaced persons (IDPs); refugees; smuggled persons; trafficked people
displacement causes 523–524
ethical issues 539
healthcare in developed countries 536–538
healthcare quality 538–539
immediate drivers 524–525
infectious disease spread 723–724
legal frameworks 528–530
mental health 533–536
mortality causes 531–536
North–South relationships 524
protection right evolution 530
public health professionals 538
public health situation 530–531
public health solutions 536
rates of 720
f
statistics of 527
voice, visibility, & agency 539
Foresight obesity system map 65
f
FPG (fasting plasma glucose) 86

frailty, older people 515
Framework Convention Alliance (FCA) 329
Framework Convention on Climate Change (UN, 1992) 281
Framework Convention on Tobacco Control (FCTC) 319, 329, 668, 765
  adolescent health 457
Framingham Heart Study 5
  heart failure 10
  stroke risk factors 10
Friends Overseas Service Committee 720
frontotemporal dementia (FTD) 548–549
frontotemporal lobar degeneration (FTLD) 148, 151
FRR (familial relative risk) 309
  f
fruits & vegetables
  cruciferous vegetables 21–22
  infectious diseases control 613
  laryngeal cancer 29
  prostate cancer 30
FTD (frontotemporal dementia) 548–549
FTLD (frontotemporal lobar degeneration) 148, 151
fungi, chain of infection 598
fusion inhibitors 202
  t

gametocytes, malaria 242
gangs, interpersonal violence 384
gastroesophageal reflux 27
gastrointestinal tract infections 169–174, 170
  t
  aetiology 170–171
  clinical features 172
    t
  diagnosis 172
  evidence-based interventions 172, 173
    t
  geographical distribution 171
    t
  global epidemiology 169–170
  medical treatment 172–173
  presentation 172, 172
    t
  preventative measures 173–174
  regional epidemiology 169–170
  transmission 600
GATS (Global Adult Tobacco Survey, China) 325
gay people *see* LGBT+ people
GBD *see* Global Burden of Disease (GBD)
GBV *see* gender-based violence (GBV)

GCS (Glasgow Coma Scale) 145
gender 415–430
  advocacy & group community initiatives 426
  alcoholic liver disease 259
  asthma 49
  cardiovascular disease 420–421, 421
    b
  child mortality & morbidity 434
  definition 416
    b
  empowerment programmes & interpersonal violence 387
  health interventions 424–426
  health issues 417–418
  health needs 416
  HIV/AIDS 419–420, 419
    f, 420
    b
  interpersonal violence 422
  intersectionality 416–417
  labour division 422
  life expectancy 418–419, 418
    t
  mental health 95–96, 422–423, 423
    b
  multiple sclerosis 154–155
  non-communicable diseases 420–421
  physical activity differences 76
  population ageing 509–560
  power hierarchies 416
  power inequity 422
  power relation changes 426, 426
    b
  power relations 418
    t
  Social Determinants of Health 415
  social roles 509–510
  suicide 96
  transformative initiatives 424 *see also* maternal health
gender-based violence (GBV) 421–422, 421
  b, 422
  b
  advocacy & group community initiatives 426
gender constructions 416
gender hierarchies 416

gender identity
definition 416
b
interpersonal violence 386
gender inequalities 417, 417
b
interpersonal violence 385–386
gender mainstreaming 417
gender norms, violence 384
gene–environment interactions 300, 644
generalized seizures 147
genes 297–298
genetic determinism 310
b
genetic discrimination 310
b
genetic epidemiology 297–315
definition 297
disease gene discovery 300–307
drug discovery & development 307–308
environmental exposure 306–307
ethical implications 310, 310
b
genomic medicine 308
importance of 307–310
LMICs 311
personalized medicine 309–310
phenotypic characterization 311–312
public health practice 297
risk communication 312
risk prediction 308–309, 309
f
social implications 310, 310
b
stratified medicine 309–310, 309
f
unexplained variations 310–311
genetic heterogeneity, hepatitis B virus 251
genetic instability, colorectal cancer 24
genetic linkage scans 301
genetic modification, bioterrorism 294
genetic polymorphisms, alcoholic liver disease 259
genetic recombination 301
genetics 297–298
Alzheimer’s disease 150
asthma risk factor 51
cancer aetiology 19–20
carcinogenesis 17
complex multifactorial diseases 299–300
COPD risk factors 47–48

diversity generation 298
DNA 297–298
environment and 300
epilepsy 148
genes 297–298
genetic variants 298
genotypes 298
inheritance patterns 298
laryngeal cancer 29
low back pain 131
multiple sclerosis 154
obesity 64–65
oral cavity cancer 29
phenotypes 298
RNA 297–298
single-gene disorders 298–299
genetic variants 298
Geneva Convention (1979) 394, 730
GeneXpert MTB/RIF test 218
GeneXpert MTB/RIF Ultra test 218
genome linkage studies 51
genome variation 298
genome-wide association studies (GWAS) 301–302
Alzheimer's disease 150
analytical issues 303
asthma 51
causal inference 303–304
COPD 47–48
lung cancer 22
oesophageal cancer 26
pancreatic cancer 28
prostate cancer 30
rare-variant searches 302–303
single-nucleotide polymorphisms 301–302, 302
b
stomach cancer 23
trait–drug matches 308
t
genomic control 304
genomic diversity, LMICs 311
genomic medicine 308

genomics
  cardiovascular disease 5–6
  electronic health records 310
genotype 298
  interpersonal violence 381
genotype imputation 303
Genuino Clandestino (Italy) 426

geography
child mortality & morbidity 434
gastrointestinal tract infections 171
t
healthcare service expenditure 688
maternal health 423
mental health 95
germ theory 567
Gerstmann–Straussler–Scheinker disease 183
gestational diabetes 85
indigenous peoples' health 481
GFATM (Global Fund for AIDS, Tuberculosis and Malaria) 753
GHPSS (Global Health Professions Student Survey) 325
GHWA (Global Health Workforce Alliance) 716
giardiasis 188
t, 617
GINA (Global Database on the Implementation of Nutrition Action, WHO) 66–67
gingivitis 115
Glasgow Coma Scale (GCS) 145
Global Action Plan, antimicrobial resistance 763
Global Action Plan for Healthy Lives and Well-being for All 760
Global Action Plan for Pneumonia and Diarrhoea 447
Global Action Plan for the Prevention and Control of Non-Communicable Diseases 2013-2020 (WHO) 75, 574
Global Action Plan on Antimicrobial Resistance (WHO) 280
Global Action Plan on Non-communicable Diseases 329
Global Action Plan on Physical Activity 2018-2030: More Active People for a Healthier World (WHO) 75
Global Adult Tobacco Survey (GATS, China) 325
Global Adult Tobacco Use Surveys (China) 330
Global Alliance for TB Drug Development 766
Global Alliance for Vaccines and Immunization (Gavi) 583, 754, 761
global burden, injuries 363–364, 363
t
Global Burden of Disease (GBD)
acute respiratory infections 174
cancer 21
diabetes mellitus 86, 88
environmental health 643, 643
b
gender 418–419
headache 144
injuries 361, 362
meningitis 177
mental health 95
musculoskeletal disorders 134, 137
substance use/misuse 337
tobacco smoking 323
typhoid fever 180
t

under-fives morbidity 432–433, 433
t
Global Burden of Diseases, Injuries and Risk Factors 2015 study 362
COPD 46
Global Commission on Health Professionals Education (2010) 711
Global Database on the Implementation of Nutrition Action (GINA, WHO) 66–67
global estimates, intimate partner violence 378
Global Fund for AIDS, Tuberculosis and Malaria (GFATM) 753
Global Fund to Fight AIDS, Tuberculosis, and Malaria 223
Global Green and Healthy Hospitals 671
Global Health Professions Student Survey (GHPSS) 325
global health research, intersectionality 471–473, 471
f
Global Health Sector Strategy (WHO) 192
Global Health Sector Strategy on Sexually Transmitted Infections 2016-2021 (WHO) 194
global health sustainability 759–770
better health & well-being 763–764
disabilities 761
emergency preparedness & response 761–763
epidemic/pandemic prevention 762–763
future work 767
innovation/data/statistics 765–766
Global Health Workforce Alliance (GHWA) 716
Global Humanitarian Assistance 728
Global Initiative for Chronic Obstructive Lung Disease (GOLD) 43, 44
b

global initiatives
alcohol consumption policies 356
child health 447–448
environmental health 640
public health strategies 661–662

globalization
alcohol consumption policies 356
forced migration 523–524
Global Malaria Eradication Programme (GMEP) 227
Global Malaria Programme (GMP) 243
global monitoring framework (GMF) 664
Global Network for Age-friendly Cities and Communities (WHO) 519
Global Network of the Society for Research on Nicotine and Tobacco (SRNT) 329
Global Non-Communicable Disease Action Plan (2013-2020) 664
Global Observatory for Physical Activity (GoPA!) 77
Global Outbreak Alert and Response Network (WHO) 282
Global Polio Eradication Initiation 587
global political economy, non-communicable diseases 575–578
Global Recommendations on Physical Activity for Health (WHO) 75
global risk scores, cardiovascular disease risk assessment 10–11
Global Road Safety Leadership Course 372
Global School Personnel Survey (GSPS) 325
Global Strategy for the Prevention and Control of Non-Communicable Diseases (WHO) 664
Global Strategy for Women's, Children's and Adolescents' Health 464
Global Strategy on Human Resources for Health Workforce 2030 (GSHRH 2030) 712
Global Strategy to Reduce the Harmful Effects of Alcohol (WHO) 351
Global Technology for Malaria 2016-2030 (GTS) 240, 241
t, 244, 244
t
Global Tobacco Surveillance System (GTSS) 325
global trends, environmental health 641–642
Global Vaccine Action Plan (GVAP)
measles eradication 590
World Health Assembly 595
Global Youth Tobacco Survey (GYTS) 325
GLP-1 receptor agonists 7
glucagon-like peptide 1 (GLP-1) analogues 262
glucocorticoid receptor (GR) 546
glutathione transferases (GST) 153
glycaemic control, diabetes mellitus diagnosis 6
glycated haemoglobin (HbA1c) 6, 86
GMEP (Global Malaria Eradication Programme) 227
GMF (global monitoring framework) 664
GMP (Global Malaria Programme) 243
GOLD (Global Initiative for Chronic Obstructive Lung Disease) 43, 44
b
Good Schools Programme (Uganda) 387
GoPA! (Global Observatory for Physical Activity) 77
Gould, George 624

governance
child health 436
public health strategies 676
government department reforms 651

governments
environmental health 649
healthcare services 693
public health strategies 660
GR (glucocorticoid receptor) 546
Grameen Foundation 756
Gram-negative pathogens 167
Gross Domestic Product (GDP) 139
gross national product (GNP) 647
group community initiatives, gender healthcare 426
Growing Old in America report 513–514
GSPS (Global School Personnel Survey) 325
GST (glutathione transferases) 153
GTS (Global Technology for Malaria 2016–2030) 240, 241
t, 244, 244
t
GTSS (Global Tobacco Surveillance System) 325
guanine 297
Guatemala, disabled people exclusion 494
Guideline for the Management of Blood Cholesterol (ACC/AHA) 13
Guideline on Cardiovascular Risk Assessment (ACC/AHA, 2014) 11
Guideline on Lifestyle Management (ACC/AHA) 13
The Guide to Community Preventive Services: Systematic Reviews and Evidence-Based Recommendations 368
Guiding Principles on Internal Displacement 530
Guillain–Barré syndrome 152
GVAP *see* Global Vaccine Action Plan (GVAP)
GWAS *see* genome-wide association studies (GWAS)
GYTS (Global Youth Tobacco Survey) 325

H1N1 influenza virus 270, 723
pandemic 2009, 270
h_2 (heritability) 299
H5N1 type A influenza virus 270, 618
bioterrorism 277, 294
emerging/re-emerging infections 276
H7N7 influenza outbreak (2003) 276
habitat modification, malaria prevention 237
Hackers for Humanity 733
Haddon matrix 366, 366
t
Haemophilus ducreyi infection (chancroid) 187
t
Haemophilus infections 51
Haemophilus influenzae infection 168
t

Haemophilus influenzae type b (Hib) vaccine 177, 178–179, 584
t, 589
child health 441
child vaccination routine 586
t
Haemophilus pneumoniae infection 177
HAIs (healthcare-associated infections) 608
HALE (health-adjusted life expectancy) 512
Handbook Series in Zoonoses (CRC) 611

handwashing
  acute respiratory infections 177
  gastrointestinal tract infections 174
  host infection control measures 606–607
Hantavirus pulmonary syndrome 617
harmonization, private support of public health 756

harm reduction
alcohol consumption 351
alcohol consumption prevention/control 353–354
medical screening 630–631

hazards
environmental health assessment 646
humanitarian emergencies 724
HbA1c (glycated haemoglobin) 6, 86
HBV *see* hepatitis B virus (HBV)
HCC *see* hepatocellular carcinoma (HCC)
HCV (hepatitis C virus) 255
HDT (host-directed therapies), tuberculosis 221
HDV (hepatitis D virus) 253–254
HDWs *see* healthcare workers (HCWs)
headache 143–145
chronic daily headache 145
cluster headaches 144
cost & public health impact 145
familial risk 144–145
genetic risk 144–145
prevalence 143, 144
risk factors 145
tension-type headache 144 *see also* migraine

health
differences 564–565
equity 565–566
gender effects 416
prisoners 546–547
research in disabled people 500
health-adjusted life expectancy (HALE) 512
Health at Work 671
Health Behaviour in School-aged Children 363

healthcare access
adolescents 459–460
disabled people 497–498
HICs *vs.* LMICs in diabetes mellitus 89
immunization programme challenges 583
internally-displaced persons 534
musculoskeletal disorders 140
older people 515
healthcare-associated infections (HAIs) 608
Health Care Financing Administration 688
Health Care in Danger (ICRC) 730
Health Care of Asylum Seekers and Refugees (Australian Medical Association) 538
healthcare services 685–701
adolescent health 459
attacks on 730–731
bias in indigenous people 483
building blocks in child health 445–446
burden of disease 687–688
child health 445
context for 685–689
cost sharing 693
demography 687
dental health improvement 122–123
developing countries 704, 704 *f*
distribution & need 566
district level structure 705 *f*
ecosystem approach 691
efficiency 695–697
efficiency of delivery 695–697
enablers 697–699
future delivery challenges 709–710
integrated *see* integrated healthcare
international differences 685
macroeconomics 686–687
needs assessment & priority setting 693–695, 694 *f*
norms & beliefs 686
older people *see* older people healthcare
patient engagement & empowerment 691, 693
patient expectations 689
public health strategies 675, 677
quality & waste reduction 688–689
state & market 689–690
strategies in 693–699
system re-organization 690–691
technological changes 688
workforce 687
healthcare workers (HCWs) 687, 698

career paths 707
child health 446
education in developing countries 706
emerging/re-emerging infections 276
forced migrants 538
hepatitis C transmission 256
new future skills 713, 713
f
responsibilities of 704–706
roles *vs.* responsibilities 703–704
self-regulation 690
shared regulation 690
vaccination programmes 595
healthcare workers training 703–708
barrier overcoming 706–707
problems with 706
technologies 707 *see also* transformative learning
Health Emergency Programme (WHO) 761
Health-For-All strategy (WHO) 641, 685
Healthier Hospitals Initiative 671
Health Impact Assessment (HIA)
environmental health impact 648
non-communicable disease 579
Health in All Policies (HiAP) 561, 648
case study 665
mental health 100

health inequality
  definition 565
  Health Needs Assessment 564–566
health inequity 565
Health Information Systems (HIS) 706
  child health 445
  disabled people 499, 501–502
health insurance premiums 693
Health Needs Assessment 559–572
  asset building 569
  capability building 569
  economics 562
  epidemiological approach 560–561
  equity–efficiency trade-off 563
  health differences 564–565
  health economics 562–563
  health systems 563
  inequality to equity 565–566
  maximising outcomes 568–569
  methods 559
  philosophy 563–564
  politics 563–564
  preventative approach 569
  techniques 561
Health of London Project 537
health professional organizations 537
Health Promoting Schools initiative (WHO) 447
health promotion 663
  disabled people 496–497
  non-communicable disease prevention 574
healthy cities 649–650
Healthy Hospital Choices strategy (CDC) 671
Healthy Latin America Coalition (CLAS) 579
healthy life expectancy, older people 512–513
Healthy Lives, Healthy People: Improving Outcomes and Supporting Transparency (UK) 660, 672 *b*
hearing impairment, older people 515
heart failure 4
  indigenous peoples' health 480
  risk factors 10
heated *tobacco* products 765
HEEG (High-Level Commission on Health Employment and Economic Growth) 712–713

hegemonic masculinity
HIV infection 420
interpersonal violence 385
mental health and 423 *see also* masculinities
Helicobacter pylori infection
cancer aetiology 19
stomach cancer 23
helminth infections 181
hepatitis, forced migrants 532
hepatitis A 188
t, 617
artificially acquired passive immunization 604
child vaccination routine 586
t
sexually transmitted disease as 191
t
vaccination 584
t, 593
hepatitis B 188
t, 249–255
cancer aetiology 19
clinical *manifestations* 251
endemicity 250
f
epidemiology 249–250
extrahepatic manifestations 251
liver cancer 22
misinformation 277
perinatal infections 250
prevention 252–253
public health impact 253
sexually transmitted disease as 190–191, 191
t
transmission 249–250
treatment 251–252 *see also* hepatocellular carcinoma (HCC)
hepatitis B vaccination 252–253, 584
t, 587
child vaccination routine 586
t
liver cancer prevention 22–23
hepatitis B virus (HBV) 249
genetic heterogeneity 251
hepatitis C 188
t, 255–257
cancer aetiology 19
clinical manifestations 256
epidemiology 255–256, 255
f
extrahepatic manifestations 256
liver cancer 22

people who inject drugs 341
prevention 257
prisoners 550
public health impact 257
sexually transmitted disease as 190–191, 191
t
transmission 255–256
treatment 256–257 *see also* hepatocellular carcinoma (HCC)
hepatitis C virus (HCV) 255
hepatitis D 253–255
epidemiology 254, 254
f
prevention 254–255
sexually transmitted disease as 191
t
transmission 254
treatment 254–255
hepatitis D virus (HDV) 253–254
hepatitis E 191
t
hepatocellular carcinoma (HCC) 22, 259
mortality 192
non-alcoholic fatty liver disease and 262
herbicides, Parkinson's disease 153
herd immunity 603
hereditary conditions, colorectal cancer 24
heritability (h_2) 299
heroin 338
herpes simplex 1 infection
mother-to-child transmission 191
sexually transmitted disease as 190
herpes simplex 2 infection
HIV infection prevention 205
mother-to-child transmission 191
prevalence 190
f
sexually transmitted disease as 190
uterine cervical cancer 31
herpes simplex virus (HSV) infection 188
t
herpes zoster 165
t
HHV-8 (human herpesvirus 8) 188
t
HIA *see* Health Impact Assessment (HIA)
HiAP *see* Health in All Policies (HiAP)
HIB integrase strand transfer inhibitors 202
t
HICPAC (Hospital Infection Control Practices Advisory Committee) 610
HICs *see* high-income countries (HICs)

high-income countries (HICs)
acute care injury prevention 371
adolescent health education 460
colorectal cancer 23–24
COPD prevalence 45
diabetes mellitus care 89
drowning prevention strategies 369
hepatitis C transmission 256
indigenous peoples' health 480
LMICs *vs.* in diabetes mellitus 88–90
mental health costs 96
musculoskeletal disorders 135
obesity 64
physical activity trends 76
polypharmacy 514
road traffic incidents 368 *see also* developed countries
High-Level Commission on Health Employment and Economic Growth (HEEG) 712–713
high-quality home services, mental health 106
Hindu society, castes & positions 474–475

Hippocrates
environmental health 639
malaria 227
Hiroshima atomic bombing (1945) 395
HIS *see* Health Information Systems (HIS)
histidine-rich protein-2 (HRP-2) 233
historical trauma, indigenous people 482–483
HIV antibody test 198–200
HIV infection 188
t
adolescent health 456
advocacy & group community initiatives 426
ageing process and 203
biological HIV vulnerability 419, 419
f
cancer 192
child health 443–444
COPD risk factors 47
drug resistance 202
epilepsy 148
forced migrant mortality 532
gender 417, 419–420, 419
f, 420, 420
b, 424
geographical *distribution* 199–200
global burden 200
f, 598
historical aspects 198, 199
f
injecting drug use 206–207
maternal health 423–424
medical male *circumcision* 193, 605
mortality 165
t
mother-to-child transmission prevention 203–204
peripheral neuropathy 152
prevalence 165
prisoners 550
rape from 379
reduction in 759
sexually transmitted disease and 192
sexually transmitted disease as 190
sexual transmission prevention 204–206
substance use disorders 340–341
treatment 201–202
t, 201–203, 420
tuberculosis 213
vaccination programmes 595
viral structure 200, 200
f

HIVNET 012 perinatal HIV prevention trial 204
HLA complex, multiple sclerosis 154
HLR (human landing rate), malaria 242
HMG-CoA reductase inhibitor drugs (statins) 9
home-based treatments, humanitarian emergencies 732

home environment
child mortality & morbidity 434
under-fives child health 433
home mental health services 106
homosexuality *see* LGBT+ people

Hong Kong
alcohol consumption control 668–669, 669
b
College of Community Medicine 677
Council on Smoking and Health 668
severe acute respiratory infection case study 661
horizontal public health strategies 662
horizontal transmission 599
hormones, asthma risk factor 49
Hospital Infection Control Practices Advisory Committee (HICPAC) 610
Hospital Preparedness Program (HPP) 289

hospitals
infection control 608–609
public health strategies 671
reorientation to primary healthcare 696
host-agent-environment model, war prevention 399, 399
f
host defense mechanisms, infectious diseases 601
host-directed therapies (HDT), tuberculosis 221
host infection control measures 600–601, 602–607
barriers 607
behavioural changes 605–607
chemoprophylaxis 604–605
eating behaviour 606
international travel 607
intravenous drug use 606
passive immunization 603–604
resistance improvements 607
reverse isolation 607
screening programmes 605
surgical intervention 605
vaccination *see* vaccination
working behaviour 606

households
environmental health deprivation 642
family composition 404
food insecurity 434
psychological toll 422
surveys on substance use disorders 342
House Subcommittee on Emergency Preparedness, Response, and Communications (HSEPRC) 288

housing
design & infectious diseases 614
disabled people 494
Howard Hughes Medical Institute 753, 754
Howard, John 543
HPA (hypothalamic–pituitary–adrenocortical) axis 546
HPP (Hospital Preparedness Program) 289
HPTN 052 trial 205–206
HPV *see* human papilloma virus (HPV) infection
HRDAG (Human Rights Data Analysis Group) 723
HRP-2 (histidine-rich protein-2) 233
HSEPRC (House Subcommittee on Emergency Preparedness, Response, and Communications) 288
HTLV-1 (human T-lymphotrophic virus type I) 19, 188
t
human behaviour, emerging/re-emerging infections 276–277, 617
Human Development Index 723
human dignity 569
human genome 298
Human Genome Project 302
b
human herpesvirus 8 (HHV-8) 188
t
humanitarian accountability 731
Humanitarian Accountability Partnership 538
Humanitarian Code of Conduct 538
humanitarian emergencies 719–737
child health 446–447
classification 721–722
definitions 721–722
emergency risk management *see* emergency risk management
epidemiology 722–724, 722
t
mortality 719
f, 723, 724
f
new developments 729
non-traditional participants 731–734
political trends 729–730
preparation & planning 724
public health needs 727–729
response coordination 726–727
spending on 731–732, 732
f
human landing rate (HLR), malaria 242
human papillomavirus (HPV) infection 188
t
cancer 19, 192
laryngeal cancer 29
lower lip cancer 116
oral cavity cancer 28–29

prostate cancer 30
sexually transmitted disease as 190
uterine cervical cancer 31
vaccination 193
human papillomavirus (HPV) vaccination 586
t, 591
oral cavity cancer/pharyngeal cancer prevention 29
human rights, mental health 97
Human Rights Data Analysis Group (HRDAG) 723
human rights violations, war 394
human T-lymphotrophic virus type I (HTLV-I) 19, 188
t
human T-lymphotrophic virus type-III (HTLV-III) *see* HIV infection
hygiene hypothesis 51
hyperglycaemia 85

hypertension
cardiovascular disease 7–8
definition 633
heart failure 10
overdiagnosis in screening 634
stroke 10
testing for 632–633
treatment 633
hypopharyngeal cancer 28–29
hypothalamic–pituitary–adrenocortical (HPA) axis 546

IADLs (instrumental activities of daily living) 514
IARC *see* International Agency for Research on Cancer (IARC)
IASC (Inter-Agency Standing Committee) 731, 762
IASFM (International Association for the Study of Forced Migration) 525
IAVI (International AIDS Vaccine Initiative) 753, 766
ICC (International Criminal Court) 720
iCCM (integrated community case management) 445
ICD *see* *International Classification of Diseases* (ICD)
ICHD (International Classification of Headache Disorders) 143, 144
b
ICRC *see* International Committee of the Red Cross (ICRC)
ICT *see* information and communication technology (ICT)
ideological power, non-communicable diseases 578
IDF (International Diabetes Federation) 86
IDMC (Internal Displacement Monitoring Centre) 534
IDPs *see* internally-displaced persons (IDPs)
IFRC (International Federation of the Red Cross and Red Crescent Societies) 536
IGRA (interferon-gamma release assays) 216, 218
IHR *see* International Health Regulations (IHR)
IIPFCC (International Indigenous Peoples' Forum on Climate Change) 484
ILO *see* International Labour Organization (ILO)
images, older people 513
IMAGE study (South Africa) 387
IMCI (integrated management of childhood infections) 177, 445
IMCO (integrated management of childhood infections) 177, 445
IMF (International Monetary Fund) 731
Immigration Removal Centres 559
immunizations *see* vaccination
immunocompromised individuals, Bacillus Calmette–Guérin 585–586
immunosuppression, emerging infectious diseases 617
impact assessment training 648–649
impaired fasting glucose 86
implementation level, emergency risk management 726
IMPROVE-IT trial 9
Improving the Health of the Public by 2040 679
inactivated complex antigen vaccines 602
inactivated toxin vaccines 602

inactivated whole-cell vaccines 221
inappropriate care, healthcare service expenditure 688–689

incentives
healthcare services 699
b
healthcare workers 707
Incident Management Process (WHO) 740
incontinence 516–519

India
health inequities case study 474–475
intersectional health research 474
indicators, environmental health 646–647
indices, *multi-risk* assessment methods 743–744
indigenous peoples' health 479–487
cancer 480–484
data inclusion 479
definitions 479
equity interventions 482–484
future inequity risks 484
gestational diabetes 481
healthcare access 481
healthcare *obstacles* 480
indigenous workforce 483–484
infections 480
injuries 481
mental health 481
non-communicable diseases 480–481
nutritional disorders 481
older people 510
present state 479–482
prisoners as 545
representation & framing 482
respiratory conditions 481
state of health 479–482
type 2 diabetes 481
indirect advertising bans, tobacco smoking 326

indirect costs
obesity 62–63, 63
t
oral disease 117
indirect transmission, infectious diseases 600

individual level
  infectious diseases control measures 602
  injury care 366–367
individual needs, Health Needs Assessment 564
indoor air pollution 50
indoor biomass exposures 47
indoor release, bioterrorism 290–291
indoor residual spraying, malaria prevention 237, 443

industry
non-communicable disease public health 579
non-communicable diseases 576–578
inequality, forced migration 523–524
infected human control measures 608–611
activity restriction 611
behavioural change 611
chemotherapy 609
hospital infection control 608–609
isolation 609–607
quarantine 610
surgical interventions 609
infectious disease agents 598–599
control measures 601
infectious diseases 163–183
acute respiratory infections *see* acute respiratory infections (ARIs)
aetiology 167, 168
t, 169
t
age-of-onset 168
t
antimicrobial resistance 182–183
asthma 50–51
burden of 163–166
cancer 19, 20
t, 599
chain of infection 598–601
children 166
chronic diseases *vs.* 623
definitions 597–598
Dengue fever 180–181
emerging/re-emerging infectious diseases *see* emerging/re-emerging infectious diseases
epilepsy 148
gastrointestinal tract infections *see* gastrointestinal tract infections
global *burden* 598
global epidemiology 167–182
historical aspects 163
host responses to 601
indigenous peoples' health 480
medical treatment 167–168, 170
t
meningitis *see* meningitis
mortality 165
t
parasitic diseases *see* parasitic infections
peripheral neuropathy 152
preventative measures 168–169
regional epidemiology 167–182
school-age children mortality 434
specific categories 166–167

spread by forced migration 723–724
typhoid fever 179–180
infectious diseases control 597–621
agent measures 615
animals 611–612
control & prevention programmes 615–616
definitions 597–598, 601
emerging infectious diseases 617–618
environment 612–614
eradication 616–617
hosts *see* host infection control measures
tools for 601–602
vaccination 584
t
vectors *see* vector control *see also* infected human control measures
infectiousness 599
infective dose 599
infectivity 599

inflammation
  carcinogenesis 17
  cardiovascular disease risk assessment 11–12
inflammatory diarrhoea 172

influenza
global burden 598
transmission 270
vaccination 13–14
informatics, healthcare services 699
b
information and communication technology (ICT)
adolescent health 461
transformational learning 715–716
information management, healthcare services 697–698
informed consent 310
b
infrastructure damage, war 394
infrastructure design, healthcare services 696
inheritance patterns, genetics 298
inherited conditions, carcinogenesis 17
injecting drug use *see* people who inject drugs (PWID)
injuries 361–375
causal model 361–362
child health 447–448
data sources 362–363
drowning 362, 364
economic & social burden 365
epidemiology 362, 362
f
falls 364
fire burns 364
global burden of 363–364, 363
t
indigenous peoples' health 481
interpersonal *violence see* interpersonal violence
prevention *see* injury prevention
risk factors for 366–367
self-harm 364–365
worldwide burden of 361 *see also* road traffic incidents (RTIs); trauma
injury phase 365
injury prevention 368–369
capacity building 371–372
programme identification 367–368
successful strategies 367
theoretical basis 365–366
innovation, global health sustainability 765–766
inorganic acid exposure 29
inpatient management, child malnutrition 439
insecticides 607–608
insecticide-treated nets (ITNs)
infectious diseases control 614
long-lasting 443
malaria prevention 236–237, 443, 607
INSPIRE 388

Institute of Economic *Alternatives* 577
Institute of Medicine (IOM)
Best Care at Lower Cost: The Path to Continuously Learning Health Care 688–689
Investing in a Healthier Future 650
institutional level, infectious diseases control 602
institutional reform, transformational learning 714, 714
t
instructional reform, transformational learning 715
instrumental activities of daily living (IADLs) 514
integrated community case management (iCCM) 445

integrated healthcare
child health 445
healthcare services 696–697
integrated management of childhood infections (IMCI) 177, 445
intellectual disabilities 492
intelligent purchasing 697–698
intentional injuries 361–362
intentional self-harm 145
Inter-Agency Committee on Tobacco (Philippines) 577
Inter-Agency Standing Committee (IASC) 731, 762
interferon-gamma release assays (IGRA) 216, 218
intermittent preventative treatment in infants (IPTi) 238
intermittent preventative treatment in pregnancy (IPTp) 237–238
Internal Displacement Monitoring Centre (IDMC) 534
internally-displaced persons (IDPs) 524, 525, 526
t
distribution of 529
f
healthcare access 534
legal frameworks 529–530
mental health 533–534
numbers of 528
International Agency for Research on Cancer (IARC) 19, 321–322
monograph 2002 322–323
tobacco smoking cessation 329–330
International AIDS Alliance strategy 426
International AIDS Conference (2000, Durban) 753
International AIDS Vaccine Initiative (IAVI) 753, 766
International Association for the Study of Forced Migration (IASFM) 525
International Centre for Migration and Health 536
International Classification for Functioning, Disability, and Health 490, 490
f
International Classification of Diseases (ICD)
alcohol consumption 350
musculoskeletal disorders 129
substance use disorders 340
International Classification of Functioning, Disability and Health 761
International Classification of Headache Disorders (ICHD) 143, 144
b

international commerce
emerging infectious diseases 617
infectious diseases control 616
International Committee of the Red Cross (ICRC) 762
forced migrant health 536
Health Care in Danger 730
International Criminal Court (ICC) 720
International Diabetes Federation (IDF) 86
International Federation of Environmental Health 653
International Federation of the Red Cross and Red Crescent Societies (IFRC) 536
International HapMap Project 302
b
International Health Commission 751–752
International Health Partnership 760
International Health Regulations (IHR) 280, 281
f, 616, 741, 761
adherence to norms 283–284
core capacity development 282
Emergency Committee 282
International Indigenous Peoples' Forum on Climate Change (IIPFCC) 484
International Labour Organization (ILO)
healthcare worker numbers 710
musculoskeletal disorders 138
International League Against Epilepsy (2017) 147
International Monetary Fund (IMF) 731
International Obesity Task Force (IOTF) 58
International Organization for Migration 536
Missing Migrants project 731
International Physical Activity and the Environment Network (IPEN) 79–80
International Physicians for the Prevention of Nuclear War 398
international policy, environmental health 650
international regulations, environmental health 651
international sanitary conferences 280
International Sanitary Council 280
International Study of Asthma and Allergies in Childhood (ISAAC) 48
International Tobacco Control (ITC) Policy 326

international travel
emerging infectious diseases 617
host infection control measures 607
internet, humanitarian emergencies 733
interpersonal relationships, injuries 367
interpersonal violence 365, 377–391
community level effects 384–386
cultural scaffolding 384–385
definition 377
early years' exposure 381–382
economic impact 377
gender 422
health consequences 379
magnitude of 377–379
masculinities 385–386
microsocial relationships 383–384
non-partner sexual violence 378
peer relationships 383–384
prevention of 386–388
propensity to use 383
risk factors 380–381
social & economic consequences 379
societal levels 386
understanding of 379–380, 380
f
witness effects 382 *see also* intimate partner violence (IPV); violence against women
interprofessional education (IPE) 712
transformational learning 715
interprofessional practice (IPP) 712

intersectionality
  case study 474–475
  definition 470–471, 471
    f
  epidemiology 469–471
  gender 416–417
  global health research 471–473, 472
    f
  processes 474
    t, 475
intersectional social epidemiology 469–477
  future work 475
intersectional social positions 472
intersectoral environmental health 649–650
intersectoral environmental health 641

interventions for public health
barriers to 679
health services 677
research & evaluation 677–678, 679
f
intervention targeting, healthcare services 695–696
intimate partner violence (IPV) 365, 377–378
global estimates 378
intravenous drug use *see* people who inject drugs (PWID)
invasive *diarrhoea* 169
Investing in a Healthier Future 650
Investing in Research, Improving Health (UK) 678
investments, adolescent health 462–464
Investor-state Dispute Settlement 577
IOM *see* Institute of Medicine (IOM)
ionizing radiation 22
IOTF (International Obesity Task Force) 58
IPE *see* interprofessional education (IPE)
IPEN (International Physical Activity and the Environment Network) 79–80
IPP (interprofessional practice) 712
IPT (isoniazid preventative therapy) 214
IPTi (intermittent preventative treatment in infants) 238
IPTp (intermittent preventative treatment in pregnancy) 237–238
IPV *see* intimate partner violence (IPV)
Iran, delayed marriage 406
irrigation, infectious diseases control 614
ISAAC (International Study of Asthma and Allergies in Childhood) 48
ischemic heart disease 473
isolation, infection control measures 609–610
isoniazid preventative therapy (IPT) 214
Israel, healthcare service provision 695
Italy, *Genuino Clandestino* 426
ITC (International Tobacco Control) Policy 326
ITNs *see* insecticide-treated nets (ITNs)

Japan, divorce 407
Japanese encephalitis (JEV)
emerging/re-emerging infections 275
vaccination 586
t, 591–592
Jenner, Edward 583
Johns *Hopkins* University, Bloomberg School of Public health 372
Joint Evaluation of Emergency Assistance (Rwanda) 537
joint external evaluation (JEE, WHO) 284
Joint Learning Initiative (2004) 716
Joint Strategic Needs Assessment (JSNA) 561
Justice Reinvestment 550

Kaposi’s sarcoma 197
Klebsiella pneumoniae infection 168
t

knowledge management systems
  forced migrant healthcare 539
  healthcare services 697–698, 699
    b
Korea, disabled health services 497
Kostov, Vladimir 291–292

labour division, gender 422
lactate dehydrogenase (LDH) 233
The Lancet, healthy cities 649–650
Lancet Commission on Adolescent Health and Well-Being (2016) 456–457
landmines 397
larval source management (LSM), malaria prevention 237
laryngeal cancer 29–30

Lassa fever
  emerging/re-emerging infections 275
  quarantine 610
Last Mile Health and Living Goods 703
late adolescence 455
late marriage 406–407
late onset sepsis (LOS) 167

Latin America
  delayed marriage & cohabitation 406
  refugees 528–529
Laveran, Antoine 227, 233
law enforcement, bioterrorism 293
LBP *see* low back pain (LBP)
LBW *see* low-birthweight (LBW) infants
LDH (lactate dehydrogenase) 233
LEaB (Life-Expectancy-at-Birth) 480

lead
monitoring 646
poisoning 642

leadership
public health strategies 675–676
transformational learning 714
LEADER trial 7
lead time, medical screening 626, 627
f
lead-time bias
cancer prevention 33
medical screening 625
f, 628–629, 629
f
Legionella infection 614
Legionella pneumophila infection 271

legislation
adolescent healthcare 459–460, 462
firearms 386
interpersonal violence 387
LGBT rights 462
road traffic incidents 369
substance use disorders 339
tobacco smoking prevention 578
transformational learning 714–715
leishmaniasis 165
t
leisure 77
b
length time bias 629, 629
f
leprosy 609
leptospirosis 276, 279
lesbians *see* LGBT+ people
Less Salt, More Life campaign (Argentina) 672
b
lethal mean restriction, suicide prevention 104
leukotriene receptor antagonists 52
levamisole 182
Lewy body disease 148, 151
LGBT+ people
legislation 462, 463
older people 510–511
life-course approach
non-communicable diseases 674
f
public health strategies 672, 673
f
life course epidemiology, gender 415
life expectancy 512
environmental health 641
gender 418–419, 418
t
geographical differences 418
older people 512
t
population ageing 509
variations in 512–513
Life-Expectancy-at-Birth (LEaB) 480

lifestyle
breast cancer 25
cancer primary prevention 31–32
non-alcoholic fatty liver disease/non-alcoholic steatohepatitis 262
obesity 64
osteoarthritis 133
limited human–human transmission, emerging/re-emerging infections 269–270
line probe assays (LPAs), tuberculosis 218
linkage analysis 301
linkage studies 300, 301
lipid control 13
lipoprotein (a) 306–307
live attenuated vaccines 602
liver biopsy 261
liver cancer 22–23
prevention 22–23
liver disease 249–268 *see also* alcoholic liver disease; hepatitis B; hepatitis C; hepatitis D; non-alcoholic fatty liver disease (NAFLD)
liver fibrosis 262
Living well for Longer (UK) 660
LLINs (long-lasting *insecticidal* bed nets) 443
LMICs *see* low- and middle-income countries (LMICs)
local government, health funding 756
local-level actions
Health Needs Assessment 560
non-communicable diseases 578–579
public health strategies 667–674, 669 *t*, 704
long-lasting *insecticidal* bed nets (LLINs) 443
long-term institutional care, older people 517–518
long-term sustainability, health funding 756
LOS (late onset sepsis) 167
low- and middle-income countries (LMICs)
acute care injury prevention 371
adolescent health 460, 461, 463–464
burns prevention strategies 370
childbirth 168
child health programmes 445
child mortality 164 *f*
COPD prevalence 45–46
dental services 123
diabetes mellitus care 89
disabled people rehabilitation 498
drowning prevention strategies 370
economics 577
gender inequality 417
genetic epidemiology 311
healthcare workers training 703
HICs *vs.* in diabetes mellitus 88–90

immunization programmes 583
infectious diseases 163, 164
f, 167, 169
t
injury data 363
maternal health 423
non-communicable diseases 420–421, 573, 577
obesity trends 61
parasitic diseases 181–182
physician tobacco smoking 331
population ageing 507
poverty in disabled people 493–494
road traffic incidents 365, 368
school-age children health 434
tobacco smoking controls 327, 331–332
typhoid fever 179
typhoid vaccination 592
violence against disabled people 494–495 *see also* developing countries
low back pain (LBP) 130–131
treatment 139–140
low-birthweight (LBW) infants 438
neonatal sepsis 166–167
low-density lipoprotein (LDL)-cholesterol 6, 8–9, 8
t
lower lip cancer 116
lower respiratory infections 165
t
low-risk drinking guidelines 350
LPAs (line probe assays), tuberculosis 218
LSM (larval source management), malaria prevention 237
lung cancer 21–22
prevention of 22
lung function, asthma assessment 44–45
Lyme disease 617
lymphadenopathy-associated virus (LAV) *see* HIV infection
Lynch syndrome 24

macroeconomics 686–687
magnetic resonance imaging (MRI) 260
major depressive disorder (MDD)
mortality 95
substance use disorders 339
malaria 227–247
airport malaria 279
cerebral malaria 231
child health 443
clinical presentation 231–232
complicated malaria 231–232

diagnosis 232–233
elimination of 242–243, 242
t
entomological surveillance indicators 242
epidemiology 232
t
eradication 243
forced migrant mortality 532
frequency measures 240–241
gametocyte carriage 242
global burden 598
global distribution 228
f
global programmes 243–244
historical aspects 227
human parasites 228–229
incidence 240–241
less-used terminologies 243
mortality 165
t, 721
f
parasite lifecycle 229–230, 230
f
pathogenesis 231
pregnancy 232, 236
t
prevalence 165, 241
severe malaria 231–232
signs & symptoms 231–232
surveillance 239–240
T3 (Test-Track-Trace) 233–234
travel medicine 238–239
uncomplicated malaria 231
vector life cycle 230–231
Malaria Eradication Research Agenda (malERA) 243
Malaria Policy Advisory Committee (MPAC) 243–244
malaria treatment 234–237
antimalarial drugs 234–235
chemoprophylaxis 237–239, 604
drug-resistance 278
pregnancy 236
prevention 236–237
rapid diagnostic tests for malaria (mRDT) 233
vaccines 239
Malawi, disabled health service access 497
male-on-male violence 421
malignant tumours 17 *see also* cancer

malnutrition
adolescent health 455
child health 439–440
chronic malnutrition 440
disabled people 494
forced migrant mortality 533
gastrointestinal tract infections 174
host infection control measures 607 *see also* diet; nutrition
MAM (moderate acute malnutrition) 439
mammography 26
management reform, transformational learning 714
man-made disasters 524
Mannheim, Karl 463
MAOA see monoamine oxidase A gene *(MAOA)*
Marburg virus 276
geographic distribution 271

marketing
  adolescent health 461
  healthcare services 689–690
  neoliberalism 575–576
Markov, Georg 289
marriage, delayed 405–406

masculinities
high-risk sexual behaviour 424
interpersonal violence 384, 385–386
social constructions of 420 *see also* hegemonic masculinity
masks, host infection control measures 607
mass drug administration (MDA) 238
master of public health (MPH) 705
Master Settlement Agreement 578
MAT *see* medication-assisted treatment (MAT)
maternal care, child health 436, 437–438
t, 438
maternal deaths, indigenous peoples 480
maternal health 423–424, 424
b
health system 423–424
interventions 424
malnutrition 440
men and 424
nutrition 434 *see also* gender
maternal mortality rate (MMR) 423, 759
maternal vaccination 595, 603–604
pertussis vaccination 589
matrices, *multi-risk* assessment methods 743
MCI (mild cognitive impairment) 149
MDA (mass drug administration) 238
MDD *see* major depressive disorder (MDD)
MDGs *see* Millennium Development Goals (MDGs)

measles
child vaccination routine 586
t
forced migrant mortality 531–532
misinformation 277
mortality 165
t, 721
f
quarantine 610
vaccination 584
t, 585, 590–591
measles, mumps & rubella (MMR) vaccine 594
misinformation 277
meat contamination 613
mebendazole 182
Mectizan Donation Program 753
Médecins San Frontières (MSF) 762
humanitarian emergencies 720
Medical Care Under Fire 730

media
adolescent health 461
alcohol consumption prevention/control 353
suicide depiction 104–105
Medical Care Under Fire 730
medical conditions, pancreatic cancer 28
medical degrees, healthcare workers 705
Medical Education Partnership Initiative (MEP) 717
medical male circumcision 193
HIV infection prevention 205, 605
medical screening 623–637
benefits 630, 631
cancer *see* cancer screening
costs 630–631
definition 623–624, 623
f
disease status classification 626
t
early detection & overdiagnosis 634–636
effectiveness of 625–626
harms 630–631
historical aspects 624–625
illustrations of 631–633
multiphasic screening 633–634
multiphasic targets 624
performance of 626–628
predictive values 627–628, 627
f
proof of 628–630
randomized controlled trials 630, 630
f
sensitivity & specificity 627
sensitivity *vs.* specificity 628
systematic process as 624
tests 626
medication-assisted treatment (MAT)
HIV infection prevention 207
infectious diseases 167–168, 170
t
medication-related diabetes 85
Meigen, J. W. 230
Melbourne Charter (2008) 100
Member States to establish Emergency operation Centres (WHO) 762
men, maternal health and 424
MENA (Middle East and Northern Africa), mortality 719–720, 720
f
Mendelian randomization studies 306–307, 306
f
meningitis 177–179

aetiology 177, 178
t
medical treatment 178
mortality 165
t
paediatric signs & symptoms 178
presentation & diagnosis 178
prevention 178–179
typhoid fever and 179
Meningitis Vaccine Project 766

meningococcal disease
child vaccination routine 586
t
forced migrant mortality 532
vaccination 593
vaccine 179
mental health 95–112
adolescent health 456, 459
advocacy & group community initiatives 426
alcohol consumption 350
determinants of 98–100
disorder prevention 101–104
forced migrants 533–536
gender 417, 422–423, 423
b
global burden of 422–423
health interventions 424
indigenous peoples' health 481
interpersonal violence 379
large-scale epidemiological studies 422–423
magnitude & costs 95–97
physical health and 423
policies 97–98
prisoners 546–547
stigma 97, 98
well-being promotion 100–101 *see also* suicide
mental health services 105–107, 106
t
social services and 105–106
Mental Health Taskforce 667
mental, neurological and substance (MNS) use, gender 422–423
Mentor Violence and Injury Prevention (MENTOR-VP) 371
men-who-have-sex-with-men (MSM)
hepatitis B 250
HIV infection & substance use disorders 341
MEP (Medical Education Partnership Initiative) 717
merozoites 229
MERS (Middle East Respiratory Syndrome) 270, 618
MESA (Multiethnic Study of Atherosclerosis) 12

metabolic syndrome
  Alzheimer's disease 150
  cardiovascular disease 6–7
  non-alcoholic fatty liver disease and 261, 262
methadone 343
methadone maintenance therapy (MMT) 343
  HIV infection prevention 207
methamphetamine 338
methicillin-resistant *Staphylococcus aureus* (MRSA) 277, 609
methylenedioxymethamphetamine (ecstasy) 338
Mezirow, Jack 713
mHealth 206
"M Health" 707
MI (motivational interviewing) 344
MI (myocardial infarction) 3–4
microbial aerosols 600
microbiological studies, pneumonia 176
microbiome 51
  analysis 206

micronutrient deficiencies
child health 435, 440, 440
t
forced migrant mortality 533
microsatellite instability (MSI) 24
microscopy, tuberculosis diagnosis 217–218
microvascular disease, diabetes mellitus 152
mid-adolescence 454–455
Mid day meal Program (2009) 670
Middle East, LGBT+ legislation 463
Middle East and Northern Africa (MENA), mortality 719–720, 720
f
Middle East Respiratory Syndrome (MERS) 270, 618

migraine
  prevalence 144
  stroke 145
migraine with aura 143–144
migraine without aura 143

migration
family size 404
forced *see* forced migration
humanitarian emergencies 722
infectious diseases control 616
population ageing 509
vaccination programmes 595
war 394
mild cognitive impairment (MCI) 149
milk sanitation 614
Millennium Development Goals (MDGs)
child health 448
disabled people 501
environmental health 641
forced migrants 538
gender 415
global health sustainability 759–760
healthcare workforce 687
infectious diseases 163
introduction of 765–766 *see also* Sustainable Development Goals (SDGs)
Milliken Institute obesity report 63
mindfulness 100
minimum age limits, alcohol consumption 354

misinformation
emerging/re-emerging infections 276–277
immunization programme challenges 583
vaccines about 277
Missing Migrants project (International Organization for Migration) 731
mistrust, emerging/re-emerging infections 276–277
MMR *see* measles, mumps & rubella (MMR) vaccine
MMR (maternal mortality rate) 423, 759
MMT *see* methadone maintenance therapy (MMT)
MNS (mental, neurological and substance) use, gender 422–423
mobile phones 732–733
MoCA (*Montreal* Cognitive Assessment) 548
modelling, intersectionality 473–474
Model List of Essential Medicines (WHO) 761
Model State Emergency Health Powers Act (2001) 287
moderate acute malnutrition (MAM) 439
modified quarantine 610
molluscicides 608
Molluscum contagiosum virus infection 188
t
Momconnect (South Africa) 424
monetary aid, humanitarian emergencies 733

monitoring systems
  anti-microbial resistance 278–279
  environmental health 646
monkeypox virus *infection* 269
monoamine oxidase A gene *(MAOA)*
  aggression 381
  interpersonal violence 383–384
monogamy, HIV infection prevention 204–205
monogenic diabetes syndrome 85
monogenic disease 300–301
mononeuropathies 151
monozygotic (MZ) twins 299
Montevideo Roadmap 2018–2030on NCDs as it's Sustainable Development Policy 665
Montreal Cognitive Assessment (MoCA) 548

morbidity
acute respiratory infections 175
Health Needs Assessment 563
non-communicable diseases 573
under-fives 432–433, 433
t

mortality
  acute respiratory infections 175
  adolescent health 457
  forced migrants 531, 536
  population ageing 509
  under-fives 431–432, 432
    f
mother-to-child transmission (MTCT)
  childhood HIV/AIDS 443
  sexually transmitted diseases 191
motivation, humanitarian emergencies 729
motivational interviewing (MI) 344
Motsepe Foundation 752
mousepox virus 294
MPAC (Malaria Policy Advisory Committee) 243–244
MPH (master of public health) 705
MPOWER package (WHO) 319, 668
  tobacco smoking cessation 330
mRDTs (rapid diagnostic tests for malaria) 233
MRSA (methicillin-resistant *Staphylococcus aureus*) 277, 609
MSA (multiple system atrophy) 152
MSDs *see* musculoskeletal disorders (MSDs)
MSF *see* Médecins San Frontières (MSF)
MSI (microsatellite instability) 24
MSM *see* men-who-have-sex-with-men (MSM)
MTBDR*sl* test 218
MTCT *see* mother-to-child transmission (MTCT)
mucous membranes, infectious disease 600
multi-agency approaches, prisoner treatment 551
multiclass combination products, HIV infection treatment 202
  t
multidrug-resistant (MDR) *Salmonella typhi* 179
multidrug-resistant (MDR) tuberculosis 752
Multiethnic Study of Atherosclerosis (MESA) 12
multifactorial disease 300–301
multifactorial exposure, environmental health 644
multihazard risk assessment 742
multimorbidity, older people 514
multiphasic screening 624, 633–634
multiple sclerosis 154–155
multiple sex partners, hepatitis B 250
multiple system atrophy (MSA) 152
multiple testing, genome-wide association studies 302
multi-risk assessment methods 743–744

multisectoral approach
  dental services 123
  humanitarian emergencies 724
mumps vaccination 594
  child routine 586
    t
  disease reduction 584
    t
Munich Security Conference (2017) 288
musculoskeletal disorders (MSDs) 129–142
  case definitions 130
  classification 129–130
  costs of 134–135
  definition 129–130
  determinants of 137, 137
    f
  future directions 138–140
  geographical distribution 135
  increase in 135, 136
    t, 137
  low back pain 130–131
  natural history 130
  neck pain 131–132
  osteoarthritis 132–133
  osteoporosis 133
  pathophysiology 130
  point prevalence 135
    t
  practice efforts 140
  prevalence 134
  rheumatoid arthritis 133–134
  risk factors 130
  treatment 130
mutations, emerging infectious diseases 617
MVA85A 222
Mycoplasma genitalium infection 187
  t
myocardial infarction (MI) 3–4
MZ (monozygotic) twins 299

NAAT (nucleic acid amplification tests) 217–218
NAFLD *see* non-alcoholic fatty liver disease (NAFLD)
Nagasaki atomic bombing (1945) 395
naltrexone 343
Namibia, disabled health service access 497
nanotechnology, tuberculosis diagnosis 219
NASH *see* non-alcoholic steatohepatitis (NASH)
National Academy of Science, alcohol consumption 351

National Biodefense Analysis and Countermeasures Center (NBACC) 289
National Bioforensic Analysis Center (NBFAC) 289
National Biological Threat Characterization Center 289
National Biosurveillance Integration Center (NBIC) 289
National Bioterrorism Hospital Preparedness Program (NBHPP) 289
National Cancer Institute (NCI) 578
National Center for Biomedical Research and Training (NCBRT) 289
National Center for Injury Prevention and Control (USA) 368
  Smoke Alarm Installation and Fire Safety Education 370
National Centre for Sports and Exercise Medicine (NCSEM) 674
National Cholesterol Education Expert Panel on Detection, Education, ad Treatment of High Blood Cholesterol in Adults 261
National Cholesterol Education Program 6
National Disability Insurance Schemes (Australia) 500
National Disability Survey 491
National *Disaster* Medical System 288
National Domestic Preparedness Consortium (NDPC) 289
National Ebola Response Committee (NERC) 727
National Environmental Health Association (NEHA) 653
National Environmental Policy Act (NEPA 1969) 648
National Health Service (NHS) 686–687
National Healthy Schools Programme (NHSP) (England) 671
National Heart Lung and Blood Institute 7
National Highway Traffic Safety Administration 368
National HIV/AIDS Strategy 676
National Institute for Health and Clinical Excellence (NICE) 695
  Health Needs Assessment definition 559
National Institute for Health Research 678
National Institutes of Health (NIH)
  Fogarty International Center 371–372
  tuberculosis economics 223
National Lung Cancer Screening Trial 22
National Pre-Conception and Pre-Nata Diagnostics Techniques (PNDT, 2006) 474
national public health strategies 667–668
National Research Council (NRC) 645–646
National Risk Register of Civil Emergencies (2017, UK) 744
  b
National Sexual Health Strategy 675
National Steering Committee on Warning and Informing the Public 745
National Survey of Drug Use and Health (US) 342
National Tobacco Control Plan 675
National Trauma Data Bank (USA) 363
National Trauma Registry (Canada) 363
natural disasters 741
  emerging/re-emerging infections 275–276
  forced migration 524
  risk characterization 743
natural ecosystems, health effects and 640
naturally acquired passive immunization 603
NBACC (National Biodefense Analysis and Countermeasures Center) 289

NBFAC (National Bioforensic Analysis Center) 289
NBHPP (National Bioterrorism Hospital Preparedness Program) 289
NBIC (National Biosurveillance Integration Center) 289
NC-005 regimen 220–221
NCBRT (National Center for Biomedical Research and Training) 289
NCDs *see* non-communicable diseases (NCDs)
NCSEM (National Centre for Sports and Exercise Medicine) 674
NDPC (National Domestic Preparedness Consortium) 289
Necator americanus infection
  chemoprophylaxis 604
  child health 443
  diagnosis & treatment 182
    t
neck pain (NP) 131–132
needle and syringe programmes (NSPs) 206–207

needs
assessment in healthcare services 693–695, 694
f
prioritization in private support of public health 754–755
negative predictive value (NPV) 627–628
neglected tropical diseases (NTDs)
child health 442–443
mass drug administration 238 *see also* emerging/re-emerging infectious diseases
NEHA (National Environmental Health Association) 653
Neisseria gonorrhoeae infection 187
t, 604
morbidity 191–192
mother-to-child transmission 191
reduction in 194
Neisseria meningitidis infection
age-of-onset 168
t
emerging/re-emerging infections 275
meningitis 177 *see also* meningococcal disease
Nelson Mandela Rules (2015) 551
neoclassical concepts, healthcare service market 689–690
neoliberalism, non-communicable diseases 575–576
neonatal antiretrovirals 444
neonatal care, child health 436, 437–438
t

neonatal infections
evidence-based interventions 167
meningitis 170
t
mortality and infectious diseases 166
t
sepsis *see* neonatal sepsis
tetanus 585
neonatal sepsis 166–167, 167
t, 439
antimicrobial therapy 170
t
NEPA (National Environmental Policy Act) 648
NEPAD (Vision and Mission, Principles, Goals, and Objectives) 666
NERC (National Ebola Response Committee) 727
Netherlands, Dunning Committee 694–695
networks, transformational learning 716
neurological diseases 143–162
dementia *see* dementia
epilepsy *see* epilepsy
frontotemporal lobar degeneration 151
Lewy body disease 148, 151
mild cognitive impairment 149
multiple sclerosis 154–155
Parkinson’s disease *see* Parkinson’s disease (PD)
peripheral neuropathy 151–152
sexual violence 381
traumatic brain injury 145–146
vascular dementia 150–151 *see also* headache
nevirapine (NVP) 201
new psychological substances (NPSs) 339

New Zealand
healthcare service provision 695
indigenous people as prisoners 545
next-generation whole-genome sequencing (NGS) 219
NGOs *see* non-governmental organizations (NGOs)
NHS (National Health Service) 686–687
NHSP (National Healthy Schools Programme) 671
NICE *see* National Institute for Health and Clinical Excellence (NICE)
nicotine replacement therapies 330–331
9/11 attacks 725
t
Nipah virus 270, 275, 617
NNRTI (non-nucleoside reverse transcriptase inhibitors) 201, 201
t
No Health Without Mental Health: A Cross-Government Mental Health Outcomes Strategy for People of All Ages 667
non-alcoholic fatty liver disease (NAFLD) 259–262
definitions 259–260
diagnosis 260
fibrosis score 260
incidence 260–261
metabolic syndrome and 261
mortality 261
natural history 261–262
outcome 261–262
prevalence 260–261
prevention 262
treatment 262
non-alcoholic steatohepatitis (NASH) 259–262
definitions 259–260
diagnosis 260
prevention 262
treatment 262
non-ambulant casualties, decontamination 746
non-camp settlements, refugees 530–531
Non-Communicable Disease Monitoring Framework (WHO) 329
non-communicable diseases (NCDs) 573–582
adolescent health 455, 456, 457
child health 447–448
comorbidities 576
definition 573, 664
determinants & risk factors 764
economic losses 573
environmental health 643
environments of risk 575–578
gender 420–417, 421
b
global political economy 575–578
indigenous peoples' health 480–481
industry power 576–578

life-course view 674
f
morbidity 573
older people 511, 511
t
policy action on prevention 574–575
political economy 576–578
premature deaths 573
preventability 573–574
public health policy 578–580
public health strategies 664, 664
b
school-age children health 434
Sustainable Development Goals 574
under-fives child health 431 *see also* alcohol consumption; food industry; tobacco smoking
non-convention refugees 529–530
non-decision-making power, non-communicable diseases 577
non-governmental organizations (NGOs)
environmental health 649
humanitarian emergencies 729
war prevention 398
non human–human transmission, emerging/re-emerging infections 269
non-nucleoside reverse transcriptase inhibitors (NNRTI) 201, 201
t
non-traditional participants, humanitarian emergencies 731–734
non-traumatic lower limb amputation 86
North–South relationships, forced migration 524
Norway, *Treatment and identification of Psychosis Study* 103–104
(#)NotATarget 762
NOURISHING framework 66–67
Now More Than Ever (WHO) 677
NP (neck pain) 131–132
NPSI (new psychological substances) 339
NPV (negative predictive value) 627–628
NRC (National Research Council) 645–647
NRTIs (nucleoside reverse transcriptase inhibitors) 201, 201
t
NSPs (needle and syringe programmes) 206–207
NTDs *see* neglected tropical diseases (NTDs)
Nuclear Non-Proliferation Treaty (NPT) 396
nuclear weapons, war 395–396
nucleic acid amplification tests (NAAT) 217–218

nucleoside analogues
  hepatitis B treatment 251
  HIV infection treatment 201
nucleoside reverse transcriptase inhibitors (NRTIs) 201, 201
  t
nucleotide analogues 251
Nuffield Provincial *Hospital* Trust 624
nurses, training & qualifications 706
Nursing Education Partnership Initiative (NEPI) 717
Nurturing Care Framework 444

nutrition
breast cancer prevention 26
child health 435
disabled people 494
disorders & indigenous peoples' health 481
double burden of 63–64
mental health 102
optimal quantity & quality 764–765
pancreatic cancer 28
South Africa 579–560 *see also* diet
nutritional neuropathies 152
nutritional switch 455
NVP (nevirapine) 201

Obama, Barack 288
obesity 57–71
adolescent health 457
adult definition 57–58, 58
t
burden of disease 62–63, 62
t, 63
t
causes 64–66
cost-effectiveness of prevention 68, 68
t
definitions 57–59
disabled people 494
disease as 58–59
epidemic of 64, 66
estimated costs 62–63
Foresight obesity system map 65
f
future trends 62
global trends 59–62, 59
b
prevalence 59–60, 59
f, 60
t, 764
prevention 66–68, 67
f
public health approaches 67
b see also overweight
obesity health outcomes 62
asthma 50
cancer 19
cardiovascular disease 9, 13
liver cancer 22
low back pain 131

musculoskeletal disorders 137
non-alcoholic fatty liver disease/non-alcoholic steatohepatitis 261
oesophageal cancer 27
Obesity: Preventing and Managing the Global Epidemic (WHO) 59
observational epidemiological studies 303
OC (oral contraceptives) 22

occupational exposures
asthma risk factor 50
COPD risk factors 47

occupations
cancer aetiology 19
emerging/re-emerging infections 276
musculoskeletal disorders 138
OECD *see* Organization for Economic Co-operation and Development (OECD)
oesophageal cancer 26–27
oestrogen replacement therapy 150
Office International d'Hygiène (OIHP) 280
Office of Healthy Schools (USA) 671
Office of Transition Initiatives (OTI) 726
OGTT (oral glucose tolerance test) 86
old-age dependency ratio 514
older people 507–522
age-friendly cities 519
contributions of 519
definition 507
demographic transition 511–512
differential ageing 513
disability & function 514–515
elder abuse 517
epidemiological transition 511–512
ethnic minorities 510
families and 409–411
frailty 515
healthcare services 687
LGBT+ people 510–511
life expectancy 512–513, 512
t
mental health promotions 101
mortality causes 511
t
multimorbidity & polypharmacy 514
musculoskeletal disorders 138
obesity definition 58
own perceptions of ageing 513–514
physical activity 74, 75
prisoners 548–549
productivity & social contribution 514
rural area 511
societal perceptions of ageing 513–514
stereotypes 514
successful ageing 513
usual ageing 513
older people healthcare 515–517
access problems 515
community-based care 517–518
long-term institutional care 517–518
primary care 517
promotion activities 515–516
speciality care 517

specific conditions 516–517
unpaid caregivers 518–519
one-child families, China 404–405
1000 Genome Project 302, 302
b
OOP (out-of-pocket) *expenses*, child health 446
OPCW (Organization for the Prohibition of Chemical Weapons) 396
operational details, public health strategies 662
operationalization, intersectional social positions 473
opioids 338
opioid substitution therapy (OST) 343
HIV infection prevention 206–207
Opisthorchis viverrini infection 19
opportunistic infections 617
OPV (oral polio vaccine) 276–277
oral cavity cancer 28–29, 116
prevention 29
oral cleaning, dental health 121–122
oral contraceptives (OC) 22
oral disease 114–118
cancer *see* oral cavity cancer
dental caries 114–114
global burden 117
impacts of 117
links with other disease 117–118
periodontal diseases 115–116, 117
trauma 116–117
oral glucose tolerance test (OGTT) 86
oral health, consumerism 124–125
oral hygiene 117
oral cavity cancer 29
oral inspection 29
oral polio vaccine (OPV) 276–277
oral rehydration salts (ORS) 441–442
oral rehydration therapy (ORT) 721
organ donation, infected human control measures 611
Organization for Economic Co-operation and Development (OECD)
binge drinking 259
environmental health assessment 645
healthcare macroeconomics 686–687
Health Care Quality 692
f
musculoskeletal disorders cost 139
older people 513
Pressure-State-Response (P-S-R) framework 647
Organization for the Prohibition of Chemical Weapons (OPCW) 396
Organization of African Unity (1951)
forced migrants 527
refugees 528
oropharyngeal cancer 28–29

ORS (oral rehydration salts) 441–442
ORT (oral rehydration therapy) 721
OST *see* opioid substitution therapy (OST)
osteoarthritis 132–133
osteomyelitis 179
osteopathies 130
osteoporosis 133
  older people 516
OTI (Office of Transition Initiatives) 726
otitis media 165
  t
Ottawa Charter for Health Promotion (1986) 641, 663
  healthcare services 685
outbreak settings, vaccination schedules 603
outcome focused plans, public health strategies 671
outdoor activities, infection control measures 606
out-of-pocket (OOP) *expenses*, child health 446
overdiagnosis, medical screening 634–636

overweight
  adolescent health 457
  causes 64–66
  disabled people 494
  global trends 59–62
  liver cancer 22
  oesophageal cancer 27
  prevalence 764
  prevention 66–68 *see also* obesity
oxazolidinones 220
oxygen therapy, pneumonia 177

Pacific Open Learning Health Net (POLHN) 717
PAHPA (Pandemic and All-Hazards Preparedness Act) 289
PAHPRA (Pandemic and All-Hazards Preparedness Reauthorization Act) 289

pain
musculoskeletal disorders 130
osteoarthritis 133
Palestine, refugees 527
Pan American Health Organization 280
Pan American Sanitary Bureau 280
pancreatic cancer 27–28
Pandemic and All-Hazards Preparedness Act (PAHPA) 289
Pandemic and All-Hazards Preparedness Reauthorization Act (PAHPRA) 289
Pandemic Influenza Preparedness Framework (WHO) 762
pandemics 270
prevention of 762–763
Papanicolaou smear (Pap test) 31
parasitic infections 181–182
cancer aetiology 20
t
chain of infection 598
diagnosis & treatment 182
t
parent–child relationships, mental health 99, 100, 102
Paris Climate Change Agreement 741
Parkinson's disease (PD) 152–154
clinical overview 152
incidence 153
prevalence 153
public health impact 154
risk factors 153–154
partial restrictions, tobacco smoking 326
Partial Test Ban Treaty (PTBT) 395
Partners for Mental Health 675

partnerships
  environmental health 649–650
  public health strategies 669–670
partner violence 387
passive immunization 603–604
pasteurization 615
Pasteur, Louis 583
PATH (Programme for Appropriate Technologies in Health) 766
pathogen genomics 311
pathogenicity definition 599
pathogenic paradigm 567
Patient Protection and Affordable Care Act (2010) 559

patients
clinician partnerships 699
b
empowerment of 691, 693
engagement of 691, 693
expectations 689
patrilineal kinship systems 404
patrilocality 475
PBTZ 169, 220
PCA (principal component analysis) 304
PCL-R (Hare Psychopathy *Checklist* - Revised) 382
PCV (pneumococcal conjugate vaccine) 589–590
PD *see* Parkinson's disease (PD)
Peace Research *Institute* in Oslo (PRIO) 446
peak expiratory flow (PEF) meter 44–45
pedestrian safety 369
peers, adolescent health 461
penetrance 298
penetrating injuries, traumatic brain injury 145
penicillin-resistant *Staphylococcus aureus* 277
PEOPLE framework 669–675
People Living With AIDS (PLWA) 198
people who inject drugs (PWID)
hepatitis B 250
hepatitis C 256, 341
HIV infection 206–207, 340, 341
infection control measures 606, 611
prevention frameworks 343
prisoners 547
performance indicators, environmental health 647
perinatal asphyxia 438–439
periodontal diseases 115–116, 117

peripheral artery disease
  prevalence 4–5
  risk factors 10
peripheral neuropathy 151–152
persistent diarrhoea 172
personal genomic risk profiles 312

personality disorders
interpersonal violence 382–383
prisoners 546
personalized medicine, genetic epidemiology 309–310
persuasion, alcohol consumption prevention/control 352
pertussis (whooping cough)
control by chemotherapy 609
mortality 165
t
pertussis vaccine 584
t, 588–589
misinformation 277
pesticides, Parkinson’s disease 153
PETRA trial 204
pharmacogenomics 308

pharmacotherapy
advocacy & group community initiatives 426
animal control measures 612
genetic epidemiology 307–308
improved response identification 307–308
infected human control measures 609
non-alcoholic fatty liver disease/non-alcoholic steatohepatitis 262
tuberculosis 216
pharyngeal cancer prevention 29
PHE *see* Public Health England (PHE)
phenethylamines 339
phenotypes 298
genetic epidemiology 311–312
HICs *vs.* LMICs in diabetes mellitus 89
Philippines, Inter-Agency Committee on Tobacco 577
philosophy, Health Needs Assessment 563–564
PHPT-2 Thai trial 204
Phthirus pubis infection 188
t
physical abuse, interpersonal violence 381–382
physical activity 73–83
cardiovascular disease prevention 13
definitions 73–74, 74
t
ecological perspective 77–78
extrinsic influences 77, 78
f
geographical distribution 80
f
guidelines for 75, 75
b
health benefits 74, 74
b
historical aspects 73
obesity 64, 66
oesophageal cancer 30
prevalence 75–76
public health strategies 78–82, 78
b
settings-based approaches 79
b
settings of 77
b
trends 76–77, 76
f see also exercise; sedentary behaviour
Physical Activity and Health campaign (WHO) 664
*Physical Activity and Sedentary Behaviour Guidelines (*Australia) 75
Physical Activity Guidelines for Americans 75

physical health
adolescent health 456
mental health and 423
prisoners 546–547
physical inactivity *see* sedentary behaviour
Physician Health Study 10
pictorial warnings, tobacco smoking 324–325
pilot cessation clinics, tobacco smoking prevention 331
pioglitazone 262
PIVOT trial (Prostate Cancer Intervention Versus Observation Trial) 632
placental disease transmission 600
plague, urbanization and 275
plaque, dental cleaning 121–122
Plasmodium falciparum 228
chloroquine-resistance 235, 278 *see also* malaria
Plasmodium knowlesi 229 *see also* malaria
Plasmodium malariae 228–229 *see also* malaria
Plasmodium ovale 229 *see also* malaria
Plasmodium vivax 228
parasitaemia 230 *see also* malaria
Plummer–Vinson syndrome 27
PLWA (People Living With AIDS) 198
PMTCT (prevention of mother-to-child transmission) 203–204
pneumococcal conjugate vaccine (PCV) 589–560
pneumococcal disease vaccination 584
t
pneumococcal polysaccharide vaccines (PPV) 590
pneumococcal vaccines 177, 589–560
child health 441
child vaccination routine 586
t
Pneumocystis jirovecii infection 617
chemoprophylaxis 604
childhood HIV/AIDS 444
historical aspects 197
pneumonia 174–175
antibiotic treatment 176–177, 176
t
diagnosis 176
environmental health 643
neonates 439
typhoid fever and 179
under-fives child health 433
pneumonic plague 610
pneumococcal vaccine 179
POC (Protection of Civilians) 378
POLHN (Pacific Open Learning Health Net) 717

policies
adolescent health 462
musculoskeletal disorders 138–139
non-communicable diseases 578–580
war prevention 398

polio
child vaccination routine 586
t
eradication target 616
reduction of 584
vaccination 584
t, 587
polishing, periodontal diseases 116
Political Declaration on Non-communicable diseases (UN) 329

politics
commitment to immunization programmes 583
Health Needs Assessment 563–564
humanitarian emergencies 729–730
non-communicable diseases 575
public health strategies 675
Polluter Pays Principle 650
environmental health 640
pollution 644
cancer aetiology 19
environmental health 641
polonium *contamination* (2006, UK) 747
b
polymerase chain reaction (PCR), malaria diagnosis 233
polymorphic markers 301
polyneuropathies 151
polypharmacy 339
older people 514
population ageing 507, 508
f
determinants of 507, 509
family effects 409, 409
f
implications of 519
specific factors 509–511
population-based approaches
follow-up in emergencies 746–747, 746
b
healthcare services 697
Health Needs Assessment 560–561
mental health promotion 100
obesity prevention 68
public health strategies 672
population growth 642
emerging/re-emerging infections 274

population health
  Health Needs Assessment 563–564
  intersectionality 473
population shifts, emerging/re-emerging infections 274–275
population stratification 304
positive predictive value (PPV) 627–628
positive youth development (PYD) 461
possible serious bacterial infections (PSBI) 167
post-traumatic stress disorders (PTSD)
  forced migrants 535
  interpersonal violence 383
  prevention 103
  prisoners 546
  rape 379
  traumatic brain injury 145

poverty
  disabled people 493–494
  eradication of 641–642
  forced migrants 538
  humanitarian emergencies 721
  interpersonal violence 384–386
power hierarchies, gender 416
PPPs (public–private partnerships) 752–754
PPROM (preterm premature rupture of membranes) 169
PPV (pneumococcal polysaccharide vaccines) 590
PPV (positive predictive value) 627–628
practice level, musculoskeletal disorders 139–160, 139
  f
practitioners, environmental health services 652–659, 652
  b
Precautionary Principle, environmental health 640
preclinical phase 625, 625
  f
predictive values, medical screening 627–628, 627
  f
pre-erythrocyte malaria vaccines 239

pregnancy
adolescent health 459
malaria 231, 232, 236, 236
t
seasonal influenza vaccination 594
sexually-transmitted infections 192
pre-industrial societies, puberty onset 454
pre-injury phase 365
premarital sexual activity 407

prematurity
  health and 438
  mortality from non-communicable diseases 573
  neonatal sepsis 166–167
premenstrual asthma 49
prenatal maternal stress 99
PrEP, HIV infection prevention 205, 207
prepayment systems, child health 446
Pressure-State-Response (P-S-R) framework 647
presumptive malaria diagnosis 233
preterm premature rupture of membranes (PPROM) 169
pretomanid 220
preventative post-exposure tuberculosis vaccines 221
preventative service provision, war 398
prevention of mother-to-child transmission (PMTCT) 203–204
Prevotella infection 51
price controls, alcohol consumption 354
primaquine 235
primary healthcare 760
  developing countries 704
  disabled people 498
  older people healthcare 517
  public health interventions 677
  reorientation from hospitals 696

primary prevention
cancer 31–33
cardiovascular disease 12
definitions 598
environmental health 641
forced migrant healthcare 536–537
mental health 102
primary producers 577–578
Prince Mahidol Award Conference (2014) 711
principal component analysis (PCA) 304
PRIO (Peace Research *Institute* in Oslo) 446
prions 183
chain of infection 599
priority setting, healthcare services 693–695, 694 *f*
prisoners 543–555
cost of incarceration 550
drug abuse 547
family members 548
future challenges 551–552
health benefits 546–547
historical aspects 543–544
ill-health reasons 545–547
indigenous peoples 545
international covenants & protection 551–552
international perspective 544–545
mental health 546–547
older offenders 548–549
physical health 546–547
population numbers 544–545, 544 *f*
public health *opportunities* 550, 550 *t*
research challenges 549
tobacco smoking 547–548
women prisoners 548
Private Finance Initiative (PFI) 754
private industry, public health support 755–756
private support of public health 751–758
activism 753
challenges 755
current giving 752
future work 757
historical aspects 751–752
need prioritization 754–755
private industry 755–756
public–private partnerships 752–754
privatization, healthcare services 690–691
privilege 471
product-based partnerships 753

product-development partnerships 753
production monopolization, alcohol consumption prevention 354
productivity, older people 514
Productivity Commission (Australia) 499–500
profit, perception of 576
Programme for Appropriate Technologies in Health (PATH) 766
programmes, war prevention 398
prohibition, alcohol consumption prevention 354
Prohibition of Child Marriage Prevention Act (2006) 474
project plans, Health Needs Assessment 560

promotion
older people healthcare 515–516
restrictions in alcohol consumption prevention 355
prophylactic appendectomy 605
prostate cancer 30–31
Prostate Cancer Intervention Versus Observation Trial (PIVOT trial) 632
prostate cancer screening 631–632
overdiagnosis 635
prostate-specific antigen (PSA) 30, 631, 632
protease inhibitors 201, 201–202
t
protection duration, vaccination 602–604
Protection of Civilians (POC, UN) 378
protein–energy deficiencies 533
Proteus infection 168
t
Protocol Against Smuggling of Migrants by Land, Sea, and Air (2000, UN) 526
Protocol to Eliminate Illicit Trade in Tobacco Products 329
Protocol to Prevent, Suppress, and Punish Trafficking in Persons (2003) 526
PSBI (possible serious bacterial infections) 167
Pseudomonas infection 168
t
psittacosis 612
P-S-R (Pressure-State-Response) framework 647
psychiatric treatments, suicide prevention 105
psychopathy, interpersonal violence 382
psychosis 492
prevention 103–104
psychosocial development, adolescents 454–455
PTBT (Partial Test Ban Treaty) 395
PTEN gene 25
PTSD *see* post-traumatic stress disorders (PTSD)
puberty onset 454
Public Health Act (1848) 639
Public Health England (PHE) 676
case study 670
Public Health Security and Bioterrorism Preparedness and Response Act (2002) 288
public–private partnerships (PPPs) 752–754
public reassurance, emergencies 747
pulmonary embolism 635
pulmonary tuberculosis 21
purchasing, healthcare services 697
purified antigen vaccines 602
PWID *see* people who inject drugs (PWID)
PYD (positive youth development) 461
pyelonephritis 179
pyrantel palmoate 182
pyrethroids 236–237

Quaker Upper Arm Circumference (QUAC) Stick 720
qualifications, environmental health officers 653
quality, healthcare services 688–689
quality-adjusted life years (QALYs)
  cost utility and 567
  gender effects 419
  Health Needs Assessment 562
quantitative multiple risk assessment 744

quarantine
complete quarantine 610
historical aspects 639
infected human control measures 610
modified quarantine 610
quinine 227, 235

RA (rheumatoid arthritis) 133–134
rabies 269
artificially acquired passive immunization 604
mortality 165
t
rabies vaccination 593
child vaccination routine 586
t
historical aspects 583
race 469–477
caste *vs.* 474
t
terminology 470
racialization 471

racism
  definition 471
  indigenous peoples' health 483
radiation, cancer aetiology 19–20
radical prostatectomy 632
Radioactive Incident Monitoring Network (RIMNET) 742

radioactive incidents
  decontamination 746
  risk characterization 743
radioactive weapons 395–396
rainfall patterns 763
randomized controlled trials (RCTs)
  medical screening 630, 630 *f*
  prostate-specific antigen 632

rape
alcohol abuse 382–383
post-traumatic stress disorders 379
poverty 384
rapid diagnostic tests for malaria (mRDTs) 233
rare-variant searches, functional validation 304, 306
b
rat control 608, 612

rationing
  alcohol consumption prevention/control 354
  healthcare services 694
RBCs (red blood cells), malaria 231
RBM (Roll Back Malaria) 227, 754
RCTs *see* randomized controlled trials (RCTs)
reablement services 518
ready-to-use therapeutic foods (RUTFs) 439–440
recentralization, healthcare services 690–691
receptors, public health risk assessment 743
recombinant antigen vaccines 602
recombinant DNA technology, hepatitis B virus vaccines 252
recovery, emergency risk management 726, 746
recruitment reform, transformational learning 715
red blood cells (RBCs), malaria 231

Red Cross
International Committee 762 *see also* International Committee of the Red Cross (ICRC)
Red Cross Code of Conduct 730, 731
humanitarian emergencies 721
reduced penetrance 298
reduction, animal control measures 611–612
Reduction of Harmful Use of Alcohol campaign (WHO) 664
re-emerging infectious diseases *see* emerging/re-emerging infectious diseases
refugees 526
t
legal frameworks 528–529
mental health 534–535
public health 530
vaccination programmes 595 *see also* forced migration
regions, dental health inequalities 118
Registered Environmental Health Specialist/Registered Sanitarian (REHS/RS) 653

registries
animal control measures 612
birth registries 447

regulation
healthcare services 690
healthcare workers 698
non-communicable disease prevention 575

regulations
  environmental health policies 650
  reforms in transformational learning 714–715
rehabilitation, disabled people 498–499

rehabilitative care
musculoskeletal disorders 130
older people healthcare 518
REHS/RS (Registered Environmental Health Specialist/Registered Sanitarian) 653
relationship-level factors, injuries 367
relative risk (RR)
migraine 144–145
tobacco smoking and 320
religious movements, alcohol consumption prevention/control 355
remote sensing, humanitarian emergencies 733
renal failure, diabetes mellitus 86
renin angiotensin 13–14
Report of the Committee on Infectious Diseases 603
reproduction number *(R_0)* 603

reproductive health
disabled people 498
families 410
forced migrants 533

research
biases in indigenous peoples' health 484
prisoners 549
public health interventions 677–678, 679 *f*
reservoirs, infectious diseases 599
residential treatment, substance use disorders 344–345

resources
  healthcare services 694
  Health Needs Assessment 562–563
  infectious diseases control 615
  war 394–395
RESPECT 388
respiratory diseases, indigenous peoples' health 481
respiratory syncytial virus infection 51
respiratory tract, infectious disease transmission 600
respondent-driven sampling, substance use disorders 342
responses, emergency risk management 726
restriction, animal control measures 611–612
reverse isolation, host infection control measures 607
re-victimization, interpersonal violence 380
rheumatic heart disease (RHD) 480
rheumatoid arthritis (RA) 133–134
rhinovirus infection 51
Rickettsiae 598
rifampicin resistance, tuberculosis 213–214
Rift Valley fever 275
rights-based framework, humanitarian emergencies 729
rights of the child 436
RIMNET (Radioactive Incident Monitoring Network) 742
rinderpest eradication 585, 616
Rio Declaration on Environment and Development (1992) 640
risk assessment for public health 742–744
  multi-risk assessment methods 743–744
  pathway of 742–743
  receptors 743
  risk characterization 743
  shelter in place 745
  source of 742
  techniques & countermeasures 745

risks
genetic epidemiology 312
medical screening 626
prediction in genetic epidemiology 308–309, 309
f
public health risk assessment 743
RNA 297–298
road traffic incidents (RTIs) 362
economic cost 365
historical aspects 368–369
LMICs 365
mortality 362, 364
prevention strategies 368–369
traumatic brain injury 145
Rockefeller Foundation 751–752, 756
Roll Back Malaria (RBM) 227, 754
Ross, Donald 227
rotavirus infection 171
t
child diarrhoea 441
rotavirus vaccine 441, 590
child vaccination routine 586
t
gastrointestinal tract infections 174
routes of transmission, infectious diseases 599–600
routine immunizations, older people 515
RR *see* relative risk (RR)
RTIs *see* road traffic incidents (RTIs)
RTS,S/AS01 vaccine 239
rubella vaccine 591
child vaccination routine 586
t
reduction by 584
t
Rules for the Treatment of Women Prisoners and Non-custodial Measures for Women Offenders (UN) 548
rural areas 710
f, 711
f
older people 511
urban areas *vs.* environmental health 642
urban migration, emerging/re-emerging infections 274
Russell Sage Foundation 751
RUTFs (ready-to-use therapeutic foods) 439–440
Rwanda, Joint Evaluation of Emergency Assistance 537

safe drinking guidelines 350
safe environments, suicide prevention 104

safe water provision, infectious diseases control 613
SAINT trial 204
sale conditions, alcohol consumption prevention/control 354
Salmonella enterica infection
  bioterrorism 277, 290 *see also* typhoid fever
Salmonella infection 188
  t
  age-of-onset 168
    t
  food sanitation 613

salt
  fluoridation 120
  stomach cancer 23
SAM (severe acute malnutrition) 439
same-sex interpersonal violence 422

sample size
gene–environment interactions 300
genetic studies 303
genome-wide association studies 302
Mendelian randomization studies 307
sampling, substance use disorders 342

sanitation
  emerging infectious diseases 617
  parasitic diseases 182
sanitation movement (1800s) 639
Sarcoptes scabiei infection 188
  t
Sarin biological attack (1995) 292
SARS *see* severe acute respiratory syndrome (SARS)
SARS-CoV2 pandemic (2020) 282–283
SASA! Intervention (Uganda) 387
satellite imaging, humanitarian emergencies 733
Saving and Empowering Young Lives in Europe (SEYLE) 97, 105
SBET (standby emergency treatment), malaria 239
SCC (squamous cell carcinoma) 26
Schistosoma haematobium infection 19
schistosomiasis 443
schizonts, *Plasmodium* lifecycle 229
school-age children 433–434, 433
  f
  health initiatives 447

schools
adolescent health education 460–475
alcohol consumption prevention/control 355
mental health promotions 101
physical activity 79
b
public health strategies 670
suicide prevention interventions 97
surveys in substance use disorders 342
science, healthcare services 699
b
scrapie 183

screening programmes
  cancer prevention 33
  host infection control measures 605
SDGs *see* Sustainable Development Goals (SDGs)
SDH *see* Social Determinants of Health (SDH)

seasonal influenza
child vaccination routine 586
t
vaccination 594
seasonal malaria chemoprevention 238
secondary injury prevention 365

secondary prevention
asthma 52
cancer 33
cardiovascular disease 12
definitions 598
environmental health 641
forced migrant healthcare 536–537
mental health 102
second-hand smoking (SHS)
adverse effects 322, 323
chemicals & carcinogens 323

sedentary behaviour
cardiovascular disease 9
definitions 73–74, 74
t
disabled people 493
ecological perspective 77–78
guidelines for 75, 75
b
health consequences 74–75
obesity 64
prevalence 76
replacement with light-intensity activity 74
settings of 77
b
trends 76–77 *see also* exercise; physical activity
seizures 147, 147
t
selection bias, rare-variant searches 303
self-determination, indigenous peoples' health 482
self-harm 364–365
self-regulation, healthcare professionals 690
self-reporting, physical activity 77
Senate Subcommittee on Bioterrorism and Public Health Preparedness (SSBPHP) 288
Sendai Framework for *Disaster* Rick reduction (2015–2030) 741, 743

sensitivity
  medical screening 626, 627
  specificity *vs.* of medical screening 628
sepsis, neonates *see* neonatal sepsis
septic arthritis, typhoid fever and 179
sequencing studies 300
SERCaP (Single Encounter Radical Cure and Prophylaxis) 243
SERPINA1 gene 47
Serratia infection 168
  t
service user empowerment, mental health services 106
SES *see* socioeconomic status (SES)
severe acute malnutrition (SAM) 439
severe acute respiratory syndrome (SARS) 272, 617–615, 723
  Hong Kong case study 661
severe community-acquired diarrhoea 173
severe malaria 231–232
severity, prostate cancer screening 631
sex definition 416, 416
  b

sexual abstinence
HIV infection prevention 204–205
host infection control measures 605–606
Sexual and Reproductive Rights 424
sexual behaviour, infection control 605–606, 611

sexual health
adolescents 457, 459
disabled people 498
sexual history, uterine cervical cancer 31
sexually-transmitted infections (STIs) 187–188
t, 187–196
antimicrobial resistance 194
cancer 192
control strategies 193
b
diagnosis 189
forced migrants 532, 533
future strategies 193–194
global initiatives 194
global management costs 192–193
global numbers 189–190, 189
f
HIV infection prevention 205
morbidity 187–188
t, 191–193
prevalence 187
prevention 189
regional initiatives 194
signs & symptoms 188
syndromic management 189–190
technological advances 194
transmission 193
viral infections 188
t, 190–193
sexual minorities, suicide 97
sexual networks 193
sexual orientation 416
b
sexual partner numbers 606
sexual violence causes 381

sex workers
interpersonal violence 378
trafficked/smuggled people 535–536
SEYLE (Saving and Empowering Young Lives in Europe) 97, 105
SF-12 (Short-Form Health Survey) 132
SGLT-2 inhibitors 7
Shanghai, Declaration on Promoting Health (2016) 663–664
shared regulation, healthcare professionals 690
shelter in place 745
Shigella infection (shigellosis) 188
t, 617
antibiotic resistance 442
shoes, infection control measures 606
short-course chemotherapy, tuberculosis 219
Short-Form Health Survey (SF-12) 132
SHS *see* second-hand smoking (SHS)
SIDS (sudden infant death syndrome) 322
Single Encounter Radical Cure and Prophylaxis (SERCaP) 243
single-gene disorders 298–299
single-nucleotide polymorphisms (SNPs) 298
cardiovascular disease 6
genome-wide association studies (GWAS) 301–302, 302
b
single risk assessment 742
skilled birth attendants 438
skin, infectious disease transmission 600

smallpox
eradication of 583, 616
historical aspects of vaccines 583
reduction by vaccination 584
t
Smoke Alarm Installation and Fire Safety Education, National Center for Injury Prevention and Control 370
smoke-free areas 327
Smoke-Free World 755
smuggled persons 526, 526
t
legal frameworks 530
mental health 535–536
smuggling, alcohol consumption 356
sneeze behaviour, infected human control measures 611
Snow, John 639
SNP Consortium 302
b
SNPs *see* single-nucleotide polymorphisms (SNPs)
SNRT (Global Network of the Society for Research on Nicotine and Tobacco) 329
SNS (Strategic National Stockpile) 288
social burden, injuries 365
social consequences, interpersonal violence 379
social constructions, masculinity 420
social contribution, older people 514

social determinants
child health 435–436, 436
b
child mortality & morbidity 434
substance use disorders 339
Social Determinants of Health (SDH)
definition 471
disabled people 493–496
gender 415
public health strategies 662
social disparities, obesity trends 60–61
Social Ecological Model 366, 367
f
social epidemiology 469–468
social equity, mental health 102
social identity 470
social implications, genetic epidemiology 310, 310
b
social justice 567
dynamic approach to 568
social learning theory 380, 383
social movements, alcohol consumption prevention/control 355
social norms, violence 384

social position
definition 470
health 471–472, 472
f
social responsibility, tobacco industry 328
societal disasters 741

societal factors
alcohol consumption and 351
belief systems 686
gender 509–510
injuries 367
older people 513–514

societal levels
alcohol consumption policies 356
interpersonal violence 386

socioeconomics
delayed marriage & cohabitation 406
family changes 403
HICs *vs.* LMICs in diabetes mellitus 89–90
tuberculosis 217
socioeconomic status (SES)
asthma risk factor 49
dental health inequalities 118
obesity 61
physical activity 76
prisoner health 545
suicide 99
women 420
Socio-Economic Survey (Cambodia) 494
sociopolitically directed violence, traumatic brain injury 145
soft tissue disorders 130
soil contamination, war 394
soil-transmitted helminths (STH) infections 443
somatic variants 298
source–pathway–receptor model 742
b

South Africa
Choice of Termination Act (1996) 424
disabled health service access 497
IMAGE study 387
Momconnect 424
non-marital childbearing & divorce effects 408–409
nutrition policies 579–580
Stepping Stones 387
Treatment Action Campaign 426

South Asia
  delayed marriage 407
  under-fives child health 431
Southeast Asia Regional Committee 711
Soviet Union, biological weapons 289
Special Assistant to the President for Biodefence (ASPR) 288

specialization
  healthcare workforce 687
  older people healthcare 517
special studies, substance use disorders 342
specific conditions, older people healthcare 516–517

specificity
  medical screening 626, 627
  sensitivity *vs.* of medical screening 628
Specific Trade Concerns (2010, WTO) 577
SPHERE 538
spirometry 43
sponsorship bans, tobacco smoking 326–327
sporozoite rate (SR), malaria 242
SPRINT (Systolic Blood Pressure Intervention Trial) 8
squamous cell carcinoma (SCC) 26
SR (sporozoite rate), malaria 242
Sri Lanka, delayed marriage 407
SSBPHP (Senate Subcommittee on Bioterrorism and Public Health Preparedness) 288
stakeholder collaboration, transformational learning 714
Standard Minimum Rules for the Treatment of Prisoners (UN) 551
standard precautions, isolation recommendations 610
standby emergency treatment (SBET), malaria 239
Staphylococcus aureus infection
  age-of-onset 168
    t
  food sanitation 613, 614
  penicillin-resistant *Staphylococcus aureus* 277

state control
  healthcare services 689–690
  neoliberalism 575–576
statins (HMG-CoA reductase inhibitor drugs) 9

statistics
genome-wide association studies 302
global health sustainability 765–766
stavudine (d4T) 201
Stepping Stones programme (South Africa) 387, 426
stereotypes, older people 513, 514
sterilization 615
STH (soil-transmitted helminths) infections 443
Stichting INGKA Foundation 754

stigma
disabled people 496
interpersonal violence 386
mental health 97, 98
stimulants, substitution treatments 343
STIs *see* sexually-transmitted infections (STIs)
stomach cancer 23
Stop TB Partnership 754
STPs (Sustainability and Transformation Plans) 564
Strategic National Stockpile (SNS) 288
strategies for public health 660–683
Africa 665–666
Canada 667
b
definitions 661–662
delivering of 675
England 667
Europe 665
global strategies 662
global to national strategy and action 668
leadership & governance 675–676
life-course approach 673, 673
f
local-level action 668–669
local strategy implementation 669–675, 669
t
national & subnational 667–668
need for 660
non-communicable diseases 664, 664
b
population-based interventions 672
reasons for 660–661
regional 665–668
social determinants of health 662
strategy implementation 668
workforce 676–677, 677
b see also health promotion
strategy implementation, public health strategies 668
stratified medicine, genetic epidemiology 309–310, 309
f
strength training, older people healthcare 515
streptococcal sore throat 609
Streptococcus mutans infection 114
Streptococcus pneumoniae infection 175
age-of-onset 168
t
meningitis 177
Streptococcus pyogenes infection 168
t, 605

stroke
incidence 4
indigenous peoples' health 480
migraine 145
overdiagnosis 635
risk measures 10
Strongyloides stercoralis infection 182
t
structural interventions, adolescent health 462–463
structural policies, non-communicable disease prevention 575
studies, intersectionality 473
stunted growth 440
subclinical cardiovascular disease assessment 12
subnational public health strategies 667–668
suboptimal development, child health 444–445

subpopulations
obesity trends 61
physical activity 76
sub-Saharan Africa
alcohol companies 577
non-marital childbearing & divorce effects 408–409
under-fives child health 431
substance use disorders (SUDs) 96, 337
administrative data 342
diagnosis definitions 340
disabled people 493
global epidemiology 338–339
harm reduction 345
historical interpretation 339
household surveys 342
infectious diseases and 340–341
interpersonal violence 382–383
interventions 342–345 *see also* opioid substitution therapy (OST)
mental health 99
mortality 341–342
partner violence 387
prevention 103
prisoners 546, 547
public health consequences 339–340
sampling strategies 342
school surveys 342
social determinants 339
special studies 342
study methods 342
substance use *vs.* 339–340
surveys 342 *see also* substance use/misuse
substance use/misuse 337–347
global epidemiology 338–339
prevalence 337
public health consequences 339–340
substance use disorders *vs.* 339–340 *see also* substance use disorders (SUDs)
substandard medicines 766
substitution, healthcare services 696
Sudan, disabled health service access 497
sudden infant death syndrome (SIDS) 322
sudden unexplained death in people with epilepsy (SUDEP) 148
SUDs *see* substance use disorders (SUDs)
Sugar Association (USA) 578
sugar intake, WHO recommendations 121

suicide
  adolescents 96–97
  community level 99
  epidemiology 96–97
  method choice 104
  prevention 104–105
  risks of 96
suicide contagion 104–105
sulfadoxine-pyrimethamine 235
sunlight, lower lip cancer 116
Superstorm Sandy (2012) 723
Surgeon General's Report on Physical Activity and Health 75
surgical interventions, infection control measures 605, 609

surveillance
anti-microbial resistance 278–279
against bioterrorism 293
emerging infectious diseases 617
emerging/re-emerging infections 273–274
HIV infection 199
infectious diseases control 615
war documentation 397
Surveillance and Treatment of Prisoners with hepatitis C (STOP-C) project 550
survey-determined malaria prevalence 241
Survey of Disability, Ageing, and Carers (Australia) 491

surveys
prisoner research 549
substance use disorders 342
susceptibility genes identification 307, 308
t
Sustainability and Transformation Plans (STPs) 564
sustainable development, environmental health 640
Sustainable Development Goals (SDGs) 741
air pollution 764
child health 448, 448
b
child nutrition 439
disabled people 501
global health sustainability 759
interpersonal violence 377
mental health 96
musculoskeletal disorders cost 139
non-communicable diseases 574
public health strategies 662
tuberculosis vaccination 586
Universal Health Coverage 686 *see also* Millennium Development Goals (MDGs)
sustainable transportation, road traffic incidents 369
Sverdlovsk, anthrax outbreak (1979) 277
Sweden, Total Consumption Model 351
swine flu virus 271
syphilis 187
t
chemoprophylaxis 604
HIV infection prevention 205
mother-to-child transmission 191
reduction in 194
Syrian war (2011) 721
systemic connective tissue disorders 129
systems-based partnerships 754
Systolic Blood Pressure Intervention Trial (SPRINT) 8

T3 (Test-Track-Trace), malaria 233–234
Taenia solium cysticercosis 148
tagSNPs 302
Tanzania, cholera *re-emergence* (1998) 271
targeted sequencing, rare-variant searches 303
target product profile (TPP), malaria eradication 243
Task Force on Community Preventative Services 368

taxation
adolescent health 462
alcohol consumption prevention/control 354, 356, 765
substance use disorders 339
TBE (tick-borne encephalitis), vaccination 586
t, 592
TBI *see* traumatic brain injury (TBI)
TBV (transmission-blocking vaccines), malaria 239
TCM (traditional Chinese medicine), malaria 227
Teach Violence and Injury Prevention (TEACH-IVP) 371
technological disasters 741

technology
  dental health 124
  dental services 123
  family changes 403
  global health sustainability 766
  healthcare services 688
  humanitarian emergencies 733–734
  physical activity trends 76–77
teenage relationships, interpersonal violence 383
Telaprevir 257
telemedicine, humanitarian emergencies 734
temperance movements 351
tension-type headache 144

tertiary prevention
definitions 598
forced migrant healthcare 536–537
injuries 366
mental health 102
Test-Track-Trace (T3), malaria 233–234

tetanus
  mortality 165
    t
  vaccination 584
    t, 588
tetanus toxoid-containing vaccines (TTCV) 169, 588
theophylline 52
therapeutic targets, clinical advances in discovery 307, 308
  t
therapeutic tuberculosis vaccines 222
Thibela TB trial 214
third-hand smoke (THS) 323
thymine 297
thyroid cancer screening 635
TIAs (transient ischaemic attacks) 4
tick-borne encephalitis (TBE), vaccination 586
  t, 592
tick monitoring, infection control measures 606
timeline, substance use disorders 342
tissue degeneration, musculoskeletal disorders 130

tobacco
chemical & carcinogens 323–324
non-smoking use 326
oesophageal cancer 27
Tobacco Atlas, tobacco taxation 327
tobacco industry 328, 576
tobacco smoking 319–335
adolescent health 457
asthma risk factor 50
cancer aetiology 19
cannabis use and 338
cardiovascular disease 9
causation *vs.* association 321
clinical applications 321
colorectal cancer 24
COPD 46–47
disabled people 492
epidemic stages 326
laryngeal cancer 29
low back pain 131
lower lip cancer 116
lung cancer 21
monitoring 325–326
mortality and 320–321, 320 *f*
numerous reviews 322
oesophageal cancer 26
oral cavity cancer 28
pancreatic cancer 27
Parkinson's disease 153
prevalence 319, 325
prisoners 547–548
public health applications 321
public risk perception 323–324
second-hand *see* second-hand smoking (SHS)
stomach cancer 23
third-hand *see* third-hand smoke (THS)
uterine cervical cancer 31
warnings about dangers 324–325, 324 *f*
women 326
tobacco smoking control/prevention 319, 325–326, 765
advertising bans 326–327
cardiovascular disease prevention 13
clinical *vs.* public health approaches 330–331
COPD prevention 48 *b*
dental health improvement 122
legislation 578
lung cancer prevention 22

media campaigns 325
nicotine dependence *vs.* habit 330
older people healthcare 516
pilot cessation clinics 331
prisoners 547–548
protection from smoke 327
quitting help 329–330
quitting *vs.* reduction 331
quitting *vs.* uptake prevention 330
taxes 327–328
tonsillectomy 605
toothbrushing 122
topically applied fluoride 120
TOPOFF 289
Toronto Charter for Physical Activity (2010) 75, 78
toxicity, environmental health assessment 644–645
toxic shock syndrome 617
toxin release, malaria 231
Toxoplasma gondii infection 605
TP53 gene 25
TPNW (Treaty of the Prohibition of Nuclear Weapons) 396
TPP (target product profile), malaria eradication 243
trade agreements, healthcare workforce 687
trade *agreements*, alcohol consumption policies 356
traditional Chinese medicine (TCM), malaria 227
trafficked people 526, 526
t
legal frameworks 530
mental health 535–536

training
environmental health officers 653
healthcare workers 698
Training for Health Equity Network (THEnet) 716–717
transcription 297
transformative learning 709–718
authentic learning 716
curricula reform 715
financing system reform 715
global movement 711–713
health professional 714
b
information and communication technology 715–716
institutional reform 714, 714
t
instructional reform 715
interprofessional education 715
leadership development & management reform 714
legal & regulatory reforms 714–715
networks on 716
reasons for 709
recruitment reform 715
stakeholder collaboration 714
transgender people *see* LGBT+ people
transient ischaemic attacks (TIAs) 4
translation 297
transmissible spongiform encephalographies (TSEs) 183
transmission-blocking vaccines (TBV), malaria 239
transovarian transmission 600
Trans-Pacific Partnership Agreement 578
transplacental immunity 603
transportation 77
b
infectious diseases control 616
physical activity 79, 79
b
trans-stadial transmission 600
traps, vector control 608

trauma
  dental health improvement 122
  oral disease 116–117
  unintentional injuries 361–362 *see also* injuries
traumatic brain injury (TBI) 145–146
  prisoners 546
travellers' diarrhoea 172
travel medicine, malaria 238–239

treatment
alcohol consumption prevention/control 355–356
HIV infection prevention 205–206
medical screening 626
Treatment Action Campaign (South Africa) 426
Treatment and identification of Psychosis Study (TIPS, Norway) 103–104
Treaty of the Prohibition of Nuclear Weapons (TPNW) 396
Treponema pallidium infection *see* syphilis
Trichinella spiralis infection 613
Trichomonas vaginalis infection 188
t
Trichuris trichuria infection
child health 443
diagnosis & treatment 182
t
trophozoites, *Plasmodium* lifecycle 229
Trump, Donald 288
TRUNCATE-TB28 220
TSEs (transmissible spongiform encephalographies) 183
TST (tuberculin skin test) 216, 218
TTCV (tetanus toxoid-containing vaccines) 169, 588
tuberculin skin test (TST) 216, 218
tuberculosis 213–226
Bacillus Calmette–Guérin 585–587
chemoprophylaxis 604
chemotherapy 609
diagnostic approaches 217–219
diagnostic pipeline 219
drug regimen building 220–221
drug-resistance 213–214, 216, 278
forced migrant mortality 532
global burden 213–214, 222–223, 598
historical aspects 213
host-directed therapies 221
incidence 214
latent infections 218
mortality 165
t
new drug development 219–220
prisons in history 543–544
progression interventions 214, 216–218
signs and symptoms 214, 215
f
surgery 609
tobacco smoking 322
transmission drivers 216
transmission measurement 216–217
transmission prevention 216
urbanization and 275
vaccines 221–222

12-step programme, substance use disorders interventions 345

twin studies
complex multifactorial diseases 299
multiple sclerosis 154
tympanostomy 605
type 1 diabetes 85
type 2A diabetes 90
type 2B diabetes 90
type 2 diabetes 85
indigenous peoples' health 481
prevalence in young 90
typhoid fever 179–180
child vaccination routine 586
t
forced migrant mortality 533
vaccination 592
typhus 543

Uganda
Good Schools Programme 387
SASA! Intervention 387
UHC (Universal Health Coverage) 686, 710, 760–761
UK *see* United Kingdom (UK)
UKPDS (United Kingdom Prospective Diabetes Study) 6
ulcerative colitis (UC) 24
ultrasound, non-alcoholic fatty liver disease/non-alcoholic steatohepatitis 260
UN Commission on Life-Saving Commodities for Women and Children 447
uncomplicated malaria 231
UN Conference on the Human Environment (UNCED), World Commission on environment and *Development* 640
UN Convention on Refugees (1951), forced migration definition 525
under-fives
child health 431–433, 432
f, 433
t
mortality rate (U5MR) 431
undernutrition 435
undetected population stratification 304, 305
f
unemployment, suicide 99
unexplained variations, genetic epidemiology 310–311
unexploded ordinance (UXO) 397
unintentional injuries 361–362
Unit 731 (Japan) 289
United Kingdom (UK)
BSE *re-emergence* (1990–1998) 271, 274
disabled people exclusion 494
environmental health officers 653
healthcare macroeconomics 686–687
healthcare service provision 695
Health Needs Assessment 560–561
Investing in Research, Improving Health 678
Million Women Study 21
National Risk Register of Civil Emergencies (2017) 744
b
public health strategies 676–677
tobacco smoking warnings 324 *see also* England
United Kingdom Prospective Diabetes Study (UKPDS) 6
United Nation Economic and Social Council (1997), gender mainstreaming 417
United Nations (UN)
Arms Trade Treaty 395
Climate Change Conference (2017) 763
Cluster System 727, 727
f
Convention on the Rights of People with Disabilities 97
Convention on the Rights of the Child 436
Convention on Trafficking 530
Framework Convention on Climate Change 281

High-Level Commission on Health Employment and Economic Growth 712–713
humanitarian emergencies 720
indigenous peoples 479
Inter-Agency Standing Committee 762
Political Declaration on Non-communicable diseases 329
Protection of Civilians 378
Protocol Against Smuggling of Migrants by Land, Sea, and Air 526
refugee definitions 524
Rules for the Treatment of Women Prisoners and Non-custodial Measures for Women Offenders 548
Security Council and children in war 446–447
Standard Minimum Rules for the Treatment of Prisoners 551
Treaty of the Prohibition of Nuclear Weapons 396
Universal Health Coverage 686
war prevention 398
youth definition 463
United Nations Children's Fund (UNICEF) 729
forced migrant health 536
United Nations Convention on the Rights of the Child (UNCRC) 462
United Nations Declaration on the Rights of *Indigenous* Peoples (UNDRIP) 479
United Nations Development Programme (UNDP) 723
United Nations High Commission for Refugees (UNHCR) 527, 527 *t*, 528 *f*
asylum seeker definitions 525–526
humanitarian emergencies 719
international protection for refugees 529
public health solutions 536
United Nations International Strategy for Disease Reduction (UNISDR) 721
United Nations Office for Coordination of Humanitarian Affairs (UNOCHA) 727
United Nations Office on Drugs and Crime (UNODC) 337
United Nations Population Fund (UNFPA) 536
United Nations Programme on HIV/AIDS (UNAIDS) 165–166, 199–200
United Nations Relief and Works Agency for Palestinian Refugees in the Near East (UNRWA) 527
United States of America (USA)
Affordable Care Act 561
alcohol temperance movement 351
bioterrorism preparedness 288–291
Bureau of Labor Statistics, Census of Fatal Occupational Injuries 362
cardiovascular disease prevalence 3
Centers for Disease Control and prevention *see* Centers for Disease Control and Prevention (CDC)
Commission on Chronic Illness 624
Committee on Refugees 536
delayed marriage & cohabitation 405
Department of Transportation, Fatal Accident Reporting System 362
disabled people exclusion 494–495
Executive Order on Children's Environmental Health and Safety (1997) 653–654
healthcare service provision 695
Health Needs Assessment 561
late marriage 406–407

National Lung Cancer Screening Trial 22
National Survey of Drug Use and Health 342
National Trauma Data Bank 363
Office of Healthy Schools 671
Office of the Director of National Intelligence 294–295
population ageing 509
prisoner numbers 544–545
Sugar Association 578
Youth Risk Behavior Survey 363
Universal Health Coverage (UHC) 686, 710, 760–761
universal prevention, mental health 102
unmanned aerial vehicles, humanitarian emergencies 733–734
unpaid caregivers, older people healthcare 518–519
upper respiratory infections, mortality 165
t
Uppsala Conflict Database 723

urbanization
Aedes aegypti 273
Dengue fever 181
environmental health 642
healthy cities 649–650
obesity trends 61
pancreatic cancer 27
physical activity 79
planning & road traffic incidents 369
population levels 710
populations 711
f
rural areas *vs.* environmental health 642
yellow fever outbreaks 273
young people 511 *see also* cities
US Army Medical Research Institute for Infectious Disease (USAMRIID) 289
US Commission for Refugees and Immigrants (USCRI) 527
US Surgeon General (USSG)
tobacco smoking 321
tobacco smoking cessation 329–330
tobacco smoking reports 322
uterine cervical cancer 31
utilitarianism, Health Needs Assessment 562
UXO (unexploded ordinance) 397

vaccination 583–596
acute respiratory infections 177, 440–441
animal control measures 611, 612
antibody responses 602
attacks on 730
f
broader impact 585
challenges to 583
child health 437–438
t, 438, 444, 448
children's recommended routine 586
t, 603
community protection 603
Dengue fever 181
disabled people 496
emerging vector-borne diseases 595
expanded programmes 585–591
gastrointestinal tract infections 174
high-risk populations 592–594
historical aspects 583
host infection control measures 602–603
malaria eradication 243
maternal immunization 595

older people healthcare 516
programme challenges 583
programmes 594
protection duration 602–609
public health impact 583–585, 584
f, 584
t
recommendations for certain regions 591–592
rotavirus 441
schedules 603
special populations 595
tuberculosis 221–222
universal access 595–596

vaccines
development advances 585, 585
t
development timeline 585
t
efficacies 603
types of 602
VADT (Veterans Affairs Diabetes Trial) 7
value-oriented payment models 697
valvular heart disease 5
variable expressivity 298
variant Creutzfeldt–Jakob disease (vCJD) 269, 274
livestock animal products 279

varicella
  mortality 165
    t
  reduction by vaccination 584
    t
varicella zoster vaccines (VZV) 594
  child vaccination routine 586
    t
variolation 583
vascular dementia (VaD) 150–151
VC (vectorial capacity), malaria 242
VCI (National Cancer Institute) 578
vCJD *see* variant Creutzfeldt–Jakob disease (vCJD)
VCTE (vibration-controlled transient elastography) 260
vector-borne diseases, climate change 763
vector control 607–608
  biological controls 608
  bite prevention 611
  chemical control 607–608
  Dengue fever 181
  environmental control 608
  malaria prevention 236–237
vectorial capacity (VC), malaria 242

vectors
infectious diseases control measures 601
injuries of 362
resistance in emerging infectious diseases 617
vehicle accidents *see* road traffic incidents (RTIs)
vehicle-borne transmission 600
Vehicle Miles Travelled (VMTs) 578
vertical public health strategies 661–662
vertical transmission, infectious diseases 599
Veteran Administration study 262
Veterans Affairs Diabetes Trial (VADT) 7
vibration-controlled transient elastography (VCTE) 260
Vibrio cholerae infection *see* cholera
villages 671–672

violence
  against children by women 385–386
  disabled people 494–495
  forced migration 525
  gender 421–422, 421
    b, 422
    b
  prevention 103
  prisoners 551
  social & gender norms 384
  traumatic brain injury 145
  women *see* violence against women
Violence Against Children Surveys (VACS) 378
violence against women 377–378
  disabled people 494
  population ageing 509
viral haemorrhagic fever 533

viral infections
asthma risk factor 51
cancer aetiology 20
t
chain of infection 599
virulence, infectious diseases 599
visibility, forced migrants 539
Vision and Mission, Principles, Goals, and Objectives (NEPAD) 666
visual impairment 515
vitamin B_{12} deficiency 152

vitamin D
  asthma risk factor 50
  colorectal cancer 24
  multiple sclerosis 154
  prostate cancer 30
VMTs (Vehicle Miles Travelled) 578
voice, forced migrants 539
VOICE trial 205
Volatility, Uncertainty, Complexity, and Ambiguity (VUCA) 709
volunteer health workers 705
Bismarck, Otto 563

vulnerable populations
environmental health assessment 644
mental health 102
mental health promotion 100
prisoner research 549

waist *circumference* (WC) 58
waist cut-points 58
t
waist–height ratio 58
waist–hip ratio (WHR) 58
war 393–400
additional violence promotion 395
air, water, & soil contamination 394
attacks on healthcare 730–731
child health 446–447
decrease in 729
documentation 397–398
education & awareness raising 398
environmental consequences 393–394
forced displacement 394
health consequences 393–394
humanitarian assistance 728–729
human rights violations 394
infrastructure damage 394
life expectancy 418–419
morbidity 393–394
mortality 393–394, 723
precipitating causes 393
preventative service provision 398
prevention 397, 398–395, 399
f
prevention policies & programmes 398
resource diversion 394–395
traumatic brain injury 145
weapons types 395–397 *see also* conflicts
warfarin 308
waste reduction 688–689
wasting 435

water
  contamination 394
  decontamination 745–746
  fluoridation 120
  pollution monitoring 646
  purification 613
water-borne diseases, climate change 763
water, sanitation, and hygiene (WASH) 441
WC (waist *circumference*) 58
weapons of mass destruction (WMD) 288
Wellcome Trust 752, 754
Wellcome Trust Case Control Consortium (WTCCC) 302
Western Europe, divorce 408
WGS *see* whole genome sequencing (WGS)
WHA *see* World Health Assembly (WHA)
What Works to Prevent Violence Against Women and Girls? Programme 386
WHO *see* World Health Organization (WHO)
whole-cell extract vaccines 221
whole genome sequencing (WGS)
  sample size 302
  tuberculosis 217
whooping cough *see* pertussis (whooping cough)
WHR (waist–hip ratio) 58
WMD (weapons of mass destruction) 288

women
child mortality & morbidity 434
prisoners 548
reproductive health 379
socioeconomic status 420
tobacco smoking 326
violence against 421
violence against children 385–386
Women Be Healthy 2 496–497
Women's Health Initiative Memory Study 150
Women's Health Initiative study 25
Women with *Disabilities* taking Action on REproduction and sexual health (W-DARE) project 498, 500
work 77
b
disabled people exclusion 494–495

workplace
host infection control measures 606
infectious diseases control 614
mental health promotions 101
physical activity 79
b
public health strategies 671

World Bank
future healthcare systems 709–710
Health Equity database 423
humanitarian emergencies 719
influenza pandemic cost 272
tobacco taxation 327
universal health coverage 760
World Cancer Research Fund 66–67
World Commission on environment and *Development* (UNCED) 640
World Drug Report (UNODC) 337, 338
World Economic Forum 703
World Food Programme (WFP) 731
World Health Assembly (WHA)
child nutrition 439
environmental health 641
Global Vaccine Action Plan 595
non-communicable disease treatment 664
tobacco smoking prevention 765
transformative learning 711
World Health Organization (WHO)
Action Plan for the Global Strategy for the Prevention and Control of Non-Communicable Diseases 664
Action Plan on Antimicrobial Resistance (2015) 281
Action Plan on mental health 97
acute respiratory infection treatment 176
t
adolescent pregnancy 459
After Action Review process 748
alcohol consumption 351–352
anti-microbial resistance strategy 279, 279
f
bioterrorism 618–619
bioterrorism concerns 293, 293
b
Child Growth Standards 435, 439
child vaccination against hepatitis B 253
Collaborating Centres network 282
Commission on Microeconomics and Health (2001) 282
Commission on the Social Determinants of Health 565
costs of essential medicine 760
Dengue fever definition 181
Department of Violence and injuries Prevention 368
diabetes mellitus definition 85
diabetes mellitus diagnosis 86
t
Eastern Mediterranean Regional Office 728
Emergency Medical Teams Initiative 761
Emergency Response Framework 740–741, 761
environmental health burden 643
environmental health definition 641

Expanded Programme on Immunization *see* Expanded Programme on Immunization (EPI) (WHO)
forced migrant health 536
Framework Convention on Tobacco Control 319, 329, 668, 765
future healthcare systems 709–710
gender mainstreaming 417
Global Action Plan for the Prevention and Control of Non-Communicable Diseases 2013–2020 75, 574
Global Action Plan on Antimicrobial Resistance 280
Global Action Plan on Physical Activity 2018–2030: More Active People for a Healthier World 75
global *burden* of infectious diseases 598
Global Database on the Implementation of Nutrition Action 66–67
Global Health Sector Strategy on Sexually Transmitted Infections 2016–2021 194
Global Network for Age-friendly Cities and Communities 519
Global Recommendations on Physical Activity for Health 75
Global Strategy for the Prevention and Control of Non-Communicable Diseases 664
Global Strategy to Reduce the Harmful Effects of Alcohol 351
global suicide rate 96
grades for emergencies 761
Health Emergency Programme 761
Health for All policy 685
health gradient 565
Health Promoting Schools initiative 447
healthy workplace model 671–672
hepatitis B geographical endemicity 250, 250 *f*
hepatitis D epidemiology 254
HIV/AIDS treatment 675
humanitarian emergency definition 721
human papillomavirus (HPV) vaccination 591
Incident Management Process 740
infectious disease mortality 164 *f*
injury prevalence 361, 362
injury prevention 371
intermittent preventative treatment in pregnancy for malaria 237
International Health Regulations 616, 761
Internation Health Regulations revisions 280
joint external evaluation 284
malaria-free country certification 243
malaria vaccines 239
measles eradication 590
medical screening 624–625
Member States to establish Emergency operation Centres 762
Model List of Essential Medicines 761
Montevideo Roadmap 2018–2030 on NCDs as it's Sustainable Development Policy 665
MPOWER package *see* MPOWER package (WHO)
neglected tropical diseases 442
Now More Than Ever 677
obesity definition 57
Obesity: Preventing and Managing the Global Epidemic 59

obesity recommendations 66
older people in rural areas 511
older prisoners 548
Pandemic Influenza Preparedness Framework 762
Physical Activity and Health campaign 664
pneumococcal conjugate vaccination schedule 589–560
pneumonia definition 175
public health 660, 676
quality of care definition 760
Reduction of Harmful Use of Alcohol campaign 664
road traffic incidents 369
rotavirus vaccine 590
rubella vaccination recommendations 591
SARS-CoV2 pandemic (2020) 283
seasonal influenza vaccination 594
sugar intake 121
transformative learning 711
typhoid fever vaccines 180
under-fives child health 431
universal health coverage 760
World Health Report on Violence and Health 377
World Malaria Report (Global Malaria Programme) 243
World Obesity Federation 64–65
obesity as *disease* 59
World Summit on Sustainable Development (WSSD) 640
World Trade Organization (WTO), Specific Trade Concerns (2010) 577
World War I, biological weapons 289
WSSD (World Summit on Sustainable Development) 640
WTCCC (Wellcome Trust Case Control Consortium) 302
WTO (World Trade Organization), Specific Trade Concerns (2010) 577

XACT 1 trial 219

years lost to disability (YLD) 135
years of life lost (YLL) 135
yellow fever (YF)
Angola 2015 outbreak 273
urbanization outbreaks 273
vaccination 586
t, 592
vaccination lapses 273
Yersinia pestis infection 604

youth
empowerment 463
engagement of 463
interpersonal violence 378
research in forced migrants 539 *see also* adolescent health
Youth Aware of Mental Health Programme (YAM) 97, 105
Youth Risk Behavior Survey (USA) 363
Youth Smoking Prevention (YSP) programmes 328

zalcitabine (ddc) 201
Zika virus infection 188
t, 273, 618
mother-to-child transmission 191

zinc deficiency
child diarrhoea 442
gastrointestinal tract infections 174
Zuckerberg, Mark 752